

MACROSCOPIC DYNAMICS IN COMPLEX SYSTEMS

Methods on collective change in social,
physical and transportation systems

Mark M. Dekker

MACROSCOPIC DYNAMICS IN COMPLEX SYSTEMS | Mark M. Dekker



Macroscopic Dynamics in Complex Systems

Methods on collective change in social, physical and
transportation systems

Copyright © 2022 Mark M. Dekker. All rights reserved.

Department of Information and Computing Sciences
Faculty of Science, Utrecht University.
Princetonplein 5, 3584 CC Utrecht, The Netherlands

ISBN: 978-90-393-7465-8
Printed by Ridderprint

Macroscopic Dynamics in Complex Systems

Methods on collective change in social, physical and
transportation systems

Macroscopische Dynamica in Complexe Systemen

Methoden over collectieve verandering in sociale, natuurlijke
en transportsystemen

(met een samenvatting in het Nederlands)

Proefschrift

ter verkrijging van de graad van doctor
aan de Universiteit Utrecht
op gezag van de rector magnificus, prof. dr. Henk Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op
11 april 2022 des middags te 12.15 uur

door

Mark M. Dekker

geboren 7 juni 1992 te Dordrecht, Nederland

Promotor: Prof. Dr. Gerard T. Barkema
Copromotor: Dr. Debabrata Panja

This research was part of the project 'Improving the resilience of railway systems', which was financially supported by Dutch Research Council (NWO), and co-supported by Nederlandse Spoorwegen (NS) and ProRail, under project number 439.16.111.

So it is not true (...) that we each should 'cultivate our own valley, and not attempt to build roads over the mountain ranges (...) between the sciences.' Rather, we should recognize that such roads, while often the quickest shortcut to another part of our own science, are not visible from the viewpoint of one science alone.

Philip Warren Anderson, 1972

Contents

Nederlandse Samenvatting	8
General Introduction	12

I Macroscopic approaches

On the identification of macroscopic states and transitions between them 26

1 Predicting transitions across macroscopic states for railway systems	32
1.1 Introduction	34
1.2 The three-step framework	37
1.3 Coupling the framework to data from the Dutch railways	41
1.4 Implications for day-to-day operations: two case studies	50
1.5 Discussion and Conclusions	52
2 Characterising neural phase-space trajectories via Principal Louvain Clustering	58
2.1 Introduction	60
2.2 Materials and methods	61
2.3 Results	68
2.4 Discussion	72
3 Cascading transitions in the climate system	78
3.1 Introduction	80
3.2 Mathematical framework for cascading tipping	81
3.3 Statistics of cascading tipping points	86
3.4 Application: the coupled AMOC-ENSO system	89
3.5 Summary, discussion and conclusions	96

II Static Network approaches

On the relation between dynamics and underlying network structures 98

4 Geographic delay characterisation of railway systems	104
4.1 Introduction	106
4.2 Methods	109
4.3 Application to European railway systems	113
4.4 Interpretation and relation to daily operations	118
4.5 Conclusion	122
5 Modelling railway delay propagation as diffusion-like spreading	124
5.1 Introduction	126
5.2 Model	129
5.3 Data and Methods	135
5.4 Results	139
5.5 Conclusions	145

6 Cascading dominates large-scale disruptions in transport over complex networks . .	148
6.1 Introduction	150
6.2 The three building blocks for large-scale disruptions.	153
6.3 The model.	155
6.4 Quantification of cascading	157
6.5 Predicting the evolution of large-scale disruptions	160
6.6 Conclusion and outlook	162

III Temporal Network approaches

<i>On the role of interaction sequences and vulnerability to spreading.</i>	<i>166</i>
---	------------

7 The hidden dependence of spreading vulnerability on topological complexity	172
7.1 Introduction	174
7.2 Entropy of temporal entanglement.	175
7.3 Results	178
7.4 Discussion and outlook.	181

8 Quantifying agent impacts on contact sequences in social interactions	184
8.1 Introduction	186
8.2 Methods.	188
8.3 Application to human interaction data	192
8.4 Discussion and Conclusion	199

9 Evaluation of interventions during the COVID-19 first wave in the Netherlands	202
9.1 Introduction	204
9.2 Methodology	206
9.3 Reproducing the first wave	211
9.4 Evaluation of national-level intervention measures.	212
9.5 Evaluation of regional intervention measures	214
9.6 Discussion and Outlook	217

General Discussion	220
Supplementary Information	234
Bibliography.	296
About the author.	313
List of Publications and Working papers.	314
Acknowledgements	316

Nederlandse Samenvatting

Het jaar 2020 zal de geschiedenis ingaan als het jaar van de coronacrisis. Nadat in het voorgaande jaar een nieuw coronavirus werd aangetroffen in Wuhan, China, duurde het niet lang voordat in Europa nationale lockdowns werden afgeroepen. Tot verbluffing van vele wetenschappers en beleidsmakers werd de ziekte in een paar maanden tijd over de hele aarde uitgespreid. Toch vindt besmetting plaats tussen individuele mensen, lokaal en zelfs tijdelijk — ver van de globale schaal waarop verspreiding zichtbaar was. De paradoxale relatie tussen individuele interacties van componenten (de *microscopie*) en fenomenen in het grotere geheel (het collectief, of de *macroscopie*) zijn zichtbaar in systemen uit allerlei wetenschapsvelden. Klassieke voorbeelden zijn te vinden in de biologie, zoals sporeeuen die prachtige zwermformaties laten zien en eenvoudige mieren die samenwerken voor nestbouw en voedsel vergaring. In ons brein zien we een dergelijke relatie terug in grote hoeveelheden zenuwcellen die als collectief leiden tot fenomenen als bewustzijn, herkenning en herinnering. Ook op een veel grotere schaal komt deze paradox terug: naast het genoemde voorbeeld van epidemieën, zijn filevorming en verstoringen in de transportsector of de bankencrisis in de financiële sector noemenswaardige voorbeelden.

In de laatste eeuw is er in de wetenschap de verschuiving gaande vanuit het reductivistisch denken, waarin onderdelen van dergelijke systemen apart worden bestudeerd, naar het systeemdenken, waarin het collectief het hoofdonderwerp wordt. Dit is met name relevant voor de genoemde paradox — bijvoorbeeld, voedselvergaring van mieren is lastig te deduceren vanuit het gedrag en functioneren van een individuele mier, en is pas zichtbaar als men duizenden mieren bij elkaar zet. Een interdisciplinaire toolbox van dergelijke methodiek is te vinden in zogenaamde *complexiteitstheorie*, waarin dit soort systemen worden beschreven als *complexe systemen*, gekarakteriseerd door grote hoeveelheden componenten, en wiskundige eigenschappen als zelforganisatie, systeemgeheugen, niet-lineariteit en netwerkstructuren. Complexiteitstheorie maakt het mogelijk om met een (vaak) natuurwetenschappelijke blik een grote variëteit aan systemen op een abstracter niveau te beschrijven en nieuwe verbanden tussen deze velden te leggen.

Dit proefschrift borduurt verder op deze stroming. Het heeft als hoofdonderwerp 'collectieve verandering in complexe systemen' — hoe dynamiek in deze systemen op de macroscopische schaal ontstaat uit de interactie van veel kleinere onderdelen. Dit is intentioneel een breed onderwerp dat niet met een enkel onderzoek beantwoord wordt.

Door in elk hoofdstuk in te gaan op een andere toepassing, krijgt elk hoofdstuk een eigen perspectief op deze vraag. Dit leidt tot een grote verscheidenheid aan toepassingen: verstoringen in transportsystemen (hoofdstukken 1, 4, 5, 6), kantelpunten in het klimaat (hoofdstuk 3), gedragstoestanden in het brein (hoofdstuk 2), interacties tussen mensen (hoofdstukken 7 en 8) en de verspreiding van infectieziekten (hoofdstuk 9). Het proefschrift heeft daarin een tweeledige focus. Vooropgesteld is de methodologische kant: de hoofdstukken bevatten nieuwe methodiek en modellen die gebruikt kunnen worden in vervolgonderzoek. Voorbeelden zijn dimensie-reductie technieken (hoofdstukken 1 en 2), nieuwe metrieken om systeem-kwetsbaarheden te herkennen (hoofdstukken 7 en 8) en niet-traditionele manieren om recht te doen aan heterogeniteit (hoofdstukken 4 en 9). Naast de methodologische kant van dit proefschrift, bevat elk hoofdstuk ook lessen voor de respectievelijke toepassingen. Dat is de tweede focus van dit proefschrift.

In elk complex systeem, hebben interacties op lokaal en ook op collectief niveau een karakteristieke tijdschaal: een ziekte als COVID-19 verspreidt zich doorgaans in het tempo van dagen en weken, terwijl een gerucht of nieuwsbericht zich op social media binnen enkele uren verspreidt. In dit proefschrift wordt een onderscheid gemaakt tussen de (collectieve) dynamica in een complex systeem, en de interactiestructuren of *netwerken* die onderliggend zijn aan deze dynamica — contactnetwerken van mensen en volgersnetwerken in de genoemde voorbeelden. Ook deze netwerken kunnen veranderen en hebben een bijbehorende tijdschaal. De genoemde relatie tussen dynamica en het onderliggende netwerk is een leidraad in dit proefschrift. Er zijn verschillende methoden om recht te doen aan deze interactie en daarom is dit proefschrift opgedeeld in drie delen. Deel I bevat methoden waarin de interactiestructuur impliciet of statistisch wordt meegenomen door dimensiereductie van alle dynamica en interactiestructuren. Deel II beschouwt systemen met een ruwweg constant onderliggend netwerk, zodat slechts de tijdschaal van de dynamica relevant is. Deel III bevat methoden om macroscopische dynamica van complexe systemen te simuleren wanneer de interactiestructuren ook (in hoog tempo) veranderen. Hieronder ga ik wat dieper in op deze drie delen.

Het eerste deel van dit proefschrift betreft het beschrijven van complexe systemen op macroscopisch niveau. Zonder expliciet de netwerken in het systeem te beschrijven, probeer ik de dynamica te vatten in een kleine hoeveelheid variabelen. Met andere woorden, de dimensie van het systeem wordt gereduceerd tot een laag-dimensionale toestandsvariabele. Vervolgens wordt het gedrag van die variabelen geanalyseerd door het definiëren van toestanden (of *states*) waarin het systeem zich kan begeven, op basis van statistiek of de fysica achter het systeem. De drie hoofdstukken in dit deel betreffen drie casussen: verstoringen in treinsystemen (hoofdstuk 1), gedragstoestanden in muizenbreinen (hoofdstuk 2) en kantelpunten in het klimaat (hoofdstuk 3). In hoofdstuk 1 wordt op basis van data van treinvertragingen een hoofdcomponentenanalyse uitgevoerd om covariante vertragingpatronen te herleiden. De dynamiek binnen de faseruimte met als assen de twee dominante hoofdcomponenten gebruiken we vervolgens om de data te clusteren — resulterende in clusters waarin het treinsysteem efficiënt verloopt, en in clusters waarin er grote verstoringen zijn. Vervolgens bespreekt het hoofdstuk een *early warning* systeem waarin we transities tussen de clusters voorspellen waarin we aantonen middels de Peirce Skill Score hier voorspellingskracht te hebben. De methodiek in hoofdstuk 2 is geïnspireerd op hoofdstuk 1, waar het is toegepast op EEG data van muizenbreinen, met als doel neurologische toestanden met een betekenis voor gedrag te identificeren.

De data is afkomstig uit een experiment over objectherkenning en is opdeelbaar in drie fysiologische delen van het muizenbrein. Het hoofdstuk laat zien dat de hoofdcomponenten in drie delen van het brein, berekend op een covariantie matrix van tijdseries die de relatieve kracht van oscillaties in verschillende frequenties laten zien, ruwweg universeel zijn tussen een elftal muizen. De resulterende clusters in een zesdimensionale faseruimte van twee dominante hoofdcomponenten per deel van het brein worden gerelateerd aan het gedrag van de muizen tijdens het experiment. Hoofdstuk 3 reduceert de macroscopische dynamiek niet op basis van statistische methoden (zoals in hoofdstuk 1 en 2), maar vanuit de fysica: middels laag-dimensionale wiskundige en natuurkundige modellen wordt het concept van kettingreacties tussen kantelpunten voorgesteld. Hierin wordt onderscheid gemaakt tussen een viertal mogelijke kettingreacties op basis van fold en Hopf bifurcaties. Met een eenvoudig natuurkundig model gebaseerd op een Stommel box model voor de Atlantische oceaan en het Timmermann-model voor de El-Niño Southern Oscillation illustreren we de potentie voor dit soort macroscopische fenomenen.

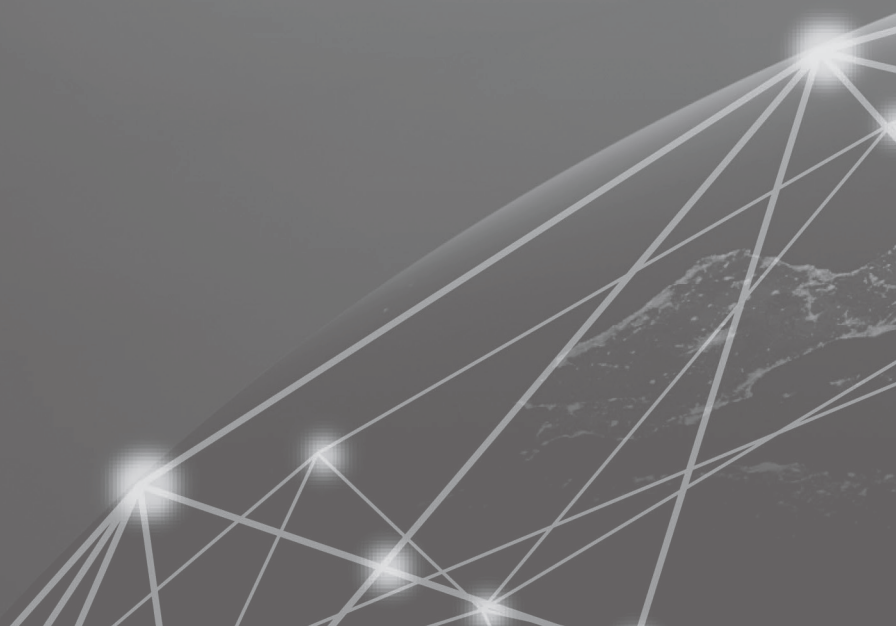
Het tweede deel stapt af van de puur macroscopische blik in deel I, en neemt de invloed van de precieze structuur van de interacties tussen de systeemcomponenten mee in de vorm van *netwerken*. Terwijl onderliggende netwerken constant blijven (of een grote tijdschaal hebben), kunnen dynamische fenomenen die gebonden zijn aan deze interacties wel sterk worden beïnvloed door de precieze netwerkstructuren. De drie hoofdstukken in deel II beschouwen hoe analyse van de netwerkstructuur gebruikt kan worden om de macroscopische dynamica beter te begrijpen. Alle drie de hoofdstukken gebruiken het voorbeeld van vertragsingsproblematiek in treinsystemen, elk vanuit een ander perspectief. In hoofdstuk 4 definieer ik clusters (regio's) in enkele Europese treinnetwerken door middel van een spectrale clusteringmethode. Deze clusters worden vervolgens geïnterpreteerd in termen van hun rol in de totale vertraging en hoe sterk hun connectie is met de andere clusters, wat gebruikt kan worden om de regio's in treinnetwerken geografisch te kunnen karakteriseren. Hoofdstuk 5 beschouwt het Belgische treinsysteem, en bevat een modelstudie met een diffusie-achtig model dat volledig afhankelijk is van de netwerkstructuur. De hoofdvraag in dit hoofdstuk is in hoeverre we vanuit een dergelijk netwerk-georiënteerd perspectief de vertraging kunnen voorspellen. Daarbij variëren we de netwerkresolutie om de nauwkeurigheid te bepalen. Met name bij grotere vertragingen, waarbij de vertragsingspropagatie minder ruisgevoelig is en ook in meerdere richtingen ontwikkelt (i.e., meer isotroop is), neemt de nauwkeurigheid van voorspellingen toe. In hoofdstuk 6 beschouwen we een specifiek mechanisme wat zich voor kan doen in zeer verstoorde treinsituaties in het Nederlandse treinsysteem: vertragsingspropagatie in de vorm van kettingreacties doordat vertraagd personeel en materieel hun overstappen en herkoppelingen niet meer halen. Door data van tijdschema's van drie logistieke lagen (treinlijnen, personeel en materieel) te verbinden, simuleren we deze situaties. Middels een vergelijking van modelresultaten met observaties, deduceren we waar dergelijke kettingreacties hebben plaatsgevonden, en ook waar verkeersleiders mogelijke vertraging succesvol hebben gemitigeerd.

In het derde en laatste deel van dit proefschrift wordt er gekeken naar systemen waarin het netwerk ook verandert. Dit is in het bijzonder van belang in de interactie van mensen: infectieziekten en informatie verspreiden zich via korte connecties die constant veranderen en herstructureren. In deel III is in het bijzonder aandacht besteed aan de rol van volgorden van interacties. In de epidemiologie is het uiteraard bepalend of een con-

tact plaatsvindt nadat je in contact bent gekomen met een ziek persoon, of ervoor. De hoofdvraag in dit derde deel van het proefschrift is hoe de structuurverandering en de bijbehorende kettingen van opeenvolgende interacties meegenomen kunnen worden om de kwetsbaarheid van het systeem te bepalen tegenover verspreidingsfenomenen. Hoofdstuk 7 beantwoordt dit door het definiëren van *entropy of temporal entanglement* dat, puur gezien vanuit de netwerkstructuur, weergeeft hoe sterk systeemcomponenten en hun interacties met elkaar zijn verweven. Als ze minder verweven zijn (i.e., lage waarden van de entropie), zal het systeem doorgaans minder kwetsbaar zijn voor verspreidingsfenomenen als ziektes, en *vice versa*. Hoofdstuk 8 is hieraan gerelateerd. In dit hoofdstuk gebruiken we empirische data van de interactie van mensen uit draagbare sensoren op een kunsttentoonstelling om te meten wie er de belangrijkste rol spelen in de verspreidingsdynamica. De resultaten, uitgedrukt in een nieuwe metriek *contact sequence centrality*, suggereren dat zogenaamde ‘*superspreaders*’ anders gedefinieerd kunnen worden dan puur degene die de meeste contacten heeft. In het laatste hoofdstuk van dit proefschrift modelleren we de eerder genoemde interactie-kettingen specifiek voor de casus van de COVID-19 verspreiding in Nederland in de eerste golf. We doen dit met een demografisch-gestratificeerd individueel-gebaseerd model dat de bewegingen van 170.000 modelagenten simuleert met behulp van mobiele telefoon data. Doordat het model demografisch en geografisch expliciet is, kan het verschillende (geïmplementeerde en hypothetische) interventies testen, waarin we kwantificeren hoe een lokale (in plaats van nationale) strategie op gemeentelijk niveau potentiële voordelen zou hebben opgeleverd. Op een methodologisch vlak laat de analyse zien wat er voor nodig is om de kennis uit hoofdstukken 7 en 8 tot uiting te brengen op een nationale schaal.

Samenvattend biedt dit proefschrift enerzijds nieuwe interdisciplinaire methodiek om collectieve verandering in een tal van systemen te analyseren, en anderzijds geeft het inzicht in systeem-specifieke vragen over deze dynamica. De algemene introductie (p. 4) bevat een achtergrond over de genoemde interactie tussen microscopie en macroscopie, en motiveert dit onderzoek op een interdisciplinair vlak. In de gelaagdheid van drie delen op basis van de typen netwerkstructuren werkt het proefschrift toe naar de vraag hoe microscopische interacties kunnen leiden tot collectieve systeemverandering, in een variëteit aan toepassingen. Aan het einde van dit proefschrift beschrijf ik enkele lessen en toekomstrichtingen voor dit type onderzoek in een algemene discussie (p. 194).

GENERAL INTRODUCTION





In 1910, the American entomologist William Morton Wheeler published his book *Ants, Their Structure, Development and Behavior*, where he proposed the novel concept of ant colonies operating as ‘superorganisms’. He observed that the collective of many ants demonstrates unexpected behaviour that is not easily explained from what is known about individual ants (Wheeler, 1910). As biologist Nigel Franks puts it, almost a century later:

“The solitary army ant is behaviorally one of the least sophisticated animals imaginable (...) If 100 army ants are placed on a flat surface, they will walk around and around in never decreasing circles until they die of exhaustion.”
(Franks, 1989)

But, as Franks also observed, if you increase the amount of ants up to half a million, they will form a ‘superorganism’ with a sophisticated ‘collective intelligence’ (Mitchell, 2009), with the abilities to form living bridges and night shelters, to construct nests and to collaborate on transporting food over large distances. Ants belong to the so-called *eusocial* type of insects (Batra, 1968): by means of a set of simple behavioural interactions expressed in chemical signals, they achieve a high biological level of social organisation, vital to their survival. Another example of extraordinary collective intelligence is the immune system in the human body: the complex interactions of B cells, T cells, and macrophages, among many other components, does not only lead to recognition and direct response to invaders, but also has a remarkable learning capacity to induce longer-term immunity — exemplified in the recent discovery of immunological memory to SARS-CoV-2 (Dan et al., 2021; Netea and Li, 2021). Describing and understanding the human adaptive immune response requires analysis beyond individual cells and involves tracking interactions of large quantities of components.

The paradox of how unintelligent beings such as ants and cells have complex collective behaviour when put together remains an active research topic today. In fact, the apparent mismatch between the functioning of individual components and the phenomenology of the whole is present in many real-world systems, far beyond entomology and immunology. Both the ant colony and the immune system are examples of *complex systems*, which are loosely defined as systems with many interacting components that, without any central control, show distinct behaviour on the macroscopic level. The latter is referred to as *emergence*. Complex systems and emergence can be found in a wide variety of real-world systems — from disruptions in transportation systems to the evolution of global pandemics such as COVID-19 and from the development of hurricanes to the polarisation on social media. This dissertation explores new avenues of studying dynamical phenomena in complex systems by proposing cross-disciplinary methods linking microscopic interactions of the system’s components to emerging macroscopic behaviour of the collective. At the same time, each individual chapter involves specific applications of these methods to describe example complex systems, also bringing contributions in their respective fields. The chapters are classified into three parts. Below, a motivation for the need of methodology at the macroscopic scale is provided from a philosophical and historical viewpoint, followed by an introduction of complex networks. This introduction ends with listing the contents of this dissertation.

Motivation for complex systems methodology

The examples of entomology and immunology reflect that macroscopic behaviour is not always easily deducible from interactions of components at the microscopic level. Approaches to filling this gap varied throughout history and reflect how collective or macroscopic behaviour is perceived in a scientific-philosophical sense. It is valuable to be aware of this history because it motivates the principles on which many of the approaches in this dissertation are built.

Perspectives on emergence

The study of complex systems is sometimes referred to as ‘complex systems science’, which is a hypernym referring to a framework and toolbox of methodologies that can be applied to many fields of science. Historically, one may not regard complex systems science as a separate scientific field in itself, but mainly as a change in perspective as a response to traditional approaches that lacked the ability to describe emergent behaviour. In fact, the paradigm change in myrmecology (i.e., the study of ants) brought by Wheeler is exemplary to illustrate different scientific views of analysing emergent behaviour. Before Wheeler, ants were studied following a *reductionist* approach (Nagel, 1998), which assumes that all system information can eventually be reduced to a small set of fundamental laws exerted on the level of individual components. *Constructionism*, arguably regarded as a corollary of reductionism, says that, using these fundamental laws, one can reconstruct and predict the system as a whole, and, subsequently, fully deduce the macroscopic emergent behaviour: e.g., a full understanding of the ant colony is obtained from understanding individual ants themselves. In contrast, *holism* says that systems should be treated as wholes to be able study their behaviour properly (Smuts, 1926): the ant colony contains new phenomenology and associated laws and concepts. In popular words, reductionism argues that the whole merely consists of the ‘sum of its parts’, while holism argues that the whole is ‘*greater* than the sum of its parts’. Whether this phrase is entirely true is up for debate.

Setting aside the semantic discussion on what exactly the ‘whole’ comprises and what the ‘sum’, an important discussion lies in the phenomenological domain: how does the behaviour of the system as a whole relate to that of individual components? In fact, Aristotle already came to the conclusion that when ‘defining’ an object — which is something he struggled with in his famous work *Metaphysics* — purely listing the object’s components is not the same as the whole itself (Aristotle, 350 BC). Centuries later, in 1926, the South African statesman and philosopher Jan Smuts, seen as one of the founders of holism, described the concept of ‘wholes’ as (meta-)physical structures surpassing their components (Smuts, 1926). Their holistic character has varying degrees, and Smuts sketches six grades of wholes, from material structures, via animals and personalities, to states and absolute values like truth, beauty and goodness. Since Smuts, many modern variations of holism have been proposed.

Involving the ideas of Aristotle and Smuts in the discussion of reductionism and holism reveals an almost philosophical aspect on the connection between the components and the whole, which has implications for the analysis of many systems, among which those discussed in this dissertation. For example: can we ultimately reduce the emergence of

human consciousness in our brain to fundamental chemistry and physics? If so, human behaviour as a phenomenology on its own (e.g., as studied in chapters 7, 8 and 9 of this dissertation) may be redundant and, instead, regarded as applied chemistry, should be expressed in fundamental laws on a more microscopic level. These different types of research are sometimes referred to as ‘intensive’ and ‘extensive’ research, respectively (Anderson, 1972). In contrast, by emphasising emergence — qualitatively different phenomenology as a consequence of differences in scale — complex system science aims to identify new laws and concepts that emerge only at larger scales. This dissertation follows this line, and studies how macroscopic dynamical phenomena may result from microscopic interactions.

There are two main perspectives on the existence and origin of emergence. The first is the idea that systems at larger scales work physically different from what can be reasoned from their mere collection of components. For example, one may argue that an increased quantity of ants changes the collective behaviour such that it becomes *literally* impossible to predict it from the behaviour of individual or small groups of ants alone. Systems as wholes are then argued to require different laws because the ones applying to their components do not apply to the whole. This argument on the (meta)physical existence of such new laws is referred to as the *ontological* explanation for emergence (Ladyman et al., 2013), and argues directly against constructionism.

The second perspective is an *epistemological* explanation for emergence, indicating that, although the same set of fundamental laws may apply to all scales of a complex system, it becomes *practically* impossible to predict the system’s behaviour from individual or small groups of components alone (Ladyman et al., 2013). Such practical limitations may be due to the existence of non-linearity, scaling laws, memory, critical transitions and other features typically emerging at larger scales in complex systems. Such reasoning is largely present in the natural sciences, e.g., in how science deals with molecular dynamics in statistical physics or the ideal gas law and turbulence in hydrodynamics. There are no ‘new’ or ‘extra’ physical laws beyond those that are already exerted on the system’s individual components, but the system as a whole has different phenomenology, functions or behaviour that are too difficult to derive from individual component behaviour — these aspects only become apparent when considering the system at large.

Neither of the two perspectives argue against new phenomenology on macroscopic scales, meaning they both support the idea that complex systems require additional understanding and descriptions at that system-wide level — forming the motivation of this dissertation. In this dissertation, I limit complex systems to structures in the (non-meta) physical domain, which are in terms of composition by definition equal to the sum of their parts, but may require methods to analyse the macroscopic properties, with particular emphasis on dynamical phenomena emerging at these scales. Hence, the title of this dissertation: *Macroscopic Dynamics in Complex Systems*.

The dominance of reductionism

Although contradictory to the aforementioned motivation of understanding emergence from a holistic viewpoint, the reductionist viewpoint has been dominant throughout a large part of scientific history. In fact, from the 16th century until the early 20th century, the ultimate goal of science was to produce a reductionist explanation of everything in

terms of a finite set of fundamental laws of physics. This viewpoint had implications on how the different fields of science were perceived: chemistry could be viewed as an applied form of these physical laws, and, in turn, biology or even psychology as applied chemistry. In fact, this hierarchical order of sciences was quite the debate in the 19th and 20th century and is referred to as Comte's Theory of Science, after the French philosopher Auguste Comte who reasoned about the impact this hierarchy would have on consensus and advancement rates in science. In line with the reductionist's approach, René Descartes, one of the founders of modern science, described his own scientific methodology as follows:

"... to divide all the difficulties under examination into as many parts as possible, and as many as were required to solve them in the best way" and "...to conduct my thoughts in a given order, beginning with the simplest and most easily understood objects, and gradually ascending, as it were step by step, to the knowledge of the most complex" (Descartes, 1637)

This is a natural way to start research: to divide larger complex problems into smaller pieces to study one by one. It also works in most fields of science, as many scientific concepts contain reductionist elements. An example is the superposition principle in classical mechanics, that states that for linear systems, the net response of two stimuli can be deduced by adding the two stimuli individually: e.g., one can determine a net force on a body of mass by determining and superposing individual forces exerted on the body.

Up to the 19th century, the dominance of the reductionist approach and hypothesised reducibility of scientific principles to a small set of fundamental laws led, via the concept of constructionism, to the idea of *causal determinism*. This idea was embodied by the thought experiment called "Laplace's demon", named after the French scholar Pierre-Simon de Laplace (Laplace, 1814): if an intelligent being (the 'demon') would know the precise momenta and locations of all items in nature at a specified moment, and if this being would have infinite computational power, it would be able to predict the past and future of everything. This idea was debunked in the 19th century by the second law of thermodynamics: entropy in closed systems never decreases and processes are only reversible when entropy remains equal. In other words, most natural processes are irreversible and time is asymmetric in those cases. Hence, the demon could not predict the past from the present. Various different types of reasoning are used to also reject future deterministic predictions from the demon, reasoned from, e.g., both quantum mechanical to information theoretical grounds. (A practical problem for Laplace's demon is found in the concept of chaos, which I address below.)

Opposing constructionism

Throughout the 20th century, the idea of a constructionist 'theory of everything' received more criticism, not only in the philosophical argument, but also through field-specific advances, such as that of the aforementioned 'superorganism' ant colony. A systematic problem of constructionism was proposed in 1972 by the American physicist Philip W. Anderson in his famous essay *More is different*, who acknowledged the reductionist idea

that emphasised the importance of understanding fundamental laws underlying every system, but challenged the constructionist hypothesis at the same time:

“The ability to reduce everything to simple fundamental laws does not imply the ability to start from those laws and reconstruct the universe.” (Anderson, 1972)

He uses the theory of *symmetry breaking* in statistical physics to illustrate how quantitative changes can lead to qualitative differentiation¹ — much like when putting half a million ants together, a superorganism will emerge, behaving qualitatively different from a group of merely one hundred ants. Symmetry breaking refers to the idea that infinitesimally small fluctuations can lead to the crossing of a critical point, determining which branch of a bifurcation is taken. This resulting state is asymmetric: going back to the previous (disorderly, undetermined) symmetric state and crossing the bifurcation again might not yield the same results. He reasoned that *“the internal structure of a piece of matter need not be symmetrical even if the total state of it is”* and *“the state of a really big system does not at all have to have the symmetry of the laws which govern it; in fact, it usually has less symmetry.”*, implying a disconnection between the system-wide symmetry and laws and those of the system’s components. Anderson believed that scale and complexity of systems yield entirely new properties, suggesting that the research of which may be as fundamental as any other.

A relevant development in this context is the discovery of *chaos* in deterministic dynamical systems, which refers to apparent random behaviour resulting from the interaction of elements. It is commonly illustrated by the butterfly effect, a phenomenon that is consequential to chaos: small changes in initial conditions result in radically different outcomes when trying to predict the state over a period of time. Or, as a quote that is attributed to the American mathematician and meteorologist Edward Lorenz, describes chaos:

“When the present determines the future, but the approximate present does not approximately determine the future.” (Lorenz)

Lorenz was a pioneer in chaos theory. He identified chaotic behaviour when searching for ways to forecast the weather, resulting in his famous paper *Deterministic nonperiodic flow* in the mid-20th century (Lorenz, 1963). Still, chaotic behaviour was actually found earlier already by Poincaré, at the end of the 19th century when studying the Three Body Problem. The concept of chaos leads to a practical problem for the aforementioned Laplace’s demon: because we can never know the precise (i.e., up to infinitely many decimals) location and momentum of every particle in the universe, we can never accurately predict the past and future. Some scholars also argue that this is a *principal* argument against the potential of Laplace’s demon (Mitchell, 2009).

¹He ends the paper with a dialogue that illustrates this principle. It concerns the American writers F Scott Fitzgerald and Ernest Hemingway in the 1920s in Paris. Fitzgerald: *“The rich are different from us”*. Hemingway: *“Yes, they have more money”*.

Complex systems science

Contemporary to the growing criticism on constructionism and determinism in the 20th century, the need for interdisciplinary approaches to both scientific and societal problems was recognised. American physical chemist George Cowan brought together a number of brilliant minds from a variety of fields, many of which were acquainted from the Manhattan project during World War II. They founded the Santa Fe Institute in 1984, aimed at disseminating the concept of a new research area, referred to as ‘complexity theory’ or ‘complex systems science’, aimed at using the potential of natural sciences to address interdisciplinary problems.

Despite the growing success of complex systems science, today, there is still no consensus on a sharp definition of a *complex system*. Part of this is because of the many different elements that may or may not make a system complex (Ladyman et al., 2013). Several aspects are already mentioned: emergence, lack of control and a large number of interactive components. Other aspects include non-linearity, self-organisation and even memory. Chaos is commonly confused with complexity. However, chaotic behaviour can often be directly deduced from precise (even simple) mathematical principles, which can result in a form of complicatedness (Ricklefs et al., 2007). The link between micro principles and macro properties can be sharply formulated for many chaotic systems, which is one of the problems in complex systems. In this dissertation, I use the following qualitative definition of a complex system by Mitchell (2009), which emphasises the generality of the concept, and how macroscopic behaviour may arise without central control:

“A system in which large networks of components with no central control and simple rules of operation give rise to complex collective behavior, sophisticated information processing, and adaptation via learning or evolution.”

Reading this definition, one realises that many systems — both in natural sciences as well as in social sciences — can be regarded as complex systems, e.g., animal brains (see chapter 2), systems involving human interaction (see chapter 8), logistic systems (see chapter 6) and parts of our climate system (see chapter 3). In a sense, it does not matter whether a system is fully ‘complex’ or not. What matters is whether one observes the paradoxical disconnection between the macro-behaviour and the fundamental laws governing the system’s individual components — i.e., the (epistemological) breaking of constructionism. If this is the case, applying complex systems methods can be a valuable asset, especially when realising that many mathematical and phenomenological principles found in complex systems turn out to be cross-applicable between disciplines.

This cross-disciplinarity is exemplified in the detection of network structures in a wide variety of these systems, which is an important aspect of this dissertation and greatly influenced complex systems methods in the past few decades. A *network* is a mathematical structure consisting of *nodes*, being the components or agents in the complex system, and *links*, reflecting the interactions between nodes. (A more elaborate introduction to networks is given in Part II.) Finding its basis in graph theory, the history of networks can be traced back to the mid-18th century, when the great Swiss mathematician Leonhard Euler proved the first theorem of graph theory: the solution to the Königsberg bridge problem. There were seven bridges in the Prussian city of Königsberg (now Kaliningrad in Russia) crossing the Pregel River, some leading to islands inside the river. The problem

was to find a route that crossed all of them once and only once. Euler abstracted the islands and both sides of the main land to ‘nodes’ as the routes within each land mass was irrelevant. Linking these nodes resulted in a mathematical graph, and using this structure he proved that no such route was possible. Although this initial problem did not involve any complex multi-component interaction, the mathematical field of graph theory it established formed the basis to analyse exactly that.

Until the late-20th century, the study of networks was mainly confined to the mathematical field of graph theory. With the discovery of important network models such as random networks (Erdos et al., 1960), small-world networks (Watts and Strogatz, 1998) and scale-free networks (Barabási and Albert, 1999), ‘network science’ started to become more wide-spread. In particular because of the increasing computational power and data availability, it became increasingly possible to study large networks in the real world, such as social networks (Kossinets and Watts, 2006; Cattuto et al., 2010), mobility structures (Wesolowski et al., 2012), protein interaction (Rual et al., 2005), and many more. One notable finding was the discovery of the scale-free nature of real networks across many disciplines, which implies that there is no characteristic scale in the network: the distribution of the node degree — i.e., the amount of links the node has — follows a power law. Today, network science is becoming a field of science in itself, involving own journals, conferences and even educational programmes.

Dynamical phenomena and networks

Central to this dissertation is the relation between the structure and interactions of a complex system and how observed phenomena are dependent on this structure. This section introduces concepts relevant to this relation. Many phenomena in complex systems concern *dynamics*. In fact, many concepts in complex systems science involve dynamical systems theory. In the context of the aforementioned relation, a distinction can be made between two types of dynamics. The first is concerned with *dynamics of networks* (see Fig. ia), which, quite literally, refers to variations of the network itself: the changing, disappearing and reappearing of nodes and links. In statistical physics, such analyses have been well-established for quite some time already in the form percolation theory, for example. The analysis of dynamics of networks leads to the concepts of resilience, robustness and vulnerability of the connectivity in network structures.

The second branch of dynamical analysis of networked systems is concerned with *dynamics on networks* (see Fig. ib), which treats the network as pathways that allow any dynamics in the system to evolve upon. However, the exact relation between dynamics on networks, and the underlying network itself is not always trivial and depends on the application. In fact, this is one of the main topics in this dissertation. Two considerations can be mentioned to illustrate potential discrepancies between dynamics on networks, and the network itself. The first consideration is potential topological anisotropy of the dynamics that is not aligned with network weights. This is the case when looking at train delay dynamics in railways (e.g. in chapter 6): delays are advected by discrete trains travelling between stations. So when a delayed train enters a station (node), and it re-departs in a particular direction (edge), the delay moves along with it. In many cases, this means that potential other edges are unaffected by this particular delay, even though they are

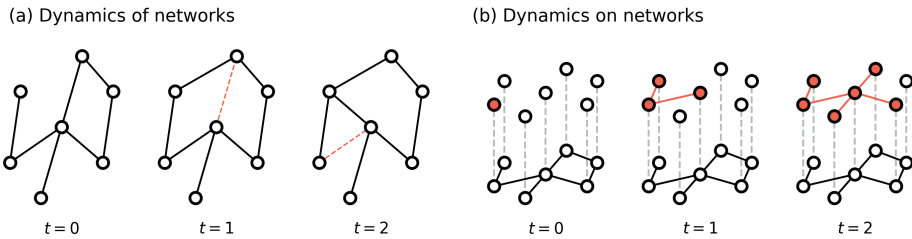


Figure 1: Illustration of (a) dynamics of networks, which involves changes in the structure of the network itself, such as the addition and removal of links or nodes, and (b) dynamics on networks, where the interaction structure of the nodes remains the same, but a dynamical process evolves on top.

connected in the network (from a diffusion-like perspective which is in fact discussed in chapter 5). Another consideration when linking dynamical phenomena to underlying network architecture is concerned with time scales and probabilistic aspects of the dynamics. Crucial connections in the topology might be probabilistically unused by the dynamics, and when the contacts in the network are much shorter than the dynamics evolves, many contacts might be meaningless, illustrating the importance of time scales. (I will come back to this in part III of this dissertation.)

Even though network architecture may not have a one-to-one relationship with dynamics on top, a crucial advantage of network approaches with respect to non-network approaches is that they explicitly account for heterogeneity in the agent connection structure on the level of individual agents, which, in many systems, is of high relevance to the dynamics. A network structure can reveal which agents and communities easily affect each other, and which do not. For illustration purposes, four distinct network structures are shown in Fig. ii: (a) an early version of the World-Wide-Web, as measured by Google in 2002 (Dasgupta et al., 2009), (b) the interaction of children in a primary school (an aggregation of two hours) (Gemmetto et al., 2014; Stehlé et al., 2011), (c) roads in California (Li et al., 2005), and (d) the citations network of high-energy physics papers on preprint server arXiv from 1992-2002 (Gehrke et al., 2003). Clearly, these networks are very different. The WWW graph (panel [a]) seems to have a layered, core-periphery structure, while the primary school (panel [b]) has a clustered structure (in fact, due to the division in classrooms). The road network in California (panel [c]) has a planar structure, because of its geographic nature, while papers in the citation network (panel [d]) seem to have almost random connections. Traditional network approaches investigate the distributions of node characteristics such as degree and centrality measures. More sophisticated is the question of clustering: can we divide these graphs up into well-separated communities? At first glance, this seems possible and useful in the primary school and the Californian road network, but less so for the WWW and the citation graph.

In this figure, the distinction between dynamics of networks (Fig. ia) and dynamics on networks (Fig. ib) can be made clear: an example of dynamics of networks is the vulnerability of the road network in panel (c): does the connectivity in the network remain in tact, even if multiple large accidents fully remove certain nodes or links from the network? Or, were particular papers crucial in the connectivity of the information flow that we see in high-energy physics papers in panel (d)? An example of dynamics on networks

is modelling the dynamics of cars and traffic congestion on the Californian roads in panel (c), which evolve on top of the network architecture (i.e., the roads), but keep it in tact.

More recently, partly due to developments in high-resolution and high-frequency sampled data in a wide variety of complex systems, a sort of ‘combined version’ of the two aforementioned network dynamics approaches has emerged: the analysis of dynamics on *temporal networks*. Temporal networks are networks of which the nodes and links change over time. Analysing the dynamics on top of temporal networks is usually more complicated and computationally demanding, but has major advantages — the obvious advantage being able to track sequences of interactions which may be crucial to the dynamics. When simulating the evolution of a disease or rumour spreading, for example, across the children in a primary school, the (non-temporal) *static* network shown in Fig. iib does not have all necessary information. Assuming no incubation time, if child *A* spreads the disease to child *B* at 9:00AM, and *B* is in contact with another child *C* at 11:00AM, the disease may spread to *C*. But if the contact between *B* and *C* is before 9:00AM, it will not, exemplifying the importance of sequences. Such temporal information on the ordering of links is not contained in a static network. Of course, sequential information is not always relevant to the dynamics: on social media, the spreading of a popular video on social media goes much faster than the changes in its pathways — i.e., the followers and friend networks — which may be assumed to be time invariant on the scale of days.

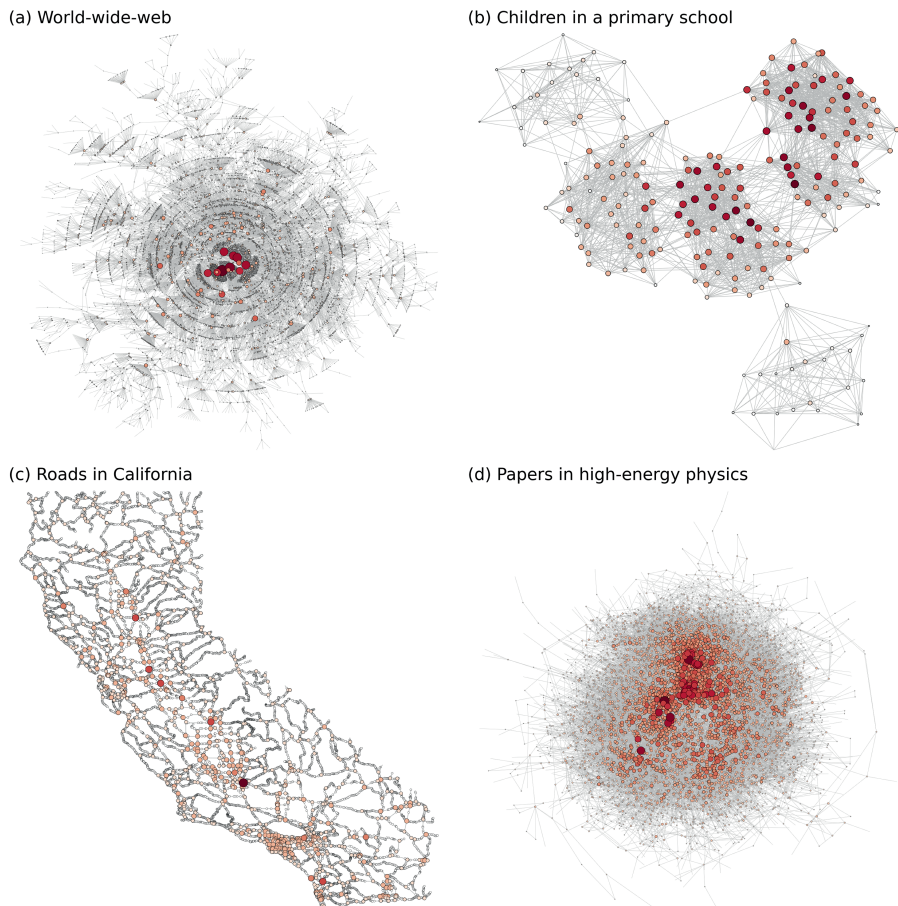


Figure ii: Examples of complex networks in real-world systems. **Panel (a):** the world-wide-web in 2002 (Dasgupta et al., 2009). **Panel (b):** children in a primary school (Gemmetto et al., 2014; Stehlé et al., 2011). **Panel (c):** California road network (Li et al., 2005). **Panel (d):** citations in high-energy physics (Gehrke et al., 2003). Node colour and size reflect the node degree. Only giant components are shown. In panels (a) and (d), due to computational limits, only 7% and 21% of the nodes are shown, respectively, which is for the purpose of this illustration not a relevant problem.

Dissertation content

The goal of this dissertation is to contribute to the complex systems methodology literature by proposing and applying new methods to analyse how macroscopic dynamics emerges from microscopic interactions. In each chapter, methods are applied to real-world problems and a large part of the contribution of this dissertation lies in explaining emergent behaviour in the respective example systems themselves, which involve a variety of systems: behavioural states in animal brains (chapter 2), disruptions in railway systems (chapters 1, 4, 5, 6), human behaviour (chapters 7, 8), climate subsystems being coupled via geophysical principles (chapter 3) and epidemic spreading of diseases (chapter 9). The nine chapters in this dissertation are subdivided into three parts, based on the nature of the methods used. The parts all contain their own (small) introduction, accompanied by an example, to guide the narrative in this dissertation. I conclude with a general discussion at the very end.

Part I: Macroscopic approaches [chapters 1-3]

The first part concerns three chapters with macroscopic approaches, referring to methodology based on system-wide variables rather than agent- or component-specific ones. In particular, these chapters concern the effort to define system-wide or macroscopic states by trying to reduce the dimension of the system to only a few variables — defining a so-called *low-dimensional* state variable — and analysing how these variables evolve with respect to each other. In chapter 1, we capture the evolution of railway delays in the Dutch railway system into only a few system-wide variables and construct an early warning scheme for the evolution of disruptive delay state. In chapter 2, a similar effort is presented in electroencephalography data from animal brains, aiming to specify patterns in the brain belonging to behavioural states. In chapter 3, we analyse macroscopic states in climate subsystems, and research critical behaviour and interactions between them. How such low-dimensional state variables should be defined, and whether they show the relevant dynamical behaviour such as states and transitions, is not a trivial task. This leads us to the topic of part I:

Research question Part I: (a) *How can we identify macroscopic states in high-dimensional complex systems?* (b) *Given macroscopic states, how can we describe transitions between them?*

Part II: Static network approaches [chapters 4-6]

In contrast to part I, where the aim was to search for (low-dimensional) system-wide variables, in part II the system's component structure is explicitly accounted for. The methods in this part focus on the role of the system's underlying network architecture in determining the evolution of dynamical phenomena. The main question addressed in these chapters is:

Research question Part II: *How can we investigate the interdependence of a system's dynamics and its underlying static network architecture?*

Intentionally, I use the example of railway dynamics in all three chapters in part II. Here, the role of the network architecture is far from trivial due to the discrete nature of evolution of delays and congestions. However, the three chapters involve three different angles to this research question. A geographic characterisation of the dynamics is given in chapter 4: can we use a network topology clustering to obtain insights in the dynamics on top? Chapter 5 focuses on prediction by using a network-based model of the dynamics, quite explicitly linking the two. It also investigates how the spatial resolution affects the predictability of the dynamics from this model. Chapter 6 focuses on mechanisms driving a rare but remarkable dynamical phenomenon in railways — that of large-scale disruptions — and aims to use that information to predict which areas of the network are at risk during such disruptions. In particular, we assess how initially local delays are able to amplify and spread towards all corners of the country-wide railway system, exemplifying the fundamental question of how microscopic interactions can lead to macroscopic behaviour of the system.

Part III: Temporal network approaches [chapters 7-9]

The final part in this dissertation concerns the architecture of *temporal* networks and its relation to dynamical phenomena. By taking into account change of the system's network over time, we can explicitly address the role of sequences in affecting the dynamics. All three chapters in this part are focused on spreading phenomena on temporal networks. In particular, the following question is addressed:

Research question Part III: *How can we quantify and influence spreading vulnerability in temporal networks?*

The first two chapters (7 and 8) in part III address this from a theoretical point of view by proposing two metrics, *entropy of temporal entanglement* and *contact sequence centrality*, which reflect the system-wide spreading vulnerability and the impact of individual agents on this, respectively. In these chapters, a variety of empirical datasets is used to illustrate the metrics, with most attention to social and human interaction networks. Chapter 9 concludes this part with a model in which the first wave of COVID-19 in the Netherlands is simulated using an advanced model involving datasets on mobility, mixing and demography, with the ultimate aim of testing the effectiveness of various governmental interventions. Also this chapter relates to the concept of spreading vulnerability by proposing techniques to limit spreading using both demography- and geography-affecting measures.

These three parts together form a narrative of a step-wise growing level of heterogeneity accounted for: while in part I, the aim was to fuse as much information as possible into a low-dimensional macroscopic view, parts II and III explicitly add information on the system's components and their (temporal) interactions. Even though the three parts contain different methodologies, they all focus on the general question of how macroscopic dynamical phenomena arise from microscopic interactions.

PART I

MACROSCOPIC APPROACHES

On the identification of macroscopic states and transitions between them



"Science has explored the microcosmos and the macrocosmos; we have a good sense of the lay of the land. The great unexplored frontier is complexity."

HEINZ R. PAGELS, 1989



Introduction to Part I

In daily life, we often express dynamical phenomena in terms of a small set of system-wide variables. For example, country economic performance is expressed in Gross Domestic Product (GDP), epidemic spreading capability using the reproduction number (R_0), and a person's physical health using, e.g., blood pressure and heart rate, all of which are in fact emerging from complex interactions. From a methodological point-of-view, when analysing phenomena at the system-wide, or *macroscopic*, scale, using a highly reduced number of variables can have important advantages in terms of understanding, applicability of analytical approaches and computational feasibility. The reduction of high-dimensional complex interactions to a small set of system-wide variables and the subsequent analysis of these variables is referred to as a *macroscopic approach* to the phenomenology of the system.

In generalised terms, macroscopic approaches focus on the description of the dynamics of a system consisting of N interactive components, referred to as *agents* denoted by a_i , where $i \in \{1, \dots, N\}$. While many variables may define the exact interactions and behaviour of the agents, the macroscopic approach involves defining a low-dimensional state variable $\mathcal{S}(t) \in \mathbb{R}^n$ that evolves over time t :

$$\mathcal{S}(t) = \{x_0(t), x_1(t), \dots, x_n(t)\}, \quad (1)$$

where $x_i(t)$ are continuous 1D time series for all $i \in \{1, \dots, n\}$ and, typically, $n \ll N$. The most important distinction between macroscopic approaches (Part I) and network approaches (Part II and III) lies in the absence of explicitly accounting for heterogeneity among the agents in terms of their interactions: the variables x_i are continuous time series, and do not explicitly account for the discrete network embedding of individual agents. The challenge of macroscopic approaches lies in finding such relevant time series x_i , especially when dealing with high-dimensional systems. The subsequent analysis of the state variable varies per system at hand, but an important branch of complex systems methods deal with the definition of macroscopic *states* of the system: configurations of the state variable that have a dynamical and phenomenological meaning. The stability of and transitions between such states are important aspects of emergence in complex systems, dealing with concepts such as system-level bifurcations and tipping points. In fact, the identification of macroscopic states and the transitions between them is the domain within this broad branch of methods where the contribution of this part of the dissertation lies, in the form of the chapters 1, 2 and 3.

To illustrate a generic macroscopic approach and how associated states are derived, let me present a typical example of a macroscopic approach to emerging behaviour – an example from epidemiology. A macroscopic description of the evolution of the epidemic is to describe the population-wide evolution of how many people are infected based on how many are already infected, and how many are not (yet) infected. One of the first macroscopic approaches to epidemiology is by Kermack et al. (1927), arguably even earlier by the British Ross and Hudson (1917) and made more explicit in Kendall (1956). These papers explored the concept of compartmentalised models, where the population is divided into ‘compartments’ in which each agent can reside and transfer from to

another, usually associated with the various stadia in an infectious disease. The interactions and nature of these agents are fully homogeneous. Numerous versions of compartmentalised models exist in epidemiology, for example by distinguishing more compartments, stratifying the population or even distinguishing topological heterogeneity among agents – which is discussed in Parts II and III. One of the simplest versions of compartmentalised models is the SIR model. The characters S, I and R refer to the possible disease stages an agent can have: *Susceptible* (S), *Infectious* (I) and *Recovered* (R). Mathematically, the symbols $S(t)$, $I(t)$ and $R(t)$ are used for the fractions of the population that belong to these respective disease stages, such that $S(t) + I(t) + R(t) = 1$. In other words, this system is normalised with respect to the number of agents N . In line of Eq. (1), this model approaches the complex problem of epidemic spreading by reducing its description to a mere three-dimensional state variable:

$$\mathcal{S}(t) = \{S(t), I(t), R(t)\}.$$

In this SIR model, the evolution of this state variable can be quantified using a set of three differential equations, with two predefined parameters: β , which is the infection rate of the disease (factoring the contact rate and the probability of contacts resulting in infection), and γ , which is the recovery rate (mortality is not explicitly included in this model). The differential equations are as follows:

$$\begin{cases} \frac{dS}{dt} = -\beta IS \\ \frac{dI}{dt} = \beta IS - \gamma I \\ \frac{dR}{dt} = \gamma I. \end{cases}$$

Analytical analysis or numerical simulation of a model like this already provides insights in the macroscopic evolution in general. Using an Euler-forward numerical integration scheme with $\Delta t = 0.1$, an initial condition of $(S, I, R) = (0.99, 0.01, 0)$, $\beta = 0.1$ and $\gamma = 0.02$, an evolution is obtained like shown in Fig. iiii. As can be seen, the model ends up in a stationary state where R is close to 1, and I and S are approximately 0 – this is an example of the aforementioned macroscopic states: for these parameter values of β and γ , the initial conditions of high- S and nonzero- I are unstable and drawn to stable state with low- I , low- S and high- R . For other parameter values, this is different. In particular, they are commonly combined into to the reproduction number $R_0 = \beta/\gamma$, reflecting the average number of new infections a single infected agent causes at the start of the epidemic (a meaning that is later less true due to the effect of a decreased S , also known as ‘herd immunity’). In panel (b), one can see how, given the same initial conditions as in (a), R_0 modulates whether the epidemic will directly diminish, or evolve. This is expressed in $I_{\max} - I_0$, i.e. the peak value in I (red in panel (a)) minus its starting value: it is near-zero for values $R_0 < 1$ and grows when $R_0 > 1$. In other words, here, two main outcome states of the epidemic are now identified, but in fact, as seen in panel (c), there are many in-between outcomes in the configuration of S , I and R possible, based on the parameter values.

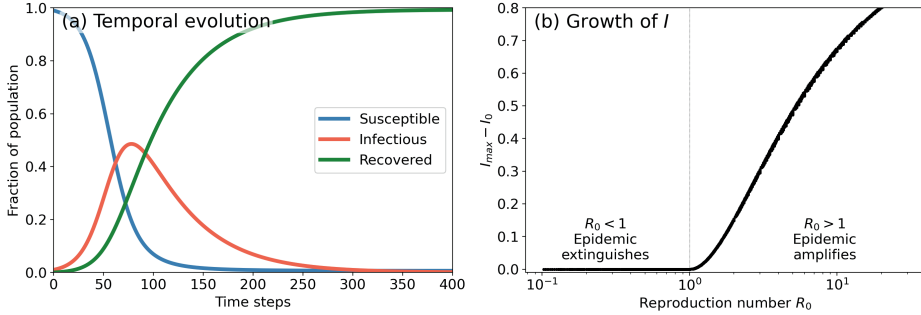


Figure iii: Illustration of the SIR model. **Panel (a):** Evolution of $S(t)$ (blue), $I(t)$ (red) and $R(t)$ (green) over time, with $\beta = 0.1$ and $\gamma = 0.02$. **Panel (b):** Peak height I_{\max} minus initial cases I_0 , which reflects the maximum growth in $I(t)$, with $R_0 = \beta/\gamma$ on the horizontal ($R_0 = 1$ is marked by the vertical dashed line).

While within the bounds of the assumptions, describing epidemic spreading in a homogeneous population can be done using the variables S , I and R and their interactions, the identification of these variables that form \mathcal{S} is not trivial in general. In fact, many complex systems comprise of many heterogeneous interactions that cannot easily be collapsed into a few variables. Identifying \mathcal{S} in a high-dimensional complex system, analysing patterns in the evolution of \mathcal{S} to define macroscopic states, and describing transitions across such states, is the focus on the following chapters 1, 2 and 3. Note that throughout these chapters, I use *transitions* in the broader sense as the ‘alteration of a state or condition to another’, rather than the more formal definition in physics relating to the transfer between two low-energy states.

In particular, in the following chapters, methods are proposed to identify macroscopic states and dynamics across them in three different real-world applications and in each case lessons are drawn from these approaches for complex systems in general. In chapter 1, the evolution of railway delays is analysed, using the example of the Netherlands. Involving numerous trains per day on a fixed and highly utilised railway network, this system is prone to the development of nation-wide disruptions, acting like macroscopic pseudo-states. The description and anticipation of such disrupted states is the aim of this chapter. The railways are a typical example of high heterogeneity and an associated high dimensionality of relevant components. The Dutch railway infrastructure contains approximately 800 nodes and 1400 railway links between them, and on top about 6000 unique train lines travel each day. We compress all delay information into two system-wide variables using a principal component analysis (PCA) approach. The resulting low-dimensional phase-space is further analysed to identify various stages of the system with respect to how severe the delays are and how close it is to disruption. In other words, \mathcal{S} comprises of the two aforementioned national principal components P_1 and P_2 :

$$\mathcal{S}_{\text{Ch. 1 - Dutch railways}}(t) = \{P_1(t), P_2(t)\}.$$

In chapter 2, we deal with the dynamics of the brain (of mice), infamously known

for its complexity and the difficulties to find structure in the data. The aim is to identify states of the brain that have a behavioural meaning. Also here, we approximate the system's state by reducing data into a smaller set of macroscopic variables, and find that the eigenvectors representing the state variable are unexpectedly similar across different mice. The state variable \mathcal{S} in this chapter reduces to:

$$\mathcal{S}_{\text{Ch. 2 - Mice brains}}(t) = \left\{ P_{1,\text{HIP}}(t), P_{2,\text{HIP}}(t), P_{1,\text{PFC}}(t), P_{2,\text{PFC}}(t), P_{1,\text{PAR}}(t), P_{2,\text{PAR}}(t) \right\},$$

where P_i for all i are also principal components, reflecting dominant covariant patterns in the residual power spectra of EEG signals. The abbreviations HIP, PFC and PAR refer to the brain regions from which the principal components are derived: hippocampus, prefrontal cortex, and parietal cortex, respectively. Analysing the reduced phase-space spun by the resulting state variable results in a set of clusters that we analyse in terms of their behavioural meaning. It turns out that from this procedure, we can identify parts of the EEG time series in which the mice are behaviourally different that in other parts of the time series.

In chapter 3, we devise a procedure of identifying stability and instability between macroscopic states when two systems are coupled together, with the application of systems in the Earth's climate. We start from a mathematical point-of-view for an analytical basis for such coupled instabilities, discussing various sorts of coupled bifurcations in low-dimensional sets of differential equations. We proceed by constructing a low-dimensional model that couples dynamics of an important ocean current in the North-Atlantic (the meridional overturning circulation, or MOC) to El-Niño-Southern Oscillation and identify stable and unstable states, both stationary and oscillatory. In this chapter, using characters T and S for temperature and salinity, respectively:

$$\mathcal{S}_{\text{Ch. 3 - Coupled MOC-ENSO model}}(t) = \left\{ \Delta T_{\text{Atlantic}}, \Delta S_{\text{Atlantic}}, T_{1,\text{Pacific}}, T_{2,\text{Pacific}} \right\}.$$


Summarised, the chapters 1, 2 and 3 are not meant to be a comprehensive overview of macroscopic complex system modelling, but rather focus on the following questions in this branch of research:

Research question Part I: (a) *How can we identify macroscopic states in high-dimensional complex systems?* (b) *Given macroscopic states, how can we describe transitions between them?*

CHAPTER 1

PREDICTING TRANSITIONS ACROSS MACROSCOPIC STATES FOR RAILWAY SYSTEMS

This chapter is published as: Dekker, M. M., Panja, D., Dijkstra, H. A., and Dekker, S. C. (2019). Predicting transitions across macroscopic states for railway systems. *PLOS ONE*, 14(6):e0217710. Supplementary material to this chapter can be found in SI A.



ABSTRACT

Railways are classic instances of complex socio-technical systems, whose defining characteristic is that they exist and function by integrating (continuous-time) interactions among technical components and human elements. Typically, unlike physical systems, there are no governing laws for describing their dynamics. Based purely on micro-unit data, here we present a data-driven framework to analyse macro-dynamics in such systems, leading us to the identification of specific states and prediction of transitions across them. It consists of three steps, which we elucidate using data from the Dutch railways. First, we form a dimensionally reduced phase-space by extracting a few relevant components, wherein relevance is proxied by dominance in terms of explained variance, as well as by persistence in time. Secondly, we apply a clustering algorithm to the reduced phasespace, resulting in the revelation of states of the system. Specifically, we identify 'rest' and 'disrupted' states, for which the system operations deviates respectively little and strongly from the planned timetable. Third, we define an early-warning metric based on the probability of transitions across states, predict whether the system is likely to transit from one state to another within a given time-frame and evaluate the performance of this metric using the Peirce skill score. Interestingly, using case studies, we demonstrate that the framework is able to predict large-scale disruptions up to 90 minutes beforehand with significant skill, demonstrating, for the railway companies, its potential to better track the evolution of large-scale disruptions in their networks. We discuss that the applicability of the three-step framework stretches to other systems as well — i.e., not only socio-technical ones — wherein real-time monitoring can help to prevent macro-scale state transitions, albeit the methods chosen to execute each step may depend on specific system-details.



1.1 Introduction

Railways are classic examples of complex socio-technical (ST) systems. Their defining characteristic is that they integrate (continuous-time) interactions among technical components and human elements/influence in their existence and functionality. In this chapter, we think of ST systems as dynamical systems, although, typically there are no laws that govern their time evolution. There exists a substantial amount of literature to model the behaviour of ST systems; e.g., on innovation (Geels, 2004; Hekkert et al., 2007), the performance of medical services (Righi et al., 2012), spread of diseases (Scarpino and Petri, 2019), agri-food systems (Dermody et al., 2018) and infrastructure (Monechi et al., 2018) or social media networks (Sobkowicz et al., 2012). Ranging from agent-based models (Dermody et al., 2018) to more analytical network-diffusion ones (Vespignani, 2012), various model frameworks have been proposed for describing their time evolution. However, to the best of our knowledge, less attention has been paid to developing data-driven frameworks to analyse their dynamical properties.

One common denominating factor for ST systems is the ubiquity of the accumulated heterogeneous spatio-temporal data: numerous points in (network-)space own a time series of measurements, constituting a ‘signal’. For railways, the subject of this chapter, this signal is the accumulated delay of trains that (should) pass by a part of the network. In epidemiology, the signal is the number of infections in a city or a region. In social networks, the signal can refer to, for instance, a video, an opinion or a Twitter hash tag. The aspects of heterogeneity, absence of physical laws and the ubiquity of spatio-temporal data can also be attributed to many other (non-ST) systems, notably in neuroscience. The natural questions that are central to all of these systems are: how do the signals evolve (e.g., development of a communicable disease to a pandemic, or large-scale disruptions in railway systems)? In particular, can one define specific macro-states for these systems, and describe transitions across such states? In this chapter we focus on these questions.

In physical systems, transitions are widely discussed topics. Therein, transitions are often associated with bifurcation points, inducing multiple equilibria or limit cycles (Strogatz, 1994; Ashwin et al., 2012; Kuehn, 2011). For the analysis of such bifurcation points, the underlying equations should (at least partly) be known, which is generally not the case for ST systems. In contrast, working with observations of such a system usually involves dealing with noise and filtering long-range correlations from time series that at first sight do not show any trend or approximation of a transition regime. The behaviour of such a system close to (standard) transitions is referred to as ‘critical slowing-down’, and can often be traced to an increased variance and autocorrelation in time (Scheffer et al., 2009; Dakos et al., 2008). More advanced techniques, especially when dealing with different time scales in the data, are degenerate fingerprinting and detrended fluctuation analysis (Kantelhardt et al., 2001; Held and Kleinen, 2004). Applications of these techniques can be found in, for example, vegetation systems (Bathiany et al., 2012) and climate variability (Thompson and Sieber, 2011).

In physical systems too, the underlying dynamics or equilibrium structure may not always be known; modelling of such systems calls for the development of data-driven frameworks. An example of such a data-driven analysis of a physical system is (Tantet et al., 2015), where two different regimes in atmospheric northern hemisphere flow were

analysed. Using a reduced phase-space obtained through principal component analysis, the dynamics of the system was therein analysed using the properties of transition matrices, which are determined using Ulam's method (Ulam, 1964). Transition matrices as part of forward integration simulation models have also been used by (van Sebille et al., 2012), approximating ocean surface circulation using buoy data to simulate the movement of marine plastics in time. In this chapter we generalise and combine some of these data-driven approaches for ST systems.

Our framework is showcased by data made available by the Dutch railways. Physical properties of a railway system are, for example, the velocity of trains, (dynamic) network capacity and the locations of switches and connections. Human elements in the system range from local dispatching and driving of trains and passengers to macro-scale decision making on the cancellation or rerouting of trains. Like in many ST systems, the interplay between the two causes prediction difficulties, with questions like: 'In what circumstances will people cancel a train?' or 'How does delay of a certain train affect others if it gets rerouted?'

For railways, it is highly relevant to be able to understand and predict the propagation of delay. If prediction were to be accurate, many disruption management techniques have been proposed in literature, involving research on the timetable adjustment (Corman et al., 2014), and rolling stock and crew rescheduling (Ghaemi, 2018; Jespersen-Groth et al., 2009).

While any day contains numerous small fluctuations, specific combinations of delayed trains and external factors can build up to severe, nation-wide disrupted situations, as described for several instances of the Dutch railways in (Dekker et al., 2021f). The winter of 2012 contained several of these nation-wide disrupted days for the Dutch railways. One of them is shown in Fig. 1.1, where a clear propagating signal of high accumulated-delay is visible. An investigation conducted by the Dutch ministry of infrastructure concluded that a series of unfortunate coincidences were the cause for the disruption on that day, with 2-3 times higher-than-usual infrastructure disruptions and delay caused by missing personnel (Nederlandse Spoorwegen, ProRail, Ministerie van Infrastructuur en Milieu, 2013). This is a classic example of the emergence of a (disrupted) macro-state, built up by micro-interactions (e.g., one train affecting the next, and so on). To the knowledge of the authors, there is no evidence for critical slowing down prior large disruptions in the railway systems, probably due to the heterogeneous and discrete nature of the variables.

Although many disturbances (caused by, e.g., accidents or ill personnel) are by definition unpredictable, the subsequent evolution of a primary delay source may be. Approaches in transportation science to investigate this can roughly be split into (a) modelling studies (for the testing of time-tables, and first-order robustness of the railway system, modelling studies are vital), and (b) data-driven studies on micro-interactions. Modelling of delay propagation is done at various levels, although they typically leave out human influence on the system, like decision making by the dispatchers. Micro-simulation modelling tools are often used to test time-tables, and simulate first-order delay propagation, involving tools like FRISO (Middelkoop and Loeve, 2006) and OpenTrack (Nash and Huerlimann, 2004). More abstract modelling studies involve analytical approaches to system equilibria (Ball et al., 2016) and first-order delay propagation (Goverde, 2010; Schöbel, 2012). Agent-based models, too, are used to simulate railway

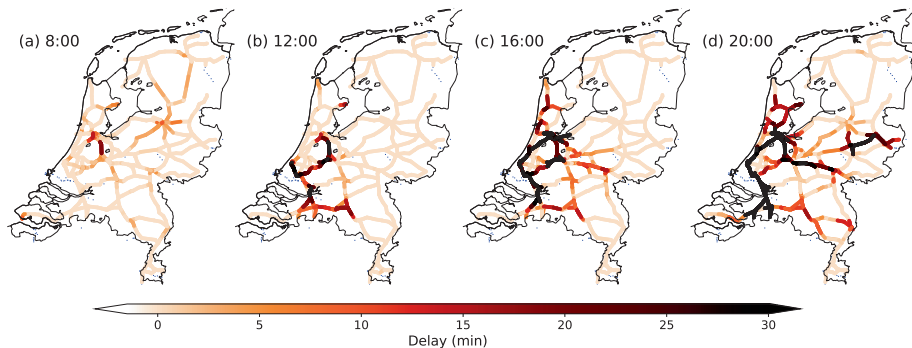


Figure 1.1: Dutch railway network in coloured lines (thin black lines depict the coastal and country borders of the Netherlands). Colouring indicates delay per segment on the Dutch railway network, at four instances on February 3, 2012. Figure similar as in Dekker et al. (2021f), which discusses this day in more detail.

systems (Gambardella et al., 2002), or gaming studies to back up the behaviour of agents (Middelkoop et al., 2012).

The key advantage of data-driven studies, in contrast to the above, is that all interactions and human influence are fully captured by the data. The main disadvantage is, of course, that it is difficult — if not impossible — to disentangle physical processes from human or case-dependent influence, specifically those processes that are robust (in many cases the same), rather than incidental (unique per case). Examples of data-driven studies in railways literature have mostly been performed at the micro-scale, e.g., the statistical estimation of specific train activities like running and dwell times (Kecman and Goverde, 2015a,b; Li et al., 2016; Şahin, 2017). Machine learning techniques like support vector machines are used to predict train arrival times from data in Serbia (Marković et al., 2015) and Italy (Oneto et al., 2017). Instances of combining data-driven and modelling approaches for railways systems exist too, helping to extract processes (like physical ‘laws’) from data (Monechi et al., 2018), which has led to the identification of delay propagation patterns for the Italian and German railways, and eventually to the modelling of the dynamics like backward propagation of delays. A recent study to predict delay propagation uses hybrid Bayesian network models, focusing on one high-speed train line in China (Lissan et al., 2019), while another combines Bayesian networks with stochastic prediction by updating the probability distribution from which future train delays are drawn (Corman and Kecman, 2018).

What is currently still missing in railways literature is a data-driven framework to analyse and predict evolutions in delay at the meso- (regional) as well as the macro- (entire system) scale (as in Fig. 1.1). With reference to the focus of this chapter, we note that most data-driven studies, being mainly focused on the micro-scale, fail to reproduce (or are made for another scope) emergent phenomena like transitions across macro-states. Given that the means to predict these major events is of high societal relevance, the framework we present in this chapter demonstrates the potential to successfully address this issue.

Our framework consists of three steps and is described in Sec. 1.2. We start with di-

mension reduction by applying principal component analysis to the data, where a small set of components is chosen based on variance and persistence in time to define a reduced phase-space. This is followed by clustering on this phase-space, resulting in a (near-automatic) identification of the ‘rest’ and ‘disrupted’ states. The third step consists of applying an early warning procedure that allows for real-time forecasting of the system, specified to predict the evolution towards the disrupted state. In the subsequent sections, we couple the framework to data made accessible by the Dutch railways: we introduce the data and the general results of the framework in Sec. 1.3, and present the results of two case studies in Sec. 1.4. We end the chapter with a summary and concluding remarks in Sec. 1.5.

1.2 The three-step framework

As mentioned above, many ST systems concern spatio-temporal data: numerous points in (network-)space that own a time series, constituting a ‘signal’. For the railways (as we will discuss in Sec. 1.3), this signal is the accumulated delay of trains that (should) pass by a part of the network.

1.2.1 Step 1 - Dimension reduction

Given the spatio-temporal data, the first step concerns capturing the dynamics of the system in a few relevant system descriptors that define a reduced phase-space (while simultaneously saving computational efforts and time). Given the spatial dimension N (i.e., the number of nodes/grid-points where system variables are measured), and temporal dimension M , the matrix D containing all data for the system has dimension $N \times M$. For many dimension reduction schemes, the averages of each time series is subtracted before D is computed.

An example of such a dimension reduction scheme is principal component analysis (PCA), which is the method we use for the railways. PCA finds orthogonal vectors, expressed as linear sums of the N time series, optimised for the portion of variance they explain. In mathematical terms, when the full time-series is involved, this entails the diagonalisation of the correlation matrix $D^T D$ ($N \times N$ matrix, superscript T denoting the transpose): i.e., expressing $D^T D = U \Sigma V^T$ (note that $U = V$, since $D^T D$ is real and symmetric). The columns of V are the eigenvectors of $D^T D$, commonly referred to as the principal components (PCs). The elements of the columns of V represent co-varying parts of the system (in space) that are known as empirical orthogonal functions (EOFs), while Σ is a diagonal matrix containing the eigenvalues of $D^T D$ or the variances of the PCs. By construction, every snapshot of the full system, i.e., the system-wide values of the time-series data at any instant of time, can be expressed as a sum of the PCs with amplitudes. In other words, the PCs define the phase-space for the system, and the dynamics of the system is then described by the time evolution of the amplitude of the PCs.

One important disadvantage of PCA is that it is built to retrieve patterns that optimise the explained variance, while in practice these patterns might not be the most relevant ones. For some systems, such as in neuroscience, high-variance patterns may simply concern uninteresting features. Further, the most interesting signals may not be

orthogonal to each other, which PCA enforces. For such systems generalised eigenvalue analysis may be more suitable (Chen and Ruan, 2009).

PCA does work for the railway system, but it is important to find those PCs that are the most suitable to capture the development of large-scale disruptions. For this reason, first of all, PCs gained from the full dataset D are not the best ones. Instead, (as we will see in Sec. 1.3) a subset of the time series that hold the data on ‘disrupted days’ is more suitable, resulting in a matrix D' with dimensions $N \times M'$, with $M' < M$. Secondly, in order to describe the system’s dynamics in a reduced phase-space, the right PCs need to be retained (and the rest discarded). We find that in choosing the right PCs for the railways, we need to consider (a) the amount of variance explained by these PCs, and (b) the persistence of their amplitudes in time. We will take these up in Sec. 1.3.

1.2.2 Step 2 - Identification of macro-states

The goal of the second step is to identify macro-states within the reduced phase-space. The macro-states are system-dependent. In case of infectious disease spreading, they may include a state with no infectious and a state involving a large-scale epidemic or pandemic. For railways, we can distinguish two types of states: ‘rest’ states, where the system largely adheres to the timetable, and ‘disrupted’ states. Depending on where in the network the delay is concentrated, the latter can take many different forms.

These macro-states can in general be approximated by identifying quasi-stationary areas in the phase-space: areas where the system is likely to remain confined up to a certain time scale. We call these areas *clusters*.

In order to identify the clusters, the reduced phase-space is split into grid cells, denoted by G . The (conditional) transition probabilities are required to group the cells into the sought-for clusters. We capture these probabilities in a transition matrix T , whose elements T_{ij} are defined as the likelihood of going from one grid cell to another within a timelag τ , i.e.,

$$(T_\tau)_{ij} = \frac{\#\{(x_t \in G_i) \wedge (x_{t+\tau} \in G_j)\}}{\#\{x_t \in G_i\}} \quad (1.1)$$

Given the phase-space probability density vector $\vec{\rho}_{t_0}$ (discretised over the grid) at time t_0 , one can then calculate the probability density at time $t_0 + \tau$ by simply operating T_τ on $\vec{\rho}_{t_0}$:

$$\vec{\rho}_{t_0+\tau} = T_\tau \cdot \vec{\rho}_{t_0} \quad (1.2)$$

For small τ , realisations would typically move towards a neighbouring grid cell or stay within the same one, which results in a sparse matrix with most nonzero elements on the diagonal or slightly off the diagonal, but for longer τ realisations would travel further in phase-space. For a given value of τ , the matrix elements $(T_\tau)_{ij}$ can then define a ‘transition probability network’ with grid cells as nodes, with the strengths of the corresponding links being determined by the transition probabilities among the grid cells. Clusters can then be identified on this network by searching for groups of nodes (i.e., grid cells) that

are strongly *intra*linked (by transition probabilities), but weakly *inter*linked.

In graph theory, there are many clustering methods to achieve this. Here we use the Louvain method (Blondel et al., 2008), which optimises modularity, defined as the fraction of the edges that fall within the given groups minus the expected fraction if edges were distributed at random (for more details, see (Newman, 2006)). This algorithm merely optimises the clusters in terms of semi-invariance for a given value of τ , but does *not* guarantee it — e.g., it does not guarantee that if the system enters a certain resulting cluster, it will not leave the cluster at a time $\tau' < \tau$, only to return to it precisely at time τ . We will return to this issue in Sec. 1.5.2.

The resulting clusters, therefore, can be used as good approximations of semi-invariant macro-states of the system, with the caveat that they need to be interpreted in terms of which states of the system they are referring to.

1.2.3 Step 3 - Early warnings

Step 3 concerns predicting transitions towards and across macro-states (throughout the rest of this section we will interchangeably refer to them as clusters).

The probability of transitioning from any grid cell i to grid cell j in time τ is given in Eq. (1.1). This can be aggregated by summing over j belonging to a given macro-state k as

$$\begin{aligned} P_{t_0+\tau}(\text{to cluster } k) &= \sum_{j \in \{\text{cluster } k\}} \bar{\rho}_{t_0+\tau}(j) \\ &= \sum_{j \in \{\text{cluster } k\}} (T_\tau \cdot \bar{\rho}_{t_0})(j) \end{aligned} \quad (1.3)$$

The result is a first (real-time) indicator of whether we expect a transition towards a cluster. However, being far away from the macro-state hardly ever results in a high ('alarming') probability when looking only small time lags τ , but it might for a longer time lag. This illustrates that for the construction of an *early warning* metric, one needs to incorporate a variable time lag.

In practice, it is important to know what the minimal lag is, given that one wants to be sure by a certain percentage that the system is remaining in a certain cluster. For example, given that one wants to be sure by a probability of 0.95 that the system remains in cluster A , an alarm needs to be given whenever the system enters a cell that allows transitioning towards another cluster B with probability ≥ 0.05 , and find out what the minimal lag is of doing so. We call this (latter) percentage the *critical probability* p_c and the related minimal lags $\tau_{\text{alarm}}(p_c)$ (different per grid cell). For the purpose of accurate prediction, a maximum time horizon t_{max} is also set depending on the estimated memory of system at hand; meaning that in our calculations, events of entering a cluster at time lags larger than t_{max} are not considered in the statistics.

Summarised, a 'warning' or alarm is given at any time if the system is likely to enter the cluster at a probability $\geq p_c$. Attached to this alarm is a time lag τ_{alarm} at which this is expected to happen.

	A	A	A	\bar{A}
	$\tau_0 < \tau_{\text{alarm}} - \epsilon$	$ \tau_0 - \tau_{\text{alarm}} < \epsilon$	$\tau_0 > \tau_{\text{alarm}} + \epsilon$	
O	FA2	H	MA2	MA1
\bar{O}	FA1	FA1	FA1	CR

Table 1.1: Various outcomes concerning the correctness of an early warning metric, based on whether an alarm is given (A) or not (\bar{A}), and whether within the time horizon the macro-state transitions (O) or not (\bar{O}). Precision in time is assessed by the predicted time lag τ_{alarm} , the actual time of transitioning τ_0 and a bandwidth ϵ . Outcomes are as in the text: various types of false alarms (FA), missed alarms (MA), hits (H) and correct rejections (CR).

1.2.4 Evaluation of predictions

In practice, as well as for the estimation of the parameters p_c and t_{max} , it is imperative to assess the forecasting skill of the early warning metric. The skill should increase for correct predictions, and reduce in case of false positives (‘false alarms’) and false negatives (‘missed alarms’).

A commonly used metric to assess the accuracy of a forecast is the Brier score (Brier, 1950), defined as:

$$S_{\text{Brier}} = \frac{1}{\mathcal{N}} \sum_{i=1}^{\mathcal{N}} (p_{f,i} - O_i)^2 \quad (1.4)$$

where \mathcal{N} is the total number of predictions made, $p_{f,i}$ the forecast probability of entering the cluster at instance i and O_i is 1 if the system enters the cluster (within t_{max}), and 0 if not. The Brier score acts like mean squared error of the forecast.

However, the disadvantage of the Brier score is that no variable time lag can be incorporated in it, while the skill should also be penalised by predicting the time lag at which the transition happens wrongly. Therefore, following (Tantet et al., 2015), we use another metric in this study called the Peirce Skill score (PSS), also known as the true skill statistic or as the Hanssen and Kuipers discriminant (Peirce, 1884; Allouche et al., 2006).

To introduce this metric, we first explain different types of correct and incorrect predictions. At every time step t , the system is either alarmed (A) or not alarmed (\bar{A}), based on the critical probability p_c and the prediction time τ_{alarm} (see Sec. 1.2.3). We check whether the system indeed entered the specified cluster between now and a specified time from now. If not, we call this a non-occurrence (\bar{O}); but if it does, we call it an occurrence (O) and the time lag at which this happens τ_0 ($0 < \tau_0 < \tau_{\text{alarm}}$). The different outcomes are schematically shown in Tab. 1.1.

Let us start with the upper row (O), when in reality the system indeed enters the specified cluster. When no alarm is given (\bar{A}), we call this bad prediction a missed alarm type 1 (MA1). Similarly, when an alarm is given, it is only correct if τ_{alarm} is close enough to τ_0 . We allow for a bandwidth ϵ around τ_0 ; if $|\tau_{\text{alarm}} - \tau_0| < \epsilon$, the prediction is correct. We call this a hit (H). If the metric predicts it too early ($\tau_0 < \tau_{\text{alarm}} - \epsilon$), we call it a false alarm type 2 (FA2), and if it is predicted too late ($\tau_0 > \tau_{\text{alarm}} + \epsilon$), we call it a missed alarm type 2 (MA2).

Next, we consider the lower row (\bar{O}) of Tab. 1.1, where the system does not reach the

specific cluster within the time horizon (t_{\max}). If an alarm is given (at any τ_{alarm}), we have a false alarm type 1 (FA1). Similarly, if no alarm is given, then we call it a correct rejection (CR). Later in this chapter, we will refer to FA2, MA2 and MA1 as ‘Missed Alarms’ (MA), and to FA1 as ‘False Alarms’ (FA).

With the above, the PSS for a given critical probability p_c is then calculated as follows:

$$S_{\text{Peirce}}(p_c) = \frac{\#H(p_c)}{\#O(p_c)} - \frac{\#\text{FA1}(p_c)}{\#\bar{O}(p_c)}. \quad (1.5)$$

One sees that the score is rewarded for correct alarms, but penalised for incorrect alarms (including rewarding the score for instances where the early warning metric correctly did not give an alarm).

If $S_{\text{Peirce}}(p_c) > 0$, there is more skill in the prediction than random prediction. Note also that by construction $-1 \leq S_{\text{Peirce}}(p_c) \leq 1$; the closer it is to 1, the more skill the prediction has.

1.3 Coupling the framework to data from the Dutch railways

1.3.1 Data description

The data for this study has been provided by the manager of the main railway network in the Netherlands (ProRail), logged at so-called service control points (SCPs). With a total of 801 SCPs spread over the entire network (of which passenger stations are a subset), they divide the Dutch tracks into 1438 smaller segments. Only passenger trains data have been considered, for a number of reasons (as follows). Freight trains are (economically) privacy-sensitive and it is therefore difficult to get a complete dataset. Further, it is only a small fraction of the total railway traffic in the Netherlands: about 5.7% of all Dutch train kilometres in 2017 was by freight trains (numbers courtesy of ProRail). Also, while passenger trains are bound to time schedules and routes, freight trains schedules differ every day. This means that the interaction of freight trains are non-systematic and therefore not really predictable. Moreover, some of the tracks of freight trains are partly separated from passenger trains (like the Dutch ‘Betuweroute’ from Rotterdam to Germany) dedicated to freight trains. This reduces their effect on the dynamics of the whole system even more. We therefore decide to focus only on passenger trains.

We work with the data of one year, from July 1, 2017 to June 30, 2018. The data includes the logging of passing trains, including characteristics of the train, but also the planned time and subsequent delay of the train at 1-second resolution. This data is aggregated to continuous time series on segments at 1-minute resolution. In short, we define delay $d_i^j(t)$ of train j on segment i at time t as:

$$d_i^j(t) = \begin{cases} 0 & \text{if } t < t_p \text{ (before the planned time)} \\ t - t_p & \text{if } t_p < t < t_r \text{ (activity is not yet realised, while it should have been)} \\ 0 & \text{if } t > t_r \text{ (after the activity was realised)} \end{cases} \quad (1.6)$$

where t_p and t_r are the planned and realised time of an activity of train j at segment i , respectively. The above definition involves the buildup of delay when a train should be at the segment while it is late, and disappears from that segment when the train exits it, giving rise to a sawtooth pattern of the delay. (Note also that once the delay disappears from a given segment, unless it happens to be the last segment of service for that train, the delay simply continues on the next segment.) We then compute the total delay $d_i(t)$ on segment i at time t , by summing d_i^j over all trains j (both directions) as:

$$d_i(t) = \sum_j d_i^j(t) \quad (1.7)$$

For more details, the reader is referred to SI A.1.1 and A.1.2. The aggregation of delay results in 1438 time series spread over the spatial Dutch railway network. Using the notation as in Sec. 1.2, this means that $M = 365 \cdot 1440 = 525600$ minutes and $N = 1438$ segments.

1.3.2 Results principal component analysis (Step 1)

The first step concerns the extraction of the most relevant principal components from the data, with ‘relevance’ referring to their suitability to best describe the evolution towards large-scale disruptions.

A commonly used metric for the severeness of delay on the network, is a day-to-day classification defined by ProRail. All days are labelled with one of four (severity) categories: green, neutral, red or black. These classes are based on the punctuality and cancellations of trains on important train tracks. ‘Green’ (46 out of 365 days) days refer to ‘quiet days’: few cancelled and delayed trains, while ‘red’ (21) and ‘black’ (6) days refer to ‘disrupted days’, that contain a lot of delays and cancellations. ‘Neutral’ (292) days are those that are neither quiet, nor disrupted days. We term the green and neutral days together as ‘regular days’. For more details, the reader is referred to SI A.1.3.

To get the relevant PCs, we choose to perform PCA only on the delay data from the disrupted days (‘red’ or ‘black’), as by definition they explain the most variance for disrupted situations (a robustness check on the PCA results is performed in SI A.2). To ascertain the relevance of these PCs, we have assessed the performance of the PCs on the entire dataset (i.e., including delay data from the regular days): for example, if the PCs calculated from the delay data on disrupted days barely explain any variance on regular days, they may not be useful for analysing the evolution of the system towards disrupted states (in that case, they would essentially only explain persistent patterns on the disrupted days).

Subsequently, the question arises regarding which and how many PCs we need to retain to construct the reduced phase-space. The metrics we use to answer these questions are the explained variance by, and persistence in time of, the PCs. The latter we define as the timescale τ_0 at which the autocorrelation function decays, which we extract by fitting a function e^{-t/τ_0} to the autocorrelation function of the PC amplitudes, scaled to its value at zero time-lag. From the combined delay data, it turns out that the first two PCs distinguish themselves from the rest (more details in SI A.1.4). This motivates our choice

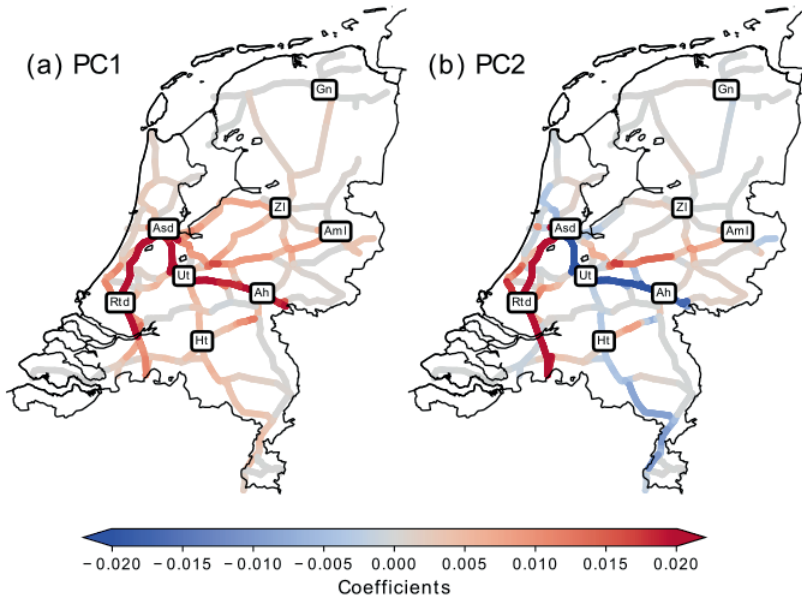


Figure 1.2: The first two relevant EOFs for the Dutch railway system, explaining respectively 13% and 16% of the variance over a full year. A running spatial average-smoothing is applied for visualisation reasons. Abbreviations refer to important passenger stations: Amsterdam Central (Asd), Rotterdam Central (Rtd), Utrecht Central (Ut), Arnhem (Ah), Groningen (Gn), 's-Hertogenbosch (Ht), Amersfoort (Amf) and Almelo (Aml).

to retain only PC1 and PC2 to define the reduced (two-dimensional) phase-space for describing the dynamics of the Dutch railway system.

Putting all the above ingredients together, our results for step 1 are as follows. The variance explained on the disrupted days by the first two PCs are 16.0% and 9.5%, respectively, while these numbers change to 13.0% and 16.0%, respectively when the variance is calculated on the full dataset, i.e., including regular days. Sorting the PCs by variance explained makes sure that we use those components that show more-or-less robust covariant patterns, rather than the effect of (incidental) individual cases.

The corresponding EOFs, i.e., the spatial plots of the individual PC elements are shown in Fig. 1.2. Segments that have high equally-signed amplitudes in these plots are parts of the network at which delay often co-occurs. Three dominating train lines can be distinguished in Fig. 1.2 (see Fig. 1.3):

- L1: The line from Amsterdam southward through Rotterdam. It is connected to Belgium by trains ultimately reaching Antwerp and Brussels.
- L2: The line from Amsterdam southeastward towards the cities of Utrecht and Arnhem. It continues towards the German cities of Düsseldorf, Köln and Frankfurt.
- L3: The line from Amsterdam eastward towards the cities of Amersfoort and Almelo. This line is connected to major German cities like Münster, Dortmund and Berlin.

These lines do not only incorporate the effect of long-distance (international) trains

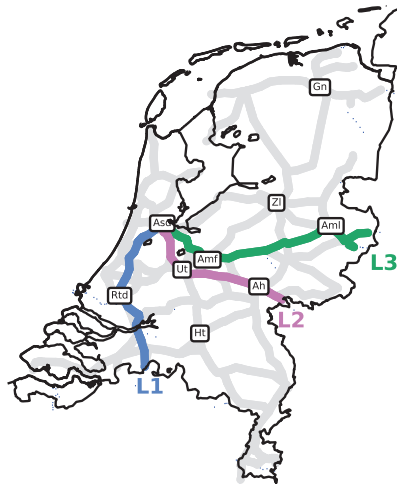


Figure 1.3: Three important lines L1, L2 and L3. Abbreviations as in Fig. 1.2.

that are prone to building up delay, but they also include busy tracks and major cities in the Netherlands, explaining the large amount of variance on these lines.

We can interpret the PCs in terms of the delay spread across these three lines, using the EOFs in Fig. 1.3. It is important to stress that delay is almost always a positive variable on any segment. A negative value anywhere at any time means that the sum of the delay of all planned trains at that time and track is negative. Trains may be early, but such occurrences are infrequent and not by a large amount of time. Note that this statement does not contradict the plots in Fig. 1.4, since the actual delay on the network at any given time linearly relates to the amplitudes of these PCs, which can be both positive and negative. However, as is visible in Fig. 1.2a, almost all coefficients of PC1 are positive, resulting in only positive PC1 amplitudes. Looking at the coefficients of PC2 in Fig. 1.2b, we see (large) negative coefficients on L2 and positive coefficients on L1 and L3.

1.3.3 Results identification of macro-states (Step 2)

Within the PC1-PC2 reduced phase-space (we henceforth refer to it simply as ‘phase-space’ for brevity) we now identify states, a process that we describe in this section. First, we compute the region of the phase-space covered by the entire year’s delay data, obtained simply as the amplitudes of the two PCs (M data points all together). This region exclusively corresponds to $PC1 \geq 0$. We then discretise this region into 123×123 grid cells on a logarithmic scale, on which we compute the transition matrix T_τ with $\tau = 30$ minutes. We follow this up by performing Louvain clustering on the transition matrix data, leading to the identification of four clusters that are shown in Fig. 1.4a in thick black boundaries. (Robustness of the Louvain clustering against the choice of grid resolution and τ has been checked in SI A.2.2). We find that each cluster in Fig. 1.4a, being so large,

contains multiple delay configurations of the full system. In order to differentiate among these, we perform Louvain clustering a second time within each cluster. This action leads to the identification of further clusters within each cluster; we call these *subclusters*. Our calculation reveals the existence of ten subclusters all together, numbered 1 through 10; these are shown in different colours and thin black boundaries in Fig. 1.4a.

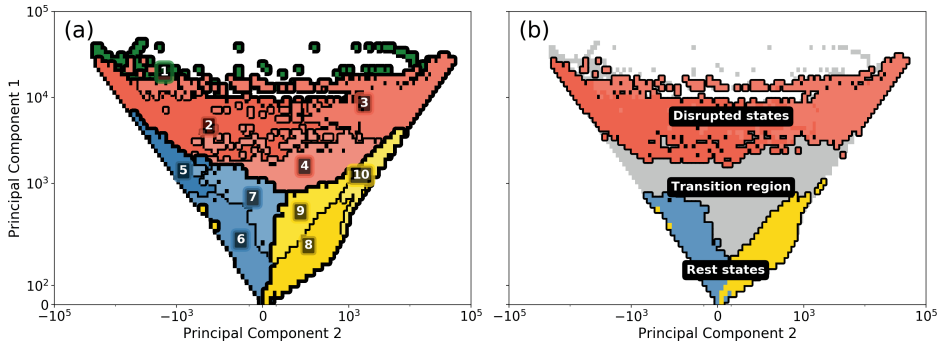


Figure 1.4: Phase-spaces showing (a) clusters in colours, and subclusters in various shades of the same colour, and (b) how we interpret various parts of the phase-space: rest states (subclusters 6 and 8), transition regions (4, 5, 7, 9 and 10) and disrupted states (2 and 3). Colours have been used only for those areas, where there are realisations. The choice of colours is arbitrary and has no relation to the classes defined in SI A.1.3.

We next interpret the subclusters in terms of macro-states they may represent. Obviously, the area around $(PC1, PC2) = (0, 0)$ — i.e., the origin — must represent the so-called rest states, since in this area the delays are small. (Indeed, every day at the beginning and at the end of service the system respectively starts from and returns to the origin.) The interpretation of the subclusters as macro-states is more subtle, and can only be ascertained by evaluating all instances of the system within each subcluster. To this end, we analyse each instance in terms of two variables: (a) amount of delay, and (b) in what respect various day-classes (as in SI A.1.3) are represented, by defining a ‘bias factor’ B as

$$B(i, j) = \frac{P(i|j)}{P(i)} \quad (1.8)$$

where $P(i)$ is the probability of a realisation to have day-label i (i.e., green, neutral, red or black), while $P(i|j)$ is the probability of a realisation to have the same label within subcluster j . We add the day-labels in this analysis since delay alone does not signify the severeness of disruption — e.g., a sharp spike in delay is not necessarily a severe event if it is resolved quickly. Also (as noted above), every day the system starts from $(PC1, PC2) = (0, 0)$; for this reason, even the black days have some realisations in the subclusters around $(0, 0)$. The bias factor reflects the prevalence of a certain day-label within each subcluster in comparison to its overall prevalence. For example, if $B_G(3)$ is less than 1, it means that green days are less frequent in subcluster 3 (of all subclusters considered). The results are shown in Tab. 1.2.

Using the numbers shown in Tab. 1.2, we provide an interpretation to the subclusters shown in Fig. 1.4b. The interpretation of the phase-space is in three parts:

(Sub)cluster	B_G	B_N	B_R	B_B	Delay
1	0	0.33	1.45	39.5	95.6
2-4	0.40	0.98	2.16	2.42	20.2
2	0.56	0.91	2.57	3.31	23.3
3	0.34	0.98	1.69	4.22	33.2
4	0.36	1.01	2.20	1.32	13.4
5-7	1.21	0.99	0.84	0.65	3.9
5	1.01	1.01	1.04	0.46	8.0
6	1.44	0.96	0.62	0.68	2.2
7	0.77	1.02	1.29	0.6	6.6
8-10	1.03	1.02	0.72	0.68	3.8
8	1.25	0.99	0.68	0.76	1.4
9	0.73	1.06	0.88	0.46	6.3
10	0.74	1.07	0.54	0.80	8.9

Table 1.2: Biases of realisations of all labels per (sub)cluster. Lower-script characters of B refer to the bias factor on ‘green’ (G), ‘neutral’ (N), ‘red’ (R) and ‘black’ (B) labelled days, as defined in SI A.1.3. The delay in the last column depicts the summed delay (in hours) over the whole network (i.e., summed over all segments) averaged over the realisations that are inside the respective (sub)clusters.

- Subclusters 6 and 8 have the highest B_G scores, indicating a strong bias on green days to be in these subclusters. Moreover, they have relatively low bias scores on other days, and by far the lowest total delays. These are characteristics of the system being ‘at rest’ and therefore we use these subclusters as an approximation of the *rest state*. This choice is also supported by the fact that these subclusters are closest to the origin: obviously PC1 and PC2 are low in magnitude when delays are small.
- The largest B_R and B_B scores and the largest total delays are found in subclusters 1 to 4. Only 0.14% of all realisations are inside subcluster 1, which accounts for only about 12 hours in the entire year. (Because of this data sparsity, we will not consider subcluster 1 separately). While subclusters 2 and 3 separate themselves from subclusters 5 to 10, subcluster 4 seems to be somewhat in between: it has a B_N score of above 1, meaning that neutral days are above averagely represented in this subcluster. Moreover, its total delay (13.4 h) is on average only slightly higher than the total delay in some of the lower subclusters [e.g., 5 (8.0 h) and 10 (8.9 h)]. We therefore choose to denote subclusters 2 and 3 as *disrupted states*. (However, when presenting the early warning results in following sections, we will include subcluster 4 for completeness.)
- The rest of the subclusters (4, 5, 7, 9 and 10) then automatically form a *transition region* between the rest and the disrupted states. Computation of conditional probabilities (not shown) of reaching one subcluster from another reveals that subclusters 5 and 10 act as transition regions for the system to move towards subclusters 2 and 3, while subclusters 7 and 9 have a lot of dynamical exchanges with subcluster 4, which in turn acts as a conduit for the system to enter subclusters 2 and 3.

From these, we conclude that the subclusters 2 and 3 are the ones of interest for studying evolution towards disrupted states. In order to get an intuition on which de-

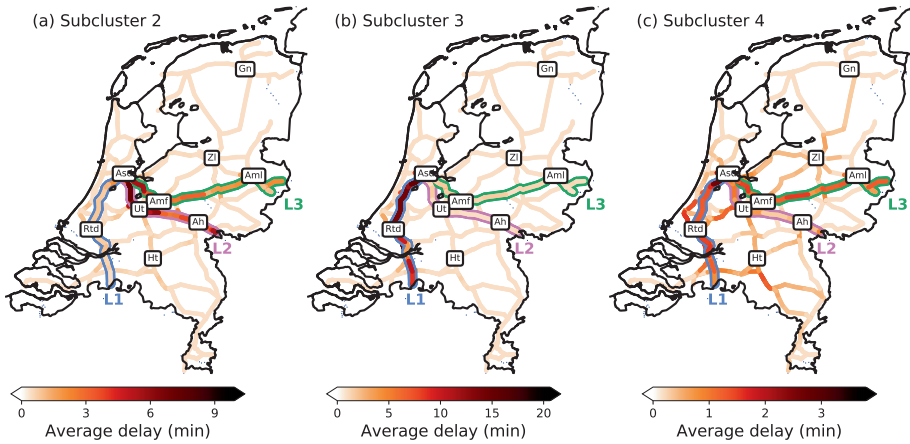


Figure 1.5: The colour of each segment shows the average delay of all realisations in subclusters (a) 2, (b) 3 and (c) 4 on the particular segment. Abbreviations as in Fig. 1.2. The lines L1, L2 and L3 are highlighted.

lay configurations of the full system they represent, the average delay per segment per subcluster is shown in Fig. 1.5 (for the sake of completeness, we also add subcluster 4). The delays seem small (~ 2 -20 minutes on important trajectories), but note that the presented results are averages over all realisations per subcluster: there are many types of delay combinations on these important trajectories, which make the signal per subcluster appear diffuse. The delay patterns in Fig. 1.5 are quite distinct from each other, and indeed allow for intuitive interpretations of the three subclusters: an upper-left region with delays as L2 and L3 (subcluster 2), an upper-right region with delays on L1 (subcluster 3), and a middle region containing less severe and more spread out delays (subcluster 4).

As the subclusters are distinctively positioned in the phase-space, we could have qualitatively anticipated the above results already from Fig. 1.2. The fact that Fig. 1.5 confirms these patterns reflects that the delay signals in the *reduced* phase-space indeed corresponds well to the delay signals in the actual realisations. Moreover, we see in Fig. 1.5 that the magnitude of the average delay patterns differ (note the colour bars). The fact that the maximum average delay in subcluster 3 is much higher than in subcluster 2 and 4 does not necessarily mean that the delays in general are longer (although that it is the case on average can be seen in Tab. 1.2); it could also mean that the delays are more consistent with where on the network they occur. In particular, one could conclude that delays in subcluster 3 are longer and more concentrated on the network.

1.3.4 Results early warning and skill score (Step 3)

Using subclusters 2 and 3 as approximations of two different types of disrupted macro-states, the prediction towards them amounts to predicting a transition towards large-scale disruptions. That is the aim of the third step.

As described in Sec. 1.2.3, the final strength of our prediction is dependent on the

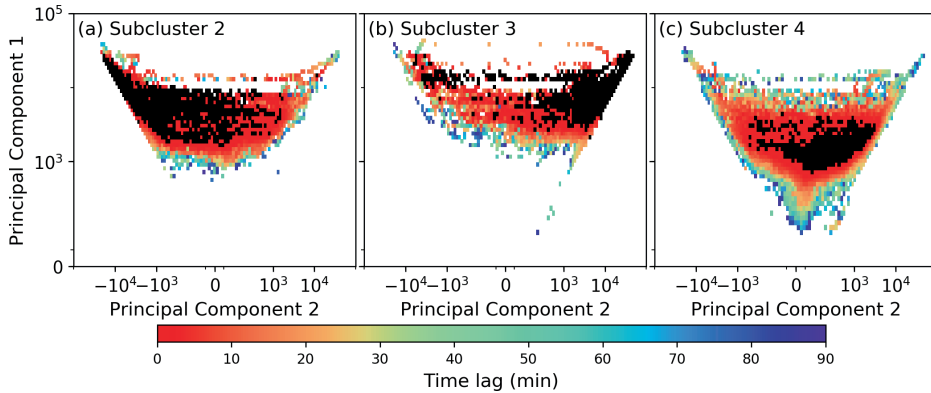


Figure 1.6: Phase-space showing (in black) (a) subcluster 2, (b) subcluster 3 and (c) subcluster 4, and the predicted time lag τ_{alarm} (in colours) for entering these subclusters for $p_c = 0.08$; τ_{alarm} values are discretized to 5 minute intervals. No colour (white) means no alarm is given from these grid cells, reflecting a time horizon t_{max} of 90 minutes.

choice of parameters p_c , ϵ and t_{max} . First, there is an optimum in the skill with varying critical probability p_c : a p_c that is too low increases the amount of FA1, but a p_c that is too high increases the amount of MA1 (or decreases H). Second, the skill is dependent on the bandwidth ϵ . The skill grows for larger ϵ , but that also means loss of accuracy in the timing of the predicted event. It is therefore a trade-off between accuracy in the time-bandwidth and accuracy in being right in predicting. The third parameter that influences the skill is the time horizon t_{max} . If t_{max} is small, the overall skill will generally increase, simply because it is easier to predict events that are imminent. Here too there is a trade-off: a small time horizon means only short-term predictions, which reduces the value of giving a warning, as there would then be little time left to prepare for taking intervening measures.

Considering the trade-off for all the involved parameters, the ideal situation would be a high skill, at a small ϵ (at least $\epsilon \ll \bar{\tau}_{\text{alarm}}$, where $\bar{\tau}_{\text{alarm}}$ is the average alarm lag for correct alarms), a large t_{max} and a large p_c . Nevertheless, a good choice of the parameters ϵ , p_c and t_{max} is important, because it modulates the skill greatly. The right choice will depend on the system and what type of accuracy the user deems important. See SI A.2.4 for the parameter sensitivities of predicting evolution towards various subclusters.

For $p_c = 0.08$ and $t_{\text{max}} = 90$ minutes, the time lags at which an alarm is given are shown in Fig. 1.6 (i.e., for every grid cell, we determine whether there is a probability of at least 0.1 to reach the specific subcluster at a time lag ≤ 90 minutes). In Fig. 1.6a we see that alarms are mostly given in the upper part of the phase-space: a total of 60% of the grid cells representing actual realisations is covered. Most alarms are given at low time lags (i.e., close to the subcluster). More long-range predictions can be made mostly in the upper-right corner and in subclusters 5 and 7 (cf. Fig. 1.4a), even up to 90 minutes. The large coloured ‘tail’ in the upper right of the figure reflects the possibility of transition between an L1-dominated delay signal (Fig. 1.5b) towards an L2/L3-dominated delay signal (Fig. 1.5a), possibly through propagation of delay via Amsterdam, the city that connects the three lines.

Subcluster	Variable	G	N	R	B	Overall
2	I	0.4	0.7	2.1	2.8	0.8
	H	0.4	0.8	2.3	3.0	0.6
	O	0.9	1.5	3.5	4.0	1.6
	PSS	0.15	0.19	0.38	0.39	0.2
3	I	0.3	0.9	1.5	3.5	0.9
	H	0.3	0.6	1.3	2.6	0.6
	O	0.6	1.3	2.6	4.0	1.3
	PSS	0.08	0.17	0.21	0.36	0.16
4	I	0.8	2.3	5.1	3.0	2.3
	H	2.0	3.6	5.3	4.2	3.5
	O	2.5	4.0	5.8	4.7	4.0
	PSS	0.25	0.32	0.38	0.47	0.32

Table 1.3: Average number of hours per day that the prediction system records a Hit (H), occurrence (O) or that the system is inside the subcluster (I) for variable severity label (for example: a value of 0.4 at row I, subcluster 2 and column G means that on average, 0.4 hours of green days are inside subcluster 2). We also present the Peirce Skill score (PSS) for each subcluster and days. Capital characters in header refer to green (G), neutral (N), red (R) and black (B) days. Parameter settings: $p_c = 0.08$, $\epsilon = 30$ minutes and $t_{\max} = 90$ minutes.

The time lags for subcluster 3 are shown in Fig. 1.6b. The number of coloured grid cells is about the same as those in Fig. 1.4a: 58% of the grid cells representing actual realisations is covered, reflecting alarms at time lags ~ 30 minute of prediction in subclusters 10 and 3. Further, alarms are given only in the upper part of the phase-space. From Figs. 1.6a and b we also see that there are a lot of system evolutions from subcluster 2 towards subcluster 3 (and *vice versa*).

For completeness, we also calculate the time-lag results for subcluster 4 (although this does not belong to the disrupted state defined in the previous subsection). The results are shown in Fig. 1.6c. As expected, it has a much more spread-out pattern of τ_{alarm} as 79% of the grid cells representing actual realisations is covered. This confirms our choice to interpret it as a transition region (cf. Fig. 1.4b): this subcluster shares boundaries with many others, and therefore easily visited by the system. The predicted times are also not mainly restricted to short lags (a more pronounced signal is visible for values between 60 and 90 minutes). Revisiting Fig. 1.4, one sees that almost no alarms are given from subclusters 6 and 8, which points to the fact that subclusters 7 and 9 (5 and 10 in a different manner) act as ‘buffer regions’, from which the system can quickly recover (refraining from transitioning towards the disrupted region).

The corresponding early warning skills, for $p_c = 0.08$, $t_{\max} = 90$ and $\epsilon = 30$ minutes, are shown in Tab. 1.3. A first observation is the difference between subcluster 4 and (each of) subclusters 2 and 3: the system is inside subcluster 4 more often in general (2.3 h per day), as well as more often on relatively ‘quiet’ days (green and neutral). This may refer to the fact the combined lines L1, L2 and L3 are often delayed together (such that PC2 amplitude is neither strongly positive nor strongly negative, and PC1 amplitude is relatively large, cf. Fig. 1.2). Further, subclusters 2 and 3 are most often reached on red and black days, coinciding with higher values for H and O.

Concerning the PSS in Tab. 1.3, we see that the system evolution towards subclusters 2-4 are the most difficult to predict on green days, while it is far easier to do so on red

and black days. On green days, these subclusters are visited intermittently and in a less structural manner, which makes predictions more difficult. The PSS scores for subclusters 2 and 3 are on average roughly equal, but with a strong difference on red days (0.38 to subcluster 2 versus 0.21 to subcluster 3). The average PSS for subcluster 4 is much higher than those for subclusters 2 and 3. In general, for these parameter settings the skill scores are roughly around 0.1-0.3. These values are larger than 0, indicating (a) that there is skill in our predictions, and further (b) that our framework could potentially anticipate approaches towards disrupted macro-states.

1.4 Implications for day-to-day operations: two case studies

To showcase the potential strength (and limits) of our framework for anticipating approaches towards disrupted macro-states, we analyse its performance on two specific days as case studies (Wednesday January 3, 2018 and Thursday April 19, 2018). The case studies are chosen based on the fact that these are ‘red days’ on which the system indeed evolved towards the disrupted state (cf. Fig. 1.4). Note that if the system would not enter the disrupted state, the performance would be calculated only on an empty-occurrence (O) set, resulting in $S_{\text{Peirce}} \leq 0$ by definition. Indeed, if we are interested in the performance of the predictions on sets with no occurrences, some other skill score could be a better choice. See SI A.3 for a discussion on a regular day, on which the system does not enter the disrupted state.

We first provide an overview of the statistics for these days in Tab. 1.4. Therein, we see — upon comparing to the total delay and number of cancellations (per minute) over the entire year — that both January 3 and April 19 have strikingly different characteristics than the system has on average. The average total delay of 5.5 hours on average days can be interpreted as 13.8 seconds delay on *every* segment at every instance of the day. Both days contain 2-3 times more average total delay, and the maximum total delay on January 3 is even about 4 times as high as on average. The number of cancelled train activities per minute is also larger, with slightly higher numbers on April 19 than on January 3.

The overall score might give an indication of the predictability for individual subclusters, suggesting that the movement towards subcluster 4 is easier to predict. This is however not the case *per se*. As explained above, an issue with the Peirce Skill Score is that for specific subclusters the score is purely negative on days when the subcluster is not reached. This leaves a bias in the skill score: independent of their predictability, subclusters that are more often reached generally have a higher PSS (we have seen earlier that subcluster 4 is more often visited).

Another important observation in Tab. 1.4 is the number of false alarms. Although the metric does penalise false alarms, a small number of correctly predicted occurrences can result in a high PSS, even when the number of false alarms is also large. Considering that for railway companies, false alarms can in practice be more destructive than missed alarms, it is important to keep track of the false alarm rate as well. Note that the false alarm rate involving subcluster 4 is quite high, which indeed points to the fact that it is easily visited and not *per se* related to large disruptions. In comparison, the false alarm rates towards subcluster 2 and 3 are much less.

Variable	Metric	All	03/01	19/04
Delay	Average	5.5 h	17.3 h	13.5 h
	Maximum	23.4 h	101.4 h	35.6 h
Cancellations	Average	2.1	6.2	6.6
	Maximum	7.7	17.0	22.2
Predictions to 2	PSS	0.20	0.37	0.31
	Max t_{alarm}	90 min	87 min	50 min
	FA%	8.9%	12.8%	21.0%
Predictions to 3	PSS	0.16	0.34	0.14
	Max t_{alarm}	89 min	65 min	86 min
	FA%	7.0%	5.2%	17.7%
Predictions to 4	PSS	0.32	0.34	0.45
	Max t_{alarm}	90 min	82 min	75 min
	FA%	35.6%	39.7%	35.6%

Table 1.4: Two case studies in comparison to the statistics over the entire year. Here, we use $p_c = 0.08$, $t_{\text{max}} = 90$ minutes and $\epsilon = 30$ minutes for the calculation of the predictability (Peirce Skill score, or PSS). Cancellations mean cancelled train activities per minute. A running mean of 30 minutes has been applied. See text for details. The alarm lag time t_{alarm} values are shown only for successful alarms ('Hits'). Statistics for all days are given in median values.

1.4.1 Wednesday January 3, 2018

Fig. 1.7 shows the system's trajectory in the phase-space on this day along with all prediction outcomes in the colors of the dots, and the corresponding evolution of total delay and cancellations per minute (see panels on the right).

Since, by construction, the system is at (0,0) in the phase-space at the start and at the end of each service day, the system is seen to be around (0,0) at early morning and late night. This coincides with a lot of correct rejections. However, already around 6:00 hrs, a strong increase in delay forces the system into subcluster 2, and it is predicted only shortly before (visible in the first blue shaded area in the upper-right panels of Fig. 1.7a). This event moves the system's position in the phase-space to the upper-left, corresponding to delays in the west of the Netherlands (lines L1 and L3). During this period, some false alarms towards subcluster 3 are also given. After this event, the system returns to 'safer areas' in the phase-space, resulting in correct rejections for both figures around 10:00 hrs. After this, delay builds up strongly towards the upper-right of the phase-space, relating to delay in the centre of the Netherlands, on line L2. This corresponds to the long period (large part of the afternoon and evening) that the system remains within subcluster 3, keeping the total delay to also remain high for a long time. During this period, some small entrances of subcluster 2 are recorded, but this is due to the existence of disconnected subcluster 2 cells scattered within subcluster 3. The system leaves subcluster 3 to return to subcluster 2 again for another hour, after which the system slowly moves back towards the origin.

Correct predictions towards subcluster 2 are made up to 87 min beforehand. We can see in the right panels that not all occurrences are predicted equally well — the entrance of the subcluster in the early morning, for example, is not predicted far ahead. Considering the prediction towards subcluster 3, there was only one time the system entered the subcluster (before it remained there for quite some time), which is predicted roughly

half-an-hour before, and on one instance there was a correct prediction at 65 min before.

The right panels of Fig. 1.7 also show the evolution of the total delay and cancellations. For cancellations we use the number of train activities (departure, arrival, short stops etc.) that were scheduled, but cancelled, per minute. This is a measure of the reduction of ‘stress’ on the system by human decision. Note that all large delay spikes are situated in either of the two subclusters, reflecting that these subclusters do indeed refer to a disrupted state. The first cancelled activities started in the early morning already, reflecting the early start of the problems on the network.

1.4.2 Thursday April 19, 2018

Similarly, Fig. 1.8 shows the system’s dynamics in the phase-space for April 19, 2018. On this day, the system mostly remains in the positive-PC2 part of the phase-space. Subcluster 2 is only reached on a few short instances, and the system never reaches the core of the subcluster (merely visiting the disconnected scattered cells belonging to it). We focus our analysis below on the instances that it reaches subcluster 3: three instances in total.

The first instance (around 9:00-10:00 hrs) seems difficult to predict (correct predictions are made only up to 35 minutes in advance), resulting in several missed alarms. This event coincides with a small but clear peak in the total delay and cancellations. The second instance (around 14:00 hrs) is better predicted, but there are still some missed alarms visible. Correspondingly, we see that this coincides with a strong increase in total delay, concentrated on L1 and L3. The system remains rather delayed after this. The third instance is in the late evening (around 23:00 hrs) and does coincide again with a strong peak in total delay, but prior to that there is another peak. This peak actually does not correspond to an entry to subcluster 3, but to (deep inside) subcluster 4. The third instance is however better predicted, up to 82 minutes in advance. The ‘quality of prediction’ is remarkable, as this event happens to be one of the strongest amplification events in the data, tripling the amount of total delay within about an hour.

Overall, the correct predictions are mainly confined to a certain part of the phase-space as most of the ‘Hits’ are found when the system is in subcluster 10. A lot of false alarms are given when the system is in subcluster 4. Missed alarms can be found far away from the subcluster of interest (mostly in the lower-left part of the phase-space), which makes sense as it is rare that the system moves so quickly from that part of the phase-space all the way towards subcluster 3. We can distinguish five delay peaks in the panels on the right in Fig. 1.8; these are indeed the instances in which the system is inside subcluster 2 or 3.

1.5 Discussion and Conclusions

1.5.1 Summary

For systems whose dynamics are poorly known or strongly heterogeneous (like in many socio-technical systems) we have developed a framework to identify macro-states, and further analyse and predict transitions across them. We have coupled the framework to a year’s data from the Dutch railways, and predict large-scale disruptions.

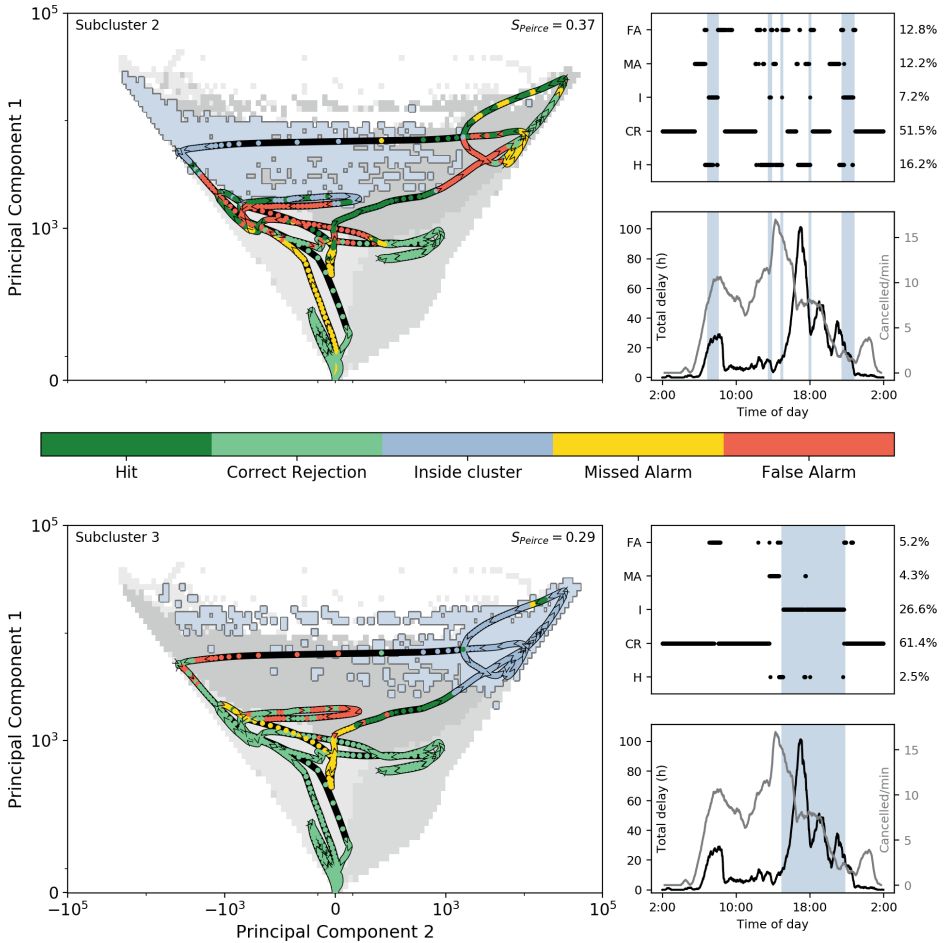


Figure 1.7: Evolution of the system in the phase space on January 3, 2018 in dots (V-shaped arrows show the direction). Dot colouring shows the predictions to (a) subcluster 2 and (b) subcluster 3, as shown by the colour bar at the bottom: ‘Hit’ indicates a correct prediction to enter subcluster, ‘Correct rejection’ means the scheme was correct not to give an alarm, ‘False alarm’ indicates an incorrect alarm due to no entrance of the disrupted subcluster and ‘Missed alarm’ means false negative (alarm should have been given). Subclusters 2 and 3 are marked in light blue (other clusters in various gray shades). On the right, for each cluster the different outcomes in time are stated including percentages and overall score (top) and the total delay (black) and cancelled activities per minute (Gray) is shown (bottom). The vertical blue-coloured columns indicate periods of time inside the subcluster. Parameters used: $p_c = 0.08$, $t_{\text{max}} = 90$ min and $\epsilon = 30$ min.

The framework consists of three steps. The first step consists of a dimensional reduction, based on identifying the relevant patterns to define a reduced phase-space, which, for the railways, has been achieved by a principal component analysis. This choice to determine the relevant patterns is based on both the amount of variance explained, as well as the persistence of the patterns in time, defined by the time-scale for the autocorrelation functions of the principal components. For the Dutch railways, we have found

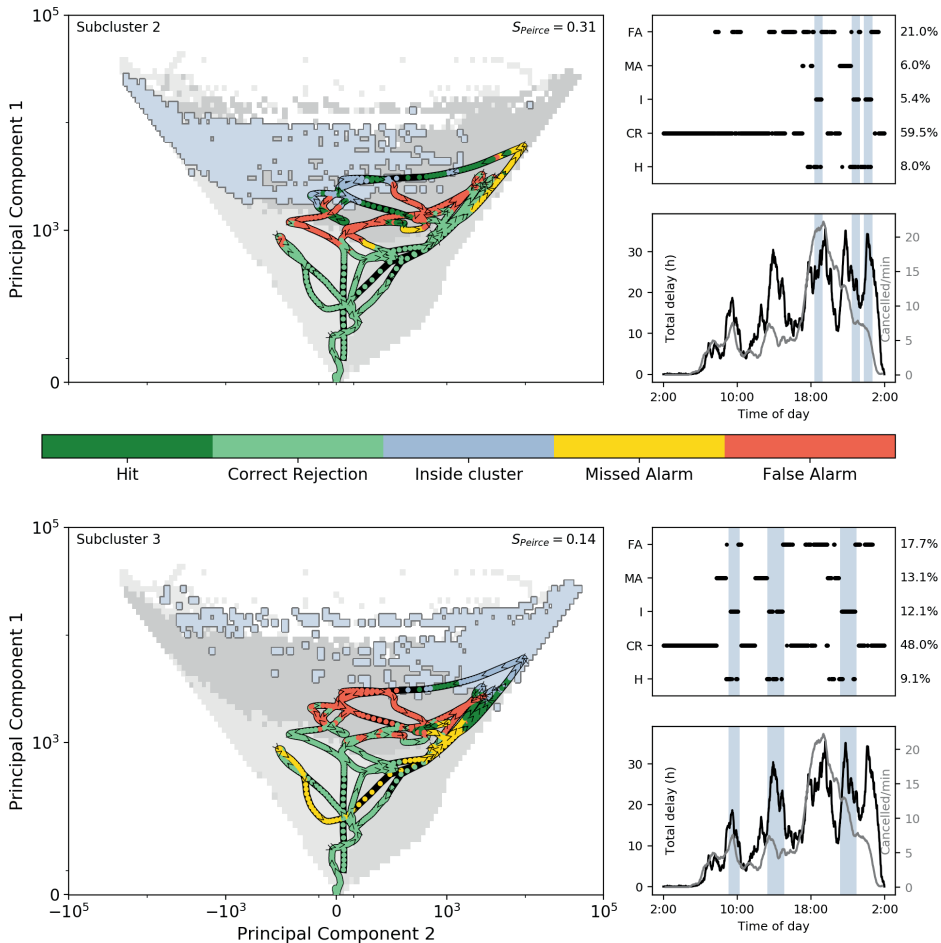


Figure 1.8: Same as in Fig. 1.7, but for the case of April 19, 2018.

that the first two principal components are the most relevant. These components reflect a combined signal of three important international railway lines.

The second step consists of defining macro-states as the (quasi-invariant sub)clusters in the reduced phase-space. We have achieved this by splitting the phase-space into grid cells, calculating the transition matrix with elements consisting of conditional probabilities of transitions between cells, and further applying a clustering algorithm. For the Dutch railways we have found 10 subclusters, which we have divided among ‘rest states’, ‘transition regions’ and ‘disrupted states’ upon analysing the realisations of the system within these subclusters. The average delay patterns per subcluster has led us to distinguish various types of disruptions — one focused on the line from Amsterdam southward via Rotterdam to Belgium, another on the lines from Amsterdam to Germany, and a third showing a combination of other patterns.

The third and final step consists of the prediction of entering specific subclusters. For the Dutch railways, the subclusters of interest correspond to disrupted states. Using conditional probabilities obtained from the transition matrix, we have devised an early warning procedure that, given a certain threshold probability, gives an alarm at a certain time lag. The skill of the alarm procedure has been analysed using the Peirce Skill score. Applying this to the Dutch railway system, we have found reasonable Peirce Skill score towards disrupted states. This reflects the potential of this framework to anticipate macro-state transitions towards disrupted states.

1.5.2 Discussion

Several aspects of our framework need to be discussed. First and foremost, for railways the delays are a combined result of (among other factors) (a) the physical interactions of trains and infrastructure (e.g. a broken train blocking a piece of track), (b) accidents and sheer coincidences (e.g. illness of crew, broken switches or trees falling on tracks), and (c) human influence on the system (passengers, crew, traffic controllers and dispatchers). All these factors cause non-systematic noise in the system. For example, the specific nature and duration of a problem involving a switch in the tracks largely determines its effect on delay, and although there are protocols bordering human decision making (at various levels) in certain circumstances, in practice every situation, person's reaction and their combined effect on the evolution of delay is unique in every situation. In a data-driven framework, all these factors (both systematic and non-systematic) are intertwined, and there is no way to disentangle them. Although this increases the uncertainty, it can be considered a strength of this analysis to not ignore the human impact on the system — after all, human influence is an integral part of the dynamics of the system. Many studies only focus on element (a), micro-simulating only train interactions, which results in an uncertainty in itself when it comes to prediction of macro-scale delay.

Related to the above is the fact that in many ST systems like railways and disease spreading, human control elements (interventions) play a role specifically in the case of macro-state transitions, both to prevent such events or to recover from such events. Especially in railways, it is difficult to distinguish whether control measures have a damping effect on the disruption's spread, or (although necessary in case of disruption) play an amplification role in spreading the disruption. In part, these control elements are correctly incorporated in the transition matrix and clustering: areas in the phase-space where the system is often stabilised (preventing amplification) can be seen as areas that are relatively safe from transitioning. However, it remains difficult to disentangle the cause versus the effect of control measures. They are inseparable from the data used in this study, exert changes on a spatial micro scale rather the macro scale used in this study, and contain discrete and sparse data that is hard to couple to the definition of the dynamic delay variable. More research is needed to develop methods that find the 'laws' underlying the system's evolution in the phase-space (e.g. during the January 3 and April 19, 2018 we have discussed above), distinguishing physical dynamics and human control.

A limiting factor for our framework is the long-term background changes that limits the usage of longer-time datasets. For example, changes in the timetable in railways, or governmental policies in economic markets, may significantly impact the dynamics in

the phase-space, and would be difficult to filter out of longer-time data. This is also the reason we only use one year of railway data in our study.

Also, it is important to note that it is difficult to distinguish development and causality *only* from the dynamics of two co-varying patterns (the EOFs in Fig. 1.2). Our framework merely describes how their amplitudes evolve in time with respect to each other. It may also be the case that when applying this to other systems, the dynamics cannot be captured well by only two principal components, requiring the usage of higher dimensional phase-space. More local and sequential interactions may be recognised before large-scale disruptions occur. Note, for example, that line L1 and L2 dominate the variance (returning in both EOF1 and EOF2), while in practice, L1 is known to be a relatively disconnected line with a large number of local (technical) problems. Incorporating more local effects might give more insight in the dynamics of the system. Using only two principal components, however, does trade precision for more significance in our statistical analysis, as the (more) local features may be quite case-dependent.

Another limitation of our framework is that the (sub)clusters found by the Louvain method are only an approximation of sets that the system is likely to remain within, but not well-defined semi-invariant sets (e.g., see (Tantet et al., 2015)). The problem with railway data is that disruptions are relatively short-lived and the probability density function across the phase-space does not show clearly defined states beforehand, but merely focuses on the area around (0,0). This forces us to divide the phase-space into subclusters as we did, while the system can move repeatedly in and out of them. This is visible in Fig. 1.7 and Fig. 1.8, where in both cases the system leaves the subclusters again in 1-2 hours (and not remaining there longer). Moreover, the transition matrix in the first place is used as a Markovian diffusion model on the phase spaces and calculated without taking into account potential memory of the system. Involving memory in the scheme may give more precise clustering and predictions, for example in the case of the depletion of a buffer in rolling stock or crew because of earlier disruption events happening on the same day.

1.5.3 Implications

Although realising the effects of non-systematic elements and limitations of the framework as described above, it can have strong implications for ST systems in practice.

In general, the framework can be used for any ST system (even for specific non-ST systems where the dynamics are *a priori* not well-known), with the important restriction is that there is enough data of some important dynamic variable. In particular, the framework is useful for heterogeneous spatio-temporal data (e.g., on a network), because it focuses on the dynamics among dominant patterns in space. For example, if one would want to use this to predict the spread of infectious diseases in a particular part of Africa (e.g. related to Liljeros et al. (2003)), one would need enough data of spreadings in the past in that area to find out the most important areas (e.g. cities) that played a role in the spreading, and to build a transition matrix within the phase-space. One could also imagine the evolution of physical systems that are prone to human interaction, for example the evolution certain populations of hunted species, which (if there is enough data) may allow the use of this framework.

Our focus on railway systems brings us to more specific implications for *railway com-*

panies. The framework can be used for real-time monitoring, in which the current position in the phase-space is tracked and the prediction scheme is used to assess the likelihood of entering the disrupted region. If practitioners would use this to anticipate a large-scale disruption, they may act to prevent it. One would start by deriving the dominant principal components. These patterns are an interesting result in themselves, depicting (anti-)co-occurrence of delay in parts of the network as parts with high coefficients (cf. Fig. 1.2). Apparently these patterns are persistent delay configurations and play an important role in the amplification of delay. There may very well be an operational or infrastructural reason behind this.

Specifically, railway traffic controllers can take into account the alarms that are given by the framework (upon construction of the reduced phase-space, followed by the identification of the states). To make the best use of it, an appropriate parameter estimation should be given, which is summarised in answering three questions:

- What time-accuracy should the prediction have? This gives the parameter ϵ : how far off is the prediction allowed to be?
- How far ahead of time should it predict, traded for accuracy? (This yields the parameter t_{\max} .)
- What risk does one want to take? In other words, up to what probability does one want to be sure that the system is not (starting to) amplifying towards a disrupted state? (The parameter p_c is 1 minus that probability.)

Upon having chosen these parameters, one can react accordingly to alarms that are given. It is important to stress that the framework's prediction time horizon is short-term (about 1-2 hours) for the Dutch system. This is both a strength and weakness. The strength is that the framework can be applied real-time, using information of the whole network (rather than delay propagation methods at the micro-scale). The weakness of the short-time horizon is of course related to the limited time left to take intervening measures.

Having discussed the implementation of the framework, the predictions still involve a significant number of false alarms. There is potential to reduce this by incorporating more operations-related quantities like details on personnel, positions of trains and more local features (rather than purely the macro-sized PCs) into the prediction scheme. Another direction of importance to railway practitioners is to make the prediction more specific in space - information about on which train line or part of the network the delay actually propagates can be valuable information to prevent the disruption's spread. Note that now, we only do predictions among the first two PCs, which combined only give a rough estimation of where the delay is currently situated. Nevertheless, we believe this chapter provides a first-order framework to analyse and predict macro-scale delay evolution in an unconventional manner.

CHAPTER 2

CHARACTERISING NEURAL PHASE-SPACE TRAJECTORIES VIA PRINCIPAL LOUVAIN CLUSTERING

This chapter is published as: Dekker, M. M., França, A. S. C., Panja, D., and Cohen, M. X. (2021c). Characterizing neural phase-space trajectories via principal louvain clustering. *Journal of Neuroscience Methods*, 362:109313. Supplementary material to this chapter can be found in SI B.



ABSTRACT

Background With the growing size and richness of neuroscience datasets in terms of dimension, volume, and resolution, identifying spatio-temporal patterns in those datasets is increasingly important. Multivariate dimension-reduction methods are particularly adept at addressing these challenges.

New Method In this chapter, we propose a novel method, which we refer to as Principal Louvain Clustering (PLC), to identify clusters in a low-dimensional data subspace, based on time-varying trajectories of spectral dynamics across multisite local field potential (LFP) recordings in awake behaving mice. Data were recorded from prefrontal cortex, hippocampus, and parietal cortex in eleven mice while they explored novel and familiar environments.

Results PLC-identified subspaces and clusters showed high consistency across animals, and were modulated by the animals' ongoing behaviour.

Conclusions PLC adds to an important growing literature on methods for characterising dynamics in high-dimensional datasets, using a smaller number of parameters. The method is also applicable to other kinds of datasets, such as EEG or MEG.



2.1 Introduction

The dimensionality of neuroscience measurements has increased drastically over the past few decades (Stevenson and Kording, 2011; Bassett and Sporns, 2017), meaning that neuroscientists are now capable of recording simultaneous activity of dozens, hundreds, and even tens of thousands of neurons. Larger and richer datasets provide new opportunities for hypothesis-testing and exploratory discovery, but also challenges in conceptualising and characterising the multivariate signals.

Most traditional data analysis methods in neuroscience are univariate or mass-univariate, such as spike counts (in single-unit studies) or spectral power (in LFP studies). Univariate means that each manifest variable – each neuron or electrode – is treated as a unique measurement point, and statistical analyses and neurophysiological interpretations are based on the functions performed by individual neurons.

Univariate methods have been and remain the backbone of neuroscience data analysis; however, larger-scale datasets might benefit from multivariate analyses that are based on identifying patterns embedded in population activity (Bartolo et al., 2020; Kriegeskorte and Kievit, 2013), which might only be weakly represented in a single neuron or electrode. Multivariate dimension-reduction methods seek to identify a low-dimensional subspace in which the most relevant activity patterns exist (Luczak et al., 2009; Trautmann et al., 2019; Le Van Quyen and Bragin, 2007). Such methods can identify task-related signals that are embedded in the patterns of correlations across the data, which might be undetectable when considering the activity of individual neurons or LFP channels. This is due to neural activity being spatio-temporally synchronised across neurons and circuits, and across multiple spatial scales (Buzsáki, 2010; Varela et al., 2001). The multivariate approach is based on the assumption that neural computations are distributed across ensembles (Buzsáki, 2010), and thus identifying spatio-temporally coherent structure in populations may reveal mechanisms that are not apparent when investigating only individual components of the ensemble.

Although univariate analyses are more common in neuroscience, there are several multivariate analyses that are standard in the field, including principal or independent components analysis, factor analysis, generalised eigendecomposition, and so on (Cohen, 2021a; Pang et al., 2016; Cunningham and Yu, 2014; Makeig et al., 2004). These methods have proven useful at identifying a small number of dimensions in which behaviourally relevant neural dynamics occur.

However, neural data are not static inside those dimensions; they ebb and flow over time with cognitive/behavioural operations, and internal brain states (Mishra et al., 2020a; Venkatesh et al., 2019; Zhang et al., 2015; Linderman et al., 2019). Characterising the time-varying trajectories inside these subspaces is often done visually, which is feasible only in 2 or 3 dimensions.

Here, we introduce a new method for identifying clusters in a low-dimensional data subspace based on spatio-temporal dynamics of continuous trajectories of neural dynamics. The method is based on a combination of principal components analysis (PCA) and Louvain clustering (Dekker et al., 2019), and involves identifying discrete pockets in the principal component (PC) space in which the neural trajectory remains roughly stationary for some period of time. We therefore term this method PLC, for Principal Louvain Clustering. Louvain clustering (Blondel et al., 2008) is a modularity optimisa-

tion algorithm applicable to weighted graphs, and has been used to characterise spatio-temporal dynamics of railway traffic (Dekker et al., 2019), clusters in air transport networks (Guo et al., 2019), diplomatic structures in formal alliance data (Walentek et al., 2021) and atmospheric states (Tantet et al., 2015).

We modified the Louvain clustering method to multisite LFP data recorded simultaneously from the prefrontal cortex, parietal cortex, and hippocampus, of awake behaving mice during novelty exploration. A spectral decomposition of the LFP via a $1/f$ -removed short-time Fourier transform was applied, and a six-dimensional (6D) PC space was constructed based on the largest two components from each region. Each time index during the recording is a ‘brain state point’ in this space, and thus successive time points create a trajectory over time. The Louvain clustering method identifies spatial clusters in which these trajectories remain roughly stationary for a period of time τ . Finally, characteristics of these clusters can be quantified and linked to the mice’ behaviour during the novelty exploration task.

2.2 Materials and methods

2.2.1 Animals, electrodes, and behavioural task

Data are from 11 Black57 background male mice. Non-overlapping results from some of these data have been published elsewhere (França et al., 2021; Cohen et al., 2021). The mice had free access to food and water. All experiments were approved by the Centrale Commissie Dierproeven (CCD), and the surgeries and experiments were conducted according to approved indications of the local Radboud University Medical Centre animal welfare body (Approval number 2016-0079).

Custom-designed and self-made electrode arrays (groups of 50 μm Tungsten wires connected to a PCB) were constructed to target three different regions of the mouse brain. There were 16 electrodes in mPFC (spread in the coordinates AP: 0.5 and 1.5; ML: 0.25 and 0.75; in three columns of electrodes in different depths -2.0, 1.5 and 1.0), eight channels in PAR (AP: -2 and -2.25; ML: 1.0 and 1.75; DV: 0.5) and eight channels in HC (AP -2.5 and -2.75; ML: 1.0 and 1.75; DV: 1.5). Inter-electrode distance was 250 μm and impedances ranged between 0.1 and 0.9 MOhm. The online reference was a metal screw placed on the skull (interparietal bone - AP: -5, ML: 1.0, DV: 0.5). Additional details of the arrays and the manufacturing process are available in França et al. (2020). These electrodes recorded data simultaneously.

For surgery, 10-16 week-old mice were anaesthetised with isoflurane (induction at 5% isoflurane in 0.5L/min O_2 ; maintenance at 1-2% in .5L/min O_2). Mice were fixed in a Neurostar Stereotaxic frame. After shaving, the skin was disinfected with ethanol (70%). Local anaesthetic xylocaine (2%, adrenaline 1:200,000 [AstraZeneca]) was injected subcutaneously at the incision site before exposing the skull. Peroxide (10-20% H_2O_2 ; [Sigma]) was applied to the skull with a cotton swab for cleaning and visualisation of bregma and lambda. Electrodes and screws were fixed onto the skull with dental cement (Super-Bond C&B). Approximately 40 minutes prior to the end of the surgery, saline and analgesic (carprofen injected subcutaneous 2.5 mg/Kg) were injected to facilitate recovery.

As illustrated in Fig. 2.1(a), the experiments involved two sets of 10 minutes, one where the mice were presented with two objects, and another one where one object was

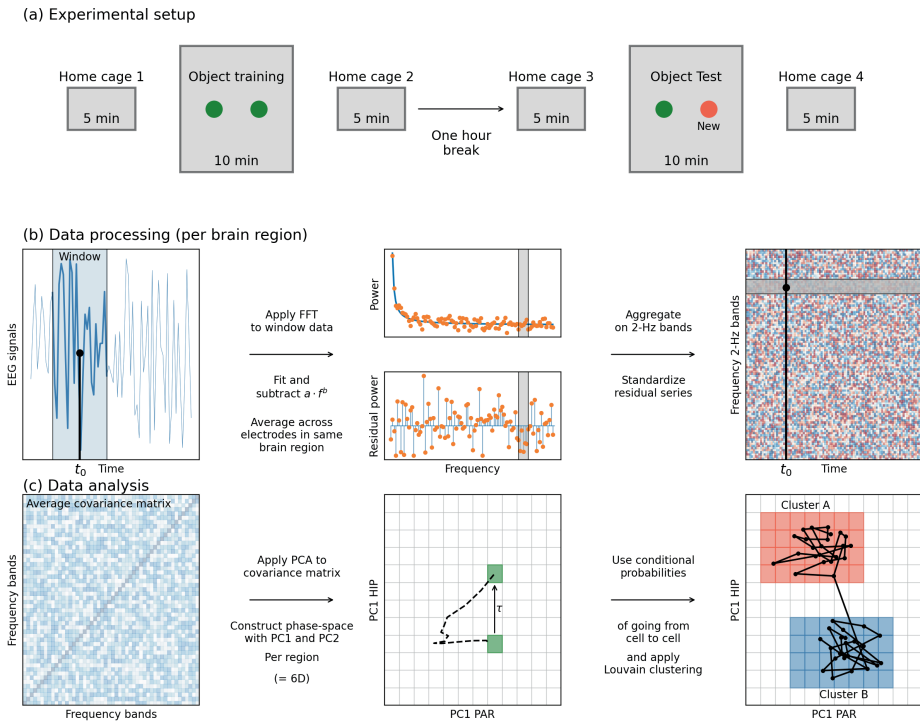


Figure 2.1: **Panel (a):** the experimental setup. Each of 11 mice was held in a home cage for 5 minutes, and then was placed in an arena with two unknown objects (green dots) for 10 minutes. We refer to this as the *training phase*. After this, the mouse was returned to its home cage and was given a break for one hour (including 10 minutes in home cages). Then, the process was repeated except that one of the old objects was replaced with a new one (orange circle). We refer to this as the *test phase*. During both the training and test phases, the mouse's LFP signals in the hippocampus, the prefrontal cortex and the parietal cortex were monitored by 32 electrodes. **Panel (b):** The raw LFP signals were binned into a window surrounding each time step t_0 (blue shaded area), from which the Fourier transform was computed. A $1/f$ function was fitted to the power spectrum, and removed to obtain the residual power spectrum around t_0 . Then, the data were aggregated into 2-Hz bands and normalised, to obtain a plot like in the upper-right, in which the value at each element is a residual power value for in 2-Hz frequency bands at each time step. **Panel (c):** The covariance matrix of the frequency bands was eigendecomposed using principal component analysis (PCA). Utilising the first two components per brain region led to a six dimensional *phase-space*, which we discretised into equal-volume cells. The multichannel spectra at each time point was localised to a cell in this space, and we computed the conditional probability of travelling from one cell to another. The Louvain method applied to the network spanned by these conditional probabilities produces clusters, defined as groups of cells where the system is likely to remain within for some time.

exchanged. More details can be found in França et al. (2020). After the experiments, the mice were euthanized for post-mortem histological confirmation of electrode location. The electrodes in PFC were distributed across anterior cingulate and secondary motor cortex. The PAR electrodes were placed among layers 2 to 5. HIP electrodes were located in the CA1 region, spread in different mice between *stratum pyramidale* and *stratum lacunosum-moleculare*.

Note that although the anatomical targets were the same in all mice, we do not assume that the electrodes have a one-to-one correspondence across mice (e.g., electrode # x in mouse 1 may not be in the same functional location as electrode # x in mouse 2).

During the recordings, the mice performed a novelty-learning task as depicted in Fig. 2.1a. For 10 minutes, they were placed in an arena with two objects that they could explore. Objects were every-day items such as a coffee mug or bath toy. This phase is called the ‘training phase’. After a 60-minute break, they were placed back in the arena for another 10 minutes, and one of the objects was replaced by a new object (orange circle in Fig. 2.1a). In between these recording sessions, mice were placed in their home cage; data from those periods are not reported here.

The mice’ real-time position was continuously monitored via a webcam sampled at 24 Hz and synchronised with the electrophysiological data. Video data were processed in DeepLabCut (Mathis et al., 2018), a software package for markerless pose estimation based on convolutional deep neural networks. 200 randomly selected frames were hand-labelled for the left ear, right ear, nose, and the beginning of the tail. The corners of the objects were also labelled. We used the ResNet-101 network with 200,000 iterations, and visual inspection was used to confirm accuracy of the marker labels in test frames.

Each video frame was given one of three labels with a corresponding variable value ζ : *non-exploration* when all of the mouse’s body markers were outside the boundaries of the object ($\zeta = 1$); *exploration* when any of the mouse’s head markers were inside the polygon derived from the object corners of the objects ($\zeta = 2$); *novel exploration* when any of the head markers were inside the polygon of the novel object during the test phase (orange dot in Fig. 2.1a) ($\zeta = 3$). Exploring the familiar object in the test session was labelled $\zeta = 2$.

2.2.2 Data processing and analysis

Electrophysiology data were acquired using Open Ephys hardware with a sampling rate of 30 kHz. Offline, data were imported into MATLAB, down-sampled to 1000 Hz (by first applying an anti-alias low-pass filter at 500 Hz and then sampling every 30 time points), high-pass filtered at 0.5 Hz, and locally referenced to the average signal from each region. This re-referencing ensured that the signals were locally generated and not volume-conducted from distal brain regions or from the online reference electrode in the skull on top of the cerebellum. The EEGLAB toolbox (Delorme and Makeig, 2004) was used for visual inspection of data quality, and for removing non-neural artefacts (such as mechanical artefacts or line noise) via independent component analysis using the jade algorithm (Cardoso, 1999), as we and others have described previously (Cohen et al., 2021; Khosrasi et al., 2019). On average, 1.7 components (out of 32) were removed per recording.

2.2.3 The PLC method

The overall aim of PLC is to characterise how multivariate brain activity clusters over time. This involves three steps that are detailed below: (1) Preparing the data via spectral time series decomposition, (2) reducing the dimensionality via PCA, and (3) identifying clusters from time-varying trajectories in the low-dimensional subspace via the Louvain method.

All analyses were implemented in Python 3.7.6 using custom-written code that relied on standard Python libraries like `numpy`, `scipy`, `pandas` and `matplotlib`, and in particular the `python-louvain` package, which is used for the clustering (<https://github.com/taynaud/python-louvain>, based on Blondel et al. (2008)). Code will be made available upon acceptance. Below, we continue with a more detailed step by step discussion of the methods.

In the first step, we implemented a time-frequency decomposition via short-time FFT. This is because the time-domain (broadband) signal comprises energy across a range of frequencies, and therefore, a spectral decomposition facilitates identifying the features of the signal.

For the second step, we reduced the dimensionality of these 50 series via PCA, retaining two components per brain region. In total, this produced six principal component time series. The third step involved clustering the 6D PCA space via Louvain method. The clusters represent specific combinations of the activation of frequency bands in the three brain regions that remain pseudo-invariant for some time, which may point to the brain to be in a certain state. In light to the experiment, we will relate these clusters to exploration behaviour of the mice.

Spectral time series decomposition

Data preparation involved obtaining a standardised time series of the relative powers of 2 Hz frequency bands from 0 to 100 Hz in 50 steps (Fig. 2.1b). We began by segmenting the data into 1-second epochs and applied the following procedure:

- Extract the power spectrum via the Fast Fourier Transform (FFT) to the data in the 1-second segment.
- Fit the function $\rho(f) = a f^b$ to the power spectrum of each 1-second segment, and compute the residual power spectrum after removing this best-fit line. This removes the ‘ $1/f$ ’ component of the spectrum, and thus allows our method to leverage the entire spectrum, rather than being biased by increased energy at low frequencies.
- Crop the resulting ‘residual’ power spectrum of this window between the frequencies of 2 and 100 Hz and average the values into bands of 2 Hz (resulting in 50 values over the entire frequency range).

The above procedure produced a matrix of time-by-frequency-by-electrodes (per mouse). Next, we averaged the time-by-frequency values across all channels within each brain region, and obtained an estimate of the dominant (‘ $1/f$ -residualised’) spectral dynamics within each brain region. To remove any biases of certain frequency bands having systematically higher power or power-variance, we z-scored the time series for each frequency.

Principal Component Analysis

We next applied PCA on the residual power spectra over time to obtain a reduced-dimensional representation of the spectral dynamics. We performed the PCA on the time-by-frequency matrices, separately per brain region. The resulting PCs (spectral modes) re-

flect linear combinations of power across frequencies that maximise the variance of their weighted combinations. In other words, the PC time series are wide-band signals, with the contribution of each frequency determined by the PC weights. Also note that these PC time series had a temporal resolution of 1 ms, because the sliding window for the spectral analysis stepped forwards at the same temporal resolution of the LFP data (1 kHz).

In practice, PCA is obtained as the eigendecomposition of the data covariance matrix (Fig. 2.1c). In order to get an aggregated result across the 11 mice, we computed the 50-by-50 average covariance matrix of the time series derived above, per brain region. Since the average covariance matrix is symmetric, the PCs constitute an orthonormal basis. The number of PCs to preserve for analyses was selected based on visual inspection of the eigenvalue spectrum (scree plot), which encodes the amount of variance explained by each PC.

Note that although we performed the PCA separately for each brain region, the data for each brain region were acquired simultaneously. This allowed us to localise the data to a reduced PC space that spans all three regions. Using these different dimensions, we constructed a 6D space with the dimensions being ($PC1_{HIP}$, $PC2_{HIP}$, $PC1_{PAR}$, $PC2_{PAR}$, $PC1_{PFC}$, $PC2_{PFC}$). While not comprising the whole principal component space, we will refer to this reduced 6D space as the *phase-space*. When the data for each mouse at every time point are projected onto this six-dimensional phase-space, it creates a brain state point corresponding to that time point in the phase space. Connecting these points sequentially in time for a specific mouse then produces the trajectory for the state of its brain in the phase-space.

Louvain clustering

Louvain clustering (Blondel et al., 2008) identifies quasi-stationary pockets of the phase-space – *clusters* – in which the data trajectory remains roughly stable for some period of time. Louvain clustering is a graph clustering method and is based on the concept of modularity, which measures the density of links within the clusters, relative to the density of links between clusters. Weights of links are taken into account in the definition of modularity. The method can be applied to find clusters based on the trajectories in the PC space by discretising the space into cells. In particular, we gridded the phase-space into 9^6 (531,441) cells, with 9 cells in each dimension, equally spread in the domain $[-12, 12]$ for all brain regions – HIP, PFC and PAR. (The value 12 as domain boundaries corresponds to the maximum absolute value of the PCs for the entire time series across all brain regions, which is just above 11.5.) We then constructed a matrix M with the following entries

$$M_{ij} = P(i \rightarrow j \text{ after } \tau \text{ steps} \mid i), \quad (2.1)$$

which denotes the conditional probability of the trajectory to move from cell i to cell j after time τ . In other words, matrix element M_{ij} stores the probability of a jump from cell i to cell j over a time interval τ . Here, we used $\tau = 30$ ms. M is a transfer matrix, and can be used to statistically simulate the time evolution of the state of the mouse brain in phase space (Dekker et al., 2019). It can also be seen as a directed network with weighted edges denoting the conditional probability of transfer from one node to an-

other in a given lag time: nodes in the network are then the phase-space cells in the phase space. Stated differently, Louvain clustering provides an algorithm to identify clusters of phase-space cells in the phase-space in which the mouse brain is (statistically) likely to remain.

The neural trajectories in the phase space are continuous, meaning that exiting one cluster means entering (or passing through) a neighbouring cluster. In contrast, a coarse time resolution or larger τ can cause the empirical trajectories to seem discrete and jump from one cell to a distant one. Further, noise can cause a trajectory to exit and rapidly re-enter the same cluster. We therefore smoothed the PC time series data using a Gaussian with a full-width-half-maximum of 30-ms (selected to match τ).

The clusters are defined by analysing spatio-temporal modularity of the neural trajectories in the phase space. In particular, clusters are regions (contiguous cells in the phase space) in which the trajectories remain roughly stable for some period of time, relative to the time it takes to transit to another cluster [Fig 2.1(c)]. This is the goal of the Louvain algorithm applied to matrix M using time parameter τ . In initial explorations with pilot data, we found that modularity generally and smoothly decreased with increasing τ . Furthermore, the locations and boundaries of clusters were qualitatively comparable with τ 's ranging from 5-300 ms. Here we selected $\tau = 30$ based on the approximate time windows of spike smoothing (Lehky, 2010). On the other hand, we acknowledge that this parameter selection is somewhat arbitrary, and it is possible that different features of the data would be highlighted by different values of this parameter. In section 2.4.3, we discuss this choice in more detail.

Cluster features

The Louvain clusters are defined based solely on the neural trajectory data. In order to characterise these clusters and relate them to mouse behaviour, we computed the following attributes for each cluster. Some of these features can also be calculated per mouse, but we computed them as aggregated over all datasets.

- *Exploration bias*, a normalised measure of the percentage of time the mice spent exploring the novel object while the brain was in each cluster. Comparing the exploration percentage while in cluster i to the overall exploration percentage yielded its *exploration bias*, defined as

$$\zeta_b(\text{cluster } i) = \frac{\text{percentage exploration in cluster } i}{\text{overall percentage exploration}}. \quad (2.2)$$

If $\zeta_b(\text{cluster } i) > 1$, then the mice explored the object for longer while in brain-state cluster i than they do on average over the whole phase-space. If this number is significantly higher than 1 across the group of mice, the mice explore more often than on average when their brain activity's position in the phase-space places them in cluster i , which, via the eigenvectors shown in Fig. 2.2, can be reconstructed as certain combinations of residual spectra being activated or covarying during exploration.

- *Absolute exploration bias*, defined as $\zeta_{\text{abs}} = \text{abs}(\zeta_b - 1)$. This is the distance of ζ_b from one, and depicts the general strength of the bias (exploration or non-explora-

tion).

- *Average residence time* of state of the mouse's brain within each cluster. Note that since we are working in phase-space, a cell in a given cluster might share boundaries with cells that do not belong to the same cluster. It is possible that noise can make the neural trajectory to rapidly exit-and-reenter a cluster within a few ms. We consider these transient fluctuations to be spurious, and we therefore introduced an additional 30-ms [aligned with the τ in Eq. (2.1)] allowance period to keep count of the residence time if the state of the brain exits and enters a cluster within this time frame. We also omit residence times of 1 or 2 ms in this statistic.
- *Interregionality*, which indicates whether the cluster has more variance in some brain regions compared to others. This indicates whether the cluster is dominated by one region, or whether it comprises dynamics from two or three regions. We defined interregionality (IR) of cluster i as

$$\text{IR}(i) = \frac{\max[\text{var}(\text{in } i, \text{HIP}), \text{var}(\text{in } i, \text{PFC}), \text{var}(\text{in } i, \text{PAR})]}{\text{var}(\text{in } i, \text{HIP}) + \text{var}(\text{in } i, \text{PFC}) + \text{var}(\text{in } i, \text{PAR})}. \quad (2.3)$$

Note that $\text{IR} \in [1/3, 1]$; the lower limit $1/3$ corresponds to the case where the variance is equally spread across all three brain regions, whereas $\text{IR}=1$ indicates that the variance is exclusively concentrated within one brain region.

- *Mouse specificity*, which indicates whether a cluster is defined from one or a small number of mice, vs. whether data from most or all mice contributed an equal number of data points. Mouse specificity was computed by first calculating the percentage $p_m(i)$ of cluster i 's data points that are from each of the 11 mice m . Then the mouse specificity is defined as $\sum_m \left| \frac{1}{11} - p_m \right|$, i.e., as the sum of the absolute differences from $1/11$.
- *Cluster magnitude*, the average distance of data points in the cluster to the origin of the phase-space. Distance to the origin was measured by computing the 6D Euclidean distance, weighing each dimension equally.

2.2.4 Statistics

Statistical robustness was checked in multiple respects. To check whether random exploration data can reproduce the same ζ_b distribution across the clusters, the behavioural time series of exploration were shuffled (without changing the brain or PC data), and the above metrics were recomputed. The shape and number of clusters themselves was checked against random permutations by shuffling the PC data, while preserving the internal 6D coordinates (i.e., only shuffling the time sequence of these 6D data points). A final statistical check was the universality of the results across the mice. To this end, we determined ζ_b for each of the eleven mice and performed a Wilcoxon signed rank test. The Wilcoxon test determines whether any sample \vec{x} is symmetrically distributed around zero. In our case, we test whether it is symmetric around one (testing $\vec{x}' = \vec{x} - 1$). It is different from the regular student's t -test in the sense that the Wilcoxon test is non-parametric – it focuses on the signed ranks, rather than the actual values of the differences. In our

case, we used it to test the hypothesis that the novel exploration was consistently in the same clusters across animals.

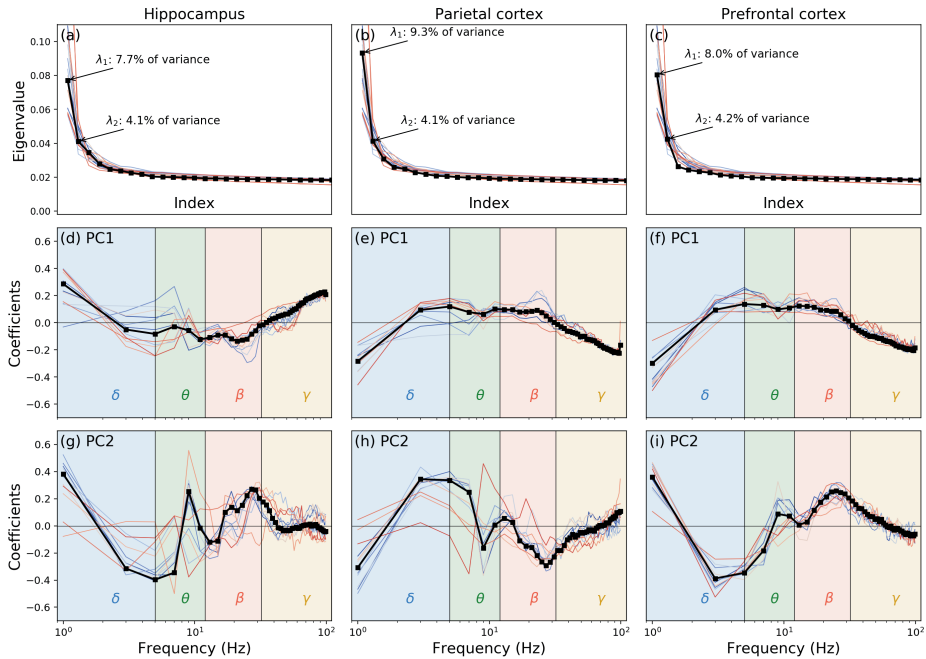


Figure 2.2: Panels (a)-(c): eigenspectra of the three brain regions. Panels (d)-(i): the top two PCA eigenvectors. For all panels: black indicates those obtained from the average covariance matrix across all eleven mice, while coloured lines indicate individual mice. The background shading in the bottom two rows indicates approximate canonical spectral boundaries for δ -, θ -, β - and γ -bands. These bands are purely for reference; the analyses was based on the full spectrum up to 100 Hz, not on individual bands.

2.3 Results

2.3.1 General

Across all eleven mice, the average percentage of time spent exploring the novel object was 16.4%. The individual numbers were 15%, 9%, 16%, 9%, 23%, 16%, 13%, 16%, 13%, 37% and 14%. The parameters a (the scaling factor) and b (the exponent) of the fits subtracted from the power spectra in each time step were $a = 0.01 \pm 0.004$ and $b = -0.36 \pm 0.09$. The averages and standard deviations were calculated over all channels, mice, brain regions and parts of the experiment (training + test).

2.3.2 PCA eigenvectors

Inspection of the eigenvalue spectra from the PCA on the residual time series per brain region and per mouse indicates that a large amount of variance in each dataset can be

explained only by two PCs (Figure 2.2). In the interest of convenience and comparability, upon visual inspection, we retained only the top two PCs from each brain region per mouse.

Remarkably, the PC eigenvectors for each mouse exhibited strong consistency across the mice, as seen in Fig. 2.2. This motivated us to compute the average covariance matrix for all the mice together, and then rerun the PCA on this average matrix. The resulting PC eigenvectors are displayed in the same figure in black. We used the latter as the reference set of axes onto which we projected the brain state data for all mice. The advantage of this approach is that it provides us with a common phase space for all mice for performing our analyses.

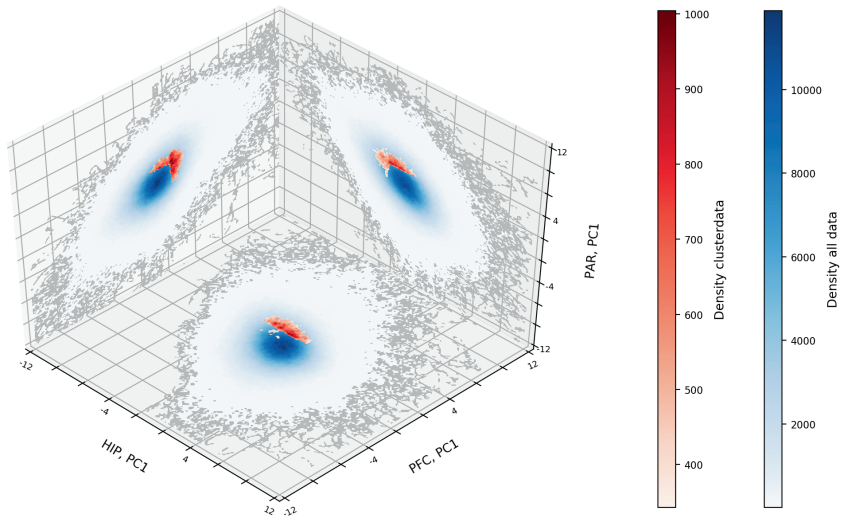


Figure 2.3: Example cluster found by applying Louvain clustering to the matrix M . The axes are the first PCs from each region. In other words, this shows only 3 of the 6 dimensions of the PC space. Blue indicates the distribution of all data, red indicates the distribution of the data of one example cluster. For visualisation purposes, the outer edges of the data clouds are marked in grey. In the background, the resolution of the Louvain clustering is marked (9×9).

2.3.3 Clusters

Projecting the brain state data onto the 6D ($PC1_{HIP}$, $PC2_{HIP}$, $PC1_{PAR}$, $PC2_{PAR}$, $PC1_{PFC}$, $PC2_{PFC}$) phase-space produces a time series of the 6D coordinates per mouse. In total, data from eleven mice produced 13,644,532 data points. Gridding the 6D space produced 2417 non-empty cells (c.f. the maximum possible 1,000,000 cells). In the calculation of the conditional probabilities for the brain state to transfer from one cell to another, used as entries in matrix M , we made sure that only movements between cells within the data of a single mouse are used. The Louvain clustering on matrix M identified 60 clusters.

An example cluster is visualised in Fig. 2.3. The axes are from the first PC of each brain region. The blue colours illustrate the distribution of all data on these axes, and the

red colours illustrate the density of the example cluster. Note that the data have higher concentration towards the origin of the phase space. This particular cluster is index #4, contains 273 cells, and comprises 529,167 data points (about 4% of the data).

2.3.4 Cluster attributes

We start investigating the general properties of these clusters by computing the metrics discussed in section 2.2.3. Fig. 2.4a shows the correlation matrix (Spearman) for the cluster metrics described in the Methods section.

We focus on the exploration bias and the absolute bias, because these two metrics are linked to the animals' behaviour. Exploration bias was largely uncorrelated to the other metrics. Absolute exploration bias, however, was negatively correlated with the amount of data: clusters with more data had exploration biases closer to 1 (see also Fig. 2.4b). This is interesting to relate to the positive correlations between absolute exploration bias and residence time and distance to the origin: Clusters in which trajectories remained for longer periods of time – and clusters that were further away from the origin of the PC phase-space – were more likely to be modulated by behaviour. One can think of these peripheral clusters are requiring more energy in the PC space to push away from the origin, and may therefore reflect brain states that deviate from the ongoing state. In contrast, the clusters located towards the centre of the PC space contain much more data per cluster, and also more transitions across clusters.

Fig. 2.4c-e shows distributions for three of the cluster properties. The residence times ranged from tens of ms to 700 ms, but were concentrated around 100 ms, with an average of 140 ms. The observation that neural trajectories tend to remain in a stable state for a few hundred ms before transitioning to another stable state is consistent with previous findings in the LFP and EEG literature (Michel and Koenig, 2018; Luczak et al., 2009; Mishra et al., 2020b). In the interregionality and mouse specificity, more spread is observed in the histograms, reflecting that some clusters were of a 'universal nature' in terms of how well they were represented across mice and across the three brain regions, but some were not. A relatively high abundance of lower values in both metrics reflects that for the larger fraction of the clusters, dynamics are shared across the animals and reflect coordination across the brain.

2.3.5 Exploration data and Wilcoxon test

To investigate the significance of the exploration bias, we computed ζ_b for all clusters, both for individual mice and for the group-aggregated data. Results are shown in Fig. 2.5a (squares for individual data and circles for aggregated data). We sorted the clusters by ζ_b of the full dataset. Around half of the clusters had below-average exploration bias. Most clusters showed individual variability across animals (note that the aggregated results can be considered a weighted average, because not all animals had the same amount of data per cluster). Panel (c) shows that clusters also varied in the total amount of data per cluster, with the clusters exhibiting strong exploration biases having overall less data.

We performed three sets of statistical evaluations on the exploration bias results. First, we tested the individual data against a bias score of 1 (corresponding to no exploration bias), which tests the consistency of the sign of the bias across animals. Only

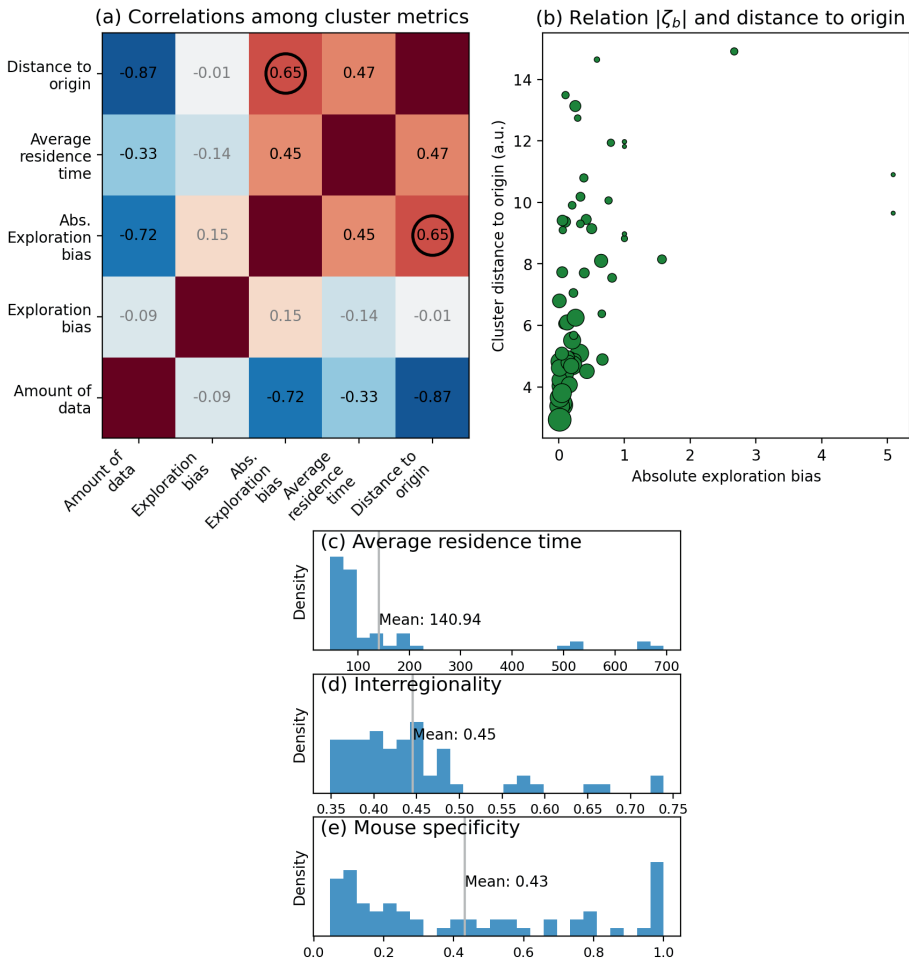


Figure 2.4: **Panel (a):** Spearman's correlation matrix of relevant variables. Negative correlations are denoted in blue, positive correlations are denoted in red. Numerical values of correlations are shown in black for those with correlation significance level $p < 0.05$, other (non-significant) correlations are shown in grey. **Panel (b):** Cluster distance to origin versus absolute exploration bias, corresponding to the encircled Spearman's correlation of 0.6 in panel (a). Scatter marker size in panel (b) corresponds to the number of data points. For reference, we denoted two clusters with their respective number of data points. **Panel (c)-(e):** Histograms and associated mean values of (c) average residence time (in ms), (d) interregionality (between 0.33 and 1, higher means more variance in a particular brain region) and (e) mouse specificity (between 0 and 1, higher means more dominated by a smaller set of mice).

animals with at least 2% of their data in the particular cluster are used in this analysis (Wilcoxon rank-sign test), as accounting for mice with smaller amounts of data may results in non-representative values of the exploration bias in these clusters. Results are shown in panel (e), with circles indicating data points of $p < 0.2$. Four clusters exceeded a $p < 0.05$ threshold. On the one hand, this is a somewhat lenient threshold and was un-

corrected for multiple comparisons, but this result is consistent with the cross-animal variability in bias scores seen in panel (a). On the other hand, with $N = 11$, a Wilcoxon sign-rank test can only achieve $p < 0.05$ with at least 10 data points in the same direction. We complemented this test using permutation testing. Visually comparing panels (c) and (e) may lead to the conclusion that the exploration significance of the clusters is driven by the amount of data in the cluster. We checked this by performing a biserial rank-correlation between the (dichotomous) significance and the amount of data, resulting in a correlation of 0.22, contradicting this idea.

The second test was based on shuffling the data to determine whether our results could have been observed in random data in the same reduced PC space. To this end, we scrambled the PC time series (which shuffles the links between phase-space cells), a re-applied Louvain clustering. Results are shown in Fig. 2.5a,d,f. Here we observed only 28 clusters with uniform amount of data and clustering bias. Note that in this process we did not shuffle the exploration data. Because the resulting clusters are now randomly situated in the phase-space, and the amount of data in these clusters is uniform, the exploration data of any mouse is perfectly spread across all clusters, resulting in a near-invariant exploration bias per individual rodent [visible in horizontal dotted ‘lines’ in panel (b)]. Note also that in all cases, the p -values of the shuffled data were larger than the p -values of the empirical data labelled as significant.

The final set of statistics involved shuffling the temporal mapping between behaviour and brain. We cut the behaviour time series at a random time point and swapped the second for the first segment, and then recomputed ζ_b for each cluster. Note that this procedure preserves both the LFP data and the behaviour data, only randomising their behaviour with respect to the LFP signal. This procedure was repeated 100 times, and the grey shaded region in Fig. 2.5(a) illustrates two standard deviations around the mean shuffled exploration bias.

Overall, these tests confirm that the clustering structure found in the data was not observed when clustering noise, and that there is individual variability in the specific clusters in which animals investigated the novel object that is not accounted for by random reshuffling of the behavioural data.

2.4 Discussion

2.4.1 Multivariate neural dynamics

Neural activity is diverse and dynamic over space, time, and frequency (Buzsáki and Llinás, 2017; Varela et al., 2001; Bassett and Sporns, 2017). Characterising these dynamics remains a major challenge in neuroscience. Here we adopted a recently developed method (Dekker et al., 2019; Tantet et al., 2015) to define low-dimensional phase-space clusters that allow for characterising and exploring spatio-temporal dynamics of neural trajectories in multi-site electrophysiology. An advantage of PLC is that it allows for a mix of linear (PCA) and nonlinear (Louvain clustering) methods to identify trajectory-based clustering in a reduced-dimensional space.

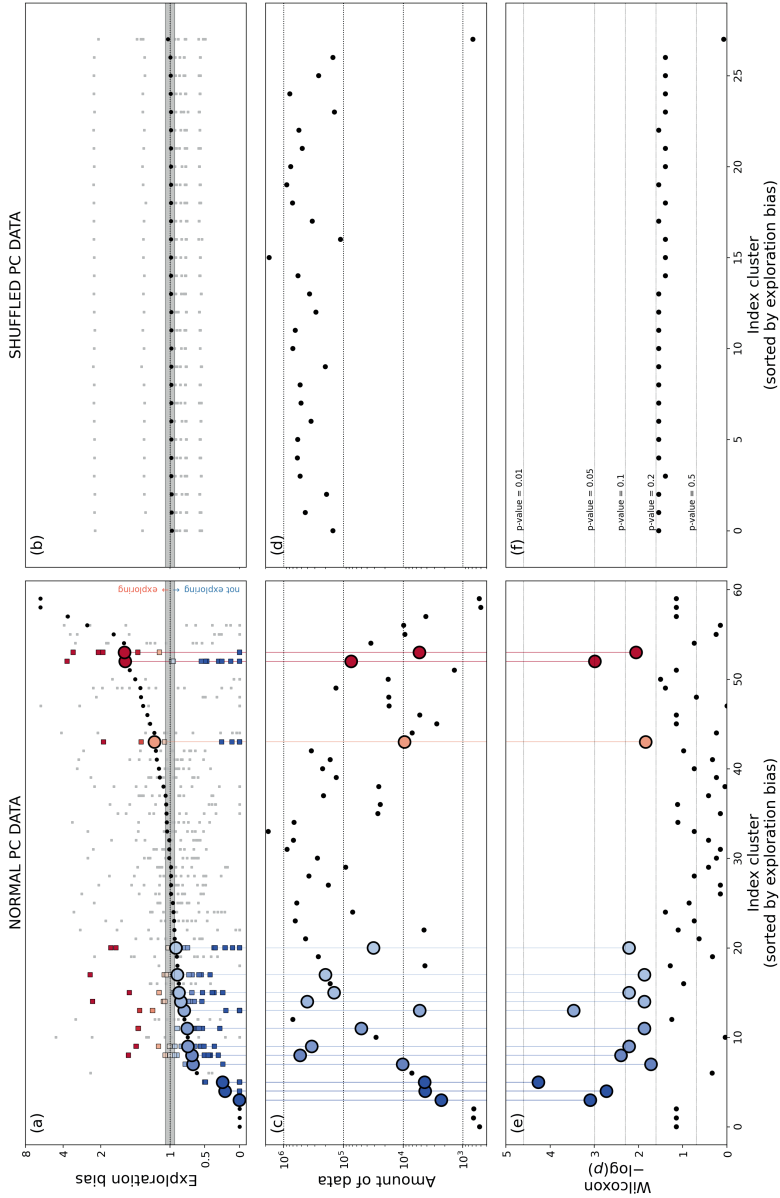


Figure 2.5: Attributes of the PC clustering (left column) and the clustering applied to shuffled PC data (right column). Exploration bias for every cluster (**panels a-b**) is compared to the cluster's respective data amount (**panels c-d**) and the Wilcoxon p -values, that test the consistency of exploration bias values across the 11 animals (**panels e-f**). Clusters with $p < 0.2$ on the left hand side are highlighted, all other clusters are denoted by black dots. Squares in panels a-b (coloured and grey) show the respective individual biases per mouse. Grey shaded areas in panels a-b indicate the estimated exploration bias range (mean $\pm 2\sigma$) among 100 times shuffling the exploration data without changing the LFP data. Note that some clusters in panels (a), (c) and (e) on the left only contain few data and non-exploration data, resulting in overlapping points in panel (a).

2.4.2 Consistency of PCs across mice

It is striking that the eigenvectors of the top two PCs were so similar across the eleven mice. This is not a trivial result due to biased selection, because we simply took the top two components rather than, e.g., picking the components that maximised a cross-mouse correlation. It is also not a trivial result of the $1/f$ -like nature of the power spectrum, because this feature was removed prior to the covariance matrix generation.

Instead, we interpret this result to indicate that there are fundamental characteristics of this cortical-hippocampal network that are (1) concentrated in a lower-dimensional space, and (2) conserved across mice. That the dynamics are concentrated in a subspace of the total data space is often reported in multichannel neural recordings (Luczak et al., 2009). This supports the notion that neural computations are distributed over space and frequency, and thus the motivation for investigating weighted combinations of data features (such as electrodes and frequencies). The similarities of the PCs across animals is likely due to intrinsic anatomical and architectural patterns that govern the local neural dynamics. On the other hand, the PCs were not fully identical across mice, and the residual variance likely comprises a combination of unique factors and sampling variability.

Whether to apply multivariate decompositions to data pooled over individuals is debated in the neuroscience literature (Calhoun et al., 2009; Parra et al., 2019; Cohen, 2021b), and leads to a trade-off between increased generalisation vs. increased sensitivity to individuals.

We decided to use the same PC space for all mice in order to have a common (for all mice) phase-space, which facilitated pooling the cluster characteristics across all mice. The mouse specificity in Fig. 2.4e indicates that there are many clusters in which most or all mice are well represented – indicating that inter-regional brain dynamics of distinct mice are situated in the same clusters – and there were also several clusters with near-1 values of mouse specificity, indicating that some clusters were unique to individual mice.

In this analysis, we have chosen to work with the top two PCs in all three brain regions. For the dataset we use here, these components turned out to be dominant in terms of explained variance, but this is not a trivial or universal result. The PLC method is therefore not strictly bound to the choice of which PCs to use. Depending on the data, it is possible that it is more appropriate to use three in each brain region, for example.

2.4.3 Interpretation of clusters and usage of Louvain clustering

Over time, the trajectory of the brain state moves around in the phase-space. These trajectories are not random, and instead remain within relatively confined ‘provinces’ for periods of time. Louvain clustering identifies these periods of relative stability as clusters. A key parameter of Louvain clustering is τ , which is the time-scale at which a trajectory is considered stationary. In our pilot analyses with data from one mouse (not shown here), we found that the clustering was stable for a range of τ values, and thus we selected $\tau = 30$ ms in the interest of consistency. It is possible that different features of the data would be highlighted for larger differences in this parameter, e.g., over hundreds of ms or seconds. Determining how trajectories and clustering are related across temporal scales is an interesting avenue for future research to explore.

Given that the PLC method is several steps away from the raw data, one might wonder whether these clusters are meaningful, or simply reflect residual noise. We addressed

this in two ways. First, Louvain clustering applied to shuffled data showed qualitatively distinct patterns of results compared to the real data (e.g., Fig. 2.4). Indeed, clustering shuffled data produced only 28 clusters, compared to the 60 clusters in the real data, even though the amount of data was the same.

Second, we examined the relationship between the cluster characteristics and exploration behaviour. The results shown in Figs. 2.4 and 2.5 suggest that clusters close to the centre of the PC phase-space were generally not modulated by behaviour, whereas clusters further away from the origin had exploration biases that deviated from 1. These findings suggest that the brain space trajectories associated with active behaviour are ‘high-energy’ states that are maintained for relatively longer periods of time.

On the other hand, there was also considerable individual variability in the exact exploration bias scores over the different animals, evidenced by the relatively large p-values in the Wilcoxon test (although one must keep in mind that with $N=11$, the Wilcoxon test can only return a $p<.05$ result if 10/11 datapoints have the same sign). Some clusters had relatively little or no data for some animals. This finding raises the question of whether the clustering should have been done on individual animals instead of on the group data. As we wrote earlier, group-defined clusters have several advantages in terms of comparability of cluster findings and universality of the spectral-temporal structure of the data. Nonetheless, these are analysis choices that users can determine on their own; our primary purpose here was to illustrate the utility and interpretability of the PLC method.

The quadratic relationship between exploration bias and distance to the phase-space origin is highlighted by the significant correlation with *absolute* exploration bias. Thus, although the direction of the exploration bias score varies across animals and clusters, there is in general more behavioural variability as the data move towards the periphery of the phase-space. It is possible that the centre of the space reflects brain states of relatively low energy or of transitions between states of higher energy. On the other hand, one must keep in mind that the outer edges of the phase-space are relatively sparse, which means that the clusters had overall less data. This can become statistically problematic in the extreme case of, for example, a cluster comprising only 50 data points from a single mouse that is exploring the object for the entire time window.

2.4.4 Limitations and future directions

The PLC method is not without limitations. As mentioned above, clusters with relatively little data, or that are dominated by a single animal, can provide cluster characteristics that may not be fully representative of group dynamics. In the present application, the higher density of data towards the centre of the phase-space means that clusters at the outer edges of the data cloud have overall less data. This is apparent in Fig. 2.5c, which shows differences in densities of clusters spanning four orders of magnitudes. We chose not to remove any clusters based on total amount of data, although we did exclude animals from group-level t-tests if those animals had too little data in that cluster.

A second limitation is the operational definition of a cluster in noisy data such as LFP. Small and fast jumps out of, and then back into, a cluster may reflect noise or a true rapid fluctuation between states. This would manifest as low residence times. We chose to filter out jump-but-return trajectories that were faster than 30 ms.

Third, the physiological meaning of PLC-derived clusters remains unknown. We spec-

ulate that the brain remains in a particular state while in these clusters, similar to the interpretation of EEG microstates (Michel and Koenig, 2018; Luczak et al., 2009; Mishra et al., 2020b). This could be investigated more in the future, for example by investigating spiking dynamics in different LFP-defined clusters. The behavioural relevance of the clusters is also unclear; it is possible that a more easily interpretable link to behaviour would result from a simpler experiment (e.g., simple visual responses in anaesthetised animals), but we choose for a more naturalistic and therefore more complex environment.


In this chapter, we focused on identifying and characterising the clusters; the data provide additional richness that we did not fully explore. For example, the trajectories exhibit changes in speed, curvature, and direction over time, and may exhibit spatio-temporal oscillations within and across clusters. These dynamics might be related to neural computations and behaviour, and might vary systematically during stimulus or motor processing (Linderman et al., 2019; Glaser et al., 2020).

The PLC method is general, and can be applied to many additional tasks or behaviour elements. It can also be applied to larger-scale measurements such as EEG and MEG. Depending on the application and the data, it may be the case that the phase-space dimensions have to be adapted, to, for example, focus on a single brain region, or more or fewer principal components.

CHAPTER 3

CASCADING TRANSITIONS IN THE CLIMATE SYSTEM

This chapter is published as: Dekker, M. M., Von Der Heydt, A. S., and Dijkstra, H. A. (2018b). Cascading transitions in the climate system. *Earth System Dynamics*, 9(4):1243–1260. Tables 1 and 2, section 3.3 (including figures) and supplementary material of the original paper are moved to SI C.



ABSTRACT

We introduce a framework of cascading tipping, i.e. a sequence of abrupt transitions occurring because a transition in one subsystem changes the background conditions for another subsystem. A mathematical framework of elementary deterministic cascading tipping points in autonomous dynamical systems is presented containing the double-fold, fold-Hopf, Hopf-fold and double-Hopf as the most generic cases. Statistical indicators which can be used as early warning indicators of cascading tipping events in stochastic, non-stationary systems are suggested. The concept of cascading tipping is illustrated through a conceptual model of the coupled North Atlantic Ocean - El-Niño-Southern Oscillation (ENSO) system, demonstrating the possibility of such cascading events in the climate system.



3.1 Introduction

Earth's climate system consists of several subsystems, e.g. the ocean, the atmosphere, ice and land, which are coupled through fluxes of momentum, mass and heat. Each of these subsystems is characterised by specific processes, on very different timescales, determining the evolution of its observables. For example, processes in the atmosphere occur on much smaller timescales than in the ocean; hence, in weather prediction the upper ocean sets the background state for the evolution of the atmosphere. Similarly, in equatorial ocean–atmosphere dynamics associated with the El-Niño-Southern Oscillation (ENSO) phenomenon, the global meridional overturning circulation can be considered a background state, as it evolves on a much larger timescale.

This notion that one subsystem provides a background state for the evolution of another subsystem is important when critical transitions are considered. In the climate system, a number of tipping elements have been identified (Lenton et al. (2008) give an overview of these), where changes in observables can occur relatively rapidly compared to the changes in their forcing near so-called “tipping points”. Examples of tipping elements are the Atlantic meridional overturning circulation (AMOC) (Stommel, 1961), the Arctic sea ice (Bathiany et al., 2016), monsoon patterns, mid-latitude atmospheric flow (Barriopedro et al., 2006), vegetation cover (Hirota et al., 2011) and more local systems like coral reefs and permafrost. When one subsystem undergoes a transition, which changes the background state of another subsystem, a transition may also be induced in that second subsystem. Such dynamical interactions which lead to coupled transitions are examples of ‘tipping cascades’ or ‘domino effects’ (Kriegler et al., 2009; Lenton and Williams, 2013).

Many tipping points have been analysed in separate subsystems, both for phenomena of the present-day climate (Lenton, 2011; Bathiany et al., 2016), and for past climates (such as the abrupt cooling of the Younger Dryas, Livina and Lenton (2007), and the desertification of the Sahel region, Kutzbach et al. (1996)). However, less attention has been paid to the interaction between transitions in different subsystems. For example, when the AMOC collapses, precipitation patterns may change such that the equilibrium structure of the vegetation cover in the Amazon rainforest is shifted (Aleina et al., 2013). This may result in another transition, concerned with forest growth or dieback. Another example is the influence of the AMOC on the trade winds (through meridional sea surface temperature gradients), which in turn influence the amplitude of ENSO. In models, a collapse of the AMOC has been found to intensify ENSO (Lenton and Williams, 2013; Timmermann et al., 2007; Dong and Sutton, 2007), although there are also other effects that would weaken ENSO (Timmermann et al., 2005).

An example in past climates is the coupling between the ocean's overturning circulation and land ice. The rapid glaciation of the Antarctic continent around the Eocene – Oligocene boundary (34 Ma) is often explained in terms of a CO₂ threshold being reached that allowed a major ice sheet to grow (DeConto and Pollard, 2003; Gasson et al., 2014). However, a two-step signal is found in the oxygen isotopic ratio, $\delta^{18}\text{O}$, which is attributed to a deep-sea temperature drop followed by the (slower) growth of the Antarctic Ice Sheet (AIS). One suggestion to explain the two-step transition is that the deep-sea temperature drop was related to a change in the pattern of the global MOC (Tigchelaar et al., 2011). The ice sheet formation is then argued to have been driven by decreasing atmospheric

CO₂ (Pearson et al., 2009). The switch in MOC (first tipping) led to the changes in the atmospheric CO₂ (e.g. Elsworth et al. (2017)) which caused the growth of the AIS (second tipping). This leads to the question of whether a cascading tipping event occurred.

In the last few years, much work has been carried out to formulate statistical indicators and early warning signals of tipping points. A system close to a critical transition shows features of a “critical slowing down” (Dakos et al., 2008; Scheffer et al., 2009; Kuehn, 2011). In the vicinity of the tipping point, the system slowly loses its ability to recover from small perturbations. This results in increased variance, auto-correlation and potentially also increased skewness and flickering (Scheffer et al., 2009). Various methods providing a specific scalar and a threshold when approaching the transition have been suggested, such as degenerate fingerprinting (Held and Kleinen, 2004; Thompson and Sieber, 2011) and detrended fluctuation analysis (DFA) (Peng et al., 1994; Livina and Lenton, 2007).

When considering cascading tipping points, the auto-correlation of two time series and their interaction need to be analysed simultaneously. Podnobik and Stanley (2008) proposed an altered form of DFA to assess the cross-correlation between two non-stationary time series and called this method detrended cross-correlation analysis (DCCA). In the computation of the fluctuation function, they used cross-covariance instead of auto-covariance and fit this to a power law. This concept is further extended by defining a coefficient ρ_{DCCA} that accounts for the auto-covariance of the individual time series (Zhou, 2008; Yuan et al., 2015). However, no statistical analysis or indicators have yet been formulated for cascading tipping events.

In this chapter, we provide a quantitative approach to cascading tipping events. We start with a mathematical framework to formulate elementary cascading tipping points (Sec. 3.2). Next, we discuss statistical metrics to analyse, diagnose and potentially predict cascading transitions, and apply them to ensemble simulations of the elementary cascading tipping points (Sec. 3.3 and SI C.2). Finally, we apply the new concepts to an example within the climate system: the potential cascading tipping mechanism between the AMOC and ENSO (Sec. 3.4). We summarise and discuss our findings in Sec. 3.5.

3.2 Mathematical framework for cascading tipping

In the climate system, tipping points are usually related to rapid transitions, where an observable in the climate system may change abruptly in a relatively short time compared with changes in the forcing of the observable. Such rapid changes usually involve transitions from one equilibrium state to another, which can often be explained using classical bifurcation theory for autonomous dynamical systems. To a certain extent, these concepts can also be applied to non-autonomous systems (so-called slow-fast systems) when the time variation of parameters can be viewed as a slow external forcing (Kuehn, 2011). They form the basics for understanding phenomena such as noise-induced tipping (Thompson and Sieber, 2011) and rate-dependent tipping (Ashwin et al., 2012).

In this section, we present a mathematical framework for simple cascading transitions, which acts as a first step towards analysing the more complex real-world transitions. We focus on bifurcation-induced tipping points, and consider two types of bifurcations that are thought to be relevant to mechanisms of abrupt changes in the climate

system; the back-to-back saddle-node bifurcation is often used to explain transitions between two coexisting equilibria (multi-stable systems), while the Hopf bifurcation can explain the appearance of oscillatory behaviour (Thompson and Stewart, 2002). Therefore, abrupt change in the system appears as a consequence of a parameter crossing a specific critical value at the bifurcation point.

A back-to-back saddle-node bifurcation (two saddle-nodes connected by a common unstable branch) generically occurs in physical systems (that have bounded states) when one parameter varies and the simplest dynamical system has a bifurcation that is described by

$$\frac{dx}{dt} = a_1 x^3 + a_2 x + \phi \quad (3.1)$$

where the a_i ($i \in 1, 2$) are constants, ϕ is a parameter, x is the state variable and t is time. There are multiple equilibria in the system if and only if $a_1 < 0$, $a_2 > 0$ and within the parameter interval $|\phi| < ((-4a_1^3 a_2^3)/(27a_1^4))^{1/2}$. In this case, the back-to-back saddle-node bifurcations occur at $\phi_c = \pm((-4a_1^3 a_2^3)/(27a_1^4))^{1/2}$. In the following, we often use the terms ‘fold bifurcation’ and ‘saddle-node bifurcation’, although in practice we only use ‘back-to-back’ saddle-node bifurcations.

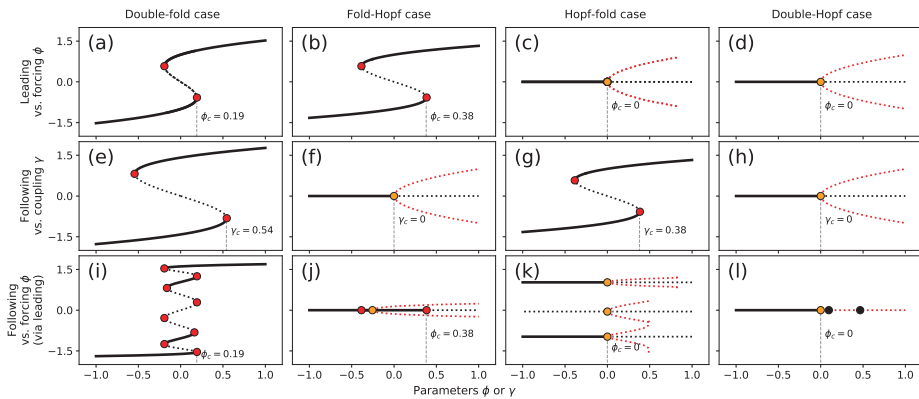


Figure 3.1: Stable (solid), unstable (black dotted) and oscillatory (amplitude red dotted) regimes of various cascading tipping types (as depicted on top of the figure). Black, orange and black dots indicate fold, Hopf and torus bifurcations, respectively. **Top panels:** leading system versus forcing ϕ . **Middle panels:** following system versus coupling γ . **Bottom panels:** (coupled) following system versus forcing ϕ . Critical values of (coming from lower branch) leading tipping (top), following tipping (middle) and the combined cascading tipping event (bottom) are marked by the grey vertical dashed line.

A Hopf bifurcation also generically occurs in physical systems, and the simplest dynamical system in which it occurs when one parameter is varied is described by

$$\begin{cases} \frac{dx}{dt} = a_1 y + a_2 (\phi - (x^2 + y^2))x \\ \frac{dy}{dt} = b_1 x + b_2 (\phi - (x^2 + y^2))y \end{cases} \quad (3.2)$$

where, again, $a_i, b_i, i = 1, 2$ are constants, ϕ is the parameter, (x, y) is the state vector and t is time. The state vector satisfying Eq. (3.2) reaches a stable periodic orbit if and only if $a_1 b_1 < 0$ and $\phi > 0$; the transition from steady to periodic occurs at $\phi = 0$. There are two other bifurcations when one parameter is varied (the transcritical and pitchfork bifurcations); however, these bifurcations are non-generic because special conditions must hold (e.g. symmetry) and so they are not considered here. Using the saddle-node and Hopf bifurcations, cascading tipping can be viewed as a combination of two coupled subsystems, where each subsystem undergoes one of these two types of bifurcations. Combining these bifurcations leads to four types of cascading tipping, which are discussed in the following.

Coupling two systems introduces a direction to the cascade and we take account of this by defining a ‘leading’ system, which during its transition changes a parameter (that is, the coupling term) in the ‘following’ system. The changing parameter in the following system can then bring the following system closer to a bifurcation point, potentially even resulting in a second transition. In this section, we only look at deterministic cases, which do not allow for noise to play a role in the tipping. Therefore, transitions in the leading system result in a changed coupling term that can lead to transitions in the following system. In bifurcation diagrams versus forcing, the bifurcation points (for deterministic systems) can overlap. However, the transitions are distinguishable in transients, because the following system always tips after the completion of the first transition. This is why we show the bifurcation diagrams of both systems versus forcing (Fig. 3.1) and an example of a transient (Fig. 3.2) for each type. They will be discussed in more detail below.

3.2.1 Double-fold cascade

The most intuitive system that has the potential to undergo a cascading tipping event is a system where both the leading and the following system have saddle-node bifurcations. Analogous to Eq. (3.1), a dynamical system containing a double fold cascade is then

$$\begin{cases} \frac{dx}{dt} = a_1 x^3 + a_2 x + \phi \\ \frac{dy}{dt} = b_1 y^3 + b_2 y + \gamma(x) \end{cases} \quad (3.3)$$

where x is the state vector of the leading system, y that of the following system, a_i, b_i ($i \in 1, 2$) are constants, and ϕ is a parameter in the leading system. The key is here that the function γ , which serves as a parameter in the following system, depends on the leading system. The most simple coupling between the two systems is represented by $\gamma(x) = \gamma_1 + \gamma_2 x$. Observe that a change in the forcing parameter ϕ can induce a transition in x , which may affect the coupling function γ such that also y undergoes a transition. We would like to emphasise that the forcing ϕ does not directly affect y but only through a change in x (which is effectively only significant when x tips).

Implementing this system with the values shown in Tab. C.1 gives insight in the system’s equilibrium structure (Fig. 3.1a) and transient behaviour (example in Fig. 3.2a). When ϕ is changed moving through the bistable regime of the leading system the coupling moves the following system through its own bistable regime (see Tab. C.1). Figures 3.1a and b show the equilibria of the leading (versus ϕ) and following systems (ver-

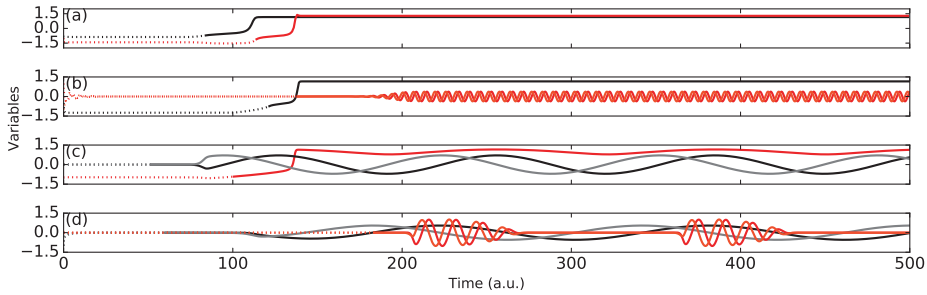


Figure 3.2: Example simulations for each cascading event type: the double-fold cascade (a), the fold-Hopf cascade (b), the Hopf-fold cascade (c) and the double Hopf cascade (d). Black and grey lines indicate the leading systems, red and orange lines indicate the following systems. Dotted lines indicate time before the critical threshold in the forcing $\phi(t)$ (black/grey) or coupling $\kappa(x)$ (red/orange) is reached, solid lines indicate the time after this. Parameter values for the modelled systems are given in Tab. C.1.

sus γ), respectively, showing the bistable regime in the centre of the figure, embedded in the saddle-node structure. Varying ϕ alters the state of the leading system, which affects the state of the following system through the coupling γ . This results in the existence of four stable states in the following system of the bistable regime of the leading system: two per state of the leading system, as shown in Fig. 3.1i. The leading system's state acts as a background condition modulating the position of the following system's equilibria; therefore, in case of transition, the leading system's state may drastically reposition the equilibria of the following system. This is intuitively visible in Fig. 3.2a, where a time series example of the dynamical system in Eq. (3.3) shows a cascading tipping event (parameters shown in Tab. C.1). When the leading system (black) is forced (by changing ϕ) to move from a bistable to a monostable regime, it transits towards a new equilibrium. During this transition, the following system (red) is affected and leaves the regime in which it had four possible equilibria; it also transits to a different state.

3.2.2 Fold-Hopf cascade

The second type of cascading tipping event involves a saddle-node bifurcation in the leading system and a subsequent Hopf bifurcation in the following system. Using analogous notation as in Eq. (3.3), the simplest system that captures this so-called fold-Hopf cascade is

$$\begin{cases} \frac{dx}{dt} = a_1 x^3 + a_2 x + \phi \\ \frac{dy}{dt} = b_1 z + b_2 (\gamma(x) - (y^2 + z^2)) y \\ \frac{dz}{dt} = c_1 y + c_2 (\gamma(x) - (y^2 + z^2)) z \end{cases} \quad (3.4)$$

where x is again the state vector of the leading system, and (y, z) that of the following system. By slowly varying the parameter ϕ (e.g., linearly as $\phi(t)$) the leading system moves through its bistable regime (see Tab. C.1 for parameter values) and via the coupling $\gamma(x) = \gamma_1 + \gamma_2 x$ forces the following system across the Hopf bifurcation point.

The bifurcation structure of the leading system of Eq. (3.4), using the parameters stated in Tab. C.1, is displayed in Fig. 3.1b. As in Fig. 3.1a, this system's bifurcation diagram shows a saddle-node structure. The following system in Fig. 3.1f, now undergoes a Hopf bifurcation and subsequent oscillatory behaviour. In Fig. 3.1j, it can be seen that increasing ϕ on the lower branch of the leading system regime will still result in a stationary equilibrium for the following system. When increasing ϕ across ϕ_c , the leading system moves towards another state (seen in Fig. 3.1b) and the coupling γ increases (across γ_c in Fig. 3.1f). An oscillation then occurs in the following system. This mechanism makes it possible for steady and oscillatory states to coexist on the right side of the Hopf bifurcation in Fig. 3.1j. An example of a time series showing a fold–Hopf cascading event is displayed in Fig. 3.2b. A transition in the leading system (black) brings the following system (red/yellow) into an unstable equilibrium that eventually leads to an oscillatory state.

3.2.3 Hopf-fold cascade

A third type of cascading event involves a Hopf bifurcation in the leading system and a subsequent saddle-node bifurcation in the following system. Using similar notation to the previous subsection, the simplest system with a Hopf–fold cascade (see Tab. C.1 for parameter values) is given by

$$\begin{cases} \frac{dx}{dt} = a_1 y + a_2 (\phi - (x^2 + y^2)) x \\ \frac{dy}{dt} = b_1 x + b_2 (\phi - (x^2 + y^2)) y \\ \frac{dz}{dt} = c_1 z^3 + c_2 z + \gamma(x) \end{cases} \quad (3.5)$$

where (x, y) is the state vector of the leading system, and z that of the following system. Again, we can slowly increase ϕ such that the leading system (x, y) crosses a Hopf bifurcation; via the coupling $\gamma(x) = \gamma_1 + \gamma_2 x$ the following system is then moved through its bistable regime such that a fold is reached in z .

Fig. 3.1c contains the typical bifurcation structure of the leading system in Eq. (3.5), containing a Hopf bifurcation separating a stationary from an oscillatory regime. The following system's equilibrium structure for varying γ is given by Fig. 3.1g. In this particular configuration, for any negative ϕ there are multiple stable equilibria, as seen in Fig. 3.1k. This makes sense, as ϕ only affects the following system via its impact on the leading system, and for negative ϕ the leading system remains constant. At $\phi = 0$, the Hopf bifurcation in the leading system is reached and (x, y) starts oscillating. The following system oscillates a little along with the leading system due to the oscillatory changing value of γ .

When ϕ increases more, the amplitude of the leading system's oscillation grows, which eventually causes γ to cross the threshold; the following system then leaves its bistable regime (be it temporarily as γ will be reduced again due to the oscillation). This forces the following system into its upper branch, as can be seen in Fig. 3.1g. The upper branch's stable oscillation ends at $\phi \approx 0.8$ (in Fig. 3.1k), because the amplitude becomes large enough for the system to swap between multiple equilibria. An example of such a cascading transition event can be seen in Fig. 3.2c, where an oscillation starts in the leading system (black/grey). A particular phase of this oscillation then causes the following system (red) to transit into the second equilibrium.

3.2.4 Double-Hopf cascade

A fourth type of cascading tipping event discussed here involves a Hopf bifurcation in the leading system and a subsequent Hopf bifurcation in the following system. Using analogous notation to the previous subsections, this double Hopf cascade is captured by the following dynamical system:

$$\begin{cases} \frac{dx}{dt} = a_1 y + a_2(\phi - (x^2 + y^2))x \\ \frac{dy}{dt} = b_1 x + b_2(\phi - (x^2 + y^2))y \\ \frac{du}{dt} = c_1 v + c_2(\gamma(x) - (u^2 + v^2))u \\ \frac{dv}{dt} = d_1 u + d_2(\gamma(x) - (u^2 + v^2))v \end{cases} \quad (3.6)$$

where (x, y) is the state vector of the leading system, and (u, v) that of the following system. If ϕ forces (x, y) such that it crosses the Hopf bifurcation point, the coupling $\gamma(x) = \gamma_1 + \gamma_2 x$ causes a crossing of the second Hopf bifurcation in (u, v) .

Figures 3.1d and h show the bifurcation diagrams of the leading and following systems, with supercritical Hopf bifurcations. The following system (in Fig. 3.1h) is stationary for many values of ϕ , up to the point that the leading system starts oscillating, which for high enough values of ϕ is large enough to make the following system cross the Hopf bifurcation. However, γ oscillates with the leading system (for $\phi > 0$). This means that oscillatory behaviour in the following system can only be expected in a particular part of the leading system's oscillation period. This interaction between the two oscillations results in torus bifurcations for particular values of ϕ . An example of a time series showing a double-Hopf cascading transition is shown in Fig. 3.2d. After a (slow) oscillation in the leading system (black/grey) has started, a (fast) oscillation in the following system (red/yellow) arises in particular phases of the slow oscillation.

3.3 Statistics of cascading tipping points

In the previous section we formulated elementary deterministic dynamical systems that can exhibit cascading tipping. In order to detect tipping events from, for example, observed time series in real systems, we need to detect whether a system is close to a critical transition. In general, systems close to critical transition recover more slowly from perturbations, which in turn increases memory in the time series. This leads to the phenomenon of 'critical slowing down' prior to bifurcation points. In this section, we simulate cascading tipping events to (a) study the statistical character of such events, (b) diagnose (post-tipping) whether tipping events are causally related and (c) take a first step towards statistical indicators for the prediction of cascading tipping events.

3.3.1 Methods for single tipping points

Several methods have been suggested for the analysis of time series to detect the approach of a single tipping point. For saddle-node bifurcations, the key features of a time series such as this is a critical slowing down. This can be investigated using standard quantities such as increasing auto-correlation (e.g. the lag-1 auto-correlation), increasing variance and increasing skewness (Held and Kleinen, 2004; Scheffer et al., 2009; Kuehn,

2011). Although critical slowing down near critical transitions is a general feature of (even chaotic) dynamical systems (Tantet et al., 2018), the standard metrics like auto-correlation at lag 1 and variance do not always provide an early warning signal (e.g. in Greenland ice core data in Livina and Lenton (2007)). Hence, more complicated indicators have been introduced, such as (i) degenerate fingerprinting (DF) and (ii) the detrended fluctuation analysis (DFA) (Held and Kleinen, 2004; Thompson and Sieber, 2011; Peng et al., 1994; Livina and Lenton, 2007). DFA is argued to be a solution to the problem that the simple lag-1 auto-correlation does not capture the approach to a transition in highly non-stationary data in long time series (Peng et al., 1994; Livina and Lenton, 2007). The latter is characterised by high auto-correlation due to the gradual increase or decrease of the system itself, distorting the signal of the critical slowing down, a problem in standard metrics and DF.

As critical slowing down implies an increasing auto-regressive behaviour in the time series prior to a transition, the memory component is increased. In general, the first step in DF is the projection of the data fields onto the leading EOF, which gives a time series x_n (Held and Kleinen, 2004). After time-equidistant interpolation and detrending of the data, in the DF method one fits the following general auto-regressive process to the time series x_n :

$$x_{n+1} = c \cdot x_n + \sigma \eta_n \quad (3.7)$$

where η_n is Gaussian white noise and $c = \exp(-\lambda \Delta t)$ the AR(1) coefficient. Here λ can be seen as the decay rate of perturbations in previous time steps. As the approaching of a bifurcation point involves an increase in memory, the value of c is presumed to increase towards one when approaching a saddle-node bifurcation point.

In DFA, one first chooses an integer window size s and divides the (cumulatively summed) time series $X(n)$ in $N_s = N/s$ segments that do not overlap, where N is the length of the time series. In every window, the best polynomial fit of a chosen order is calculated. A quadratic polynomial is used here. The squared deviation from this quadratic polynomial for every window is summed, resulting in a measure of the auto-covariance fluctuating around the fit:

$$F^2(v, s) = \frac{1}{s} \sum_{i=1}^s [X((v-1)s+i) - x_v(i)]^2 \quad (3.8)$$

with X the detrended time series and x_v the best polynomial fit in segment v . Then, an average is taken over all segments to obtain the fluctuation function $F(s)$:

$$F(s) = \sqrt{\frac{1}{N/s} \sum_{v=1}^{N/s} F^2(v, s)} \quad (3.9)$$

which depends solely on s . The long-range auto-correlations can now be recognised by fitting the fluctuation function to a power law and looking at the resulting DFA-exponent α , according to

$$F(s) \propto s^\alpha \quad (3.10)$$

For $\alpha \leq 0.5$, there is no long-term correlation and the fluctuations are indistinguishable

from white noise. However, when $\alpha > 0.5$, there are long-term correlations present and for $\alpha \geq 1.5$ the system has reached a bifurcation point. In the simulations analysed here the DFA scaling exponent is fitted explicitly for every (moving) window.

3.3.2 Methods for cascading tipping points

Cascading tipping involves two systems with their own bifurcation structure and their proximity towards bifurcation points. Although the leading system may be close to tipping, the following system might still be far from its bifurcation point and needs the critical transition of the leading system to even come close to this point; this makes the behaviour of the following system more prone to noise, less dependent on the leading system and less auto-correlated prior to the first tipping. This is why the general measures for single tipping events cannot be used, nor can regular cross (Pearson's) correlation. The reason for this is that the following and leading systems do not have a one-to-one relationship (that is, weakly Pearson correlated), but are rather coupled through specific parameters, which is only visible in long-range correlations.

When approaching a cascading tipping point, the long-range cross-correlation between the leading state x and the following state y is expected to increase. The state x becomes more auto-correlated and is less susceptible to noise, and therefore influences y through the coupling in a more robust way. To find long-range cross-correlations, the so-called 'detrended cross-correlation analysis' (DCCA) method was developed (Zebende, 2011; Podnobik and Stanley, 2008; Zhang et al., 2001; Zhou, 2008). Instead of using the auto-covariance to calculate the fluctuation function, as is used in Eq. (3.8), one uses the cross-covariance as follows:

$$F_{\text{DCCA}}^2(v, s) = \frac{1}{s} \sum_{i=1}^s [(X((v-1)s+i) - x_v(i)) \cdot (Y((v-1)s+i) - y_v(i))]^2 \quad (3.11)$$

The symbols in Eq. (3.11) are similar to Eq. (3.8). With this function, one can calculate the fluctuation function and subsequent power law scaling coefficient (Podnobik and Stanley, 2008; Zhang et al., 2001), similar to Eq. (3.9).

A variation on this method was proposed by Zebende (2011) and involves the ratio between F_{DCCA}^2 and F_{DFA} of the two systems. Specifically, one chooses a certain segment size s and computes:

$$\rho_{\text{DCCA}} = \frac{F_{\text{DCCA}}^2}{F_{\text{DFA}\{x\}} F_{\text{DFA}\{y\}}} \quad (3.12)$$

which measures the level of the long-term cross-correlation between variable x and y ; the quantity ρ_{DCCA} has values between -1 and 1.

There are several a priori limitations of using the power law scaling coefficient and ρ_{DCCA} . First of all, the results are sensitive to choices of the maximum segment size (s) and the window size, induced by the noise in our simulations. Making the window size too large decreases the possibility of seeing any development prior to the tipping points, as windows including the tipping event itself are biased by strong auto-correlation and

the (tipping) trend in the data. However, making the window size too small increases the uncertainty in the exponential fit. Similarly, adding small segments co-determining the exponential fit makes the method prone to noise, while larger segments are limited by the window size and computation time. In our simulations, window sizes of 120 data points, and segments sizes between 10 and 60 within those windows were used to calculate the fluctuations per segment size and to calculate $F(s)$ (Eq. (3.9)). More research is needed to find the optimal values for the window and segment sizes, and to potentially reduce this limitation in the analysis of cascading transitions. Another limitation of using the power law scaling coefficient and ρ_{DCCA} is that the way that the two systems are coupled (e.g. linearity, with or without temporal lag) affects how an evolution in the leading system affects its cross-correlation with the following system in both magnitude and time. Finally, to observe trends in these metrics, the signal in the cross-correlation between the two systems has to overcome the (partly mutually independent) noise. However, close to critical transition, the recovery from noise actually decreases, making the often subtle *change* in the detrended cross-correlation harder to distinguish from noise. These limitations may make the detrended cross-correlation metrics less useful in applications, but trends in the detrended cross-correlation metrics might still act as early warnings for cascading transitions, as shown in the next section.

In SI C.2, we provide detailed simulations to test the cross-correlation indicators DCCA and ρ_{DCCA} , in combination with before and after-ratios of auto-regressive metrics. The cross-correlation indicators seem to be sensitive to window and segment sizes and from this analysis alone, their potential as early warning indicators is still inconclusive.

3.4 Application: the coupled AMOC-ENSO system

In this section, we find an application of the concept of cascading tipping (the fold-Hopf case). This application reflects that cascading transitions are not only a purely mathematical concept, but do occur in idealised physical models. Here, we consider cascading tipping in a model that couples the Atlantic Meridional Overturning Circulation (AMOC) and the El-Niño-Southern Oscillation (ENSO).

3.4.1 Background

To demonstrate and quantify the coupling between the AMOC and ENSO, we use output from global climate model simulations. In the ESSENCE project (Ensemble SimulationS of Extreme weather events under Nonlinear Climate change) several simulations were performed using the ECHAM5/MPI-OM coupled climate model, including so-called ‘hosing’ experiments (Sterl et al., 2008), where fresh water is added around Greenland to mimic ice sheet melting.

From these climate model simulations we used two ensembles; the first is the “standard” experiment, where greenhouse gases evolve according to observations and from the year 2000 onward follow the SRES-A1b scenario (experiment name SRES-A1b). The second ensemble is the same as the standard experiment but has additional freshwater input ($1 \text{ Sv} = 10^6 \text{ m}^3/\text{s}$) around Greenland from the end of the year 2000 onward (experiment name HOSING-1). The HOSING-1 ensemble contains a classical hosing experi-

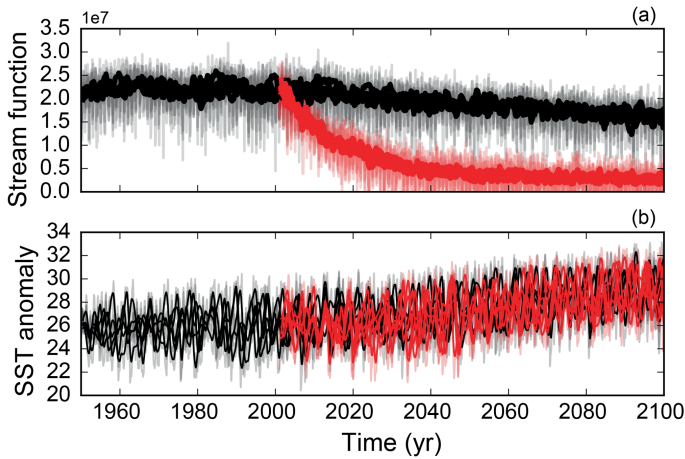


Figure 3.3: Panel (a): Evolution of the five standard SRES-A1b runs (black) and five HOSING-1 runs (red) in terms of the overturning stream function. Panel (b): NINO3.4 of the standard SRES-A1b ensemble (black) and the HOSING-1 ensemble (red). Shaded thin lines indicate monthly means, thick lines indicate the deseasonalised values.

Time period	Ensemble	Variable	Value
1950-2000	Standard SRES-A1b	Mean	25.86 ± 0.046
	Standard SRES-A1b	Variance	1.705 ± 0.447
2001-2100	Standard SRES-A1b	Mean	27.51 ± 0.032
	Standard SRES-A1b	Variance	2.581 ± 0.112
2001-2100	HOSING-1	Mean	27.27 ± 0.053
	HOSING-1	Variance	3.21 ± 0.42

Table 3.1: NINO3.4 statistics (of deseasonalised data) for the different ensembles. The uncertainty stated is the standard deviation among the five runs within the ensemble. It is visible that in the case of a collapsed overturning, El-Niño intensifies more than without a collapsed overturning. Values in bold represent increases in the variability of NINO3.4.

ment, following the procedure from Jungclaus et al. (2006). Five runs of each ensemble are chosen, specifically runs 041–045 of the HOSING-1 and runs 021–025 of the SRES-A1b ensemble. The temporal resolution used is monthly data between 1950 and 2100. The spatial fields are on a curvilinear grid, with 40 vertical levels in the ocean. We use deseasonalised data because we are interested in interannual variability, not in seasonal variability, as El-Niño is associated with these timescales. We use the maximum of the Atlantic meridional overturning stream function at 35°N as an AMOC index, and the NINO3.4 index as an ENSO index, which is the average SST over the region 170°W - 120°W by 5°S - 5°N.

The results for the evolution of the AMOC and ENSO are shown in Fig. 3.3. It is clearly visible that the AMOC decreases strongly in the hosing experiments, by approximately 85%. Tab. 3.1 compares statistical properties for the time intervals before and after 2001 (which is the year at which the hosing starts). We use the non-anomaly statistics, as this

gives us information about the differences in the mean. We do note that we only use five runs per ensemble, which means that the uncertainty is not statistically robust. We only present it in Tab. 3.1 to give the reader an idea of the range of the variables among the different runs.

It is visible from Tab. 3.1 that the variability of NINO3.4 increases (bold numbers) if we compare the 1950–2000 and 2001–2100 periods. This increased variability is visible in both the standard and the HOSING-1 runs. However, the variability is much stronger in the HOSING-1 experiment, indicating that the decrease of the AMOC indeed has an amplifying effect on ENSO. The large difference between the standard and hosing runs suggests that the NINO3.4 index changed in the hosing experiment, potentially as a consequence of the decrease of the AMOC. As the first tipping is artificially induced (without any measurable critical slowing down), and the models used here are far more complex than the simple dynamical systems we discussed in previous sections, the question of whether cascading tipping is actually occurring in these runs is beyond the scope of this chapter. We only use this data to justify the coupling of AMOC and ENSO.

Several mechanisms have been suggested in the literature regarding the coupling between the AMOC and ENSO. The first mechanism is concerned with oceanic waves. A colder North Atlantic creates density anomalies that induce the southward propagation of oceanic Kelvin waves (along the American coast) across the Equator. In western Africa, this energy radiates as Rossby waves towards the north and south, which induces Kelvin waves to move along the tip of southern Africa into the Indian Ocean, which eventually reach the Pacific. Consequently, the eastern equatorial Pacific thermocline deepens on a decadal timescale. This deepening has a weakening effect on the amplitude of ENSO (Timmermann et al., 2005).

The second mechanism is concerned with the trade winds. Cooling in the northern tropical Atlantic (due to AMOC weakening) induces anticyclonic atmospheric circulation (Xie et al., 2007) that intensifies the northerly trade winds over the northeastern tropical Pacific. This leads to a southward displacement of the Pacific intertropical convergence zone (Zhang and Delworth, 2005) and generates a meridional SST anomaly due to anomalous heat transport and the wind–evaporation–SST feedback in the Pacific. Also, Dong and Sutton (2007) found an atmospheric coupling through Rossby waves sent into the northeast tropical Pacific. This is in line with Dijkstra and Neelin (1995), who argued that part of the contribution to the zonal wind stress, τ_{ext} , arises from processes outside the tropical Pacific. The result of the wind-stress coupling between the two systems is an intensification of ENSO, and this mechanism is argued to be stronger than the coupling through oceanic waves (Timmermann et al., 2005).

3.4.2 Models and coupling

To study the possible cascading transition through the wind-stress coupling, we use a conceptual model. For the AMOC, the classical Stommel box model (Stommel, 1961) is used. It consists of a polar (subscript p) and an equatorial box (subscript e), both with a temperature T and salinity S coupled by a density-driven flow rate. The state variables are then defined as $\Delta T = T_e - T_p$ and $\Delta S = S_e - S_p$. The time evolution of these variables

is as follows (Cessi, 1994):

$$\begin{cases} \frac{dT}{dt} &= -\frac{1}{t_r}(\Delta T - \theta_0) - Q(\Delta\rho)\Delta T \\ \frac{dS}{dt} &= \frac{F_s}{H}S_0 - Q(\Delta\rho)\Delta S \end{cases} \quad (3.13)$$

where the first term in the temperature equation refers to relaxation towards a background temperature, and the second term refers to density-driven meridional transport. Specifically, t_r is the surface temperature restoring timescale and θ_0 is the equator-to-pole atmospheric temperature difference. $Q(\Delta\rho)$ is the transport function, which is calculated from a diffusion timescale and the meridional density gradient $\Delta\rho$. In the salinity equation, S_0 is a reference salinity, and H is the ocean depth. The parameter F_s is the freshwater flux, which can be used as a bifurcation parameter. The stream function

$$\Psi = \gamma_0 \Delta\rho / \rho_0 = \gamma_0 (\alpha_T \Delta T - \alpha_S \Delta S)$$

represents the strength of the AMOC, with $\gamma_0 > 0$ a flow constant, ρ_0 a reference density and α_T, α_S the thermal and haline expansion/contraction coefficients.

For the El-Niño–Southern Oscillation, we use the conceptual model as proposed in Timmermann et al. (2003). This model has a state vector consisting of the temperature of the western Pacific T_1 , the temperature of the eastern Pacific T_2 and the thermocline depth of the western Pacific h_1 . The model finds is based on the Zebiak and Cane (1987) ENSO model, with a two-strip and two-box approximation, and a shallow-water model for the upper ocean with a fixed mixed layer depth.

$$\begin{cases} \frac{dT_1}{dt} &= -\alpha(T_1 - T_r) - \frac{u(T_2 - T_1)}{L/2} \\ \frac{dT_2}{dt} &= -\alpha(T_2 - T_r) - \frac{w(T_2 - T_{\text{sub}})}{H_m} \end{cases} \quad (3.14)$$

where $1/\alpha$ a typical thermal damping timescale, T_{sub} the temperature below the mixed layer, H_m and L the depths of the mixed layer and basin width, respectively, w upwelling velocity and u atmospheric zonal surface wind being linear to wind stress: $u/(L/2) = \epsilon\beta\tau$ and $w/H_m = -\zeta\beta\tau$. The parameters ϵ and ζ refer to the strength of zonal and vertical advection (bifurcation parameters).

The subsurface temperature T_{sub} is parametrised as

$$T_{\text{sub}} = T_r - \frac{T_r - T_{r0}}{2} \left[1 - \frac{\tanh(H + h_2 - z_0)}{h^*} \right] \quad (3.15)$$

where h_2 is the east equatorial Pacific thermocline depth (calculated as the deviation from a reference depth H), z_0 is the depth for which w becomes its characteristic value and h^* is the sharpness of the thermocline. The thermocline depths are calculated as follows:

$$\begin{aligned} h_2 &= h_1 + bL\tau \\ \frac{dh_1}{dt} &= r(-h_1 - \frac{bL\tau}{2}) \end{aligned} \quad (3.16)$$

where b is the efficiency of the wind stress τ to drive the thermocline tilt. For further details, we refer to Timmermann et al. (2003). In the Stommel–Timmermann models, we

use the standard parameter settings, as given in the references, unless stated otherwise.

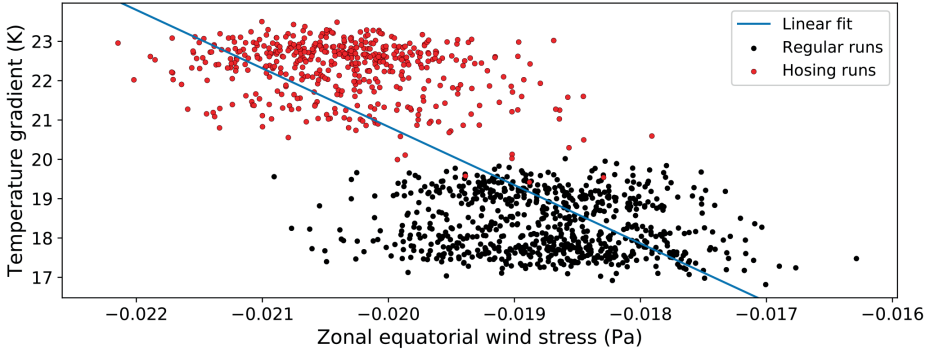


Figure 3.4: Zonal equatorial wind stress versus the Atlantic temperature gradient. Data from the ESSENCE (Ensemble SimulationS of Extreme weather events under Nonlinear Climate changeE) project are used, where the black dots refer to five members of the standard ensemble with SRES-A1b forcing (1950–2100 period), and the red dots refer to five members of the HOSING-1 ensemble where in 2000 a freshwater perturbation is applied (i.e. 2000–2100 period). A 5-year running mean is applied, yearly averages are shown. The zonal equatorial wind stress is defined here as the average zonal wind stress over the 0–10°N latitudinal band. The Atlantic temperature gradient is defined as the difference between the SST in a northern box (50–60°N, 50–20°W) and a southern box (0–20°N, 45–20°W). The blue line indicates a linear fit.

We couple the AMOC and ENSO models through the relation between the Atlantic meridional temperature gradient (in the Stommel model) and the Pacific zonal wind stress (in the Timmermann model); the Pacific zonal wind stress mechanism, which is the more important of the two aforementioned mechanisms, is found in the literature, and described in the previous section. Even in a simplified model, the relation between wind stress and meridional temperature gradient is physically justified: thermal wind balance indicates a direct connection between the adjustment of wind stress to changes in the meridional temperature gradient. In the Timmermann model, the zonal wind stress τ is expressed as

$$\tau = \frac{\mu(T_2 - T_1)}{\beta} \quad (3.17)$$

with μ/β parameters that control the influence of the zonal temperature gradient on the wind stress, set to be $0.02 \text{ Pa}\cdot\text{K}^{-1}$. We add an external wind stress term τ_{ext} that is dependent on the meridional temperature gradient in the Atlantic ΔT , i.e.,

$$\tau = \tau_{\text{ext}}(\Delta T) + \frac{\mu}{\beta}(T_2 - T_r) \quad (3.18)$$

with a negative relation between τ_{ext} and the Atlantic meridional SST gradient ΔT as we know from the literature described above (a stronger positive ΔT results in stronger easterlies and thus a negative τ_{ext}). Note that both the Pacific wind stress τ and specifically its external part τ_{ext} should always be negative. The total wind stress is negative because this area (at low altitude) is strictly dominated by easterly winds, and τ_{ext} is negative because, through the meridional temperature gradient, it reflects the influence of the zonal mean

Hadley cell on the equatorial Pacific. Physically, the Hadley cell only induces negative zonal wind stress in this region.

In the coupling (3.18), we fix β and vary μ as the coupling parameter. For τ_{ext} we take a linear relation as follows:

$$\tau_{\text{ext}} = \alpha_{\tau} \Delta T + \gamma_{\tau} - \tau_0 \quad (3.19)$$

where all coefficients are constant over time. The parameters α_{τ} and γ_{τ} can be estimated from the ESSENCE data as discussed in Sec. 3.4.1; τ_0 reflects the constant part in the zonal mean wind stress, which we subtract because we are interested in the contribution of changes in the meridional overturning. Using five ESSENCE runs per ensemble for both the standard forcing and hosing-experiment, respectively, ΔT is computed as the absolute difference between the SST in the North Atlantic area ($50 - 60^{\circ}\text{N} \times 50 - 20^{\circ}\text{W}$) and the equatorial Atlantic region ($0 - 20^{\circ}\text{N} \times 45 - 20^{\circ}\text{W}$). For the wind stress τ_{ext} , the zonally integrated wind stress averaged over the $0 - 10^{\circ}\text{N}$ region is taken. In Fig. 3.4, 5-year running means of annual averages are plotted for the hosing simulations (in red) and the standard simulations (in black). Clearly, τ_{ext} decreases as ΔT increases, meaning that when the AMOC collapses (larger ΔT) the wind stress τ_{ext} becomes more negative and the external part of the trade winds increases. However, we also note that the spread in the simulation data is large, which can, in part, be attributed to internal variability present in the simulations. The coefficients α_{τ} and γ_{τ} were found to be (from a linear fit) $-0.000376 \text{ Pa} \cdot ^{\circ}\text{C}^{-1}$ and -0.0119 Pa . By looking at the ΔT regime in Fig. 3.4, τ_0 is chosen to be the wind stress at 19°C : $\tau_0 = \alpha_{\tau} \cdot 19 + \gamma_{\tau} \approx -0.0190 \text{ Pa}$. This results in a final quantised expression for the coupling:

$$\tau_{\text{ext}} \approx -0.000376 \cdot \Delta T + 0.00715 \quad (3.20)$$

3.4.3 Results

The AMOC model's bifurcation diagrams are shown in Figs. 3.5a and b, and clearly display a saddle-node structure. For an interval of values of the freshwater flux F_s the system has multiple equilibria, whilst for other values only one equilibrium remains. This means that when we are in the high- Ψ branch and F_s is large enough, the system can make a transition to the low- Ψ branch. This is depicted by the blue arrow in Fig. 3.5b.

The bifurcation diagram of the ENSO model with τ_{ext} as a parameter is shown in Fig. 3.5c. The bifurcation diagrams become much simpler than in the original Timmermann et al. (2003) model, the reason for this is extensively discussed in Dijkstra and Neelin (1995). Figure 3.5d shows the influence of μ for the position of the oscillatory regime: on each branch, two Hopf bifurcations can be found and the μ value of the first Hopf bifurcation decreases with more negative τ_{ext} . This indicates that a Hopf bifurcation can be crossed if τ_{ext} is decreased, while μ is kept constant. In other words, for the right value of μ , the eastern Pacific SST starts oscillating (El-Niño 'intensifies') when the easterly external wind is increased. For the coupled model, we use $\mu = 0.00146$.

Using τ_{ext} to couple the AMOC and ENSO models, we performed simulations with $\Delta t = 0.1 \text{ days}$ and the Runge–Kutta fourth-order integration method. To initiate the col-

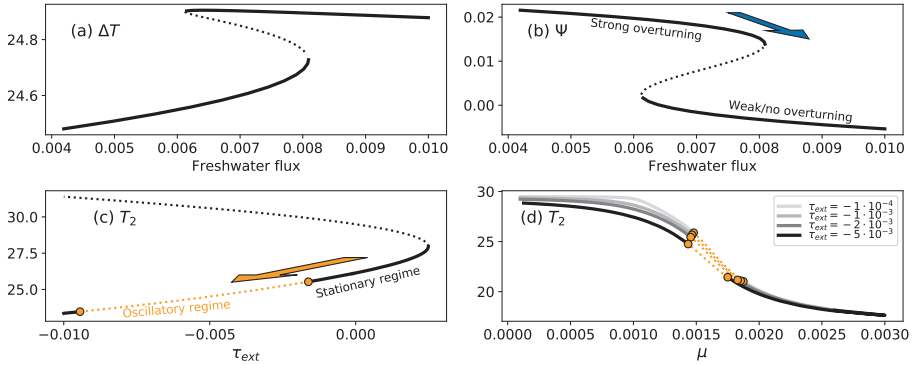


Figure 3.5: Bifurcation diagrams and forward runs of the Stommel (a, b) and Timmermann (c, d) models. The blue arrow indicates the collapse of the overturning circulation, which (negatively) amplifies the external zonal wind stress in the Pacific τ_{ext} , such that the system enters an oscillatory state. The orange arrow indicates subsequent tipping in the following (ENSO) system. **Panel (a):** Meridional temperature gradient equilibria versus freshwater flux. **Panel (b):** The non-dimensional stream function versus freshwater flux. These figures show the multiple states of the overturning. **Panel (c):** Eastern equatorial Pacific SST versus τ_{ext} (for $\mu = 0.00145$), showing a regime where the system is stationary and a regime where the system is oscillatory; **Panel (d):** eastern equatorial Pacific SST versus μ for different values of τ_{ext} . Orange dots indicate Hopf bifurcation points, and orange dotted lines indicate oscillatory regimes. Black and grey solid lines indicate stable equilibria, and black dashed lines indicate unstable equilibria.

lapse of the overturning, a freshwater forcing F_s is applied in the form of a step function:

$$F_s = \begin{cases} 0.006 & \text{if } t \leq 500 \text{ y} \\ 0.01 & \text{if } t > 500 \text{ y} \end{cases} \quad (3.21)$$

Using the coupling of Eq. (3.20), we attain the results shown in Fig. 3.6. The exact quantification of the coupling partly modulates which effect the collapse of the AMOC has on ENSO. For the chosen coupling, the collapse of the overturning leads to the crossing of the first Hopf bifurcation point in the following system, and an oscillation starts growing. As is visible in Fig. 3.3, the relation between ΔT and τ has quite some spread, which implies a large uncertainty in the values of α_τ and γ_τ . We would like to stress that the regime in which the Hopf bifurcation is crossed is dependent on multiple variables, such as these coupling parameters. Running the forward integration of the coupled model for values between $\alpha_\tau = -0.00041$ and $\alpha_\tau = -0.00033$, uncovers the fact that (*ceteris paribus*), for higher (lower) values of α_τ , the oscillation indeed becomes weaker (stronger), finally resulting in a disappearance of the oscillation at $\alpha_\tau \approx -0.000335$ at a time step of 0.25 days.

Despite the parameter sensitivity, this is a typical illustration of the fold–Hopf cascading behaviour discussed in earlier sections. This re-enforces the possibility that cascading transitions are possible in real physical systems.

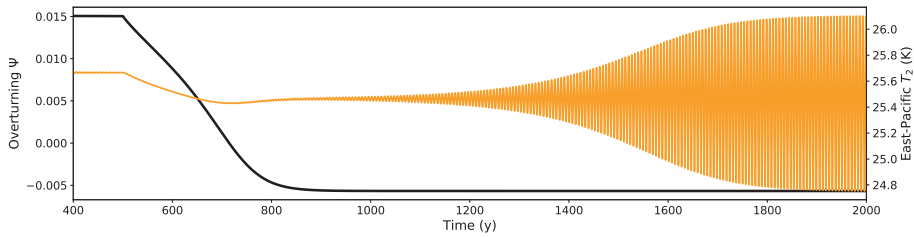


Figure 3.6: Simulation run of the coupled Stommel–Timmermann model for different model configurations, where the collapse of the overturning flow function (black) leads to the crossing of a Hopf bifurcation in the eastern equatorial Pacific SST (orange). Parameter values as in Timmermann et al. (2003), with $\mu = 0.00146$.

3.5 Summary, discussion and conclusions

In this chapter, we introduced the concept of cascading tipping, which can occur when a transition in a leading system alters background conditions for a following system that then undergoes a transition. We presented a mathematical framework around this concept, using generic bifurcations (saddle-node and Hopf) in both leading and following systems. Four types of deterministic dynamical systems with the possibility for cascading events were formulated, including the double-fold cascade, the fold–Hopf cascade, the Hopf–fold cascade and the double-Hopf cascade. In all cases we assumed a linear coupling between the following and leading system. The double-fold coupled system has previously been considered in another context (Brummitt et al., 2015), where it was also noted that not all subsystems undergo tipping (‘hopping’) in systems with more than two coupled fold cascades. Moreover, stochastically coupled multi-stable systems have been considered in networks, where different types of domino effects can occur depending on the synchrony of the transition in the different network nodes (Ashwin et al., 2017; Creaser et al., 2018). Here we only consider two coupled systems, but allow different types of bifurcations, and the systems are physically coupled in a directional way.

We discussed statistical indicators and analysis tools for cascading tipping points. Indicators for cascading tipping points are found in detrended cross-correlation analysis (DCCA) and a special case of extrapolation using the DFA of the following system. These tools were applied in simulations involving both the double-fold and fold–Hopf cascades. The increased variance, AR(1) and DFA scaling exponent are clearly found in each case of single tipping. The cross-correlation indicators (DCCA and ρ_{DCCA}) did not evolve much throughout the time series, which indicates their insensitivity with respect to the proximity to single tipping points. Several limitations on the use of these variables have been mentioned. However, it seems that these metrics are highly sensitive to window and segment sizes, which means that their potential as early warnings of cascading transition events is inconclusive. The ratios of auto-regressive metrics before and after the first transition seem to be a stronger warning of cascading transitions. More research is needed to exactly quantify these metrics.

The concept of cascading tipping was applied to study the behaviour of a model describing a link between the Atlantic meridional overturning circulation (MOC) and ENSO. We modelled this using a coupling between the Stommel (1961) model and the Timmer-

mann et al. (2003) ENSO model by a meridional temperature gradient-dependent term in the external wind stress of the ENSO model. Through analysis of the bifurcation diagrams and simulations, a cascading tipping event is indeed possible within this model in the form of the fold–Hopf cascade. Obviously, both models are highly idealised and more detailed models of both AMOC and ENSO are needed to demonstrate the occurrence of such a cascading transition in the climate system.

A potential example of a double-fold cascade, which was not further treated here, could be the impact of a bistable MOC on the (bistable) land ice formation on the Antarctic continent. In this case the coupling exists through the atmospheric CO₂ concentration, which depends on mixing and circulation in the ocean while strongly determining the existence of an ice sheet (DeConto and Pollard, 2003). During the Eocene–Oligocene transition, where a large ice sheet grew on Antarctica, a two-step signal is observed in the deep-sea $\delta^{18}\text{O}$ ratio, suggesting two abrupt transitions. Using a box model by Gildor and Tziperman (2000), Tigchelaar et al. (2011) showed that a two-step signal can be produced by (first) a MOC transition which changes the CO₂ concentration meaning that a transition occurs in the land-ice model. Although from a physical perspective, this is a potential example of a cascading transition, we make no claim about whether such a transition likely occurred at the Eocene–Oligocene transition. Here, more detailed models are also needed and transition are expected to be more complicated (Tantet et al., 2018).

These two applications reflect the relevance of this chapter. There are likely many cases in which these cascading events occur in climate and therefore highlight the importance of the topic. Future research will point out whether these events are likely to happen in the future climate and whether these effects also occur in fields other than climate science. Of course, this chapter covers the very basics of deterministic cascading events. However, one can imagine a wide range of phenomena if more complicated transitions between attractors are considered and when noise is included. For example, when a leading chaotic system is coupled to a deterministic following system with a saddle-node bifurcation structure, a slight change in the chaotic attractor may change the background conditions for the following system such that it undergoes a transition. An application here may be the effect of a mid-latitude atmospheric jet on the Atlantic MOC. We hope that this chapter will stimulate more research on the various types of cascading tipping and also on the development of well-suited indicators and early warnings of such events.

PART II

STATIC NETWORK APPROACHES

On the relation between dynamics and underlying
network structures



"The central task of a natural science is to make the wonderful commonplace: to show that complexity, correctly viewed, is only a mask for simplicity; to find pattern hidden in apparent chaos."

HERBERT A. SIMON, 1978



Introduction to Part II

Part I described methods to describe macroscopic behaviour of systems with N interacting agents a_i ($i \in \{1, \dots, N\}$) in a lower-dimensional state variable $\mathcal{S} \in \mathbb{R}^n$, with $n \ll N$. Crucial to the potential of describing a system's behaviour in such a low-dimensional state variable is a certain extent of homogeneity among the agent's interactions. If the interactions vary strongly across the population, aggregation of them might not be possible or may not show the complex behaviour observed in reality, as small may have a profound effect on the macroscopic behaviour. One example of heterogeneity can be found in to whom people interact, which is highly relevant for dynamical phenomena like fake news or epidemic spreading: people may form many interactions within ethnicities, age groups and geographical regions, but fewer between such groups. Another example is differences in how many interactions each agent has. In short, accounting for the interaction structure explicitly accounts for an additional level of detail, and is commonly done in the form of *networks*, consisting of *nodes*, representing the system agents, and *links*, representing interactions between them, defined by scalar weights w_{ij} for each agent pair (a_i, a_j) , with $i, j \in \{1, \dots, N\}$. Higher absolute weights imply stronger interactions.

The analysis of the network structures as static objects is a large and rapidly growing research field in itself, but not the scope of this dissertation. Instead, I focus on what the structure of these networks can tell us about the dynamics playing out on top of them. To do this, it is important to distinguish *static networks*, in which the network structure underlying the dynamics in a system remains invariant over time, from *temporal networks*, when this underlying network structure does evolve over time. Systems with static underlying networks are discussed in Part II, and temporal networked systems are discussed in Part III. In this light, it is also important to separate the system's dynamics from the system's underlying interaction structure (see Fig. ib on p. 12). They can be intimately linked: when playing the game of tag, a (Boolean) contact between two agents implies full transmission of the status of being 'it': dynamics and interaction structure are very closely linked. But the underlying structure and the dynamics on top are not necessarily one-to-one connected: for example, in epidemiology, the disease spreading is bound by, but not defined by the connections between agents that can infect each other, because of the presence of incubation time scales and probabilistic transmission.

To illustrate how network structure impacts dynamics of the system, I proceed with the example of epidemiology, which was also addressed in the introduction to Part I. In this case, rather than assuming a homogenised mixing and interaction of the population (which was the case in Part I), two non-uniform underlying static network structures are distinguished that describe the interactions. The first network contains the interactions of children at a primary school (Gemmetto et al., 2014; Stehlé et al., 2011) also shown in Fig. iib, and the second contains interactions of attendees at a scientific conference (Cattuto et al., 2010; Stehlé et al., 2011). For both datasets, two hours of contact data is used to generate the networks, and an SIR model is simulated on top of these connections as follows. Instead of having a state variable consisting of three system-wide variables of susceptibles $S(t)$, infectious $I(t)$ and recovered $R(t)$, each agent now attains these stages individually. With a transmission probability of β and a recovery rate of γ , the disease is transmitted probabilistically upon contact. The results are shown in Fig. iv.

Eyeing the panels (a) and (c) already indicates a strong difference in social structure: while the primary school children are quite segregated into classroom groups, the mixing of conference attendees is much more intertwined. Also when looking at the degree distributions in panel (b), we see that the primary school children have an approximate Poisson-shaped degree distribution, while the conference attendees show a larger spread: having a few attendees with very large (up to 70) unique contacts, but also many with less than 10 (reflected in the tiny dot sizes in the periphery in panel (d)). In node colours, the panels (a) and (c) show the moment at which agents on average become first infected. Clearly, more peripheral children (in (a)) and conference attendees (in (c)) are affected later, on average. In particular, we see the clustered pattern in the network of panel (a) back in the node colours: if nodes in any cluster are affected, quickly most others in such well-connected parts are as well. Even though the average degree of the primary school is higher, and the intra-cluster connections are strong, the spreading in the scientific conference grows more rapidly, reflecting the effect of the exact structure of the network, favouring fast spread through the centre nodes. If one would have focused on the average degree (indicated in panels (b) and (d)), one might have pointed to the primary school as the system with the fastest spread. Clearly, the network structure underlying the spreading dynamics plays an important role in its speed and direction. This illustrates the motivation behind part II of this dissertation.

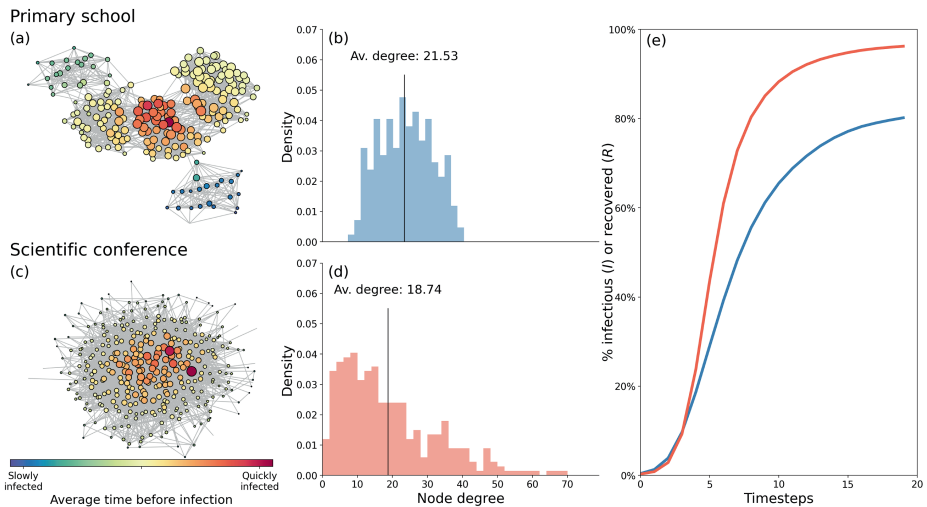


Figure iv: SIR model applied to two different network structures: a primary school (same as in Fig. iib) (Gemmetto et al., 2014; Stehlé et al., 2011), and a scientific conference (Cattuto et al., 2010; Stehlé et al., 2011). **Panel (a, c):** network structures, where node colours represent the time at which the agent (child or conference attendee) is infected, on average in the ensemble of simulations. Red is early, blue is late. **Panel (b, d):** degree distributions for the primary school (blue) and the scientific conference (red). Average node degree is annotated by the vertical line. **Panel (e):** evolution of the % agents that are either infectious or recovered. Here, $\beta = 0.1$ and $\gamma = 0.02$ and an ensemble of $10N$ is used, where each agent is used as patient zero 10 times.

Part II comprises the three chapters 4, 5 and 6 that all focus on the dynamics of delays in railway systems. The three chapters approach this problem from three different

angles, all concerning the link between the underlying infrastructure and the dynamics playing out on top. Chapter 4 involves the analysis of the network itself, weighted by spatial heterogeneities that impact delays, such as train frequency, travel times and delay change statistics. By creating these weighted graphs for four European countries (the Netherlands, Italy, Germany and Switzerland), an international comparison was made in several network aspects. In particular, a spectral clustering algorithm is applied to identify dynamical clusters in the network. These clusters are interpreted in terms of role in the total delay and their connectivity to the rest of the network, which points out region differences.

In chapter 5, the concepts in chapter 4 are extended to a model that simulates the Belgian railway delays. The resulting model is of a diffusion-like type: it evolves from a node to all its neighbours, weighted by the model aspects. This is in contrast with many traditional railway models, which are discrete, more clearly relating to the discrete nature of delays and trains that transport them. On the other side, diffusion-like models have advantages relating to their simplicity, low computational demand and easy calibration. The chapter discusses how the model increases performance when coarse graining the network, resulting in reduced network topologies consisting of interacting clusters rather than individual stations. The main difference with chapter 4 is its scope: while chapter 4 focuses on geographical characterisation and the role of subregions in the larger system dynamics, chapter 5 focuses on prediction. In particular, we address what the predefined timetable and the network can tell us about how the dynamics evolve, while also using geographical information.

The final chapter in this part, chapter 6, aims to find the causes and mechanisms of large-scale disruptions, rather than focusing on their simulation. In contrast to chapters 4 and 5, we propose a discrete-event model. In particular, we focus on how different dynamical layers, i.e., the train service lines, the (physical) rolling stock and the crew, can cause a cascade of delays that rapidly affects a large geographic area — a phenomenon that can be a main driver of the observed large geographic spreading of delays in the Netherlands when delays are severe enough. We show how this discrete-event model can be used to diagnostically infer causes of past delays and show the impact of mitigation measures of dispatchers changing the schedules of the aforementioned dynamical layers.

While chapters 4 and 5 explicitly search for the link between dynamics and the underlying network (through topological analysis of the network and using it to explicitly simulate the dynamics), chapter 6 reasons from a different perspective. There, delays are not transported continuously as determined by (weighted) network aspects, but transported in a strongly anisotropic manner, fully favouring particular directions of spread, neglecting possible other links in the network. That is, the delays follow the train line routes, with certain additional spreads to other routes caused by transfers of rolling stock and crew members. In other words, in chapter 6, delays do not spread in all possible directions in the network. The realisation of this potential disconnection between network and dynamics is in the case of railways already evident by the non-applicability of general network metrics like degree and centrality to be descriptive for its dynamics, as is also touched upon in the chapters 4-6 and at the end of this dissertation. Summarised, the overarching question in these chapter is the following:

Research question Part II: *How can we investigate the interdependence of a system's dynamics and its underlying static network architecture?*

CHAPTER 4

GEOGRAPHIC DELAY CHARACTERISATION OF RAILWAY SYSTEMS

This chapter is published as: Dekker, M. M. (2021).
Geographic delay characterization of railway
systems. *Scientific Reports*, 11(1):1–13.
Supplementary material to this chapter
can be found in SI D.



ABSTRACT

Railway systems provide pivotal support to modern societies, making their efficiency and robustness important to ensure. However, these systems are susceptible to disruptions and delays, leading to accumulating economic damage. The large spatial scale of delay spreading typically make it difficult to distinguish which regions will ultimately be affected by an initial disruption, creating uncertainty for risk assessment. In this chapter, we identify geographical structures that reflect how delay spreads through railway networks. We do so by proposing a graph-based, hybrid schedule and empirical-based model for delay propagation and apply spectral clustering. We apply the model to four European railway systems: the Netherlands, Germany, Switzerland and Italy. We characterise these geographical delay structures in the railway systems of these countries and interpret these regions in terms of delay severity and how dynamically disconnected they are from the rest. The method also allows us to point out important differences between these countries' railway systems. For practitioners, such geographical characterisation of railways provides natural boundaries for local decision-making structures and risk assessment.



4.1 Introduction

Transport systems provide a core function to our society, moving passengers and goods around the globe to allow for global trade, business and leisure. Efficient infrastructure and mobility has been found important for economies to grow (Wbcsd, 2009). However, at severe economic costs, the efficiency of many such systems is regularly affected by perturbations and subsequent spread of disruptions. Examples can be found in trade networks being affected by epidemics (Ivanov, 2020), earthquakes (Inoue and Todo, 2019) or cyclones (Shughrue et al., 2020), energy and internet problems due to power outages (Buldyrev et al., 2010), railway disruptions due to natural hazards (Dekker et al., 2021f, 2019; Bhatia et al., 2015; Ludvigsen and Klæboe, 2014) and many more. The economic costs of such events, also on smaller scales, motivate scholars and practitioners to investigate the robustness and resilience of these systems to perturbations (Pagani et al., 2019; Nogal et al., 2016; Ouyang et al., 2012), and to understand and predict subsequent evolution of perturbations (Goverde, 2010; Dekker and Panja, 2021). This chapter focuses on understanding the structure of the evolution of such perturbations in a subset of transport systems: railway systems.

Railway systems involve the on-time geographical transport of passengers or goods, utilising resources (*assets*, like physical trains or crew) according to a predefined schedule, from, via and towards nodes (e.g., stations) in a network, along certain routes or tracks (edges). While the on-time dynamics are described in a pre-defined schedule, of interest here are the temporal deviations from the schedule, which are referred to as *delays*, calculated as the executed time minus the predefined time — in situations with no delay at all, all resources run on time and no delay is present; hence, no perturbation or associated dynamics is present. What sets apart railway systems from many other transport systems, is their dependence on a detailed pre-determined system, in some countries even down to the particular assets used per activity (Dekker and Panja, 2021). This attribute, in combination with usually high utilisation of existing capacity commonly causes perturbations that are an interesting (and necessary) topic of study. In the remainder of this chapter, when we talk about ‘dynamics’ of railway systems, we actually refer to the spreading and change of *delays* rather than the scheduled movement of the assets — note that the latter is usually a well-optimised plan that contains, by construction, little interacting elements (although the timetable itself may be prone to cascading-like phenomena that in turn affect the delay dynamics (Dekker and Panja, 2021)).

The research on transport delays can be split into multiple parts. A first part focuses on understanding the origin of initial perturbations, involving research on natural disasters as mentioned above, but also on smaller-scale fluctuations of activity delays that may give rise to systemic delay generation (Wen et al., 2019; Dekker et al., 2021f; Fleurquin et al., 2013; Kecman and Goverde, 2015b). Oftentimes, these studies are case-specific, due to the heterogeneous (and often external) nature of such initial perturbations. Another line of transport delay research focuses on prediction of delays: the simulation of how a currently delayed situation will evolve to a (future) situation. Many kinds of approaches to this problem exist: micro simulation tools including a high level of detail on each edge (Middelkoop and Loeve, 2006; Nash and Huerlimann, 2004), larger-scale stochastic and analytic models both on the planning and realisation side (Goverde, 2010; Schöbel, 2012; Fleurquin et al., 2013; Monechi et al., 2018; Dekker et al., 2021d), and

in more recent years, machine learning or data-driven studies are contributing to the field (Dekker et al., 2019; Kecman and Goverde, 2015a; Marković et al., 2015; Oneto et al., 2018). A third branch of research involves the understanding of delay dynamics on a more structural level: characterising the nature and (semi-)universal 'laws' without the specific aim of prediction. For example, finding macroscopic nation-wide patterns in railway delay (Dekker et al., 2019), characterising mechanisms of how trains pass delay onto other assets (Monechi et al., 2018), how the interaction between crew, rolling stock and line planning can lead to delay cascades (Dekker and Panja, 2021).

This chapter contributes to the third branch of transport (or railway) delay literature and focuses on the large spatial scale that is usually associated with these systems. We aim to identify geographic regions that act as subsystems, partially independent in terms of delays, and these clusters are reinterpreted in terms of their role in the global system, which add to the understanding of the general structure of these often nation-wide systems. The identification of such regions contributes to the overall understanding of the dynamics of delays, providing answers to questions like 'where is delay generally generated, propagated through and attracted?' and 'what regions are dynamically near-isolated and may be treated as such?'. These questions are useful for practitioners to not only help in prediction methods, but also for more strategic decisions on infrastructure planning and updates of future timetables.

Identifying substructures in complex systems such as railways is commonly done using clustering algorithms, with applications ranging from physics to ecology (Dam, A. van et al., 2021), climate (Tantet et al., 2015) and even epidemics (Scarpino and Petri, 2019). The wide variety of applications comes with a multitude of clustering methodologies. Graph-based clustering is often based on random-walks and modularity-optimisation principles (Newman, 2006), like the famous Louvain clustering method (Blondel et al., 2008), or on spectral properties of core graph matrices (Luxburg, 2007). A famous data-based clustering method is K -means (MacQueen, 1967), while many other methods exist. It is important to note the difference between identifying substructures in the network topology alone, and doing so by accounting for spatial structures in dynamic processes happening on top of the network topology. In this chapter, we focus on the latter, using spectral clustering. The main reason for choosing spectral clustering as opposed to any modularity-optimisation tool is that spectral modes and the shape of the eigen-spectrum reveal more than just the detection of communities (as we see later in Fig. 4.3 of this chapter). But indeed, there are important advantages of modularity optimisation, as well, such as the automatic optimisation of the number of detected clusters.

Also in transportation literature, clustering techniques have been applied, e.g., to assist real-time management, operations and decision making (Chen, 2014; Yang et al., 2017; Cerreto et al., 2018; Kadir et al., 2018). Other papers apply clustering tools on statistical variables to identify general states in the system (Xia et al., 2012; Lin, 2019; Dekker et al., 2019). Even though compartmentalising the geographical system based on observed or simulated delay patterns is less common, it is not new, e.g. concerning the identification of communities in cargo ships (Kaluza et al., 2010), assessing topological properties in the Swiss railway network (Erath et al., 2009) or quantifying resilience in the Indian railways (Bhatia et al., 2015). However, combining data on the dynamics of the (delay of the) system, with data on the infrastructure has rarely been done in railway literature. This, and the subsequent interpretation of those found communities, is the goal

of this chapter. It is important to emphasise that mere (topological) clustering of only the infrastructure, neglecting all other dynamical and operational information, would provide clusters that have less of an operational meaning: delay does not diffuse equally along directions that are topologically equivalent. The relation between more general diffusion dynamics among subregions in a networks has been investigated theoretically before (Siudem and Holyst, 2019), but to our knowledge, this has never been applied to characterise dynamic structures in railway systems.

The chapter is structured as follows. The model and the associated clustering method is discussed and applied to a fictitious transport network in section 4.2. In section 4.3, we apply the methods to real data of four European railway systems and show the results of the model and clustering. We interpret the clusters in terms of connectivity and role in the country's delay in section 4.4. We end with several conclusive remarks in section 4.5.

4.2 Methods

In this section, we build a graph-based model for delay dynamics based on infrastructure and empirical train delays. We start by separating the underlying infrastructure from the delay dynamics that happens on top. The underlying infrastructure, consisting of *nodes*, being stations or departure/arrival locations, and *edges* as the railway tracks between them, is assumed to remain invariant in this analysis — different from, for example network perturbation analyses or node-failure transport problems (Bhatia et al., 2015; Buldyrev et al., 2010; Pescaroli and Alexander, 2016). We henceforth use the term ‘nodes’ and ‘stations’ interchangeably. The aim of this chapter is to provide insights in the geographic aspects of delay propagation. To this end, we distinguish factors that result in spatial non-uniformity: e.g., edges with a higher frequency of trains, fewer parallel tracks and re-routing options, are more prone to propagate or amplify delay than other edges. Rather than purely looking at the topology of the underlying infrastructure network, it is such non-uniformity that defines weights of the edges that largely impact the resulting spectral clustering. We note that there are also other types of non-uniformity in railway systems that are not directly related to geography that are outside of the scope of this chapter.

We distinguish two types of spatial non-uniformities: (1) those consequential to traffic flows, as found in the timetable, and (2) those that are due to other effects, which we infer from delay statistics. Traffic flows are static properties that affect delay propagation and can be determined from the system's timetable, e.g., (planned) running times and resource travel frequency. These factors are used to imply what portion of existing delay at a station is propagated in each possible direction. Spatial non-uniformities unrelated to the timetable, in contrast, relate to dynamic factors affecting delays and are derived from delay statistics rather than from the static (non-delayed) timetables. These factors act as multipliers: if, from the timetable, we expect delay to be transported in a certain direction, we use statistics to estimate whether it amplifies or dampens. Below, we build a model where we define both of these factors.

4.2.1 Spatial non-uniformity in traffic flows

Spatial non-uniformity in traffic flows comes from within the system's internal predefined properties, like the timetable. The intuition is that edges with more traffic propagate delay more easily and busier routes involve more congestion. In particular, delays are carried along with trains in the direction of their trajectories and nodes are more easily affected by a neighbouring node's delays if the link between them contains high-frequency traffic, a mechanism that is commonly used in delay propagation models (Monечи et al., 2018; Dekker et al., 2021d). We capture the spatial non-uniformity in traffic flows by identifying the proportional direction of delay propagation from any node towards each of its edges, using the relative frequencies f of trains. Another spatial non-uniformity in the timetable concerns the fact that in some areas, the edges are very short, leading to short running times τ . Combining these two factors of spatial non-uniformity in traffic flows results in a single edge weighing factor α (see Fig. 4.1a) for each edge ij between nodes i and j (i.e., $i, j \in \{1, \dots, N\}$ where N is the number of stations, and i and j make a direct connection in the railway network):

$$\alpha_{ij} = \frac{f_{ij} \cdot \tau_{\min}}{\sum_{j'} f_{ij'} \cdot \tau_i} \quad (4.1)$$

where τ_{\min} is the system-wide minimum running time. The factor τ_{\min}/τ_i therefore reflects changes in the weights of α of all edges of node i based on its running times relative to the minimum.

4.2.2 Spatial non-uniformity due to other effects

Besides spatial non-uniformity in traffic flows, there are also factors outside from the timetable that increase or decrease delays, depending on the area in the railway network. Examples of such factors are a large number of block signals, fewer parallel tracks, speed limits, high volume of passengers (delaying boarding times), decreased vision and increased chances of infrastructure problems or trees falling on tracks. In our model, we derive these factors in an aggregated way from data by comparing delays of trains before and after crossing every edge. Specifically, we determine the edge weighing factor β_{ij} for such spatial non-uniformities for each edge ij as follows:

$$\beta_{ij} = \frac{\langle D_{\text{arr}}^{ij} \rangle}{\langle D_{\text{dep}}^{ij} \rangle} \quad (4.2)$$

where $\langle D_{\text{dep}}^{ij} \rangle$ and $\langle D_{\text{arr}}^{ij} \rangle$ denote the observed average departure and arrival delays of trains moving from node i to j (i.e., across edge ij), respectively. In other words, it quantifies how much delay changes along this edge: $\beta_{ij} < 1$ indicates that delays are, on average, decreased, while $\beta_{ij} > 1$ indicates an average increase of delays when passing through edge ij . (We take the averages in both the numerator and the denominator in Eq. (4.2) to prevent near-zero arrival or departure delay times from strongly altering the β value of the edge.) Examples and the intuition of β are illustrated in Fig. 4.1b.

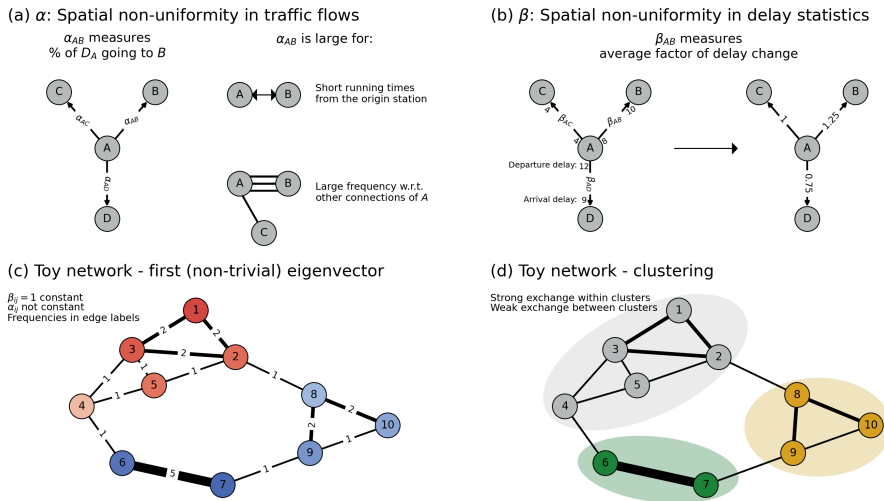


Figure 4.1: Panel (a): Illustration of the spatial non-uniformity in traffic flows α . Note that α_{ij} also contains a term concerning the running time of trains from i , but the proportions of all outgoing links from i remain in tact; hence the description of α_{ij} as the fraction of delay in i going to j . Panel (b): Illustration of the spatial non-uniformity in delay statistics β . Panel (c): First non-trivial eigenvector of an example network where we assumed β constant, and α only variant because of changing transport frequencies (denoted in edge width and labels). Eigenvector coefficients are shown in node colours, from blue to red. Panel (d): Clustering of the example network, where $K = 3$.

4.2.3 Model

The multiplications of the spatial non-uniformity factors α and β result in weights of edges that reflect how delay geographically spreads through a railway network. This can be viewed as a delay propagation model, although we never use it to actually simulate delays — we use it to identify communities. The intuition of this model is as follows. Consider two stations I and II, connected by edge ij . Given an initial delay of D_0 at station I, it can be deduced from the timetable that a fraction α_{ij} of D_0 is propagated towards II — the rest, $(1 - \alpha_{ij}) \cdot D_0$, either remains at I or travels towards other stations. During the propagation of $\alpha_{ij} \cdot D_0$ towards II, the delay is increased or decreased on edge ij by a factor β_{ij} . In other words, the delay arriving at II in the next time step is $\alpha_{ij}\beta_{ij}D_0$. We generalise this reasoning by constructing the matrix M by element-wise multiplication of the factors α and β for any pair of nodes i and j in the railway network:

$$M_{ij} = \begin{cases} \alpha_{ij}\beta_{ij} & \text{if } i, j \text{ are directly connected} \\ 1 - \sum_k \alpha_{ik}\beta_{ik} & \text{if } i = j \\ 0 & \text{elsewhere} \end{cases} \quad (4.3)$$

As we are mainly interested in the dynamics and direction of delay propagation, rather than the absolute values of delay, we assume our model to conserve delay. This requires

the matrix M to become row-stochastic, which is realised by defining the diagonal elements as in Eq. (4.3) to be $1 - \sum_k \alpha_{ik} \beta_{ik}$. We refer to the matrix M as the *dynamical* matrix, which, given a delay vector $\vec{D}(t)$ at time t , gives the delay vector at the next time step $t + \Delta t$ through the following multiplication:

$$\vec{D}(t + \Delta t) = \vec{D}(t) + \Delta t \cdot M \cdot \vec{D}(t) \quad (4.4)$$

where Δt refers to a time step, which is arbitrary and not of concern for the research in this chapter. Both α and β are dimensionless and we only use M to find out *structural* properties of the railway systems rather than for explicit delay modelling — i.e., even though Eq. (4.4) illustrates an interpretation of M , we do not use this equation.

4.2.4 Properties of M

Many important model properties can be derived analytically from the row-stochastic matrix M , and we focus our research to the properties of this matrix. (Note that M contains averaged, time invariant, entries through the definitions of α and β and therefore does not depend on time.) Any delay simulation done using M eventually ends up with a ‘trivial’ delay vector with equal entries: $\vec{D}_i(t) = \vec{c}$ at all nodes i , where $\vec{c}_i := c = \frac{\vec{D}(t_0)}{N}$ with N the total number of nodes. The trivial vector reflects that, given enough time, delay is spread in all corners of the graph (assuming it being connected). The question of how delay spreads in the *transient* is of interest to us, i.e., before this trivial solution. One way of identifying structural properties of a graph-based model is to look at the spectral modes. The intuition behind this is to find geographical delay patterns \mathcal{D} that are persistent, i.e. slow to diffuse. Such patterns reveal natural boundaries that delays might not easily cross and geographical divisions into regions where delay is easily exchanged. In 2.5, we discuss how we use these spectral modes to define clusters. Mathematically, the problem of finding spectral modes \mathcal{D} is defined in the eigenproblem of matrix M :

$$M \cdot \mathcal{D} = \lambda \mathcal{D} \quad (4.5)$$

with eigenvectors \mathcal{D} and eigenvalues λ . The closer λ is to 1, the slower the delay decay in a (relative) delay distribution fixed by \mathcal{D} (this follows directly from Eq. (4.5)) — i.e., the more persistent is the geographical delay pattern \mathcal{D} . In particular, there is one solution with $\lambda = 1$ such that $M \cdot \mathcal{D} = 1 \cdot \mathcal{D}$, because M is row-stochastic. This is the aforementioned solution with constant coefficients, and corresponds to the first eigenvector. This does not provide any insight in the dynamic structure of the system, and is referred to as the ‘trivial’ eigenvector. The attribute of these eigenvectors being persistent, points to dynamical connections among nodes with approximately equal coefficients in an eigenvector. In other words, the coefficients of the eigenvectors can be used to find clusters of nodes that are have a dynamical connection, as we follow up on in the next section.

4.2.5 Clustering

The clustering is based on the coefficients of the eigenvectors. The defining property of eigenvectors of M with eigenvalues close to 1 is that their coefficients are approximately

conserved under multiplication with the matrix M , which through Eq. (4.4) defines the change of any delay vector. In other words, geographic delay distributions following the distribution of coefficients in an eigenvector are amplified over time. As illustrated in Fig. 4.1c, it therefore makes sense to cluster eigenvector coefficients based on their value. A single eigenvector provides interesting information already, which is why we propose a spectral approach in this chapter. Combining multiple eigenvectors, however, allows us to do the actual clustering. Let us assume that we are searching for K clusters (the question of choosing K is addressed later). A common method of finding these clusters is by creating a K -dimensional embedding based on the first $K - 1$ non-trivial eigenvectors (as the first eigenvector is constant and does not add any information), and applying a K -means algorithm to this space (Dam, A. van et al., 2021; Luxburg, 2007; Shi and Malik, 2000; Ng et al., 2002). Indeed, this results in K clusters that are based on the respective differences between eigenvector coefficients. There are many more clustering methods, with various advantages and disadvantages. The advantage of K -means clustering is that it is one of the most well-known and intuitive Euclidean distance-based clustering methods. The disadvantage of K -means is that many implementations of the algorithm are not deterministic, and that K is not automatically defined.

We derive an appropriate value of K based on a procedure referred to as the ‘eigen-gap heuristic’ (Luxburg, 2007), which is based on the eigenvalue spectrum. High (near-1) eigenvalues correspond to relatively persistent — and thus to us important — eigenvectors [. Sudden ‘jumps’ in the eigenspectrum therefore point to a group of more important eigenvectors (those with higher λ) and the rest of the spectrum (those with lower λ) and can thus be used to distinguish which eigenvectors are therefore of interest. Assuming an equal number of clusters one can distinguish with this set of eigenvectors (although the trivial eigenvector is not useful in this analysis), the largest ‘eigen-gap’ in the eigenspectrum defines K . Having a maximum number of desired clusters (in the remained of this chapter, we use $K_{\max} = 15$), results in:

$$K = \max_{i=2}^{15} \{\lambda_{i-1} - \lambda_i\} \quad (4.6)$$

Summarised: we start with the eigendecomposition (Eq. (4.5)), then from the eigenvalue spectrum we determine K , we continue by constructing the $K - 1$ -dimensional embedding with the first $K - 1$ non-trivial eigenvectors (the minus-1 stems from excluding the trivial eigenvector), and apply K -means to the coefficients in this space. This results in K clusters.

4.2.6 Toy example

We illustrate the model and associated clustering in a fictitious transport system consisting of 10 nodes and 14 edges in Fig. 4.1c-d. Trains go from node to node in a networked manner and their (bidirectional) frequencies are denoted in numbers on each edge, creating geographical differences in α . The factor β is assumed here to be constant: $\beta = \beta_0 = 1$. The resulting M matrix gives the first non-trivial eigenvector as displayed in colours in the nodes in Fig. 4.1c. By eye already, one can distinguish the red coloured nodes (coefficients > 0) from the the blue coloured nodes (coefficients < 0), which also

makes sense dynamically: even though connections between nodes 4 and 6, and between 2 and 8 exist, they are much weaker than the interconnections between nodes 1-5, and subsequently do not bring the coefficients (i.e., of 4/6 and 2/8) of the first eigenvector close together. It turns out that for this system, $K = 3$. The resulting clustering is found in Fig. 4.1d, confirming our observations by eye on the coefficient separation in Fig. 4.1c: the algorithm groups the interconnected region of 1-5. It also distinguishes nodes 6 and 7 from 8-10, as a result from the strong connection between 6 and 7. While this is only a fictitious toy network, in the following sections, we apply the same algorithm to real and much larger transport networks.

4.3 Application to European railway systems

We apply the algorithm to data of nation-wide railway systems of four European countries: the Netherlands, Italy, Germany and Switzerland. We have chosen these systems based on their relative comparability: railways in the United States, for example, have a strong emphasis on cargo transport (in contrast to European railways, having more emphasis on passengers) affecting frequency and regularity of the timetables, and even various topological aspects. Another example is the Chinese railway system, differing from European systems in terms of scale: having fewer stations (per unit area) and much longer running times. Hence, for the illustration of the methods in this chapter, we focus on four railway systems that are relatively comparable, but still have smaller cross-differences. We start this section by elaborating on the data itself and topological properties of these systems, after which we present the spectral results of the M matrix for every of these four systems. We end this chapter with showing the resulting clusters.

4

4.3.1 Data

We utilize operational data from the Dutch, Italian, German and Swiss railway systems, including data on infrastructure, schedules (used to determine the values of α_{ij} for every edge ij) and realized delay data (used to determine the values of β_{ij}). Details on the source and cleaning of the data can be found in SI D. The data contains departure and arrival times, locations and their delays, per unique train number, along with infrastructure information on stations, their connections and longitude-latitude data. We obtain daily average variables like frequencies and running times by looking at periods of 16 to 31 days (depending on the country, see SI D). Frequency is computed as average number of trains per hour by summing the daily number and dividing by 24. A minimum value of 1 minute running time (in hours) is taken: i.e., lower average running times are approximated to 1 minute. In Tab. 4.1, an overview of several static properties of the railway systems is given. Insights in the networks themselves is supported by also showing two network metrics: the average *degree*, which refers to the average number of links each node has, and the average *betweenness*, which is the fraction of shortest paths between all pairs of nodes that pass through a respective node, averaged over all nodes.

In Tab. 4.1, we can see several differences across the four systems. When interpreting these numbers, it is good to emphasise that the distance between nodes and the level of detail varies across the four datasets: for example, in the Netherlands, the data is more

Variable	Netherlands	Germany	Switzerland	Italy
# connected nodes	658	5815	1346	2201
# edges	1438	15764	3827	6683
Average degree	2.26	2.90	2.56	3.26
Average betweenness	0.048	0.0052	0.0095	0.013
Average daily # train activities per node	173	20.54	102.35	22.47
Average daily # unique service lines per node	10.8	2.2	10.3	2.8
Average α^*	0.201	0.117	0.132	0.067
Average β^*	1.009	0.974	0.837	0.991

Table 4.1: General overview of the four countries assessed in this chapter. Node numbers are determined after removal of nodes that are not connected to the giant component. *Here, we show the average *non-diagonal* α and β values.

detailed the level of passenger stations: there are sensors near the tracks (also outside of stations) that log whether trains are passing by. This level of resolution is higher for the Dutch case than in the other sets, affecting static properties of the system like reducing average running time between edges or average degree, for example. (This resolution difference will not affect cross-system comparisons in later figures.) But even when taking this note into account, we conclude that the German railway system is clearly the largest, in terms of both nodes and edges. The low average degree and high average betweenness in the Netherlands can be explained by the fact that it includes many degree-2 nodes sequential on a line (by construction, but partly also due to the high resolution of the data), rather than having larger hubs that are interconnected.

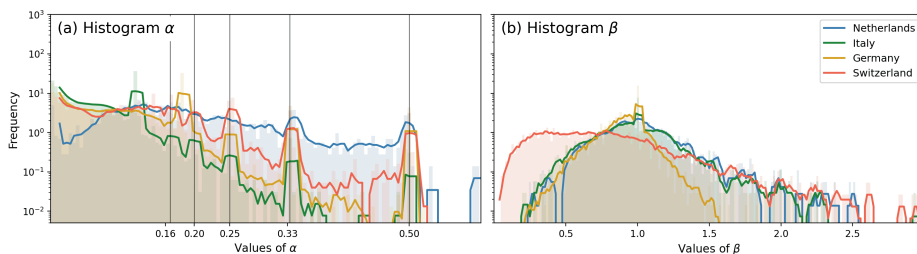


Figure 4.2: Normalised histograms of occurrence of α and β values in the four analysed countries, excluding diagonal elements of the α and β matrices. In panel (a), the vertical lines indicate fractions of 1, marking α values that are associated with proportional train frequencies or integer-minute running times.

From a dynamical perspective, the values about service lines and train activities are more relevant. Service lines are defined as single-direction trips from a starting station to an ending station (commonly at the other side of the country), crossing various in-between stations. These lines are denoted with a unique number and are important advectors of delay (Dekker and Panja, 2021), which is why their structure is important for railway dynamics. The smaller systems (Switzerland and the Netherlands) clearly dominate with respectively 10.3 and 10.8 unique lines per node, in comparison to Germany and Italy, having values of 2.2 and 2.8 respectively. This reflects the dynamically denser nature of the Dutch and Swiss railway systems, and is confirmed in the higher number of daily train activities per node (102-173 for the Netherlands and Switzerland as opposed

to 20-22 for Italy and Germany).

The average α and β values also have an important meaning. They are proxies for the density of trains — modulated by frequency and running time — and statistical delay increases (on edges), respectively. A relatively high α in the Netherlands and, to lesser extent, Switzerland depicts relatively high train density between stations and tracks of degree-2 nodes to prevent scheduled fragmentation of delay spread. The β values are like a railway edge-performance metric: the lower, the less delay is increased along its edges. Interestingly, the dense, highly utilised Swiss railway network performs best when looking at these numbers, having an average β of 0.84. (Note that for a full performance comparison among these railway systems, one should statistically correct for a number of factors like delay changes *within nodes*, which is not included here.) The histograms of α and β of the four countries are shown in Fig. 4.2a and b, respectively. In several countries, α values correspond to integer-fractions of one: 1/2, 1/3, 1/4, etc., reflecting integer-minute running times on tracks where only 1 line is traveling, or the respective proportions as calculated from the relative train frequencies (see Fig. 4.1a). This is most notable in Switzerland and the Netherlands, where we expect such short tracks (of 1, 2, 3, 4 or 5 minutes running time) and many degree-2 nodes to exist, because of the high density of these railway systems. The Italian system has more lower values of α than the other countries (with the Dutch system having the fewest), corresponding to either tracks connected to nodes with trains in many (other) directions, or tracks with large average running times. In panel (b), the distributions of β are wrapped around 1, corresponding to no significant average change in delay on these tracks. Furthermore, we see that Switzerland has more β values lower than 1 in comparison to other countries — the latter two mainly having values of β close to 1.

4.3.2 Spectral results

From the values of α and β , we determine the matrix M using Eq. (4.3). Fig. 4.3 shows the first non-trivial (i.e., second) eigenvectors of the four countries in panels (a)-(d), and the associated eigenspectra in panel (e) (with a normalised horizontal scale).

The coefficients of the first non-trivial eigenvector in all four countries show a dipole-like structure. In the Netherlands (Fig. 4.3a), the coefficients depict a north-south gradient, highlighting a northern region (near Groningen) in negative coefficients and the south (towards Maastricht) in positive coefficients, with a better connected centre in between, including cities like Rotterdam and Amsterdam. The Italian coefficients (Fig. 4.3b) also show a north-south gradient. Germany (Fig. 4.3c) shows a clear separation of the area south of Frankfurt (including cities like Karlsruhe and Stuttgart) in red. Apparently, this area may have persistent delays that are less easily exchanged with the rest of the country. The Swiss graph (Fig. 4.3d) shows, like in the Dutch and Italian case, a geographic gradient: from east to west, highlighting the south including the cities of Geneva, Lausanne and Sion on the west.

Figure 4.3d shows the eigenspectra of the four countries, with a normalised horizontal axis to compare the spectra independent of network size. The eigenvalues are relatively high. The slow decrease of the eigenspectra indicates that these are not strongly disconnected clusters. The relatively faster decrease of the Dutch eigenspectrum reveals that in the weighted network spun by M , the delay changes in Dutch railway system is

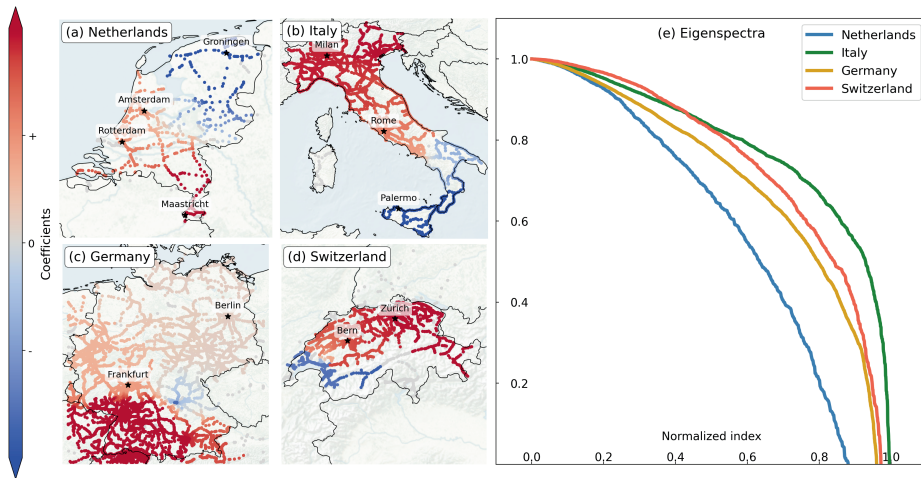


Figure 4.3: Panel (a)-(d): First non-trivial eigenvectors of the M matrix, for the four European railway systems: (a) the Netherlands, (b) Italy, (c) Germany and (d) Switzerland. As absolute values of the coefficients are not relevant (only their relative magnitude to other ones in the same country are relevant), the colourbar and coefficients are shown normalised. Several important cities are annotated by black stars. Panel (e): Eigenspectra of the four countries, with normalised horizontal axis.

more strongly dominated by the first few components, as opposed to others. This reflects that the first K eigenvector are better suited for compartmentalising the Dutch railways into subregions, than they are for other countries.

The (simple) dipolic structure of these eigenvectors and the fact that there are many high-value eigenvalues in the spectrum both point to the need of multiple eigenvectors to obtain a more refined view of dynamic structures in these networks.

4.3.3 Clustering results

Using the eigengap heuristic and the eigenspectra in Fig. 4.3d, we determine K — the number of clusters to search for, and the dimension of the embedding (see Sec. 4.2.5). Subsequently, we apply the K -means clustering algorithm to the eigenspace and find the clusters shown in Fig. 4.4, with the geographic locations of the clusters in panels (a)-(d) and abstracted networks (including delay exchange in the arrow widths) in panels (e)-(h).

In general, a total of 10, 11, 9 and 10 clusters are found in the railway systems of the Netherlands, Italy, Germany and Switzerland, respectively. By nature of the planar (2D-) structure of railway networks, the clusters are geographical intraconnected regions. In the Netherlands (Fig. 4.4a), the country is cut up in mostly equal-sized regions, with exceptions of smaller clusters (2, 5, 9) at the edges of the country (e.g. in the rural areas of Zeeland in the south-west, Limburg in the south, and a poorly connected part in the north-east) — probably as a result of service lines being less frequent and more disconnected from the centre of the country in these areas. Easily spotted are the central cluster 1, including the major cities of Amsterdam and Utrecht, and cluster 3, including Rotterdam, together including most of the so-called ‘Randstad’, the most urbanised and

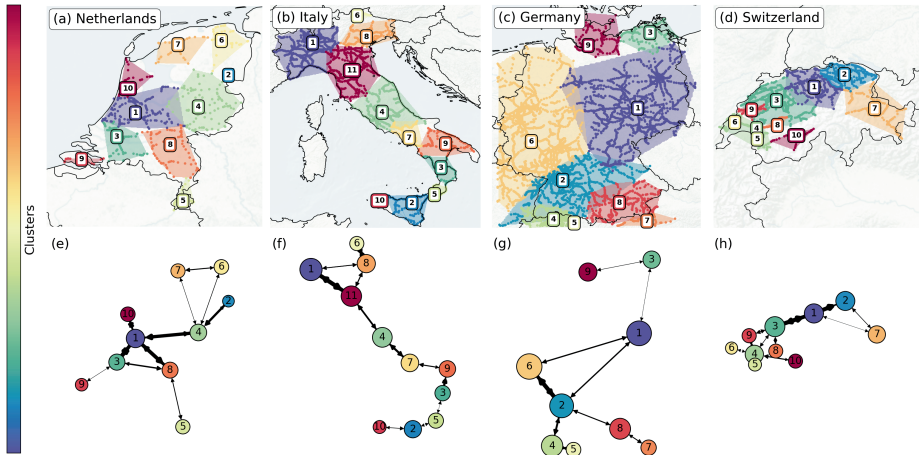


Figure 4.4: Panel (a)-(d): Clusters in the four European railway systems as found using K -means on the embedding built by the first $K - 1$ non-trivial eigenvectors of the matrix M (where K is found using the eigengap heuristic). In particular, K equals respectively 10, 11, 9 and 10 for these four countries. **Panel (e)-(h):** Abstracted versions of upper panels, reflecting the clusters and their connections. Arrows indicate connectivity between clusters in terms of α and β — the elements of M . The arrow width depicts the total sum of elements of M directed from one cluster to another. Node sizes reflect number of railway stations in each cluster.

transport-heavy area of the Netherlands. The clustering in Italy (Fig. 4.4b) also subdivides the country into more-or-less equal areas, with some exceptions of small clusters in the far north and south (clusters 5, 6, 10). Three notable clusters can be identified, including important Italian cities: 1, being the largest with Milan and Turin, 11, including most of Tuscany and Bologna, and 4, including Rome. The island of Sicily is only weakly connected to the rest via a train ferry at the city of Messina, and the island is further subdivided into two other clusters. Germany (Fig. 4.4c), in contrast, is subdivided in only 9 clusters (although being the largest railway system among these four) that are not equal in size: clusters 1, 2, 6 and 8 add up to than 85% of the country, including several foreign tracks. Cluster 6 almost perfectly coincides with the German federal states ('Bundesländer') of Nordrhein-Westfalen, Hessen and Rheinland-Pfalz, while cluster 1 covers most central-eastern states (these Bundesländer are also used in operations by the dominant railway company, Deutsche Bahn). Several much smaller clusters can be found in the south: clusters 4, 5 and 7, even being foreign (non-German) clusters and hence not of importance here. The large size of the clusters reflect that railway transport on the tracks are generally well distributed, travel long distances, and are less regional. The Swiss partitioning (Fig. 4.4d) results in rather small, intricately connected clusters (as is also visible in the abstracted graph in Fig. 4.4h). Several major urban areas can be recognised like the Geneva-Lausanne area (cluster 4) and the area around Bern (cluster 4), while the area around Zürich is divided into two clusters (cluster 1 and 2).

In the abstracted plots (panels (e)-(h)), the arrows and their widths provide information on the general flow of delay, proxied by the elements of M : thick arrows mean more (and larger) elements of M between these clusters, indicating stronger pathways of delay propagation as dictated by the α and β values, while small arrows mean the opposite.

The Netherlands has a relatively well connected dynamical core in its centre: exchanging most delay among the four clusters 1, 3, 4 and 8, containing several major transport corridors for both passengers and freight. The Italian plot indicates that most exchange is in the centre and north, between clusters 1, 8, 11, 4 and 7, reflecting touristic and urbanised areas, possibly leading to well-connected railway operations. The German clusters are mostly connected in the west and south: forming a strong bond between 2 and 6 in particular, reflecting some international transport and core lines from Cologne and Frankfurt to Stuttgart. The core connections of the clusters in Switzerland are mostly along the northern border, connecting the cities of Bern and Zürich.

4.4 Interpretation and relation to daily operations

The clusters reflect regions where delay is expected to be propagated within the cluster, and less so towards outside of the cluster, incorporating the two factors of spatial non-uniformity. This section is devoted to interpret and understand the relevance of these clusters in terms of the delay and daily operations. We start by defining two metrics that allow such interpretation and apply them to the clusters found in Fig. 4.4. Subsequently, we compare the four countries and their clusters in terms of these metrics.

4.4.1 Metrics to interpret clusters

To find the operational and dynamical meaning of the clusters found in Fig. 4.4, we first define a few relevant quantities that we later combine into two cluster-characterising metrics. We attribute any train's delay to its *departure* location. The first quantity is the total delay D_{total} from stations in the cluster. Second, we determine the internal delays D_{int} : delays of trains departing and arriving within the cluster. Third, exported delays D_{exp} of trains departing from within the cluster, but arriving in another, are computed. And fourth, the imported delays D_{imp} of trains departing from another cluster, arriving inside the analysed cluster. The following relation holds: $D_{\text{total}} = D_{\text{int}} + D_{\text{exp}}$. With these quantities, we compute two metrics which allow for easy interpretation of the clusters: (1) their fraction of the country-wide delays, measured by the *cluster severity* and (2) their dynamical (dis-)connectedness to other clusters, proxied by the *cluster independence*.

Cluster severity $S(n)$ for any cluster n is the ratio of D_{total} of cluster n to the average D_{total} over all clusters:

$$S(n) = \frac{D_{\text{total}}(n)}{\frac{1}{N} \sum_i^N D_{\text{total}}(i)} \quad (4.7)$$

with N being the total number of clusters. So, if $S(n) > 1$, the cluster covers an above-average part of the delays in the country, and vice versa. This does not incorporate cluster size, meaning that if all stations cover equal amounts of delay, larger clusters immediately have larger values of $S(n)$. We have chosen to not account for cluster size to make $S(n)$ a property of the cluster as a whole, make it better interpretable and relate to the practical use of the metric: small clusters might otherwise attain very high values of $S(n)$ while in practice not that dominant in delay.

Cluster independence $I(n)$ for any cluster n is defined as the ratio between delays exchanged internally in the cluster, w.r.t. the delays exchanged with other clusters:

$$I(n) = \frac{D_{\text{int}}(n)}{\sqrt{G(n)} \cdot (D_{\text{imp}}(n) + D_{\text{exp}}(n))} \tag{4.8}$$

where $G(n)$ is the number of nodes in cluster n . The factor $\frac{1}{\sqrt{G(n)}}$ is included to counteract the bias that larger clusters automatically have more internal delays ($D_{\text{int}}(n)$) than they exchange with neighbouring clusters ($D_{\text{imp}}(n) + D_{\text{exp}}(n)$). We assume here that the total internal delay grows with the number of stations in the cluster (e.g., $D_{\text{int}}(n) \propto G(n)$) and that the delay exchange grows with the square root of that (e.g., $D_{\text{int}}(n) \propto \sqrt{G(n)}$) — much like the area of a circle grows with the radius squared, but the circumference merely grows with the radius.

Note that we do account for the cluster size bias in $I(n)$, but not in $S(n)$. This might seem inconsistent. However, $S(n)$ is already normalised by the average value of total delays in the clusters, while $I(n)$ is not a normalised value. The quotient of internal delays versus delay exchange (i.e., $I(n) \cdot \sqrt{G(n)}$) will vary greatly and therefore, adding $\sqrt{G(n)}$ as a normalisation factor benefits the interpretation of $I(n)$ to be this quotient relative to the cluster size. More specifically, the resulting interpretation would be that $I(n) = 1$ for all clusters where the internal delays are exactly $\sqrt{G(n)}$ as large as the delay exchange, allowing a cross-comparison of large and smaller delays. Note that we proxy (and refer to) dynamical ‘connectedness’ with $I(n)$, which formally only measures the amount of delay exchange with surrounding clusters. In the text, this terminology is interchangeably used.

The metrics $S(n)$ and $I(n)$ both surround values of 1: values lower than 1 depict clusters that cover a less-than-average delays and are well-connected, respectively, and values higher than 1 reflect clusters with more-than-average delays and that are less connected. This allows to split the $I(n) - S(n)$ plane into four parts, that can be used to interpret the clusters. Four cluster categories can be distinguished, as displayed in Tab. 4.2: named for easier reference Type A, Type B, Type C and Type D clusters.

	$S(n) < 1$	$S(n) > 1$
$I(n) > 1$	Type A: less delay exchanged with other clusters small amounts of originating delay	Type B: less delay exchanged with other clusters large amounts of originating delay
$I(n) < 1$	Type C: more delay exchanged with other clusters small amounts of originating delay	Type D: more delay exchanged with other clusters large amounts of originating delay

Table 4.2: Cluster categories in the $I(n) - S(n)$ plane.

4.4.2 Metric results

The values of $I(n)$ and $S(n)$ for all clusters n across the four countries (in colours) are plotted in Fig. 4.5 (a few clusters with too little data to estimate $I(n)$ and $S(n)$ well are left out). The numbers in each circle refer to the numbers in Fig. 4.4.

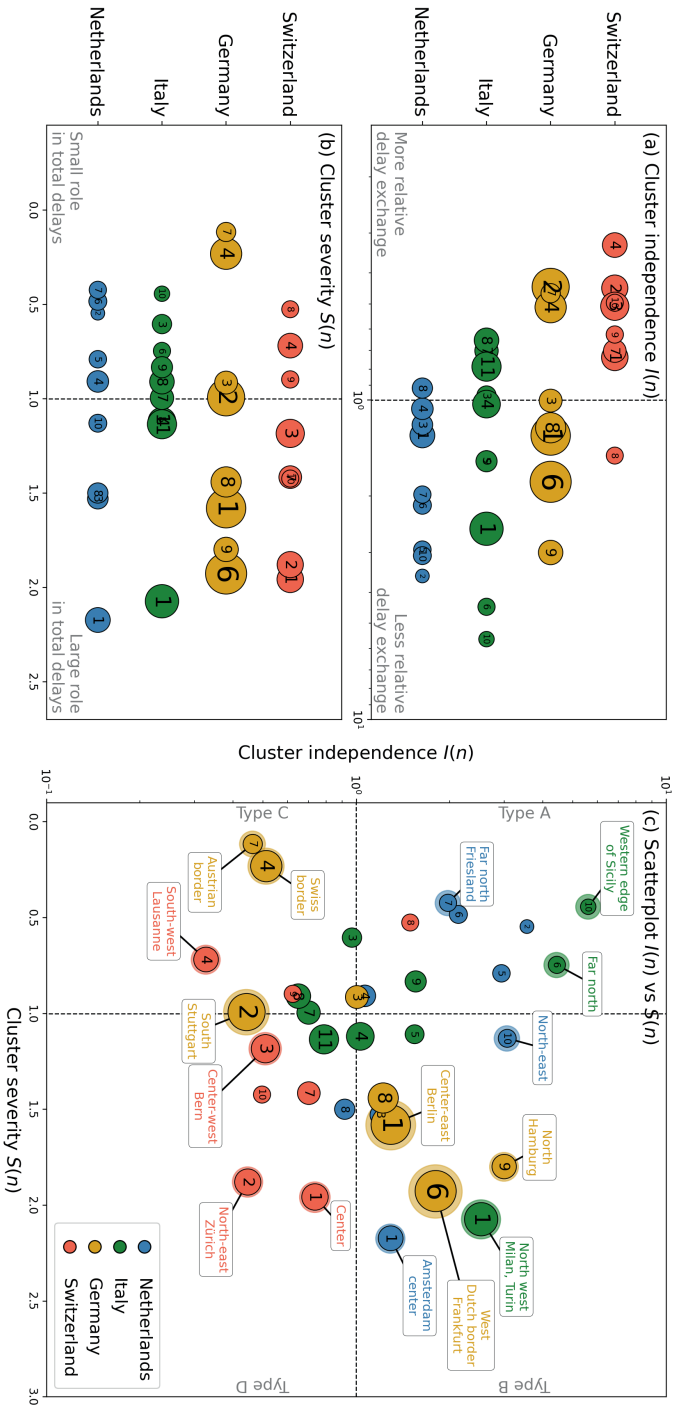


Figure 4.5: Cluster independence $I(n)$ and cluster severity $S(n)$ of the clusters across the four analysed countries. **Panel (a)** - **(b)**: $I(n)$ and $S(n)$, respectively, on the horizontal and countries on the vertical. **Panel (c)**: Scatterplot with $I(n)$ versus $S(n)$. Several notable clusters are annotated in text and their dots are emphasised. Dot size reflects the number of stations inside the cluster. Clusters with only a very small amount of data (e.g. due to being across the border, involving only international trains) are not included in this figure. Numbers in the clusters refer to numbers in Fig. 4.4.

The panels (a) and (b) allow for both an in-country comparison of the clusters as well as a cross-country comparison of the spread of the $I(n)$ and $S(n)$ values. In panel (a), the immediate observation is the relatively low $I(n)$ values in Switzerland in contrast to relatively high values in the Netherlands, with Germany and Italy approximately in between. Apparently, many clusters in the Swiss system exchange more delays with other clusters than they have internally. This may be a consequence of the fact that this relatively small country has many trains connecting clusters, rather than travelling in a separated space. This points to a potential source of vulnerability: delay can quickly spread across the country. Looking more closely to the cluster numbers within each country, we can find mostly urban and central areas having low $I(n)$ values and regions place at the edges of the country having higher $I(n)$ values. Examples are the area of Zürich, Bern and Lausanne (numbers 2, 3, 4) having low- $I(n)$ in Switzerland, while the Milan region (1) and several regions at the edges (Sicily and far north) in Italy have much higher $I(n)$ values. In the Netherlands, the highest $I(n)$ values are found in rural areas in the north and south. An exception seems to be Germany, where clusters 1 and 6 are so large that they convey well-separated, but urban areas.

The $S(n)$ values in panel (b) are by definition spread around the value of $S(n) = 1$, because $S(n)$ is normalised by its average (Eq. (4.7)). Still, the spread provides valuable information on these countries. While the Swiss clusters are relatively well spread across the average, the German systems have a few clear outliers on both sides: the southern border (4 and 7) clearly having the smallest fraction of delays, whereas the area around Köln (6) has the largest fraction of delays. Both Italy and the Netherlands show many clusters in the low- $S(n)$ domain, with a few outliers with high $S(n)$ values, mostly in the bigger cities — Amsterdam, Utrecht and Rotterdam (1 and 3) in the Netherlands, and Milan (1) in Italy.

In panel (c), the clusters are shown in the combined $I(n) - S(n)$ plane. The top-left quadrant (Type A clusters) mainly comprises Dutch and Italian rural clusters, such as the far north or south in these countries: the areas of Friesland and Limburg in the Netherlands and Sicilian and northern areas in Italy. The rural nature of clusters in this category is no surprise, as their periphery-located position in the network make them usually less connected with the rest (i.e., high $I(n)$) and less prone to delays (i.e., low $S(n)$). The Type B clusters ($I(n), S(n) > 1$) are mainly found in Germany, with the busy, well-connected areas around Berlin, Köln and Hamburg. One other notable cluster is the urban Milan region in the north of Italy (cluster 6), clearly separated from other Italian clusters. At the bottom-left we see the Type C clusters, containing several German, Swiss and Italian clusters that are quite well connected, but play less of a role in the total delay. The Type D clusters (bottom-right), are mainly found in Switzerland, interpreted as well-connected clusters that also play a large role in the total delay. The Swiss Type D clusters are the relatively urban regions in this country around Zürich and Bern.

Summarised, comparing countries in the $I(n) - S(n)$ plane reveals that Switzerland generally has low values of $I(n)$, implicating the strong connectivity of this railway network. Germany is not really represented in the Type A and Type D categories, and is well represented in Type B. This may reflect that German delays are usually quite compartmental: large delays may arise but these large clusters mainly keep these delays within. The Italian and Dutch clusters show a large spread for values $S(n) < 1$, and several clusters around $I(n) = 1, S(n) > 1$. There seem to be a few clusters in both of these countries

that determine a disproportionately large amount of the total delay (high $S(n)$), but these clusters have an average connection with the rest of the country (unlike several German clusters). Overall, the patterns in Fig. 4.5 seem to relate to operational characteristics, topological embedding in the general railway network and even urban differences across the clusters.

4.5 Conclusion

To find geographical delay structures that have a dynamical meaning in railway systems, we developed a graph-based model and proposed a method to use spectral properties of this model to characterise the railway systems' delay dynamics geographically.

In particular, combining spatial non-uniformity factors that can be derived from the timetable and from statistics, all model information is encapsulated in the dynamical matrix M whose spectral properties can be used to compartmentalise any transport system's geographical network into clusters. We analyse four European nation-wide railway systems: the Netherlands, Italy, Germany and Switzerland. Both infrastructural and operational data are used, revealing dynamic properties of the four countries (Tab. 4.1) — making an interesting comparison in itself — and an optimal partition. The clusters vary in size and are connected in different ways to the nation-wide networks, resulting in the identification of core, central clusters and peripheric, near-disconnected clusters. The operational meaning of the clusters is expressed in two variables: cluster independence $I(n)$, reflecting the dynamical (dis-)connectedness of the cluster to the rest of the country, and cluster severity $S(n)$, reflecting the fraction of delay the cluster is responsible for. This leads to the classification of four cluster types, showing that high- $S(n)$ values are usually obtained by the more urban and dense clusters (e.g., high values relating to busy areas like the region around Milan, Zürich, Frankfurt and Amsterdam), and that $I(n)$ is partially distinguishes rural regions from central regions, but is partially also affected by how a country is operationally handled.

Throughout the chapter, various comparisons have been made between these four countries, revealing the Netherlands and Switzerland to be dynamically dense railway networks (i.e., having high utilisation of their tracks), while Germany and Italy are much larger and sparser networks. Shorter steps between (logging) stations make the edges on average small in the Netherlands, with the longest edges in Italy. The clusters in Switzerland show low values of $I(n)$, depicting a strong interconnectedness among them. The Italian $S(n)$ is dominated by the Milan region, with almost all other regions having $S(n) < 1$. The upper-right corner of the $I(n), S(n)$ plane is dominated by several German clusters near Hamburg, Berlin and Köln.

The matrix M consists of quantities α and β that are based on average quantities: single values per edge of the railway networks. It should be noted that, in reality, many ingredients of these quantities, such as running times, train frequency (in Eq. (4.1)) and the delay statistics in Eq. (4.2) may vary significantly over time and per train. Also the spatial non-uniformity factor β derived from delay statistics is dependent on seasonality and weather. Working with average quantities limits the potential of simulating specific instances of delays. Additionally, the model does not incorporate node-specific aspects built in the timetable, such as dwell time supplements and buffer times — the model

rather focuses on spatial non-uniformity on the edges to assess the directions and speed of delay propagation. These limitations assert that this model should not primarily be used for prediction purposes, but rather to gain time-average and system-wide understanding of geographic delay evolution.


Railways, by schedule, have long-stretched service lines going from one side of the country to the other. Per definition, this means that perturbations quickly pass through large parts of the country, as is also reflected in the high eigenvalues of M (Fig. 4.3). The structure of the dynamics therefore does not favour partition much, and care should be taken into separating them too strongly in the interpretation of Figs. 4.4 and 4.5. Still, compartmentalisation is an important topic in railways, in particular when it comes to defining regions to subdivide operations and dispatching. The proposed clustering should therefore be viewed more as a statistical average, or a composite of dynamic modes, rather than a full geographical segregation of delays. For future research, the connection and strength between the found clusters should be compared with delay propagation on a larger spatial scale: the relation between regional effects and nation-wide effects is still an important unresolved topic in transport literature. This work brings us one step closer to a solution.

For both scholars and railway practitioners, these results shape deeper understanding of how their railway systems work, how they differ from each other, and how regions with these systems each play a unique role. While the country-by-country statistics in Fig. 4.2 and Tab. 4.1 add to system-wide insights, the clustering in Fig. 4.4 may help practitioners to find boundaries when aiming to subdivide countries into smaller operational regions. Connections among the clusters may inform railway operators of statistically average directions of delay flows. And the characterisation of the clusters by $I(n)$ and $S(n)$ in Fig. 4.5 provides insights in how analogies between countries and regions can and cannot be made, and how different regions play different roles. For scholars, the construction and subsequent clustering of the M matrix is a simple procedure and can be generalised to many other transport systems, beyond railways alone. While delay is a quantity strictly bound towards a predefined transport schedule, in theory many other dynamic variables on networks can be analysed in the same manner. We believe that the presented methodology and results for the European countries contributes to deeper understanding of these systems, and we hope that this chapter ignites more research in the relation between regional effects and nation-wide effects in transport systems.

CHAPTER 5

MODELLING RAILWAY DELAY PROPAGATION AS DIFFUSION-LIKE SPREADING

This chapter can be found as a preprint: Dekker, M. M., Medvedev, A. N., Rombouts, J., Siudem, G., and Tupikina, L. (2021d). Modelling railway delay propagation as diffusion-like spreading. *arXiv:2105.06111*. Supplementary material to this chapter can be found in SI E.



ABSTRACT

Railway systems form an important means of transport across the world. However, congestions or disruptions may significantly decrease these systems' efficiencies, making predicting and understanding the resulting train delays a priority for railway organisations. Delays are studied in a wide variety of models, which usually simulate trains as discrete agents carrying delays. In contrast, in this chapter, we define a novel model for studying delays, where they spread across the railway network via a diffusion-like process. This type of modelling has various advantages such as quick computation and ease of applying various statistical tools like spectral methods, but it also comes with limitations related to the directional and discrete nature of delays and the trains carrying them. We apply the model to the Belgian railways and study its performance in simulating the delay propagation in severely disrupted railway situations. In particular, we discuss the role of spatial aggregation by proposing to cluster the Belgian railway system into sets of stations and adapt the model accordingly. We find that such aggregation significantly increases the model's performance. For some particular situations, a non-trivial optimal level of spatial resolution is found on which the model performs best. Our results show the potential of this type of delay modelling to understand large-scale properties of railway systems.



5.1 Introduction

Railway systems are of vital importance for transporting passengers and goods. The trains in these systems travel via predefined schedules that allow for highly efficient utilisation of the routes and tracks. Temporal deviations from such scheduled operations are commonplace. They take the form of *delays* and decrease the system's efficiency. Small delays are often absorbed by built-in buffers and therefore do not have effects on larger scales (Zieger et al., 2018; Dekker and Panja, 2021). However, from time to time, logistic disruptions — often caused by external factors like weather — lead to congestion or even a large-scale stand-still, with detrimental costs to society and economy (Ludvigsen and Klæboe, 2014; Tsuchiya et al., 2007; Büchel et al., 2020; Dekker et al., 2021f).

The above shows the importance of better understanding of delay propagation and its prediction. A large variety of delay propagation models exists, and the choice of the approach depends on a number of questions related to, among other factors, the spatial focus, availability of data and the delay severity. For example, when aiming to accurately predict the delay in a geographically confined area, there are high-performing statistical models (Kecman and Goverde, 2015b; Li et al., 2016). However, such statistics generally only work accurately in circumstances where delay is not too severe — as per definition these highly delayed scenarios are exceptional. Also, when upscaling to larger areas, long-range interactions and associated correlations come into play which may be difficult to account for when using average statistics. Larger scales and more highly delayed scenarios are therefore often analysed with machine learning or big-data approaches (Dekker et al., 2019; Oneto et al., 2018), but at the cost of understanding cause-and-effect or fine spatial resolution. Alternatives to such purely data-driven methods can be found in models where mechanisms of delay propagation are explicitly implemented. For example, Monechi et al. (2018) analysed the German and Italian railways and found a set of ‘laws’ that drive the spreading of delays, analogous to epidemic spreading models.

Of course, the infrastructure networks underlying the dynamical processes in any of the mentioned models play an important constraining role. However, this information is already embedded in the schedules and therefore less discussed in the context of delay simulations. However, the role of railway network topology is addressed by various scholars in relation to resilience properties (Sen et al., 2003; Erath et al., 2009; Bhatia et al., 2015). Most models are based on the schedules of the railway system, commonly using trains as agents that have the potential to carry delays. The perspective of delays as a properties of discrete trains or events can be found in many analytical models (Goverde, 2010; Gambardella et al., 2002; Büker and Seybold, 2012; Harrod et al., 2019; Dekker and Panja, 2021), using either deterministic or stochastic techniques to derive future delays from past information. Because of the abundance of this perspective in existing delay propagation models, we refer to the view of delays as properties of discrete trains or events as the ‘traditional view’. In contrast, one could also view delays as variables associated not to trains, but to the nodes (stations) and edges of the railway network, which stay in the same position. How delay spreads between these nodes does not have to be described in terms of discrete trains and events, but instead a description may rely solely on general (or even system-wide) quantities such as the network topology and schedule. One can make the analogy of fluid dynamics: while traditionally, delays are treated as Lagrangian particles (i.e., following the trains as the fluid carrying the particles), we propose

to treat delays from an Eulerian point of view (i.e., determining incoming and outgoing delays in a fixed spatial frame). This is also discussed in Dekker and Panja (2019). This is the basis of the model proposed in this chapter.

The traditional view of delays as discrete quantities of explicitly modelled trains or events is useful because it allows for tracking expected routes of delays along the train's trajectories explicitly. In other words, given that you know that delay is in the system at location A , it is unlikely to spread in all possible directions from A , but more likely to follow a particular direction that is dependent on which trains are exactly affected. One only knows this direction if discrete train units (and their trajectories) are explicitly included in the model. But there are also disadvantages of such models. One limitation is that many such models rely on many statistics in addition to mere schedule information. For example, if trains A , B and C are simulated explicitly, the interactions of all their events and relative magnitudes of their delays have an impact on each other's delays. These relations need to be well studied using for example neural networks (Oneto et al., 2017, 2018) or probability updating (Corman and Kecman, 2018; Berger et al., 2011). Another limiting consideration of treating delays as discrete quantities is the spatial scale. In confined systems, the mechanisms of delay propagation and their parameters can be well-defined, as in Kecman and Goverde (2015b). Defining all such interactions on a country-wide scale is generally more more complex, due to potential long-range correlations.

In this chapter, we propose to treat delays not as bound by discrete trains or events, but rather as continuously spreading across the infrastructure network. The spreading between nodes of the network is weighted by properties of the system. The intuition behind these models is that on average — in a 'mean-field approximation' — these parameters drive the overall direction of delay propagation. We refer to this way of treating delay propagation as 'diffusion-like spreading'. Small-scale accuracy is traded for larger scale accuracy: when looking at a micro-scale or individual trains, we expect this non-traditional way of dealing with delays to be less accurate than more detailed models, but on a large scale, we expect the performance of such a model to increase. As is shown in section 5.2, the model contains only simple schedule information (e.g., train frequencies and travel times) rather than complicated statistics, and all model information is embedded in a single matrix, which makes analysis of the system's properties easy. The mentioned reasons motivate us to write this chapter on delay propagation as a diffusion-like spreading mechanism. We apply our proposed model to the Belgian railways as a case study to discuss when and how it is advantageous to use such models.

An important aspect of delay propagation in general is the spatial scale and resolution of the analysis. High resolution ('micro-scale') modelling allows for explicit simulation of infrastructure capacity issues, the role of speed gradients or the identification of station-specific properties, for example. Low resolution, but large-scale ('macro-scale') modelling captures the impact of long-range interactions related to resource allocation (Dekker and Panja, 2021), the impact of long train lines (Dekker et al., 2019) or other system-wide properties. Many models lie between these extremes. Diffusion-like models should typically be regarded as having a lower resolution but working well on a larger scale, because of the earlier mentioned trade of small-scale accuracy for larger-scale accuracy. Spatial resolution is often expressed by treating railway infrastructure as a network, consisting of nodes (geographical locations) and edges (connections between them). At the highest spatial resolution, the nodes are certain control points in

stations and tracks, where train activities are logged (Dekker and Panja, 2021). More commonly is a slightly aggregated version of this, namely the more coarse passenger stations (Bhatia et al., 2015; Goverde, 2010; Monechi et al., 2018). Lower resolutions are obtained when constructing regions that correspond to groups of stations — so-called ‘clusters’, on which we elaborate later. Larger geographical areas in lower resolutions combine existing delays from higher resolutions and are treated as one unit. Choosing the correct level of spatial aggregation is an important consideration to make when assessing the viability of the diffusion-like model.

When discussing spatial aggregation, it is important to define how higher levels of aggregation are derived from the lower ones. In particular: how do we join stations and tracks together into larger and coarser regions? A large amount of complex network literature is devoted to this question of *clustering*, and clustering methods come in many forms in various applications (Fortunato and Hric, 2016). For example, clustering methods focusing on the structure of the underlying networks indicate how topology leads to a natural aggregation of nodes into larger groups. This can be quantified by the so-called modularity of the partition, first proposed by Newman (Newman, 2006). Various clustering algorithms based on modularity optimisation exist, such as the Louvain method (Blondel et al., 2008).

Spectral clustering focuses on properties of the eigenspace of the Laplacian or model-relevant matrices. A third common method for clustering any — also non-networked — data is K-means (Steinhaus, 1956; MacQueen, 1967), which defines centroids and groups nodes based on their respective distances to these cluster centroids (also known as Voronoi iteration (Lloyd, 1982)), given a definition of ‘distance’ between nodes. This method has been used in the context of transportation before, albeit mostly to characterise statistical space (rather than actual stations and physical space) (Kadir et al., 2018; Cerreto et al., 2018). An important aspect of K-means, in contrast to for example the Louvain method, is that it requires the specification of the number of desired clusters (K) up front, which can be both advantageous and disadvantageous. However, the freedom of choosing K turns out to be useful when analysing our diffusion-like delay model. This, together with the fact that K-means is a well known and commonly used method, motivates us to use K-means with geographical distance to cluster the stations in this chapter. By choosing the number of clusters, we vary the spatial aggregation level. We will compare the performance of the diffusion-like model on each of these levels.

In summary, the aim of this chapter is to discuss the usefulness of treating delay propagation as a diffusion-like spreading mechanism. We propose a model doing so in section 5.2. We apply the model to the example case of the Belgian railways and discuss the data and methodology for this in section 5.3. Section 5.4 discusses the results of a toy model, its performance on different types of disrupted situations, and the overall performance of the model. Here we discuss in what cases the diffusion-like aspect of the model is beneficial and what we can learn about the Belgian railways using this framework. We end with a summary and several conclusive remarks in section 5.5.

5.2 Model

In this section we introduce our diffusion-like model. We start by defining the delay variable and set up the equations that describe its evolution over time. We continue by discussing how this model can be generalised to any spatial scale. For a detailed derivation of the model, see SI E.1. Tab. 5.1 summarises the variables and parameters of the model.

5.2.1 General concepts

The main idea behind the model is to define the delay on fixed locations, and to describe the evolution of this delay distribution over time using macroscopic parameters such as train frequencies and travel times. While delays are inherent attributes of trains (i.e. agents), we aggregate the delays on passenger stations (i.e. nodes), as the impact of disruptions can mostly be felt at level of stations rather than being a problem of individual trains. This aggregation of delays onto stations means that we lose some of the finer details on which delays belong to which train. However, it will allow us to use tools for studying dynamical processes on networks: a delay is associated to each node, and its evolution is determined by the coupling of nodes through edges. For ease of notation, we will use the terms ‘station’ and ‘node’ interchangeably even though some nodes are actually junctions and not stations. We denote the delay of a station i at time t by $D_i(t)$. This variable is defined as the sum of the delays of all trains that are moving *towards* station i at time t :

$$D_i(t) = \sum_{T \in \mathcal{T}(i,t)} d_T(t), \quad (5.1)$$

where $\mathcal{T}(i, t)$ is the set of trains moving to station i (i.e. the very next station they cross will be i , whether they stop there or not) at time t and $d_T(t)$ denotes the delay carried by train T at time t . We consider two ways in which the value of D_i can change over time:

1. A train, which was previously moving towards another station j , reaches j and is now moving towards i . Therefore, its delay is now added to D_i .
2. A train, which was moving towards i , reaches i and either moves further towards another station or ends its trajectory. Therefore, its delay is removed from D_i .

The delay of station i at the next time step — we refer to this as $D_i(t + \Delta t)$, with Δt being the time step size — is dependent on the delays in various locations at the previous time step, not only $D_i(t)$. Thus, we write the relation between the delays between two consecutive time steps using a delay vector $\vec{D} = (D_1, D_2, \dots, D_N)^T$, where N is the total number of nodes:

$$D_i(t + \Delta t) - D_i(t) = \underbrace{F_{1,i}(\vec{D})}_{\text{New incoming trains towards } i} - \underbrace{F_{2,i}(\vec{D})}_{\text{Arrival of trains at station } i} \quad (5.2)$$

with $F_{1,i}$ describing how the delay at station i changes over a time step Δt by means of the first term above (the addition of delay), and $F_{2,i}$ likewise by the second term above

(the removal of delay). In the next section we express both these functions $F_{1,i}$ and $F_{2,i}$ in terms of several parameters and $\vec{D}(t)$. An illustration of the model and its terms is given in Fig. 5.1.

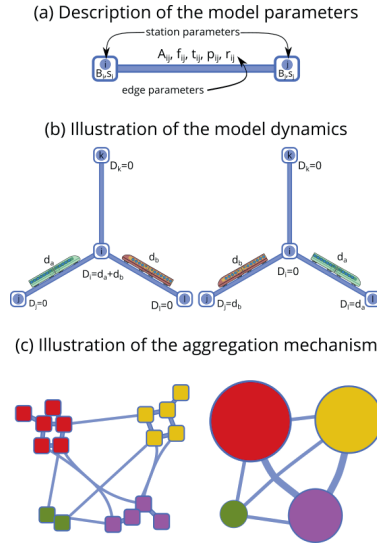


Figure 5.1: Model visualisation: (a) station and line (edge in the network) dependent parameters, (b) illustration of the two mechanisms behind the delay dynamics dynamics, i.e. the appearance of new trains with their delays and the departure of the already included ones, (c) an example of the network aggregation.

5.2.2 Diffusion model equations

The first term ($F_{1,i}$) sums the delays carried by all trains that start moving towards i in the interval $[t, t + \Delta t]$:

$$F_{1,i}(\vec{D}) = \sum_{\substack{\text{Trains } T \text{ that started moving to } i \\ \text{at } \tau \in [t, t + \Delta t]}} d_T(\tau).$$

This sum can be rewritten as a sum over the neighbours of i . By making a number of assumptions, like approximating the fraction of trains in each direction by the relative frequency (a full derivation can be found in SI E.1), we can rewrite the delay of a train moving to a station j as function of the delay of that station D_j and express how many of the trains arriving at a station j continue to i . This leads to:

$$F_{1,i}(\vec{D}) = \Delta t \sum_{j \in \mathcal{N}_{\text{in}}(i)} p_{ji} B_j D_j(t). \quad (5.3)$$

Here $\mathcal{N}_{\text{in}}(i)$ is the set of stations j that have an edge to i . The parameter p_{ji} is the probability that a train that reaches station j will continue towards station i , and is computed as follows:

$$\begin{aligned}
 p_{ji} &= P(\text{to } i | \text{from } j) \\
 &= P(\text{to } i | (\text{from } j \text{ \& do not end at } j)) \cdot P(\text{do not end at } j) \\
 &= \frac{f_{ji}}{\sum_{\ell \in \mathcal{N}_{\text{out}}(j)} f_{j\ell}} \cdot \left(1 - \frac{\text{Probability that train}}{\text{has end station at } j}\right) \\
 &= r_{ji}(1 - s_j),
 \end{aligned} \tag{5.4}$$

where $\mathcal{N}_{\text{out}}(j)$ is the set of stations to which there is an edge from j . The value of p_{ji} is equal to a multiplication of two factors. The first (denoted by r_{ji}) is the probability that if a train reaches j and it does not end its journey there, it will then continue towards i . Note that we consider this probability to be independent of where the train came from: we do not consider any memory in this process. The value is calculated as the frequency of trains going from j to i divided by the frequency of all outgoing trains from j . The second factor in Eq. (5.4) (denoted by $1 - s_j$) is the probability that the train does not end at j — s_j itself is the probability that a train that arrived at j ends its journey there, for example because it is the terminus. The variable B_i in Eq. (5.3) is a station-dependent parameter, defined as

$$B_i = \frac{\sum_{\text{edges } e \text{ to } i} f_e}{\sum_{\text{edges } e \text{ to } i} f_e t_e} = \frac{\sum_{\ell \in \mathcal{N}_{\text{in}}(i)} f_{\ell i}}{\sum_{\ell \in \mathcal{N}_{\text{in}}(i)} f_{\ell i} t_{\ell i}}, \tag{5.5}$$

where f_e denotes the frequency of trains on edge e , and t_e corresponds to the time a train takes to cross edge e . The parameter B_i has units of time^{-1} and can therefore be interpreted as a rate. The inverse of B_i is the average time of edges towards i , weighted by their frequency. A high value of B_i corresponds to a station with incoming short edges with high frequency. Intuitively, B_i can be thought of as a station's train turnover rate. The second term of Eq. (5.2) ($F_{2,i}$) counts the delays of trains that reach station i and therefore remove their delays from D_i . We express $F_{2,i}$ as follows (for details see SI E.1):

$$\begin{aligned}
 F_{2,i}(\vec{D}) &= \sum_{\substack{\text{Trains } T \text{ that reached } i \\ \text{at } \tau \in [t, t + \Delta t]}} d_T(\tau) \\
 &= \Delta t B_i D_i(t).
 \end{aligned} \tag{5.6}$$

The term only depends on the delay $D_i(t)$ at station i at the previous time step, and the previously mentioned parameter B_i . The delay loss at a station can be interpreted as an exponential process with rate B_i . The contributions F_1 and F_2 are expressed in terms of the delay state vector \vec{D} and in terms of various railway parameters (summarised in Tab. 5.1). Filling in these two terms into Eq. (5.2) gives the full expression for the evolution of the delay D at any station i :

$$D_i(t + \Delta t) - D_i(t) = \Delta t \left[\sum_{j \in \mathcal{N}_{\text{in}}(i)} p_{ji} B_j D_j(t) - D_i(t) B_i \right].$$

We can simplify the sum over the neighbours of i by using the railway network's ad-

jacency matrix A , which has entries $A_{ji} = 1$ if there is an edge from station j to station i and entries zero elsewhere:

$$\frac{D_i(t + \Delta t) - D_i(t)}{\Delta t} = \sum_j A_{ji} B_j D_j(t) p_{ji} - D_i(t) B_i.$$

Here, the sum goes over all nodes j . This equation can be written in matrix form using \vec{D} as a column vector. Moreover, we can take the limit $\Delta t \rightarrow 0$. This leads to the expression

$$\frac{d\vec{D}(t)}{dt} = \mathbf{G} \cdot \vec{D}(t). \quad (5.7)$$

The above equation contains the core model matrix \mathbf{G} , an $N \times N$ matrix defined as follows (δ_{ij} is the Kronecker delta):

$$G_{ij} = A_{ji} p_{ji} B_j - \delta_{ij} B_j. \quad (5.8)$$

All of the dynamics of the model are encapsulated in the matrix \mathbf{G} .

Variable	Description
$D_i(t)$	Delay at station i and time t
A_{ji}	Adjacency matrix of the railway network
B_i	Train turnover rate of station i
$d_T(t)$	Delay carried by train T at time t
f_{ij}	Train frequency from stations i to j
\bar{t}_{ij}	Average travel time from station i to j
p_{ji}	Fraction of trains to j that continues to i
r_{ji}	Fraction of trains to j that continue to i if they do not end at j
s_j	Fraction of trains that end at station j
$\mathcal{T}(i, t)$	Set of trains moving to station i at time t
$\mathcal{N}_{\text{out}}(j)$	Set of stations to which there is an edge from j
$\mathcal{N}_{\text{in}}(j)$	Set of stations from which there is an edge towards j
N	Number of stations
δ_{ij}	Kronecker delta ($\delta_{ij} = 1$ if $i = j$, and 0 otherwise)

Table 5.1: Overview of the model variables and parameters.

5.2.3 Model aggregation to clusters of stations

In this chapter, we aim to describe how well our model describes real delay propagation patterns. One variable in this analysis is the level of spatial aggregation at which we simulate the model. In the previous section we explained the model where each node of the network consists of a single station or junction. However, the same principles can be applied to a network where nodes correspond to a group of such stations. The method we use to group stations into clusters is explained in Sec. 5.3.4. Here, we discuss how the model parameters for the full-resolution model based on individual stations can be translated into a lower resolution version. The discussed aggregation process is very sim-

ilar to network of networks idea known in the networks literature (Gao et al., 2011; Kivelä et al., 2014; Siudem and Holyst, 2019).

Above, each delay variable D_i corresponds to one node of the network by transforming delays on trains to delays on stations via Eq. (5.1). This is already a form of coarse-graining the delay dynamics. Now, we assume that the original railway network of N stations is divided into K clusters (or groups of stations). We indicate stations with lowercase letters (i and j) and clusters with uppercase letters (I and J). The clusters naturally form a network: an edge between clusters I and J exists if there is at least one station i in I and one j in J such that there is an edge between i and j in the original network. We define the function \mathcal{C} from stations to clusters such that $\mathcal{C}(i)$ is the cluster to which station i belongs. Let $D_I(t)$ denote the total delay of all trains moving to any station in cluster I at time t , either from inside the cluster, or coming from other clusters. An equation for the evolution of this delay can be derived in the same way as we did above for stations. The delay D_I can change when trains start towards any station in this cluster, or when trains arrive at a station in this cluster. The main difference with the non-clustered case (above) is the fact that in the clustered case, self-loops in the network appear. This is because trains moving to a station in a cluster — and thus adding to the cluster's delay — can reach that station, and then continue to another station in the same cluster, again adding to the cluster's delay.

While the equations in the clustered case are the same as in the non-clustered case, the parameters such as frequencies and travel times are now defined on edges between clusters. We explain a method to express these cluster parameters in terms of their non-clustered counterparts (i.e., the ones in Tab. 5.1). We start with the total frequency f_{IJ} and weighted averaged travel time t_{IJ} of trains between two clusters I and J . We define them as

$$f_{IJ} = \sum_{i \in I} \sum_{j \in J} f_{ij} \quad (5.9)$$

$$t_{IJ} = \frac{\sum_{i \in I} \sum_{j \in J} t_{ij} f_{ij}}{\sum_{i \in I} \sum_{j \in J} f_{ij}}. \quad (5.10)$$

These definitions are intuitive: the total frequency of trains between two clusters is the sum of the frequencies on edges going from a station in the first to a station in the second cluster. The travel time is the weighted average of the travel times of the edges going from the first to the second, weighted by their frequency.

Next, we need to define the stopping probability s_I for a cluster I . In order to do this, we define the station parameter q_i as the probability that a train which arrives in the cluster $\mathcal{C}(i)$, arrives at station i . In a way, it indicates how important station i is in its cluster, measured by the total frequency of all incoming trains to that station. The quantity is approximated as follows:

$$q_i = \frac{\sum_{j \in \mathcal{N}(i)} f_{ji}}{\sum_{j \in \mathcal{C}(i)} \sum_{\ell \in \mathcal{N}(j)} f_{\ell j}}. \quad (5.11)$$

Note that the values q_i are weights of stations whose sum is one. Each station is weighted by the frequency of incoming trains. Next we use the quantities q_i to estimate

the stopping probability s_I for cluster I :

$$s_I = \sum_{i \in I} s_i q_i. \quad (5.12)$$

One can interpret the stopping probability formula using the following formula:

$$s_I = \sum_{i \in I} P(\text{stops in } i | \text{arrived in } i) P(\text{arrives in } i | \text{arrives in } I)$$

Using this approach, we can set up a model for any clustering of the original network using only the parameters of the full network. We can thus compute the matrix \mathbf{G} (Eq. (5.7) and (5.8)) for each clustered case. We denote such matrices of the clustered model by \mathbf{G}_c .

An additional possibility, which we do not discuss further, is to define a clustered model directly, without relying on the parameters of the full network. In this case, the frequencies, average travel times and stopping probabilities need to be directly measured from data.

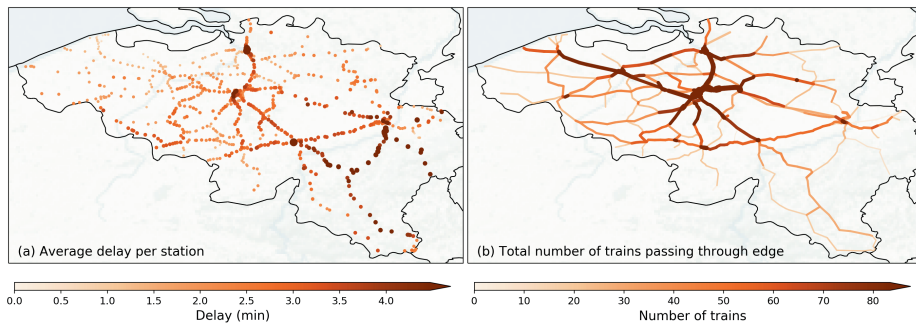


Figure 5.2: Panel (a): Average delay per train in November 2019, shown at every node. Panel (b): Average number of trains passed through the edge on April 11th, 2019 (taken as an example day). Only passenger trains are used when calculating these numbers.

5.2.4 Model considerations

There are a number of important assumptions we used in our model (see also SI E.1). Because we aggregate delays from trains onto stations, we lose a lot of details, such as origin-destination information of trains. In the derivation of the model, a delay ‘arriving’ at a station is subsequently spread out and propagated to all neighbours of that station, based on a fixed weighting of the outgoing edges. However, in real railway systems, there is a high correlation between where the delay comes from and where it goes to, and memory effects can be important. Our model is expected to work better on lower spatial resolution, on scales where a lot of trains and train routes contribute to the dynamics of a single node, such that trains picking a random direction constitute a decent approximation to the real dynamics, which on the detailed level is inherently schedule-based and not random. Furthermore, the delays in our model are treated as variables smoothly varying in time and space. In reality, delays which are localised in space are of a discrete

nature: a single train can be delayed, and when the train has ‘passed’ a station, the delay suddenly disappears from this station. This means that the time series of $D_i(t)$ in reality has a lot of jumps, namely every time a train reaches this station or starts towards it. In the model, $D_i(t)$ is smoothly varying. Another important consideration is that the model only propagates delay and removes delay from the system — it does not add any new delays. Moreover, the only mechanism by which delays are removed from the station is when a train ends its trajectory, which is encoded in the parameters s_i . An underlying assumption is thus that each train keeps its initial delay until it has reached its final stop. In practice, of course, delays are constantly generated, often due to small noise-like incidents or other (delayed) trains blocking platforms or tracks, and in more exceptional cases due to new disruptions. Moreover, trains can lose some delay by travelling faster or because of scheduled buffer times at stations, which is not included in our model. For these reasons, we will only compare the results of our model to data of days with a large disruption: by focusing on a time point with a large amount of delay and analysing its dissipation through the network, we minimise the effects of smaller stochastic delays, which are expected to contribute less to the dynamics in these situations. A final limitation we would like to mention is that in our model, the finite travel time of trains and their location on an edge is lost: in our assumptions, a train’s delay counts fully towards the next station’s delay, wherever the train is on an edge towards that station. For small time steps, this means the train’s delay also counts immediately to the propagated delay further on in the network, even if in reality the train would still need more time to cross the edge.

Next to the limitations mentioned here, our model also has clear benefits: a compact description (the matrix \mathbf{G}), the fact that it is linear and thus amenable to analytical study and the straightforward generalisation to lower spatial resolution. We discuss advantages of the model throughout, and at the end of this chapter. Some of the limitations mentioned above directly stem from our choice for a network-based, diffusion-like model. It is one of the aims of this chapter to investigate whether our model, and its built-in potential for spatial aggregation, can reproduce the dynamics of delay propagation observed in a real railway system.

5.3 Data and Methods

We apply the model to the Belgian railway system as an example. We chose the Belgian railway system for multiple reasons. Being a West-European country, Belgium has a rather dense and strongly utilised railway system with over 100 m of lines per km², being one of the world’s densest national railway systems (International Union of railways, 2020). In contrast to, for example, the American or Chinese railways (both have about 10-25 m of railways per km²). Additionally, freight and high-speed trains make up only a small fraction of the total railway transport in this country. These aspects require more complex scheduling in the Belgian case, and it implies a more interesting delay evolution to use as an example. Another reason for analysing the Belgian railways is the availability of data, which is discussed below. A discussion on the international relevance of the results is given in the conclusions.

5.3.1 Data and Pre-processing

We use the open data provided by Infrabel, the service company of the Belgian railway network (Infrabel, 2021). The data contains geographical information on railway stations and the physical railway lines, recorded tracks of passenger trains with details on scheduled and realised departure and arrival times of their activities on each station or junction, as well as associated delays. The time stamps and delay data are in seconds. We use data from all Belgian passenger railway activities between January 2019 until May 2020. The data covers an average number of 3600 daily unique trains on business days, and 2200 on weekends or holidays.

The first step is to reconstruct the graph of the Belgian railway network. First we add all stations as nodes in our graph. We get the edges by mapping the geographical locations of railway stations onto geographical shapes of railway lines and every two stations are connected together if and only if there is a line connecting them without intermediate stations. The geometry of railway lines is more intricate than simple edges between stations, since there exist places of splits and merges of multiple lines. We implement these by adding so-called “junction” nodes along the lines.

The dataset contains all railway stations, which except of passenger train stations include merchandise platforms, technical depots, carwashes, etc. Passenger trains tend to skip those intermediate platforms and the passage information is not recorded. In order to bypass this limitation, in case when there is no edge between two consecutive stations in the track record, we assume that the train follows the shortest path between them. Delay accumulation or reduction is then evenly spread across the intermediate stations along that path.

There are two kinds of passenger trains available in the data that can be characterised by the proportion of skipped stations along the track: 1) local trains, which usually circulate at shorter distances and stop at every station along the path and 2) intercity trains, which circulate at larger distances and skip some intermediate stations. We exclude from the analysis the intercity trains that skip a significant portion of stations along the track (usually these are international trains) and extra trains that run *ad hoc* on a specific day. The number of disregarded trains is less than 3-5% of the total data. We further use the notion of a railway graph and a railway network interchangeably.

The reconstructed network and two important delay statistics are shown in Fig. 5.2. The graph contains 822 stations and 972 edges. Because the network has mostly a line-like structure, 78% of all stations have degree 2 and the average degree is 2.19. In panel Fig. 5.2a, we show the average delay of trains travelling towards stations in November 2019. A general trend from small average delays in the north-west to larger average delays in the south-east is visible, with the cities of Antwerp (north) and Brussels (centre, the capital) also having rather high average delays. Panel Fig. 5.2b colours the edges of the network with the average number of trains per day that crosses them. Several lines between the large cities of Bruges, Ghent, Brussels and Antwerp stand out.

We use the recorded tracks to estimate the model parameters. In particular, we calculate the edge parameters f_{ij} and t_{ij} and the station parameters s_i (see Tab. 5.1) for each month separately. Within a month we aggregate all frequency and temporal counts for each day of the week. Moreover, for each day we keep separate counts for six 4-hour periods of the day. For each station j this leads to the estimation of parameters s_j as the

average fraction of arriving trains that end their trajectory at station j , and f_{ij} and t_{ij} , the average frequency and average passage time of trains going from station i to j . For simulations of disrupted situations, we use the parameters obtained for the month, day of the week and period of the day corresponding to the timing of the peak delay on the disrupted days. If not mentioned otherwise, we use $\Delta t = 30$ seconds in all results in this chapter. Simulations of the model were coded in Python. For the clustering discussed in section 5.3.4, we used the KMeans function of scikit-learn. The data and code is publicly available.

5.3.2 Disrupted situations

As discussed in the introduction, we expect diffusion-like models to be of most interest to study large-scale delay propagation: e.g., general directions of delay evolutions — as individual delays will be predicted erroneously due to lack of trajectory information of delays by abstracting away from individual trains. Therefore, we focus our model analysis on days in which such large-scale delay propagation can be assumed important, namely where the delays were severe. In contrast, when delays are small, they dissipate quickly and will not spread much — making identifying large-scale spread of delay of less interest. Another reason why we focus on days with severe delays is that understanding such days is of great importance to railway companies to be able to handle such situations well. We refer to days with severe delays as ‘disrupted days’. A list of disrupted days is obtained by looking at the peak in the total delays (i.e., delay summed over all nodes at any given moment in time) of every day in the dataset, and taking the 50 days with the highest peaks. The exact dates in this list are given in SI E.2. Throughout the rest of the chapter, we initialise our simulations at the peak in total delay of these disrupted days, i.e. we determine the delay on each station at the time of peak delay and use this as initial vector of delays. The model captures the spread and dissipation of existing delays, so these simulations will capture the propagation of the delays present at the peak — making such an initial point most interesting. Also, we reason that after the moment of highest total delay, the relative importance of newly generated (i.e., non-captured) delays is small as compared to existing delays.

5.3.3 Quantifying model performance

When assessing the model’s performance to reproduce reality, we focus on whether the model reproduces the correct direction of delay evolution, rather than simulating exact values well. There are a number of reasons for this. First, when aiming to understand large-scale propagation of severe delays — which is the aim of this model — accurately tracking the positioning of delays (rather than the exact values) through space is already very important information to practitioners. Analysing such directional trends of delays provides us with information on how the system works, absolute values of delays are not always necessary for that. Second, in severely delayed circumstances, numeric performance comparison can quickly become biased by several high spikes in delays: particular trains being up to one hour delayed, compared to an average delay of a couple of minutes in the rest of the network. And third, our model was not designed to capture small, stochastic variations in the delays. However, such delays are always present in the

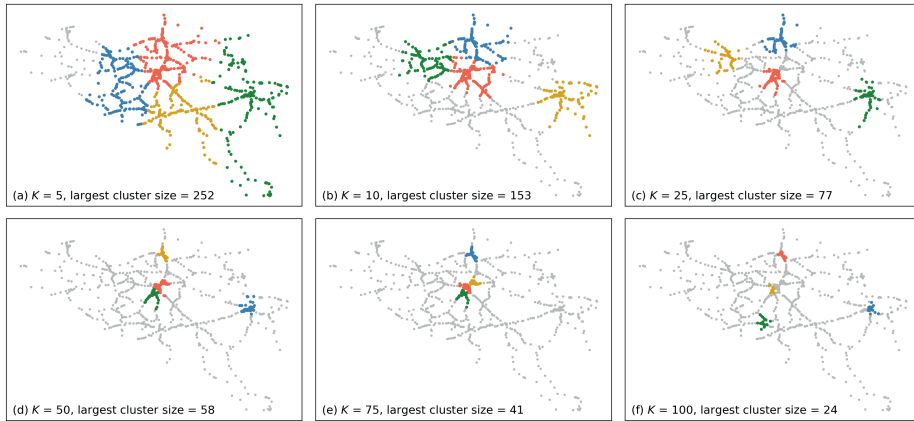


Figure 5.3: Clustering results for (a) $K = 5$, (b) $K = 10$, (c) $K = 25$, (d) $K = 50$, (e) $K = 75$ and (f) $K = 100$. For visibility purposes, only the four largest clusters are coloured (in the order of red, blue, green and yellow). Stations not belonging to any cluster are coloured grey. Cluster size is measured by the number of associated stations. Largest cluster sizes are denoted in panel labels.

data, which means that one will never get a good quantitative fit, even if the model would be a perfect representation of the propagation of existing delays. For these reasons, we use Spearman's correlation coefficient ρ to measure the model's performance. This metric is based on the rank of the variables, i.e., it assess monotonic relationships rather than linear relationships (which is the case, for example, for Pearson's correlation coefficient). We denote the observed delays at all stations at time t by $\vec{D}_{\text{obs}}(t)$: a vector with delay entries per station. Likewise, we define a simulated delay vector $\vec{D}_{\text{sim}}(t)$. We denote the vector containing the ranks of the stations based on their delays by $r(\vec{D}(t))$, using either observed or simulated delays. Then, Spearman's correlation coefficient at time t is given by:

$$\begin{aligned} \rho(t) &= \text{Pearson}\left(r(\vec{D}_{\text{obs}}(t)), r(\vec{D}_{\text{sim}}(t))\right) \\ &= \frac{\text{cov}[r(\vec{D}_{\text{obs}}(t)), r(\vec{D}_{\text{sim}}(t))]}{\sigma_{r(\vec{D}_{\text{obs}}(t))} \sigma_{r(\vec{D}_{\text{sim}}(t))}} \end{aligned}$$

In the next section, we use this metric to compare the model performance on different levels of spatial aggregation. In the clustered case, the vector $\vec{D}_{\text{sim}}(t)$ will not have N elements (the number of stations), but $K < N$, the number of clusters. Each component of the vector is the total delay in one cluster. We want to compare this with the observed data on N stations. This observed delay vector therefore also has to be aggregated on the K clusters, by summing the delays of the stations belonging to the same cluster. To be able to compare the Spearman's correlations across simulations with different K , the observed and simulated delay vectors of dimension K are de-aggregated towards dimension N (i.e., equally distributed across each cluster's stations), such that we always compute Spearman's correlation on vectors of length N , even if the model was simulated on

the network of clusters.

5.3.4 Clustering

When referring to ‘level of aggregation’, we mean the spatial resolution of the model. Full resolution would mean using all stations as entities in the model (i.e., no clustering), and lower resolutions involve clusters or groups of stations as entities in the model. Section 5.2.3 describes how we translate node and edge parameters towards a lower resolution. Here we discuss the means of clustering itself: the process of grouping stations in an appropriate manner. Many of such clustering methods exist, and we have chosen to use K-means (Steinhaus, 1956; MacQueen, 1967) on the spatial coordinates of the nodes in the Belgian railway network (longitude, latitude). We do this for the following reasons. First of all, we aim to create groups of stations that are adjacent to each other. Although the way we use K-means does not explicitly incorporate network topology, it does make sure that the groups of stations are convex (i.e., there is no station from cluster *A* in the middle of cluster *B*), since the railway network is an inherently spatial network. This geographic basis for the groups also makes them easier to interpret. Another important reason for using K-means is that we can choose K — the desired number of clusters — which we can vary to get different levels of aggregation to assess the model performance with.

We vary K between a minimum number K_{\min} and maximum number K_{\max} of clusters, which in this case we set to be 3 and 100, respectively. Note that values of K_{\min} lower than 3 are excluded because of the resulting coarseness of the resulting model, and values of K_{\max} higher than 100 are excluded because they result in many single-station clusters. The K-means algorithm starts with an initial set of K points (‘centroids’) and assigns all stations to the closest centroid. Each centroid now corresponds to a cluster of stations. Next, the centroid coordinates are redefined as the average of all the stations in its cluster. This process is then iterated (reassigning stations to closest centroid, updating centroid coordinates) until it converges to a point where the centroids do not change anymore.

The resulting clusters for several values K are shown in Fig. 5.3. In each plot, the four largest clusters in the network are shown in colours. Observe the small size of clusters in the $K = 100$ case, motivating the $K_{\max} = 100$ threshold. We can also see that the largest clusters (in terms of number of stations) for high values of K are situated around the major cities of Brussels, Antwerp and Liège. This can be explained by the fact that these cities contain numerous smaller railway stations that are geographically close together, while in more rural areas like the south and west, the station density is much smaller. Urban areas are thus expected to contain larger clusters for relatively large values of K .

5.4 Results

In this section, we show the dynamics of the model and compare it with the data for a number of disrupted days as example. Then, we apply the model to toy examples to illustrate in which circumstances this model works well and in which it does not. We end by discussing the overall performance on all 50 disrupted days.

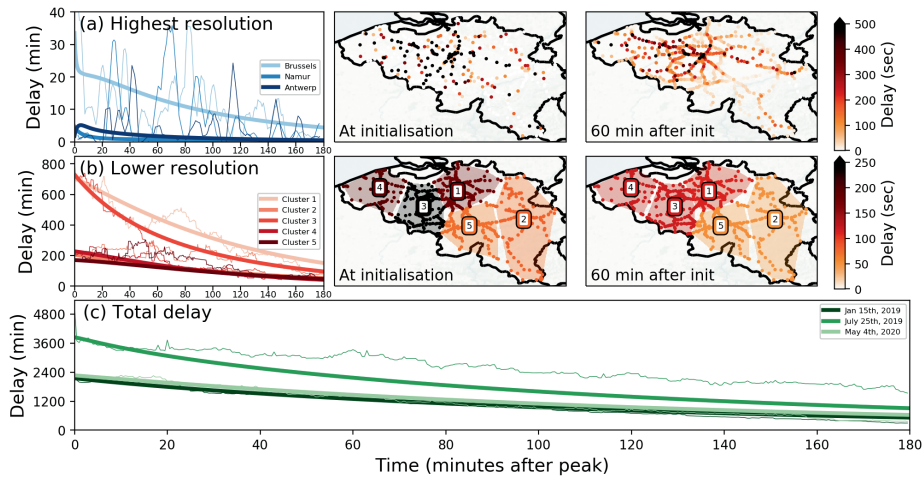


Figure 5.4: Example simulations at various resolutions. **Panel (a):** highest (non-clustered) resolution simulation of Jan 15, 2019, initialised at the peak delay (18:11). The delay evolution over time of three example stations is displayed in blue lines (Brussels, Namur and Antwerp). The spatial situation of simulated delays at initialisation (middle) and 60 min after initialisation (right) are also shown. **Panel (b):** Simulation of the same day, but at a spatially aggregated level of five clusters. Red lines in the left panel show the temporal evolution of delays for each cluster. Again, the middle and right panels indicate spatial delay distribution at initialisation and 60 min after initialisation. For clarity, the clusters are shaded in the background. **Panel (c):** Total delay evolution of three example bad days: Jan 15, 2019, July 25, 2019 and May 4, 2020, all initialised at their peak delay moments. All maps show delays in seconds, with a cut-off at 500 and 250 seconds respectively, as higher delays were rare on these instances.

5.4.1 Example simulation

We start by looking at a few example disrupted days. We start with Jan 15, 2019, which had a peak delay at 18:11. Initialising the non-clustered ('highest resolution') version of the model at this moment, we simulated the delays up to three hours after the peak. The delay evolutions at three major stations (Brussels, Namur and Antwerp) are displayed in Fig. 5.4a. It is clearly visible that the simulated delay time series is much smoother than the real time series, which has strong jumps as a consequence of the discrete nature of trains: either delayed trains are going to those stations (i.e. delay > 0), or not (i.e., delay = 0). This is also visible in the maps in the upper row of this figure: at initialisation time, the delays are distributed very discretely across the network (centre-top panel). The model diffuses the delay across the network after 60 min (right-top panel). In this figure, we can clearly see one assumption on which the diffusion-like model is based: it assumes that delay is spread by a very large number of trains, and that it travels to all other adjacent stations instantly (albeit weighted into small fractions). Of course, in reality this assumption does not hold.

In panel (b) of Fig. 5.4 we take the exact same day, but instead of modelling at full resolution, we cluster the network into five clusters and redo the analysis. We observe that, by aggregating over the many trains present in each cluster, the jumps in delays visible in panel (a) become less pronounced: the real delay evolution curves per cluster in panel

(b) are more smooth. The general trends of the real delay curves in each respective cluster resembles the simulation quite well, even though some undulations are visible. For example, increases in delays in cluster 1 around 70 minutes and in cluster 5 around 40 minutes after initialisation are visible. This is the result of newly generated delays. Snapshots of the spatial distributions of the delay are shown in the maps on the right. They show how delay dissipates and is transported across the five clusters. A quick comparison by eye between the highest resolution maps (top panels) and these lower resolution maps indicates the resemblance.

Panel (c) of Fig. 5.4 shows the evolution of the total delay in the system (both simulated and real) in three example disrupted days: Jan 15, 2019, July 25, 2019 and May 4, 2020. One can see that the total delay on Jan 15 and May 4 are simulated quite well over the whole three hours, but the real total delay on July 25 quickly overshoots the simulated curve — pointing towards the effect of newly generated delays.

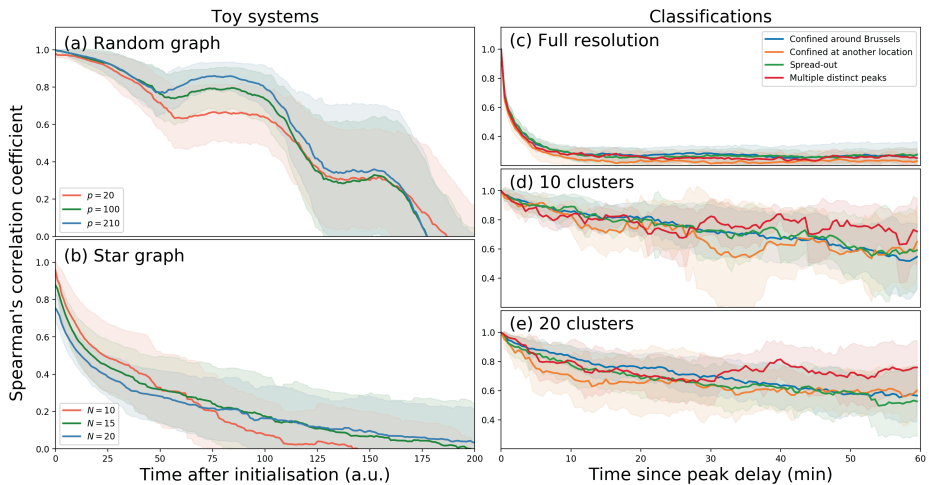


Figure 5.5: Model performance in toy systems and across classifications of disrupted days. **Panel (a):** Model performance of the random toy system for three different values of the number of lines p (see SI E.4 for details). **Panel (b):** Model performance of the star graph toy system for different values of the number of nodes N (keeping p fixed, see SI E.4 for more details). **Panel (c)-(e):** Model performance along time since peak delay of the disrupted days, averaged within each classification (see SI E.2 for details), for the model at (a) full resolution, (b) using $K = 10$ and (c) using $K = 20$ clusters. Averages are shown in lines, shaded areas indicate the range of one standard deviation from the average.

5.4.2 Toy model

In this subsection we introduce two toy systems that allow us to study more fundamental properties of the model from section 5.2. They are explained in more detail in SI E.4. The toy systems represent implementations of the model for networks with very simple topologies: random networks and star networks. Numerous other toy systems can be thought of, but we specifically compared these because they can test the model performance under different levels of the density of lines and connectivity of nodes. As for

the real data we measure the model performance using the Spearman's rank correlation coefficient.

Fig. 5.5a shows the performance of the model on a random graph-topology (with 15 nodes and 20 edges). We vary the number of lines p from 20, to 100, up to 210 (which is the maximum number of unique pairs in a 15-node connected graph). It is clear that the model performance decreases over time. At first, the different values of p do not matter: the model performance decreases slightly due to the fact that in the model, delays are instantly spreading to various directions further in the system, while in these systems (and in reality), delays need to arrive at next stations first (carried by trains) before moving onto next stations. This leads to a discrepancy. As soon as the first trains arrive at next stations (around $t = 50$), their delays contribute to delays on new edges where the model already predicted a small part of it to be. For small numbers of lines (i.e., low p), the specific direction the train is going is very important. When there are many lines (i.e., high p), all combined 'chosen' directions of the trains approximate the frequency and other attributes put in the model. In other words, the model approximates reality better for densely used lines. And this seems to be visible: high p (blue line) starts deviating positively from the red line after $t = 50$. At much later points, the initial delays start arriving at their ending stations, which collapses the correlation down to much lower values.

For the star graphs (see panel (b) of Fig. 5.5), where we fix the number of lines $p = 50$ we take a look on the dependence on the number of nodes. The number of nodes do not seem to matter much, but it is clear that the star graph indicates much smaller correlations than the random graph. Although this is merely an example system, we intuitively expect that as soon as trains start driving towards the centre, other delays (as a consequence of the diffusion-like nature of the model) are simulated to be at each of the connected nodes, quickly limiting the correlation.

The above toy systems reflect that our model works better for denser networks with the higher number of train lines.

5.4.3 Classification of disrupted days

We now investigate whether the initial geographical delay distribution has an effect on the accuracy of the diffusion-like model. For this, we distinguish four categories across the 50 disrupted days, classified by eye based on the delay patterns on the peak delay moments. SI E.1 discusses this classification in more detail and also shows the delay maps. The first and largest group (25 days) contains the situations where almost all of the delays are localised near Brussels, the capital city of Belgium and important railway hub. In Belgium, train lines between east and west and north and south respectively all pass through Brussels, which makes it an important factor in the delay dynamics in the railway network. The second group (7 days) contains situations where the delay is also localised, but on a different location than Brussels. The third group (5 days) contains those situations with multiple locations with high delays. Finally, we consider the group of stations (13 days) where the delay is not localised but instead spread out over a large region.

As before, we perform simulations with as initial condition the peak delay distribution. In Fig. 5.5c-e we show the evolution of the Spearman's correlation over time, averaged per group. We show this for different spatial resolutions. We find that there are

no clear differences between the groups. The situations with delays localised near a city which is not Brussels (shown in orange) seem to perform a bit worse than the others, but we should be very cautious in interpreting this: the variation within a group is very large, as shown in the shaded areas.

The fact that there is no clear difference in model performance between groups could indicate that the spatial localisation of the disruption is not a good determinant of the accuracy of our diffusion-like model. The obvious question then is: is there a better measure, or characteristic, which can distinguish between different disrupted situations and indicates whether a diffusion-like delay spread is warranted? We plan to explore this in future work.

5.4.4 Overall performance

We now turn to the overall performance of the model over the 50 disrupted days. On each of these days, we determine the peak in the total delay and simulate the delays up to two hours after this peak. We then compare what really happened throughout these two hours to what we simulated by computing the Spearman's correlation coefficient ρ at each time point (see Eq. (5.13)). We do this for each number of clusters K ($3 \leq K \leq 100$). The average correlations per K and t over the 50 disrupted days is shown in Fig. 5.6a. It is clear that in general, the higher t , the lower the correlation. This is intuitively correct: the higher the time after initialisation, the more the model will start differing from reality, for example due to new incoming delays in the real data that are not captured by the model or errors the model that grow with time. In the same panel, we see that higher number of clusters K also decreases the correlation, which is less obvious. On the one hand, information is lost when coarse graining: for lower values of K detailed information on the positioning of the trains is put together into larger clusters, which may reduce their simulation quality. On the other hand, the diffusion-like spreading is presumably more accurate when looking at a larger scale (lower K), since on these scales the discreteness of delays is averaged out in the data, too. Interestingly, panel (a) also shows bands of K values with near-equal correlations: up to $K = 8$, the correlations seem to be more or less the same (very high), at least up to $t = 45$ min. The second band of near-equal correlations is between $8 \leq K \leq 17$, followed by a more gradual decay of correlations with K , but a sharp decrease in those correlations at $K \approx 27$. One reason for these sudden correlation decreases could be a strong rearranging in the clustering at those K values: e.g., in Fig. 5.3, panels (a) and (b), one can see that for $K = 5$, Brussels is at the border of the red cluster, while at $K = 10$, it is in fact in the middle of a cluster. Such rearranging can be quite sudden from one value of K to another. In contrast, the slow decrease in correlation *within* those K -bands can be related to a slow change in the clustering structure.

Panel (b) in Fig. 5.6 shows the correlation ρ 40 minutes after model initialisation, as function of K , for each individual disrupted day. Clearly, these curves seem less gradual as the average displayed in panel (a). In fact, changing K by 1 may impact the correlation up to 0.5 in some exceptional cases — specifically when K is small (which makes sense as the clustering structure changes rapidly around these values). The average is displayed in black, and the gradual decrease with K is visible. Red, thin lines indicate days on which the maximum correlation is at values $K \geq 9$, which is counteracting the overall pattern we see that the correlation keeps increasing with decreasing K — on these seven days, the

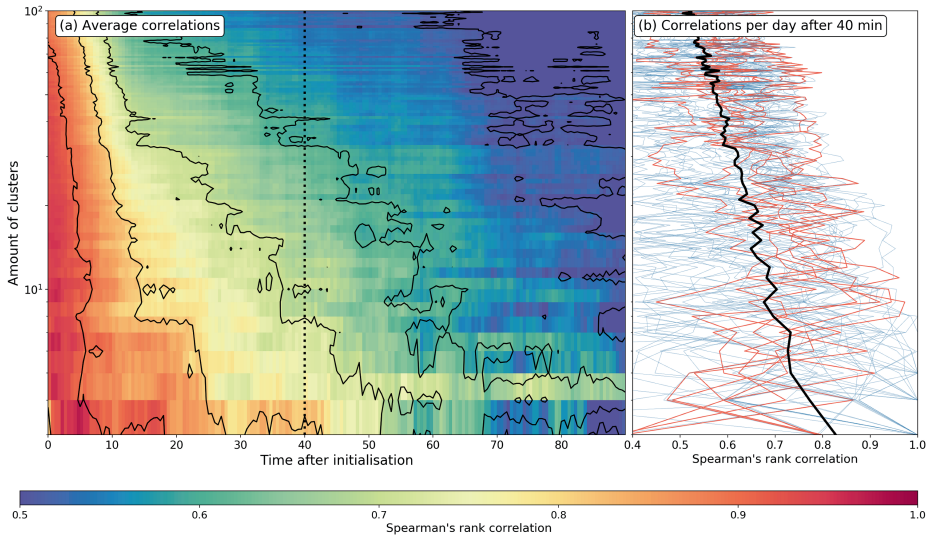


Figure 5.6: **Panel (a):** Average Spearman's rank correlation coefficients in colours averaged over the 50 disrupted days, for various values of K (vertical) and time points after model initialisation (horizontal, in minutes). The contours indicate the levels 0.5, 0.6, 0.7, 0.8 and 0.9. The vertical dashed line corresponds to 40 minutes after initialisation, which is used in the other panel. **Panel (b):** Spearman's rank correlation coefficients at 40 min after initialisation. Individual days are split into days that have their maximum at values of $K < 9$ (in blue) and those that have their optimum performance at values of $K \geq 9$ (in red). The black line indicates the 50-day average.

model performs best on a non-trivial level of K , between 9 and 11. SI E.3 discusses this non-trivial level (i.e., $K = 10$) more in-depth and potential reasons why this clustering configuration may be optimal in some circumstances. The blue lines indicate all other days (where optimal correlations are found with very low values of K).

5.4.5 Model discussion

Our results show that modelling delay as a diffusion-like spreading phenomenon clearly has limitations: on the scale of individual stations (Fig. 5.4a), the discrete nature of delays, accompanied with periodic undulations and sharp changes in delay is not simulated at all. Also, delays bound to trains usually travel in a specified direction, which is not captured by our model (which is a weighted form of unidirectional spreading). The diffusion-like-spreading assumption corresponds to the view that the delay propagation is based not on single trains carrying a delay, but on many tiny trains, all going in various possible directions with their portion of the delay, which are randomly chosen, weighted by the variables in the model (e.g., frequency and travel times). Thinking about the diffusion-like model in this way motivates the use of coarse graining to improve the model. Qualitatively we show this in Fig. 5.4b and c and quantitatively this is discussed in Fig. 5.6: a clear increase in performance is visible when comparing results from the clustered version of the model to clustered data. Still, there is a loss of correlation with

simulation time: the further from the initialisation time, the more new trains and delays enter the system and change the real delay time series, unaccounted for by the model. It is important to note that the latter will always be a caveat for delay propagation models due to the inherent stochasticity of delay generation.

The toy systems we tested are meant as illustrations to indicate what the model benefits from: average patterns. As soon as the network is dense and there are many trains travelling on it, the real delays spread roughly along simple statistics like train frequencies, which are the basis of this model. But as soon as the network becomes more sparse, especially when it becomes tree-like, the correlation drops.

Our discussion on the classification of disrupted situations showed no clear differences in model performance based on the initial condition, at least as far as its localisation (which is the basis of the classification) is concerned. However, our approach was naive: we classified the disrupted days by eye into four groups. We cannot conclude that there are no other, better metrics that do distinguish situations in which the delays spread in a more diffusion-like manner than in other situations — something we do find in the toy examples. Hence, we propose to investigate such metrics further in future work.

The high performance at low values of K implies that a coarse resolution is better suited for these type of models. The disadvantage of that is the loss of detail. Also, as shown in the toy examples, there are cases we can think of that are not suitable to be modelled well by the model: high sparsity of trains increases the discrete nature of delays and decreases the applicability of the mean-field approximation. Another example where these models have low accuracy is when the delays are mainly governed by stochasticity, and not by propagation dynamics. This is the case in situations where the overall delay level in the network is low. Such situations are difficult to capture well in many delay propagation models, in fact.

We propose therefore that the model presented here finds its niche in the problem of simulating the propagation of severe delays on a large scale. In such circumstances, the exact magnitude of delays at fixed positions is not always of most interest, while the general trend, speed of delay decay and direction of the overall bulk of delay are of high importance. Such information is well retrievable from the clustered model. In fact, this model is arguably very suited to analyse these large-scale dynamics and how they depend on network topology and high-level parameters such as train frequencies. All information of the system's dynamics is embedded in the \mathbf{G} -matrix in Eq. (5.8) — a single matrix that can be analysed using spectral methods to investigate its eigenproperties, for example. Another advantage of this model is its simplicity. Using only a small set of parameters that are easily retrievable from the schedules (which can usually be found online for any European railway system), one can model the whole railways with a single simple differential equation (Eq. (5.7)).

5.5 Conclusions

In summary, we devised a model that simulates delays as a diffusion-like spreading phenomenon. The intuition is that, on average, the direction and dissipation of delay relates to aspects of the schedule such as train frequencies and travel times. We apply the model to the Belgian railways and investigate its strengths and weaknesses. In particular, we

find that the model performance increases sharply when coarse graining by grouping stations together into clusters. We conclude that this model is mainly of use when working on larger scales and aiming to identify system properties related to delay dissipation and general directions of delay propagation, rather than accurate individual prediction.

We have illustrated the workings and performance of the model using the Belgian railway network as an example. Our framework, however, is general and could be applied to any country. Nevertheless, there are some international differences that should be taken into account. One aspect which may influence the performance of the model in different countries, is the position of important railway hubs. Hubs, such as large cities, are expected to play an important role in the delay propagation dynamics. In Belgium, these hubs are well distributed across the country, apart from the more rural areas in the south-east and far-west. This means that, in our geographic clustering, the hubs usually fall into different clusters. The Netherlands, in contrast, has its most important cities concentrated in the west of the country. In a model with a small number of clusters, it is possible that many of these hubs end up in the same cluster, which may have an unwanted effect on the model's performance and usefulness. For this reason, it might be important to consider other clustering methods.

Not only geographic differences should be considered when trying to extrapolate these results internationally. Various factors impacting delay propagation vary from country to country, like policy, protocols, infrastructure details and delay statistics in general (Schipper and Gerrits, 2018). However, we argue that the increase in model performance when coarse graining is robust to these changes, because the reason for it is of a more theoretical nature: diffusion-like spreading captures average delay fluxes, which are more prominent in clustered systems.

Applying the model to other countries is straightforward, since its ingredients are general and easily obtained: train turnover rates, frequencies, travel times and adjacency matrices are readily derived from network architecture and railway schedules. A natural extension of our work would thus be to compare the network models for different countries and explore the properties in different spatial resolutions.

All of the model's dynamics are essentially derived from the matrix \mathbf{G} , also for the clustered versions. Exploring the spectral and topological properties of the weighted network that \mathbf{G} describes and relating those properties to the dynamics of the railway system are of interest in our future work. It is possible that a few simple metrics, derived from this matrix, could be used for a quick international comparison of railway networks.

Despite its simplicity, this model already gives us a tool to better understand railway network dynamics. However, there are multiple improvements to suggest for future work. For example, it is possible to add noise to the model, to account for the generation of new delays. The magnitude and distribution of stochastically generated delays at different stations can possibly be derived from the data. It would be interesting to see how noise may lead to an 'equilibrium' delay distribution, in contrast to the highly disrupted situations we considered in this chapter.

Once delay is generated in a real railway system, there are mechanisms and feedback loops that can amplify it or mitigate it. Such mechanisms are not present in the model, but would be a valuable addition. Such feedbacks can be nonlinear, complicating the model but possibly generating new dynamics. Finally, our model is 'first-order': the spread of delay is determined only by where it is, not where it came from. Including

a memory mechanism will probably increase the accuracy of the model, but may also make it more complex. One way to do this is to not consider delay on stations, but instead on edges. In such a version of the model, the directional information of trains is partially kept, contrary to in our station-based model.

Diffusion-like spreading is researched in many other fields other than transport literature. In particular, it is well established that the vulnerability versus perturbations of networked systems including social, epidemiological and engineering systems depends on (any quantification of) the modularity (Newman, 2006): a more modular dynamical structure prevents large-scale spreading (Balcan et al., 2009; Dekker and Panja, 2021). Determining the modularity from the \mathbf{G} -matrix can be an interesting next step, as the \mathbf{G} -matrix does not only incorporate topology, but weights each edge by features from the schedules.

The results in this chapter are not only of interest for modellers, but also for railway practitioners. First, the model output can provide insights into system-wide properties, like delay decay and general directions of delay propagation. Second, it is easy to use and all information is embedded in the \mathbf{G} matrix. For example, practitioners might be interested in how isolated regions are from each other: the off-diagonal elements of the \mathbf{G} matrix at the appropriate level of coarse graining reflects how strongly regions are connected, i.e. how much delay flows from one region to another. From an operational point of view, optimal levels of clustering like those seen for the red curves in Fig. 5.6b (see also SI E.3) can be used to categorise situations, issue protocols or form threat assessments in terms of delays.

We hope that the model itself and the results of the application to Belgium motivates researchers and practitioners to vary the spatial aggregation level to non-trivial levels. We believe these diffusion-like models can offer useful insights on how aspects such as network structure, basic schedule parameters and spatial resolution affect the delay propagation through a railway network. Ultimately, such models can lead to a better understanding of railway delay dynamics.

CHAPTER 6

CASCADING DOMINATES LARGE-SCALE DISRUPTIONS IN TRANSPORT OVER COMPLEX NETWORKS

This chapter is published as: Dekker, M. M. and Panja, D. (2021). Cascading dominates large-scale disruptions in transport over complex networks. *PLOS ONE*, 16(1):e0246077. Supplementary material to this chapter can be found in SI F.



ABSTRACT

The core functionality of many socio-technical systems, such as supply chains, international trade and human mobility, concern transport over large geographically-spread complex networks. The dynamical intertwining of many heterogeneous operational elements, agents and locations are oft-cited generic factors to make these systems prone to large-scale disruptions: initially localised perturbations amplify and spread over the network, leading to a complete standstill of transport. Our level of understanding of such phenomena, let alone the ability to anticipate or predict their evolution in time, remains rudimentary. We approach the problem with a prime example: railways. Analysing spreading of train delays on the network by building a physical model, supported by data, reveals that the emergence of large-scale disruptions rests on the dynamic interdependencies among multiple 'layers' of operational elements (resources and services). The interdependencies provide pathways for the so-called *delay cascading* mechanism, which gets activated when, constrained by local unavailability of on-time resources, already-delayed ones are used to operate new services. Cascading locally amplifies delays, which in turn get transported over the network to give rise to new constraints elsewhere. This mechanism is a rich addition to some well-understood ones in, e.g., epidemiological spreading, or the spreading of rumours and opinions over (contact) networks, and stimulates rethinking spreading dynamics on complex networks. Having these concepts built into the model provides it with the ability to predict the evolution of large-scale disruptions in the railways up to 30–60 minutes up front. For transport systems, our work suggests that possible alleviation of constraints as well as a modular operational approach would arrest cascading, and therefore be effective measures against large-scale disruptions.



6.1 Introduction

Socio-technical systems such as supply chains, (inter)national trade and human mobility provide pivotal support to modern societies. Even though each one operates in its own intricate ways that are typically tuned to highly optimised benefit-to-cost ratios, their core functionality involves transport over geographically-spread, complex network architectures. Of particular interest are those situations in which — at detrimental costs to societies and economies — initial perturbations spread through a significant part of the network, leading to ‘large-scale disruptions’, i.e., near system-wide standstill of transport (Shughrue et al., 2020; Inoue and Todo, 2019; Dekker et al., 2019). The initial perturbations are often triggered by exogenous events: indeed, well-documented are the world trade and supply chain disruption events that have been caused by natural disasters (World Economic Forum, 2011), and the ongoing COVID-19 pandemic (Ivanov, 2020). Related research focuses on risks associated with critical infrastructures, service elements, alternative scenarios planning (Jaffino et al., 2020; Dong et al., 2019; Hechmann et al., 2015; Bhatia et al., 2015; Ghaemi, 2018) and case studies of (potential) disasters (Pant et al., 2016; Ivanov, 2020). On the one hand, the dynamical intertwining of many heterogeneous operational elements and agents, network connectivity, and geographical spread are oft-cited generic factors responsible for the susceptibility of these systems to disruptions (Schipper and Gerrits, 2018; Dekker et al., 2021f; Gu et al., 2020; Pagani et al., 2019; Vespignani, 2010). On the other, combined with a lack of reliable empirical data, the same factors also contribute to hinder the understanding, anticipation and prediction of systemic build-up of large-scale disruptions.

In order to address this problem, here we consider a prime example network transport system: railways. In most countries, railways witness large-scale disruptions — manifested by near system-wide train service delays — multiple times a year (Dekker et al., 2021f). Large-scale disruptions in railways lead to heavy economic damage by hindering cargo and passenger transport (Tsuchiya et al., 2007). Conversely, investments in robust railway systems may prompt economic growth (Politico, 2019; International Transport Forum (ITF), 2009). These especially hold true for densely urbanised countries in Europe, wherein railway transport, planned at a high density and frequency to match the extent of urbanisation, has developed to be an inextricable asset for societal and economic well-being, and is projected to be even more intensely developed due to rising demands and sustainability goals. While scattered delays are commonplace, they do occasionally build up to a large — near system-wide — scale, as illustrated for Italy, Germany, the Netherlands and Switzerland in Fig. 6.1.

All cases in Fig. 6.1, except Italy, were initiated by extreme weather: storms ‘Niklas’ (2015), ‘Friederike’ (2018) and ‘Burglind’ (2018, ‘Eleanor’ in the English nomenclature). *Exogenous* triggers like these — be it weather conditions, power outages (Dekker et al., 2021f), accidents or even earthquakes (Tsuchiya et al., 2007) — are typical for the onset of problems, but the consequences were driven by the system’s *internal* dynamics, propagating and amplifying delays to near system-wide scale. Indeed, in the year July 2017 - June 2018, 29 days were marked as strongly disrupted days for the Dutch railways, with most of them occurring in November, December and January. This is no coincidence, even though railway companies use icing protection, adjusted timetables and many other precautions to prevent cold weather affecting their performance. Statistics

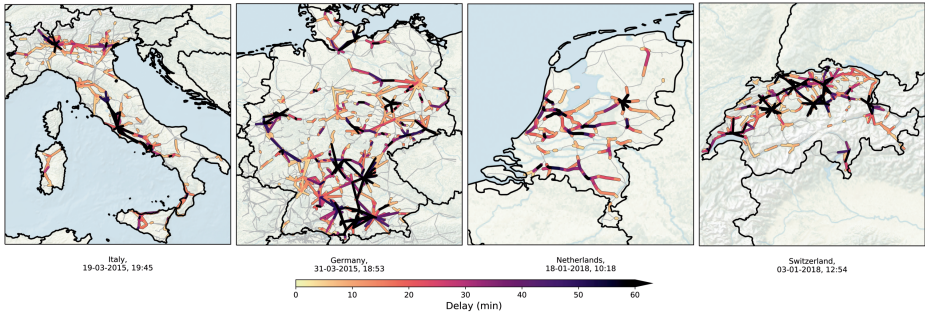


Figure 6.1: Examples of large-scale disruptions. Railway delays for strongly disrupted situations in four European countries (shown only are the delays larger than two minutes in colours; see SI F1 for data description and sources). **Panel (a):** near-simultaneous occurrence of several problems in the Italian railways in March 2015 — a major one around Rome, affecting mostly intercity trains, and one between Milan and Venice. **Panel (b):** effect of cyclone ‘Niklas’ (31 March, 2015) on the German railways. In particular, a specific train near Pegnitz (center-south) was severely damaged by a fallen tree and the rooftop of the Munich station was destroyed, along with multiple smaller incidents across the country. The high risk of more accidents and delays caused the Deutsche Bahn to cancel most of its train activity throughout the day, leaving passengers stranded in major cities like Hannover, Frankfurt, Kassel and Berlin. **Panel (c):** aftermath of storm ‘Friederike’ in January 2018 in the Netherlands, coinciding with an accident in the north of the country. Fallen trees and damaged overhead lines made the fire department force the Dutch railways to close at multiple stations — resulting in no train activity between the end of the morning and 14:00. A combination of the many disruptions with the lack of resources overview limited the possibility of mitigating delay at crucial corridors. The smaller scale and high density of the railway system in the Netherlands can be recognised also in Switzerland [**panel (d)**], where in January 2018 (coinciding with storm Burglind/Eleanor in the north-west of Europe) a strong disruption in near Zürich (north) rapidly propagated towards the rest of the country.

like these, notwithstanding the wide variety of operations management for railway systems, highlight the *generic* aspects for the build-up of large-scale disruptions: while the initial *primary* delays, caused by external events, could possibly have been quickly resolved, systems’ internal issues cause new *secondary* delays, converting an initial locally-confined problematic event to an amplified near system-wide disruption.

For transport and logistic systems at large, considerable effort has been invested in identifying risks associated with criticality aspects of infrastructure in situations of hazards like in Fig. 6.1 (Bhatia et al., 2015; Zhu and Goverde, 2019) and how to deal with disruptions in terms of rescheduling (Ghaemi, 2018; Schipper and Gerrits, 2018; Nielsen et al., 2012). Related literature aims to understand and predict the evolution of delays in transport systems, both under regular and disrupted circumstances. Most delay evolution models, however, focus on regular circumstances and predict how delay fluctuations develop using high-resolution statistics obtained from particular incidents or scenarios (Zhu and Goverde, 2019; Wei et al., 2015), or from particular stations (Kecman and Goverde, 2015b; Li et al., 2016) or lines (Huang et al., 2019; Janić, 2018; Wen et al., 2019). These models come in various forms, mainly in the context of air and railway transport: analytical (Goverde, 2010), agent-based (Gambardella et al., 2002), stochastic (Büker and Seybold, 2012; Meester and Muns, 2007; Pyrgiotis et al., 2013; Campanelli et al., 2014) and purely data-driven (Oneto et al., 2017; Dekker et al., 2019). A second relevant branch of transport literature focuses on robustness and vulnerability aspects, such

as definitions of transportation resilience (Leobons et al., 2019), perturbations in the network topology (Bhatia et al., 2015; Lordan et al., 2015), and data-based analyses on how the systems are connected (Guo et al., 2019; Pagani et al., 2019). Specifically note that the above references include studies of multiple types of transport systems: airways, railways, supply chains, and even the analogy with freight truck and cargo ship transport. All these systems share the common feature of schedule-based transport, where disruptions may lead to subsequent delay of other transport units (Chung et al., 2017; Ludvigsen and Klæboe, 2014; Helbing, 2013; Fleurquin et al., 2013). This effect, universal to many transport systems, is often commonly referred to as ‘cascading’ of delay; in this chapter we analyse its contribution to large-scale disruptions for railways.

In the context of these models, which use the language of networks and dynamics, it is crucial to make the distinction between (a) dynamics *on* networks, where dynamics of a certain state variable evolve on top of, and is thus bound by (a time-invariant) network topology, and (b) dynamics *of* networks, i.e., involving (dynamic) links that can (re)appear, disappear or change weights (e.g., Cuadra et al. (2015); Simonsen et al. (2008); Buldyrev et al. (2010); Bhatia et al. (2015); Albert et al. (2000); Holme et al. (2002)). In transport literature, the study of delay propagation generally conveys dynamics of the former type, while literature on infrastructure and resilience generally focuses on dynamics of the latter type.

Even though the above two paragraphs indicate transport literature as an active and broad field of study, there are still many unknowns, such as how the system as a whole evolves during disruptions, e.g., following an exogenous trigger. Due to heterogeneity in terms of space, time, human interactions and externalities that impact the system, the existing models typically lack accuracy and predictability (of evolution dynamics) in cases other than ‘regular’ (i.e., non-disrupted circumstances, or are purposed to simulate very specific scenarios, e.g., particular types of disruptions (Ghaemi et al., 2018) or geographical areas (Middelkoop et al., 2012; Wen et al., 2019). The contrast between disrupted and regular circumstances is manifold (and are visible in the results in this chapter, e.g., Fig. 6.4). Under regular, non-disrupted circumstances, delays are generally small and are of a less interactive nature with other delays, which allows for the applicability of data averages and a linear or local view (Kecman and Goverde, 2015b). The decreased interactivity of delays is a result of the fact that the schedules (or *timetables*) contain built-in buffers that, in case of delay, prevent the delay from affecting other transport units due to limited capacity of, e.g., platforms, tracks, or exchange of resources. Additionally, isolated delays can also be easily mitigated by human control. However, in case of (multiple) disruptions, the severity of delays exceeds the buffer and mitigation capacity and start building up and affecting other transport units, making the analysis rather complex (Dekker et al., 2021f). For railways, Fig. 6.1, provides a visual feel of delays building up and spreading across large spatial and temporal scales. The figure suggests that neither methods under regular circumstances, nor the specific incidents or scenarios covered in the existing literature are applicable to the underlying delay propagation mechanisms.

In this chapter, we focus on finding generic delay propagation mechanisms that do apply to situations as in Fig. 6.1. Indeed, Fig. 6.1 prompts us to conceptualise the intertwining of the many heterogeneous operational elements and agents in terms of stochastic processes playing out on a complex (infrastructure) network that remains invariant in time. We do so by constructing a structure with multiple layers of dynamics on a fixed

infrastructure network — each layer of dynamics conveying the movements of a certain type of resource or service required for the system's operations. This conceptualisation reveals that the interdependencies among the resources and services give rise to pathways for the delay cascading mechanism, which gets activated when, constrained by local unavailability of on-time resources, already-delayed ones are used to operate new services. Cascading amplifies delays locally for both resources and services, which in turn transport the amplified delays geographically over the network to give rise to new constraints elsewhere — describing phenomena we see in strongly disrupted situations like in Fig. 6.1. Building the above concepts into a physical model leads us to reveal that the emergence of large-scale disruptions in networked transport requires three building blocks: constraints, cascading and transport. Not only does this chapter bring new understanding of the evolution of disruptions, but the data-based interlinkages between transport resources and resulting cascading effect has also never been quantitatively shown in transport literature.

The model allows us to extract the key (delay-)amplifying role played by cascading, and to also predict the evolution of large-scale disruptions in the railways up to 30-60 minutes up front. We note here that the concept of cascading in itself is not new. Cascading in other dynamical systems like the Earth's climate (Dekker et al., 2018b), or in network science in general is an active topic of research. On the one hand, in dynamics *of* networks, cascading has resulted in concepts like 'cascading failure', depicting the loss of connectivity in a network by a sequential removal of nodes and/or links (Simonsen et al., 2008; Pescaroli and Alexander, 2016; Bhatia et al., 2015; Buldyrev et al., 2010). On the other, in dynamics *on* networks, such as innovation diffusion (e.g., in the work of Watts (Watts, 2002)) cascading refers to when an adoption, a rumour, an opinion or an infection process spreads through the entire network. In our work, the notion of cascading presented refers to interlayer spillover effects within the multiple layers of dynamics *on* a fixed infrastructure network, and it simply cannot be described by a diffusion-like model. Our work therefore stimulates us to rethink and contributes to broaden our horizon to spreading dynamics on complex networks (we will return to this discussion in section 6.6).

The chapter is structured as follows. In section 6.2 we explain the cascading mechanism and how three ingredients may lead to large-scale disruptions: constraints, cascading and subsequent transport, and illustrate using a case example. In section 6.3 we formulate a model from these building blocks. In section 6.4, using the model, we quantify the role of cascading in driving large-scale disruptions. We investigate the performance and predictive power of the model in section 6.5. We conclude the chapter in section 6.6 with a discussion on the broader outlook.

6.2 The three building blocks for large-scale disruptions

We start by pointing out the generalities of transport systems. They are based on a certain (infrastructure) network: e.g., rails and stations for railway transport, airports and airline services for air transport, and highways and cities for car transport. *Nodes* in a rail network correspond to stations between which *resources* (trains, personnel, scheduled lines) move on *edges*, which are the tracks between stations. In this chapter, we treat the

network as fixed or time-invariant, and study the system dynamics that is taking place on the network. The data for the Dutch railway system can be found in a repository on Open Science Framework (<https://osf.io/tps4r/>). Data of other national railway systems, used only in Fig. 6.1, were found partially via another paper (Monechi et al., 2018), i.e., from the following websites: OpenDataCity for German data (<http://www.opendatacity.de/>), ViaggiaTreno for the Italian data (<http://www.viaggiatreno.it/>) and OpenTransportData for the Swiss data (<https://opentransportdata.swiss/>). Details on time intervals and data considerations can be found in SI F1.

As stated above, crucial to understanding the evolution of delay in case of a disruption is the impact and interactions of operational resources and agents. Using the example of railways networks, we conceptualise these elements by distinguishing three ‘layers’ of dynamics on the network, each of which has its own operator-specified schedule (i.e., dynamics) but is dependent on the schedules of the other layers: (a) a (train) service layer, containing the planned services, of which information is made publicly available via the timetable, (b) a rolling stock resource layer, involving the physical train units used to run the services, and (c) a crew resource layer, containing the personnel required to operate the trains. (Although we focus on railways, analogous formulation and dynamical layering is applicable to other transport systems.) For railways, resources and services couple by the condition that at least one rolling stock unit and at least two crew members (one driver and one conductor) are needed to run a service, leading to *dynamic couplings* — interdependencies — that link these layers along the (train) service routes. These interdependencies serve as the potential pathways for delay cascading in the following manner. Under regular circumstances, the layers act (largely) independently due to built-in local spare resource capacities and scheduled buffer times between service activities. However, new delays are generated when, constrained by local unavailability of on-time resources (i.e., when delays exceed buffer times and no spare resources are available), already-delayed ones are reused to run new services, activating ‘delay cascading’ (i.e., a process of delay generation at specific network nodes). Subsequently, these newly generated delays then get geographically transported along the service routes, and possibly create a similar constrained situation at some other node. (Most train services do not cross national borders, making cascading mainly a national problem, seen in Fig. 6.1.) In this manner, constraints, cascading and subsequent transport of delays reinforce each other to make a localised delay perturbation amplify towards a large-scale disruption — this is exactly why we refer to these elements as the building blocks of large-scale disruptions.

Before we build these concepts into a physical model and quantify cascading, let us start with defining delay itself. All services and resources have planned and realised activity times; the planned ones constitute the *predefined schedule* of the railway operator. For a train service, every (discrete) activity a — in the forms of departures, arrivals or passings-by — can have a nonzero delay value $d(a) = t_{\text{real}}(a) - t_{\text{planned}}(a)$, i.e., the activity is executed at a realised time t_{real} later [$d(a) > 0$] or earlier [$d(a) < 0$] than the planned time t_{planned} . The change in delay of a train service’s activity a with respect to its previous activity is referred to as the *delay jump* $\delta(a)$. For example, consider a crew member coming from service X that needs to transfer in 10 minutes (i.e., the buffer time) to another service Y. Given that he is 16 minutes late from service X, the buffer time is exceeded and part of his delay from service X is transmitted to service Y if no replacement crew is

available: it will cause a delay jump of 6 minutes for service Y's next activity. Delay jumps in such situations can be suppressed by the use of spare resources or by buffers built in the schedule. Although affected by numerous factors, many large delay jumps are caused by delays cascading from one layer to another via resource-service interdependency [as illustrated in Fig. 6.2(a-c)].

6.3 The model

The model is built on the following premise: if a service line, rolling stock unit or crew member is delayed, then the delay is transported along the rest of its route, minus the buffers. We develop three variants of the model, all in the same vein of linking the full system's resource schedules as discrete events: (1) the 'monolayer model', containing only the train services, wherein we explicitly incorporate the transport of delays along the service routes (it has no cascading), (2) the 'bilayer model', where we link the rolling stock layer to the monolayer model, and (3) the 'trilayer model', where we add the crew layer to the bilayer model. In the bi- and trilayer models delay also propagates *cumulatively* via rolling stock and crew: delay (minus built-in buffers) is passed on from one activity of a resource to the next activity performed by the same resource. The model we build has similarities to existing max-plus transport models (Goverde, 2010), but is novel in (a) incorporating both rolling stock and crew layers, and (b) utilising real schedules for all resources at the full-system scale (see SI E.2 and E.3 for details).

Following the prescribed schedule of the railway operator, the model is formally described as follows. Consider activity a of a certain train service line. Then consider the recent past activities $\{a'\}$ of all resources (activities performed by the personnel, rolling stock units, and service line activities) used to execute a prior to a . Then, having denoted the delay for activity a' by $d(a')$, the buffer between activities a' and a by $\beta(a', a)$ and the Heavyside theta function by \mathcal{H} [i.e., $\mathcal{H}(x) = 1$ if $x > 0$, and 0 otherwise], the model calculates $d(a)$ as

$$d(a) = \begin{cases} \max_{\{a'\}} \{\mathcal{H}[d(a') - \beta(a', a)]\} & \text{if } \{a'\} \text{ is non-empty} \\ 0 & \text{otherwise.} \end{cases} \quad (6.1)$$

In our related works, we have also experimented with an added noise term ζ to this model, allowing us to analyse the sensitivity of the results to noise. In general, it turns out that the cascading mechanism has a much stronger impact on the delay evolution than (Gaussian distributed) noise because cascading copies and amplifies existing delay rather than creating new (noisy) delay from scratch. In this chapter, however, we do not include the noise term, lest avoiding confusion between model noise and interpreted noise in the real data, as in Fig. 6.4 (see also SI E.3). Initialised at some time t_0 , i.e., being constrained to the train operator's schedule predefined at t_0 , the model propagates initial delays to future ($t > t_0$) ones via Eq. (6.1).

For clarity, let us provide an example calculation of $d(a)$. Imagine a train activity a that is ran by a train service s . Its previous activity a_{prev} had 30 seconds delay (no buffer), and we are interested in the delay of activity a . One of the rolling stock unit used for

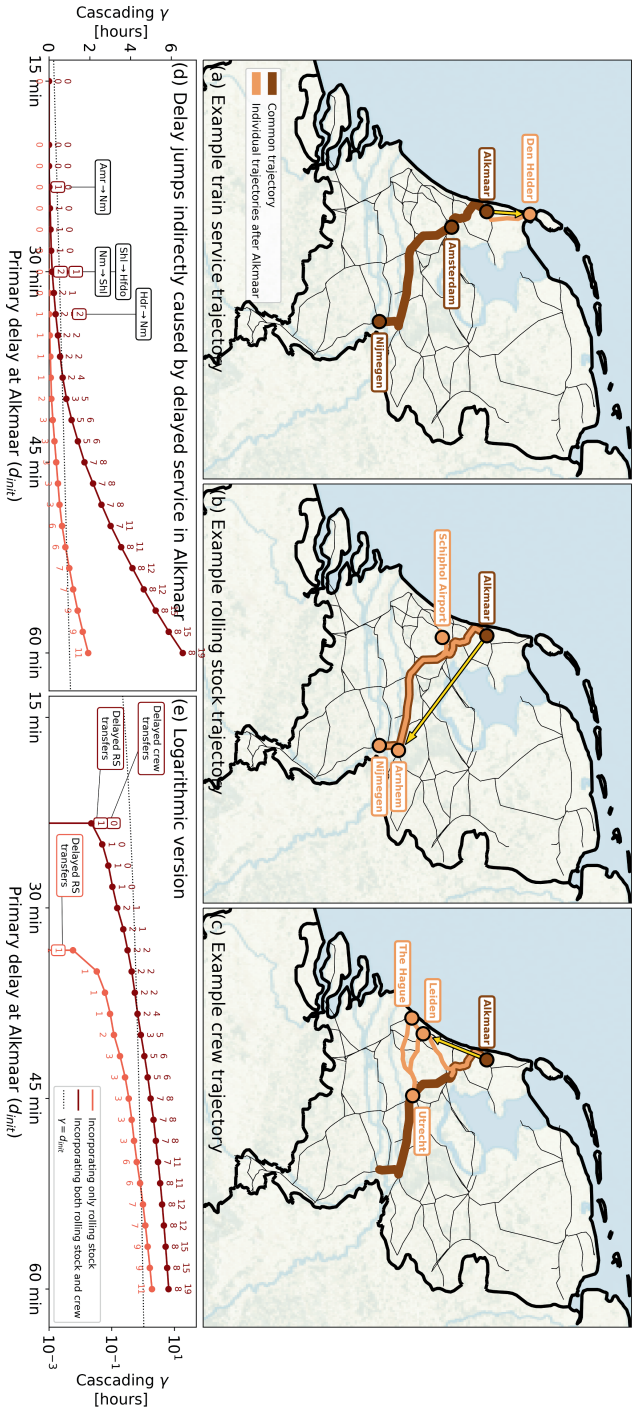


Figure 6.2: Illustration of delay cascading mechanism in the Dutch railways. Panels (a)–(c): Routes in the Dutch railways of (a) the train service 3028 from Nijmegen to Alkmaar via Amsterdam, (b) a rolling stock unit used in part of this service and (c) a crew member (partly) executing this service. The schedules for December 3, 2017 are used; a day selected randomly from the dataset. Dark brown lines mark the route they share between Nijmegen and Alkmaar, after which they go their separate ways — marked in light brown lines and yellow arrows. While the service continues along its original rolling stock — via Schiphol Airport back to Nijmegen shown in panel (b). Panel (c) shows that the crew member transfers towards another service to the south-west — via Amsterdam to Leiden and The Hague, proceeding via Utrecht to Leiden and eventually ending in Utrecht. If service 3028 would have an initial delay $d_{init} > 0$, so does the rolling stock and crew executing the service; meaning that if scheduled buffer times for the resource transfers in Alkmaar would be exceeded by the delay (and no replacement resource would be available), then the subsequent services of these resources would become delayed as well. In other words, service 3028’s delay will potentially be transmitted to other services, and subsequently carried to other geographical regions. Panel (d) shows the $d_{init} \cdot \gamma$ plot: γ remains zero for $d_{init} < 23$ minutes as the entire delay is absorbed by scheduled buffer times for resource transfers. However, with $d_{init} = 23$ minutes, a first resource delay overcomes the transfer buffer and adds delay to another service; namely the rolling stock unit in Alkmaar, going towards Nijmegen. As d_{init} grows, more and more transfer buffers are overcome and delay is added to many service lines throughout the country. Panel (e) contains the same information as in panel (d), for a log-linear plot, revealing the near-exponential increase of γ as a function of d_{init} . Abbreviations used in the panel depict Alkmaar (Amr), Nijmegen (Nm), Schiphol Airport (SH), Hooftdorp shunting yard (Hd) and Den Helder (HdH). The dotted black lines in panels (d–e) correspond to $\gamma = d_{init}$; γ becomes larger than d_{init} for $d_{init} \geq 38$ minutes.

a came from elsewhere, where it had 300 seconds delay, with a buffer of 120 seconds (for recombining into service S). Two crew members transferred to service S , one (member I) with 720 seconds delay (with buffer 600) and one (member II) with 540 seconds delay (with buffer 600). Table 6.1 summarises these numbers. The set $\{a^i\}$ in Eq. 6.1 contains four activities, of which the buffers (third column) should be subtracted from their delays (second column), to get the potential contribution (fourth column) the delay of their combined activity, $d(a)$. The corresponding delay jumps $\delta(a)$ are calculated by comparing the resulting $d(a)$ for each model to the delay of the previous activity of the service $d(a_{\text{prev}})$. In the monolayer model, we obtain $d_{\text{mono}}(a) = \max\{30\} = 30$ (such that $\delta_{\text{mono}}(a) = 0$), in the bilayer model $d_{\text{bi}}(a) = \max\{30, 180\} = 180$ (such that $\delta_{\text{bi}}(a) = 150$) and in the trilayer model $d_{\text{tri}}(a) = \max\{30, 180, 120, 0\} = 180$ (such that $\delta_{\text{tri}}(a) = 150$).

Resource	Delay d	Buffer β	Potential contribution
Train service	30	0	30
Rolling stock	300	120	180
Crew member I	720	600	120
Crew member II	540	600	0

Table 6.1: An example calculation of delay propagation in case of resource transfers. All values are stated in seconds.

6.4 Quantification of cascading

In Fig. 6.2(d-e) we plot the *delay cascading metric* γ , defined as the cumulative sum of all resource transfer-related delay jumps, i.e., $\gamma = \sum_a \delta(a)$, consequential to some initial delay d_{init} due to planned service 3028 on December 3, 2017 at the city Alkmaar. Given a predefined buffer time, the higher d_{init} is, the more buffers are exceeded; i.e., the more positive delay jumps occur, and the more γ increases. Stated differently, the ‘tighter’ the buffer times are planned — often the tendency of benefit-to-cost optimisations — the more prone transport and logistic systems would become to large-scale disruptions.

In real-time operations, cascading like this will of course not go unnoticed, as railway dispatchers will take mitigation measures in real time. Rescheduling of rolling stock and crew, and service cancellations constitute mitigation measures; in effect, for $t > t_0$, they simply alleviate (some of) the constraints imposed on the model by the operator’s schedule defined at t_0 . To briefly showcase how the model reproduces delay build-up in a real situation, let us now use it to evolve an actual delay snapshot of the *entire system*, and compare the results to the real data. To this end, we choose December 11, 2017 as a case study, another day with a severe blizzard (similar to the Dutch case shown in Fig. 6.1). On this day, a code red weather alert was issued from 12:00h onward, in anticipation of which, the Dutch railways used an adapted schedule. Notwithstanding, the west and centre of the country got disrupted to the point that almost no train traffic was possible around Utrecht and Amsterdam in the afternoon. Having initialised the model with the system snapshot at $t_0 = 19:00\text{h}$ — meaning also that the model is constrained to the operator’s predefined schedule at 19:00h — we compare the model predictions for

the evolution of delay to the observed data at 20:00h in Fig. 6.3. The aim of this figure is to illustrate whether our model reproduces amplification (in terms of delay magnitude) and geographic spread of delay well, when we initialise the model at a certain moment of time. Similarities between reality [panel (a)] and the model output [panel (b)] can be found in the centre and east of the country, whereas differences can be seen in the south and north (such differences affect the model's performance, analysed later). The total delays [panel (c)] for the mono- and bilayer models are seen to decrease quickly, while the trilayer model predicts the (increasing) evolution of the total delay rather well for the first 90-120 minutes. Note that the trilayer model prediction considerably overshoots the real data after 21:00h in panel (c). [In section 6.4, we will demonstrate that the period 19:00-21:00 happens to coincide with a large number of mitigation measures (Fig. 6.4a), suggesting that the dispatchers have mitigated much of the delay that the model predicts (this is consistent with the overestimation of the delay by the trilayer model)]. Still, the overall better performance of the trilayer model is logical, since in terms of modelling the entire system's dynamics, this model is the most complete, and therefore it captures the fullest extent of the cascading effects. For this reason, we henceforth exclusively consider the trilayer model, referring to it simply as 'the model'.

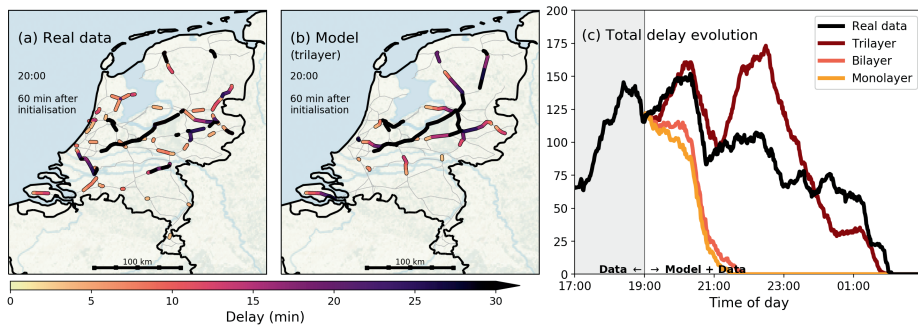


Figure 6.3: Model simulations of a real situation. Comparison of simulation and observed data on December 11, 2017. All models were initialised with the system snapshot at 19:00h. **Panels (a) and (b)** show the spatial distribution of delay at 20:00h in the real data (a) and simulation outcome of the trilayer model (b). **Panel (c)** shows the total delay evolution in time for the observed data and simulation outcomes of the three models. The trilayer model predicts the total delay well up to 120 minutes, after which it decays while in reality the total delay increased again. The differences between the monolayer model and the bi-/trilayer models stem from delay cascading, built in the latter ones. Only delays larger than 3 minutes are shown for visualisation purposes.

In order to quantify the role of delay cascading in large-scale disruptions, we resort to a comparison of the delay jumps predicted by the model prediction and the operational data, as extracting this directly from the operational data is not possible since no cascading-related information, such as explanations for mitigation measures and cascading events, are logged. By initialising the model every 15 minutes between 06:00h on a day and 01:00h on the next, we first compute *all* model-predicted delay jumps with 15 minutes lead time. This lead time is chosen based on a high performance of the model (quantified below by C in Fig. 6.5c-f). We then heuristically classify them, upon comparing with operational data, in the following six categories: cascading due to (I) crew and

(II) rolling stock transfers, mitigated cascading due to (III) rescheduling and (IV) cancellations, (V) other larger incidents (delay jumps > 10 minutes), and (VI) net noise (i.e., all other unaccounted for delay jumps), as follows (details in SI E.3). If there was a positive delay jump at a rolling stock/crew transfer point in the operational data, and the model indicated a cascading event, then the delay jump is accordingly categorised as I or II. Other large and small delay jumps in observed data are categorised as V and VI respectively. Lastly, large delay jumps indicated by the model but not found in the operational data, upon cross-checking with the latter, are attributed to mitigation measures by rescheduling (III) or cancellations (IV). Note that category III concerns crew only; we lack real-time rolling stock rescheduling data since the Dutch railways infrastructure company ProRail does not record them. Further, the distinction between categories V and VI is artificial: it is primarily meant to demonstrate the occurrences of ‘large’ vs. ‘small’ incidents across day classes in Fig. 6.4(b).

For clarification purposes, let us provide an example for the delay jump classification. Consider a train that departs from station A with 600 seconds delay, while earlier, upon arrival at this station (i.e., its previous logged activity), it had 180 seconds delay, resulting in an observed delay jump of $d_{\text{obs}} = 600 - 180 = 420$ seconds. With 15 minutes lead time, the model however happens to produce $d_{\text{sim}} = 360$ seconds delay. Since $d_{\text{obs}} > 0$ and $d_{\text{sim}} > 0$, we check the resource schedules, and find that the train re-departed from station A with a different crew than it had upon arrival at station A, and one of these crew members was delayed. We therefore mark the observed delay jump of 420 seconds as category I: cascading due to crew. Carrying out this procedure for all activities and associated delay jumps in the data, and adding their contribution together (per time window) results in the data that are shown in Fig. 6.4.

Figure 6.4(a) shows the delay jump contribution of these categories for our case study day December 11, 2017 (note: this is the same day as in Fig. 6.3); the respective contributions of these categories to the delay jumps can be seen to fluctuate heavily through the day—this is in fact typical for any day. In particular, many mitigation measures are found on this day, mostly in the evening, after 18:00h. By scaling up to four days each for four unique day-classes as used by the Dutch railways (Black, Red, Neutral and Green days), we show the relative contribution of each delay jump category in Fig. 6.4(b). Quantitative details on this classification, can be found in SI E.1. ‘Black’ days refer to days with many service cancellations and severe delays, while ‘Green’ days refer to days with barely any cancellation and only small delays (with ‘Red’ and ‘Neutral’ in between). For communication reasons we keep the label names as is used by railway practitioners.

Shown in approximate absolute hours Fig. 6.4(b) are the aggregated contributions of each of the categories. When deriving the role of cascading from this figure, it is important to realise that categories I and II (cascading through crew and rolling stock) indicate a baseline portion of cascading; those that were detectable using the heuristic described two paragraphs above. The sizeable magnitude of mitigation categories III and IV in Fig. 6.4(a) indicates that many cascading-induced delay jumps were prevented by railway dispatchers. Further, categories V (‘other large incidents’) and VI (‘net noise’) cover a combination of externalities, coincidences, but also more cascading (which could not be captured by the heuristic) and secondary effects of cascading. It is for this reason that we identify the aggregated contributions of categories I-IV as the *minimum* (average) cascading metric γ for real-time operations. Figure 6.4(b) shows that γ decreases with the

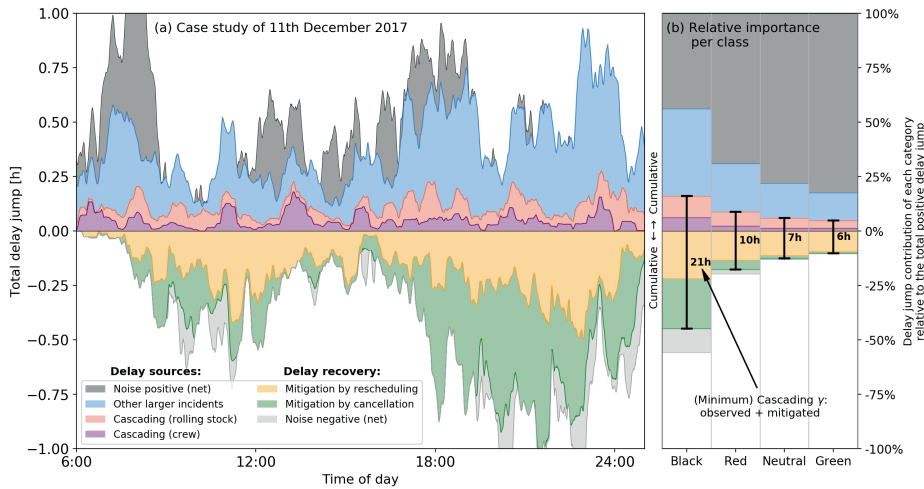


Figure 6.4: Origins of delay jumps. Total (summed) delay jumps, sorted by various mechanisms that cause them, and subdivided in 'labels', for **Panel (a)** one case study (December 11, 2017), and **panel (b)** their proportions averaged over four days each for four day classes ('Black', 'Red', 'Neutral' and 'Green') for the Dutch railways. The origins of the delay jumps were identified by comparing observed data to model output. Magnitudes [panel (a)] and relative magnitudes [panel (b)] were calculated for time windows of 5 minutes, with a 30-minutes smoothing window used for display purposes. Four types of delay jumps that act as *delay sources* are distinguished: (I) delay cascading due to crew transfers (purple), (II) delay cascading due to rolling stock transfers (red), (V) other larger incidents (blue) and the positive part of (VI) net noise. Three types of delay jumps that act as *delay recovery* are distinguished: (III) mitigated cascading due to rescheduling (yellow), (IV) mitigated cascading due to cancellations (green), and the negative part of (VI) net noise. The positive part and the negative part are plotted separately in a cumulative sense — up- and downward, respectively. In panel (b), the total identified delay cascading (observed plus mitigated) is highlighted in hours.

day-severity decreases: on Black days, we find an average of $\gamma = 21$ hours, while on Green days, this is only $\gamma = 6$ hours. The relation between γ and day-severity suggests that cascading indeed becomes increasingly important in more disrupted situations. (Also, the mitigation measures devised by the Dutch railways are evidently seen to be quite effective.)

6.5 Predicting the evolution of large-scale disruptions

Finally, we address the matter of predicting the evolution of large-scale disruptions using the model. How accurate the model predicts this evolution is referred to as the 'model performance'. The relevance of model performance is twofold: for validating the model, and for providing (early) warnings to predict the evolution of disrupted situations in real-time operations. For the former, comparing model performance across multiple types of days ('Green', 'Black', etc.) may identify the cases wherein cascading, built in the model's mechanisms (i.e., the three building blocks from section 6.2), indeed played a dominant role in delay propagation (in cases when the performance is high), or when other (excluded) factors played a significant role (in cases when the performance is low).

In general, prediction by any delay propagation model suffers from two limiting factors: (i) large, poorly predictable, new incidents external to the system (captured partly by the blue and the grey areas in Fig. 6.4), and (ii) mitigation measures (yellow and green areas in Fig. 6.4) that alleviate constraints on the model imposed by the operator's predefined schedule at initialisation time t_0 . By leaving these out by construction, our model outputs how delay would build up in their absence [this, in fact, is the qualitative explanation for the discrepancy between Figs. 6.3(a) and (b), as remarked earlier]. The model's prediction accuracy therefore gets limited to a certain lead time ($t - t_0$) — the longer it is, the more negative influence (i-ii) have on its performance. We therefore define a *predictability horizon* ($t - t_0$), for which the condition $C_{t_0}(t) = c$ holds for some c . Here $C_{t_0}(t)$ is the cosine of the angle ϕ , obtained from the dot product between the system-wide real and model-determined departure delay vectors $\vec{D} = [d(a_1), d(a_2), \dots, d(a_n)]$, obtained by aggregating all the train services in the time window $W(t, \Delta t) \equiv [t - \Delta t/2, t + \Delta t/2]$ with window size $\Delta t = 30$ minutes:

$$C_{t_0}(t) \equiv \cos[\phi_{t_0}(t)] = \frac{\vec{D}_{\text{real}}(t) \cdot \vec{D}_{\text{model}, t_0}(t)}{|\vec{D}_{\text{real}}(t)| |\vec{D}_{\text{model}, t_0}(t)|}. \quad (6.2)$$

We use this quantity as the *model performance metric*, with values between 0 (performing poorly) and 1 (performing perfectly). Although cosine similarity (like many other correlation metrics) does not incorporate absolute delay magnitudes when comparing model output to observed data since they are scaled out in Eq. (6.2), the quantity C does take into account the spatial distribution as well as the corresponding relative magnitudes of delays (of which larger delays are of particular interest) over the entire network.

The lead time should in principle correlate well to the cosine similarity modulo the absence of the two limiting factors (i-ii). While we have little means to quantify (large external) incidents, we can quantify mitigation measures by $P_{t_0}(t)$, defined as the ratio (determined using the schedule at t_0):

$$P_{t_0}(t) = \frac{\text{executed crew actions scheduled to be in } W(t, \Delta t)}{\text{all crew actions scheduled to be in } W(t, \Delta t)}. \quad (6.3)$$

(Ideally one should also account for rolling stock rescheduling in the interval $[t_0, t]$, but as pointed out earlier, we do not have data on live rescheduling of rolling stock.) Analogous to the predictability horizon, one can also define the *mitigation horizon* ($t - t_0$) for which the condition $P_{t_0}(t) = p$ holds for some p .

Using the definitions (6.2-6.3), we first investigate the predictability horizon for our case study day December 11, 2017 (as in Fig. 6.3 and Fig. 6.4a). Starting at 6:00h on the day, the model is initialised every 6 minutes, run forward, and both C and P are calculated at every time point (at 1 minute resolution) up to 01:00 AM on the next day for each run. The results are shown in Fig. 6.5(a); both the horizontal and vertical axes display time of day. The diagonal represents $t = t_0$, on which $C = P = 1$ trivially holds. Using the time-stamps displayed on the horizontal axis, in Fig. 6.5(a) we plot $P_{t_0}(t)$ in blue at every point on the left of the diagonal — e.g., for the little black circle — where t_0 is the model initialisation time-stamp of the point, and t is the crew activity time-stamp, corresponding to the intersection point of a horizontal line (shown by the right arrow) from that point and the diagonal. A similar process is followed for plotting $C_{t_0}(t)$ in red on the

right of the diagonal, but this time using the time-stamps displayed on the vertical axis. The blue and the red contours for several values of c and p then respectively correspond to the mitigation and predictability horizons; they respectively measure how far ahead in time the crew schedule remains invariant enough, and how far up front in time the delay situation can be predicted. For example, the large distance of the P and C contours to the diagonal at $t = 16:00\text{h}$ (following the arrows) means that the crew (that were working around 16:00h) were barely rescheduled prior to 16:00h, and the model also performed well to predict the delay state at 16:00h. Similarly, distance of the P -contour to the diagonal dropped considerably after 16:00 hour, resulting in large amounts of cascading mitigation, as in Fig. 6.4a.

We expect that the two horizons to track each other for large-scale disruptions (the larger P is, the less the constraints at model start-time t_0 are alleviated, meaning that the higher the model performance C ought to be). The relation between the two is explored in the C - P diagrams in Fig. 6.5(b-c) in two different ways. The scatter of points in Fig. 6.5(c) denote the C and P values for five different colour-coded lead times ($t - t_0$) at hourly values of t , for the same four large-scale disruption (Black) days as in Fig. 6.4(b). While there is a large scatter in Fig. 6.5(c), their averages, plotted in circles marked with lead-time values, indicate a near-linear relation between C and P for lead time up to an hour. The same is seen in Fig. 6.5(b), wherein the contour bands $C_{t_0}(t) = 0.35 \pm 0.1$ and $P_{t_0}(t) = 0.7 \pm 0.05$ from Fig. 6.5(a) on December 11, 2017 are also seen to track each other well with an offset all through the day. Finally, Fig. 6.5(d-f) we compute the C - P diagrams for the other days as in Fig. 6.4(b). For a given lead time, we find two distinct trends with increasing day-severity: (i) P decreases significantly [ostensibly to facilitate mitigation measures, as seen in Fig. 6.4(b)], (ii) notwithstanding that, the model performance C — the cosine similarity — improves (in particular, the 30-min lead time averages improves from 0.43 and 0.51 on Green and Neutral days to 0.59 on Black days). This is indeed counterintuitive, but is explained by the following. The delay spreading mechanisms are better captured by the model on Black (severely disrupted) days in contrast to the Green days, for which delay spreading is dominated by incidents and noise events external to the model [covered by the blue and the grey areas in Fig. 6.4(b)]. This corroborates our central message — namely that delay cascading dominates large-scale disruption events: despite more mitigation measures, more buffers are systemically exceeded on Black days than on, e.g., Green days.

6.6 Conclusion and outlook

We show that large-scale delay evolution during railway disruptions emerge from the complex interactions of resources and services. Central to large-scale disruptions are the processes of delay amplification and spreading on the (fixed or time-invariant) infrastructure network. We find that these processes require a dynamical interplay among three building blocks: (a) constraints (the required resources are delayed), (b) cascading (this delay is passed on secondary services), and (c) transport (the secondary services spread the delay across long spatial and temporal time scales), wherein cascading playing the key role for delay amplification. Although we consider only railway systems here, the conceptual similarity in resource allocation and schedule-based dynamics in other

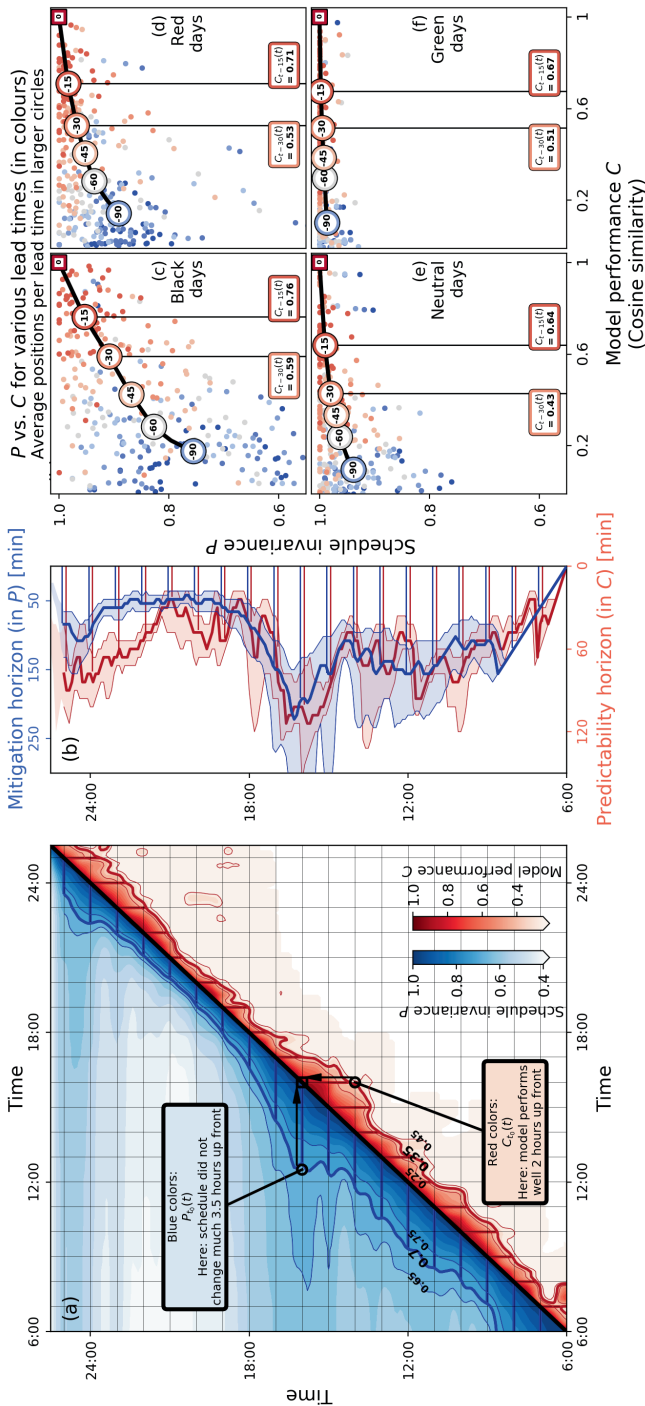


Figure 6.5: Predictability performance of the model. Panel (a), above diagonal: Mitigation measure $P_{t_0}(t)$ (blue), where both the model initialisation time t_0 and crew activity time t are read from the horizontal axis. Panel (a), below diagonal: Model performance $C_{t_0}(t)$ (red), plotted similarly to $P_{t_0}(t)$, but times are read from the vertical axis. See also text for details. The contours $C_{t_0}(t) = 0.35 \pm 0.1$ and $P_{t_0}(t) = 0.7 \pm 0.05$ are marked respectively in blue and red lines. The data in this panel is smoothed using a Gaussian-averaging for visualisation purposes. Panel (b): Horizons for P (blue) and C (red) for the same three values of c and p as in panel (a), measured as the horizontal and vertical distances to the diagonal of the P and C contours, respectively. Panel (c)-(f): Model performance $C_{t_0}(t)$ (horizontal) versus crew schedule invariance $P_{t_0}(t)$ (vertical) on multiple instances of the day ($t = 11:00h, 12:00h, \dots, 22:00h$), with various lead times (t_0 up to 1.5 hours before t), for the Green, Neutral, Red and Black days also analysed in Fig. 6.4 — 16 days in total. Colours depict lead time: red indicates small lead times (i.e., predictions are closely up front), blue indicates large lead times. Averages on 15 min intervals are shown in large circles, with extra emphasis on the C values of these averages for 15 min and 30 min lead times (marked at bottom of each panel). Calculations of $C_{t_0}(t)$ and $P_{t_0}(t)$ are performed at $t_0 = 6$ -minutes time resolution [15 minutes for (c)-(f)], with a 30-minutes window around t .

systems like airways and logistic systems suggests that cascading can be an important driver to large-scale disruptions in these systems as well.

Although, in this chapter, we focus on railway transport with a wider outlook to transport and logistic systems at large, our work connects to a wider class of dynamical processes taking place on networks. In the introduction we have already pointed out that the problem we study here — spreading phenomena on fixed (time-invariant) railway infrastructure network — belongs to the type of dynamics *on* networks. Well-studied examples of this type are diffusion-related phenomena on networks (e.g., in various (bio)chemical reaction-diffusion, or activation-inhibition processes running on networks (Nakao and Mikhailov, 2010; Diego et al., 2018)), network epidemiology (Brockmann and Helbing, 2013), and the spreading of rumours and opinions over social networks (Watts, 2002; Centola et al., 2007). Even though one can make a distinction among these regarding nature-made or human-made systems subtypes, the spreading mechanism we identify here, *viz.*, delay cascading across multiple layers of resources and service elements, is a rich (and supported by real life data) addition to literature of dynamics on networks. The constraint-cascading-transport mechanism of (delay) spreading, which we derive here using the example of railway transport, adds to the understanding of the spreading dynamics on complex networks. While we hope that our work will stimulate a wider search into spreading phenomena on complex networks, we also note that the multi-layer coupled dynamics can also be formulated as a temporal network (Holme and Saramäki, 2012), which is however beyond the scope of the the current work.

Finally, for the specialist field of transport research, we foresee two effective measures to arrest cascading, potentially averting large-scale disruptions. First, the introduction of ample spare resource capacities and buffer times, to ensure that local constraints do not get easily activated. The trilayer model pinpoints precisely where vulnerabilities lie within the railway operator's schedule: e.g., in Fig. 6.2 we show that delays in the Dutch city of Alkmaar (north-west) at a particular moment in time has the potential to cause delays in the city of Nijmegen (east) at a later point in time: a form of building long-range causal correlations in the system. Spare crew in Alkmaar would not only prevent this specific spread, but upscaling the model's results to long-term statistics would also provide any transport system's dispatcher with data to pinpoint ideal locations and quantities of spare resources — an analysis we plan to execute in the future. (Having said this, we acknowledge that optimising the system for benefit-to-cost ratios will limit the possibilities of having spare resources.) A second measure to prevent cascading is a modular design approach for the transport functionalities. Modular designs (like the Danish railway system), by definition, do not reuse resources from one area to another, which prevents delay cascading between these regions. Modular design approaches have been considered in the context of operational rescheduling during disruptions (Potthoff et al., 2010; Nielsen et al., 2012). We hope that this work prompts new research into the trade-offs and complexity of how to design transport systems resilient to disruption spreading.

PART III

TEMPORAL NETWORK APPROACHES

On the role of interaction sequences and vulnerability
to spreading



"I think the next century will be the century of complexity."

STEPHEN HAWKING, 2000



Introduction to Part III

In Part II, I discussed the interaction between dynamics in a complex system and its underlying network architecture. In chapters 4, 5 and 6, the example of railway systems is taken. In this example, it is natural to assume a fixed, constant underlying network architecture: even though the railway infrastructure is subject to changes in tracks and stations, these changes are slow with respect to the dynamics on top of the network. For example, while it may take years to build new tracks and stations, the delay dynamics in railway systems happen in a matter of hours, up to a day (because of the reset of the resources and schedules). This reveals an important concept to dynamics in complex systems in general: the concept of *time scales*. I distinguish two relevant time scales for this dissertation — a third will be added later. The first is the time scale of the dynamics (τ_{dyn}), referring to a typical period of time on which the dynamics evolve. The second is the time scale of the change in agent interactions, i.e., in the network architecture (τ_{netw}). Clearly, for the railways: $\tau_{\text{dyn}} \ll \tau_{\text{netw}}$, which allows us to assume an invariant network architecture ($\tau_{\text{netw}} \rightarrow \infty$), which greatly simplifies the analysis. This inequality, motivating the use of static networks, typically holds for urban transportation systems (hence the focus on railway systems in chapters 4-6), but also for, e.g., many hydrological systems, where the water flows much faster than rivers and lakes change, plant vein structures, in which nutrients are distributed faster than the growth of plant veins, and the spreading of a popular video on social media, which happens much faster than the networks of friends and followers change.

However, in many complex systems, this inequality does not hold: network architecture may evolve at similar time scales as the dynamics of interest ($\tau_{\text{dyn}} \approx \tau_{\text{netw}}$), or even faster ($\tau_{\text{dyn}} > \tau_{\text{netw}}$). In such cases, the changes in the network architecture needs to be taken into account explicitly by the use of *temporal networks*. Topologically, temporal networks are composed of a set of networks, each representing all agents (nodes) and their interactions (links) at a specified time or time interval. These ‘snapshot’ networks are connected in the temporal dimension by connecting identical agents at different time steps. This results in two types of links: those representing actual interactions at any time (interval), and those that represent a temporal connection of an agent. The time period between two network snapshots should of course typically be of the order of τ_{netw} to represent the changes in the network architecture well. However, oftentimes, this is not the case because of empirical constraints. For example, in human interaction data (e.g. those used in chapter 8), the time between contacts may be a few seconds, while the sampling time may be measured in minutes. Also for practical purposes, a sizeable and perhaps constant sampling time may be beneficial, which may not represent reality best. Hence, one can speak of a third relevant time scale relating network architecture changes to dynamics on top: the time scale of sampling the network architecture τ_{samp} . Despite potential limitations due to τ_{samp} being larger than τ_{netw} , temporal networks have important advantages when compared to their static (non-time varying) counterparts when aiming to describe dynamical phenomena.

The importance of including temporal changes in the network architecture becomes especially apparent when dealing with spreading phenomena, which typically involve an important memory part and a high necessary resolution of interaction sequences. Let me highlight this in an example of, again, a simple model of disease spreading using the

primary school example also used in Figs. iib and iva. The same SIR model as in the introduction to Part II is used here, but on different network structures. Note that in reality, a disease typically evolves from agent to agent on a timescale of days — e.g., for SARS-CoV-2, the incubation and infectious periods are found to vary between a couple of days up to a few weeks (He et al., 2020) — while the empirical networks used here span just over a couple of hours. This means that $\tau_{\text{dyn}} \gg \tau_{\text{netw}}$, which would be a problem if we would use this network architecture for more than just this illustration. I chose a single agent as the one being infected at $t = 0$ to initialise the disease spreading ('patient zero'). The results are shown in Fig. v, for three cases of the underlying network structure. In the first case (panel [a]), the static network is used — the fully aggregated network over 2 hours of primary school children interactions, which does not involve any time-varying network architecture. This does include a disconnected cluster, which is therefore never reached by the disease. In five time instances, the evolution of the disease across this network is shown in node colours, reflecting the average status across an ensemble of 300 simulations. The resulting evolution of the total % agents that are either infectious (I) or recovered (R) is shown to the right in purple. As suspected, the yellow and red colours spread first in the highly connected clusters (in fact, corresponding to classrooms) close to patient zero, while later also progressing to children further in the network. In panel (b), the network architecture is changed across five time intervals (each of 6 time steps, each being an aggregate over 24 min in the empirical dataset). Clearly, the number of connections in each network snapshot is much smaller, also limiting the spread: both visible in the node colours as well as in the brown line in the right panel being much lower than the purple one. Overall, the individual brown lines are quite confined — even though the spreading of the disease is stochastic and the network architecture is changing over time. To illustrate the importance of sequences in spreading phenomena, I shuffled the first four network architectures, but kept the last one the same for visual comparison. The results, shown in panel (c), dramatically change: in the $t = 24$ panel, we see that the disease has spread much further in the disconnected components to the upper right and lower right than in panel (b). Also, when looking at the results the right panel, the results vary much more than in panels (a) and (b): three main branches of outcomes can be distinguished, as marked by the arrows. Note that the eventual outcomes of branches 2 and 3 are also quite different from the outcome in panel (b), while the same network architectures are use, but merely shuffled.

A number of concepts can be learnt from the example above, that are relevant to the focus on chapters 7, 8 and 9. First, the impact of a time-varying network architecture in general: clearly, the network snapshots in panel (b) are much different from those in panel (a). Moreover, the outcomes of the very same spreading model are dramatically different (i.e., the purple versus the brown lines), resulting from the fact that the agents simply do not have all their unique interactions at the same time, throughout the simulation period, which, if true, would of course quicken the spread — as visible in panel (a). Comparing panel (b) to (c) also reveals that *sequences* matter: shuffling order of contacts may have a large effect on the outcome of the spreading dynamics. And third, specific attention can be given to the right panel of (c), where the different branches reflect that particular (stochastic) infections early on may greatly influence the general direction of the spreading later on — in this case, leading to a difference of 30%-point of the population.

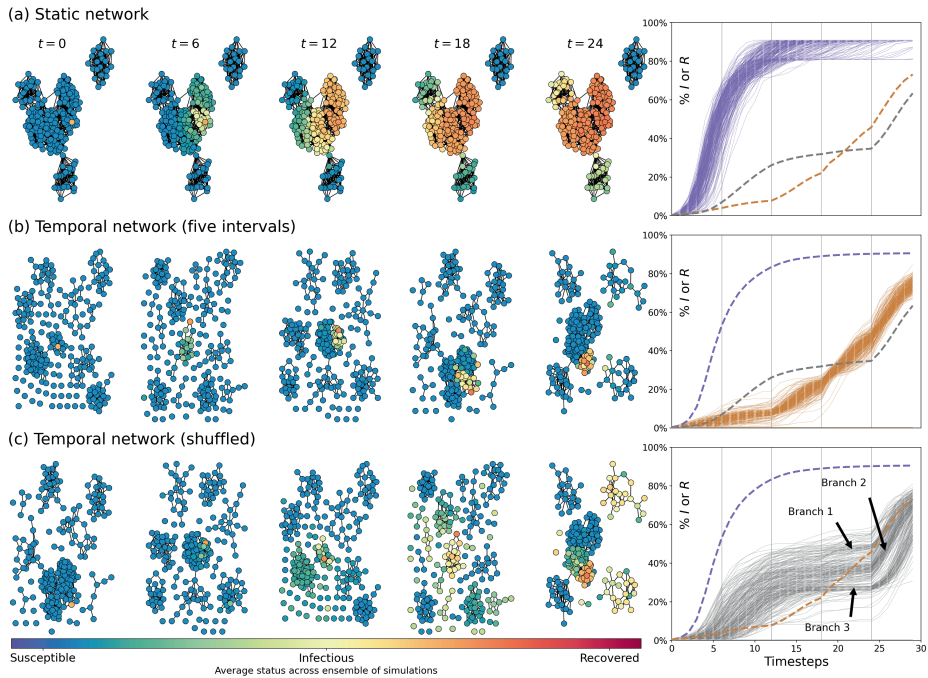


Figure v: Illustration of how (temporal) network architecture affects dynamics on top by simulating an SIR model on the contacts in a primary school (same dataset as in Figs. iib and iva) (Gemmetto et al., 2014). **Panel (a):** Static network of interactions over two hours. Node colours indicate the average status across an ensemble of 300 simulations using the same initialisation with a single agent. The average status is calculated via averaging end statuses of the simulation (0: susceptible, 1: infectious, 2: recovered). In the right panel, the % agents that are either infectious (I) or recovered (R) is given by purple lines, indicating individual runs using the network structure on the left. For reference, the grey and brown dashed lines show averages of the other approaches. **Panel (b):** Same as panel (a), but for a temporal network approach, where five intervals are used, reflecting the network architectures shown in this panel, each representing the structure of 6 consecutive time steps. The network architectures correspond to the real empirical intervals in the two hours used in (a), i.e., 24 min each. The brown lines in the right panel show the evolution of the fraction I or R . **Panel (c):** Same as panel (b), but where the network snapshots are shuffled (apart from the final one, to allow visual comparison). Grey lines in the right panel show the evolution of the fraction I or R . All network components are shown, not only the giant components (unlike before). Vertical lines in the right panels indicate the moments at which network architecture is changed (i.e., borders between intervals of the temporal network).

These concepts — the impact of time-varying network architectures in general, interaction sequences, and the importance of specific connections to the general outcome — are the focus of this part and will be addressed in the following chapters. Chapter 7 proposes a method to analyse the impact of time-varying structures on its vulnerability to spreading phenomena, much like the one in Fig. v, and quantifies this in the novel metric *entropy of temporal entanglement*. The calculation of this metric involves the transformation of grouped agent interactions to *events*, which can be linked through time (in an ‘event map’) by checking whether two events contain at least one shared agent. Tracking these events explicitly addresses sequences of interactions among the

agents. The strength of this metric lies in that it does not require any simulation model, but rather focuses on the information in the temporal networks themselves. We show that this topology-based metric is useful to describe dynamics on top of the network.

In some sense, chapter 8 is an application of chapter 7, using empirical human interaction data from an art fair in Amsterdam 2020. Like in chapter 7, it utilises events to track interactions and their sequences over time. The aim of the chapter is to quantify which agents would be most crucial to a hypothetical spreading dynamics happening through the interactions in this art fair. In colloquial words, we aimed to identify behavioural ‘superspreaders’ by taking into account the sequences of their interaction. While identifying important spreaders from mere network degree would have been trivial, unexpectedly, degree is not necessarily a dominant factor when incorporating sequences of interactions. We quantify the agent’s impact on the spreading by *contact sequence centrality*.

The final chapter of this dissertation, chapter 9, involves the simulation of the first wave of SARS-CoV-2 infections in the Netherlands, with the ultimate aim to test the government’s intervention strategies. While a much more detailed model than the more theoretical principles in chapters 7 and 8, it utilises the principles of temporal networks, sequential information and the effect of specific infections to the outcome at the same time. The model used in this chapter simulates 170,000 agents, moving across the Dutch municipalities in an hourly resolution, meaning that every hour, the network architecture underlying the spreading dynamics of SARS-CoV-2 changes. Particular interest is given to the closure of schools, interventions involving people’s mobility and locally adjusted interventions — reflecting on targeted approaches to reduce the spreading vulnerability as discussed in chapters 7 and 8.


Summarised, the overarching question addressed in Part III is:

Research question Part III: *How can we quantify and influence spreading vulnerability in temporal networks?*

CHAPTER 7

THE HIDDEN DEPENDENCE OF SPREADING VULNERABILITY ON TOPOLOGICAL COMPLEXITY

This chapter can be found as a preprint: Dekker, M. M., Schram, R., Ou, J., and Panja, D. (2021e). The hidden dependence of spreading vulnerability on topological complexity. *arXiv:2107.01651*. Supplementary material to this chapter can be found in SI G.



ABSTRACT

Many dynamical phenomena in complex systems concern spreading that plays out on top of networks with changing architecture over time — commonly known as temporal networks. A complex system's proneness to facilitate spreading phenomena, which we abbreviate as its 'spreading vulnerability', is often surmised to be related to the topology of the temporal network featured by the system. Yet, cleanly extracting spreading vulnerability of a complex system directly from the topological information of the temporal network remains a challenge. Here, using data from a diverse set of real-world complex systems, we develop the 'entropy of temporal entanglement' as a novel and insightful quantity to measure topological complexities of temporal networks. We show that this parameter-free quantity naturally allows for topological comparisons across vastly different complex systems. Importantly, by simulating three different types of stochastic dynamical processes playing out on top of temporal networks, we demonstrate that the entropy of temporal entanglement serves as a quantitative embodiment of the systems' spreading vulnerability, irrespective of the details of the processes. In being able to do so, i.e., in being able to quantitatively extract a complex system's proneness to facilitate spreading phenomena from topology, this entropic measure opens itself for applications in a wide variety of natural, social, biological and engineered systems.



7.1 Introduction

Networks, consisting of system elements (agents) and their interactions by nodes and links respectively, have proved to be effective tools for analysing complex systems. For a large variety of them, notable emergent dynamical phenomena of interest concern spreading, effected by individual agents playing the role of carriers and transmitting to others as they interact, e.g., sharing information through conversations (Cattuto et al., 2010; Stehlé et al., 2011), passing of signals among animals (Almaas et al., 2004; Yosef and Regev, 2011) and of infectious disease pathogens (Stehlé et al., 2011; Firth et al., 2020), purveyance of (fake) news (Wang et al., 2019), synchronicity in neuronal spikes (Klausberger and Somogyi, 2008; Rakshit et al., 2018), and cascading dynamics in socio-technical systems (Badie-Modiri et al., 2020; Mancastropa et al., 2019; Masuda and Holme, 2020; Li et al., 2020a). When the agents' identities, and the precise sequences and timings of their interactions over a given time interval are compiled together into a temporal network (Holme and Saramäki, 2012; Riolo et al., 2001; Isella et al., 2011; Perra et al., 2012; Li et al., 2017, 2020a), it becomes evident that spreading is actually a dynamical process taking place on top of the network.

From this, it seems natural to expect that the topological complexity of a temporal network will have a profound influence on the system's proneness to facilitate spreading phenomena — we abbreviate the latter as its spreading vulnerability (quantified later in this chapter) — e.g., the formation and dispersion of epidemiological bubbles or echo chambers will dictate *how fast, or efficiently*, pathogens or (fake) news may spread. However, establishing a quantitative relation between the two, from which the system's spreading vulnerability can be extracted directly from the topological complexity of the temporal network featured by the system, remains a challenge. The key obstacle stems from the heterogeneities in the agent interactions intrinsic to temporal networks in the temporal domain — there simply are no rules to guide which agent may interact with which others, when, for how long and in which precise sequence — that need to be suitably combined with the topological complexities that are already present in the network of agent-to-agent contacts at any given time. Note that the latter, i.e., the agents' contact network at any given time, is in fact a *static* network, for which methods to deal with heterogeneities are already well-established in the forms of degree distribution, various forms of centrality measures and community detection algorithms (Girvan and Newman, 2002).

Here we develop an insightful quantity to measure topological complexities of temporal networks: we name it the *entropy of temporal entanglement* (the choice for this name is justified below, and it bears no connection to entanglement entropy for quantum many-body systems). Using real-world temporal networks data, and simulating three different dynamical processes on top of them, we show that this entropic measure not only allows for a quantitative comparison of vastly different complex systems, but crucially, it also bears a clean relation to the system's spreading vulnerability, irrespective of the details of the processes. In other words, this entropic measure is able to directly, and cleanly, extract a complex system's proneness to facilitate spreading phenomena from its topology.

We coin the term 'entanglement' because of the following. Real-world temporal networks often follow a 'discrete time convention', wherein the agents' interactions, denoted by agent-to-agent contacts, are sampled at some fixed interval (which we denote as τ_s):

We then probe the topology of a temporal network, consisting of N agents, using random walkers hopping along agents' temporal paths forward in time. (Probing network topology using random walkers is a common practice.) The entire procedure is parameter-free, and is formally described in SI G.1.1. We summarise it here for m events at time t in three steps; a rendering of them for the example system in Fig. 7.1(a) with $N = 8$ and $m = 2$ at $t = 0$, and the corresponding calculations are shown in Fig. 7.1(b-c). (1) Using Roman and Greek letter indices to denote agents and events respectively, we construct the $N \times m$ 'agent-to-event propagator matrix' $Q^{(\text{in})}(t)$, with matrix element $Q_{i\alpha}^{(\text{in})}$ denoting the probability for a random walker to start at agent i prior to t and reach event α at (integer) time t in one hop. (2) Similarly, we construct the $m \times N$ 'event-to-agent propagator matrix' $Q^{(\text{out})}(t)$, with matrix element $Q_{\alpha i}^{(\text{out})}$ denoting the probability for a random walker, to start at event α at time t and reach agent i subsequent to time t in one hop. (For both steps, the hops are thus coupled to the sampling time interval τ_s .) (3) Finally, we construct the $N \times N$ 'agent-to-agent propagator matrix' as

$$P(t) = Q^{(\text{in})}(t) \cdot Q^{(\text{out})}(t). \quad (7.1)$$

Upon extending this procedure to finite (integer) interval Δt leads us to the product matrix $\wp(t, \Delta t) = P(t) \cdot P(t+1) \cdot P(t+2) \dots P(t+\Delta t)$ that similarly contains the full entanglement information in the finite time interval $[t, t+\Delta t]$. By construction, the matrix element $\wp_{ij}(t, \Delta t)$ is the probability of a random walker starting at agent i prior to time t to end up at agent j subsequent to time $(t+\Delta t)$ following agents' temporal paths. Using these probabilities, we define the entropy of temporal entanglement over the interval $[t, t+\Delta t]$ as

$$S(t, \Delta t) = -\frac{1}{N \ln N} \sum_{i,j} \wp_{ij}(t, \Delta t) \ln \wp_{ij}(t, \Delta t). \quad (7.2)$$

The $N \ln N$ factor ensures $S \in [0, 1]$.

Construction of the entropy of temporal entanglement that respects the precise agent-to-agent contacts within events [such as the absence of a direct contact between agents 1 and 4 at $t = 0$ in Fig. 7.1(a), while all the others are in direct contact in that event] follows a line similar to the above, albeit it is slightly more involved. The corresponding entropy of temporal entanglement S_c , formally derived in SI G.1.2, is also parameter-free and fully determined by the topology of the network. In SI G.1.2 we argue that with decreasing sampling time interval τ_s , S_c approaches S , the entropy of temporal entanglement as calculated in Eqs. (7.1-7.2) in terms of the connected components of the agents' contact networks, which we henceforth adhere to.

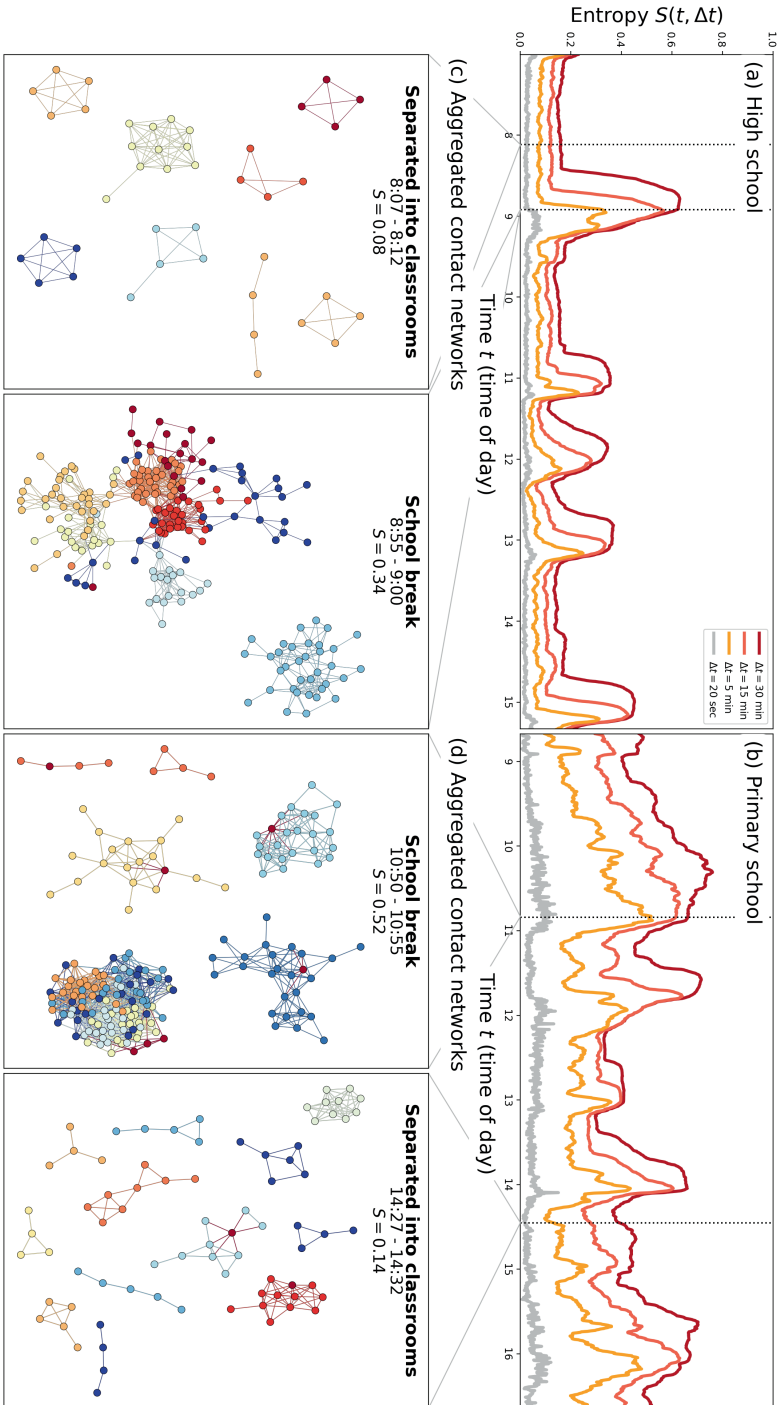


Figure 7.2: Entropy of temporal entanglement for two real-world temporal networks: contacts among high school (Thiers13 dataset day 2, left column) and primary school (LyonsSchool dataset day 1, right column) students. **Panels (a) and (b)** $S(t, \Delta t)$ for four values of Δt : 20 seconds (grey), 5 minutes (orange), 15 minutes (red) and 30 minutes (darker red) — for any value of t , note that $S(t, \Delta t)$ is monotonically increasing with Δt — x-axes denote the time of day. Vertical dashed lines correspond to 5-minute intervals, expanded in panels (c) and (d), chosen to reflect one high-entropy and one low-entropy situation per panel. **Panels (c) and (d)** Aggregated contact networks within the 5-minute intervals (only connected components with at least 4 students are shown), demonstrating high (resp. low) connectivity in high- S (resp. low- S) cases. Metadata reveal that the situations correspond to school breaks and ongoing classes respectively (students belonging to the same classroom are represented by the same node colour).

7.3 Results

7.3.1 Interpretation and properties

Three insightful attributes of $S(t, \Delta t)$ are critical for being able to extract a complex system's proneness to facilitate spreading phenomena. These are: (1) $S(t, \Delta t)$ represents the agents' mixing (i.e., mingling) propensity in the interval $[t, t + \Delta t]$ (we explore this further in 7.3.2 below), (2) it respects the *temporal sequence* of agent interactions, and (3) in building up the sequences forward in time, it automatically incorporates *causality*: note that a spreading phenomenon in the past can influence its future dynamics, but not *vice versa*. While (2-3) can be gleaned from how φ is built from the P -matrices, (1) follows from that $\varphi_{ij}(t, \Delta t)$ is the probability of a random walker starting at agent i prior to time t to end up at agent j subsequent to time $(t + \Delta t)$ following temporal paths: the more distinct temporal paths there are to trace a given agent i back from any agent j , the more nonzero φ_{ij} elements there are. These attributes further translate to the following useful properties of $S(t, \Delta t)$. (a) An agent, who never interacts with another in the interval $[t, t + \Delta t]$, has zero contribution to $S(t, \Delta t)$. This, in fact, is the basis for why small $S(t, \Delta t)$ corresponds to a highly fragmented temporal network, i.e., large number of disjoint components in the interval $[t, t + \Delta t]$, possibly with many agents separated from each other throughout the interval. (b) It allows us to quantify the contribution of an individual agent to entanglement, which we take up in SI G.1.3. (c) Typically, an increasing Δt would imply more nonzero φ_{ij} 's: more distinct temporal paths become available to trace a given agent i back from any agent j . This implies that for any value of t , $S(t, \Delta t)$ is monotonic in Δt ; e.g., as can be seen in Fig. 7.2(a-b); for a fixed value of Δt however, $S(t, \Delta t)$ is non-monotonic in t . [There are other fundamental properties of $S(t, \Delta t)$ that one expects entropic measures to satisfy, e.g., (i) as the agents keep interacting, $S(t, \Delta t) \rightarrow 1$ as $\Delta t \rightarrow \infty$, as observed in Figs. 7.3(a-b), and (ii) if we divide N agents into two groups A and B of sizes N_A and N_B respectively, then $(N \ln N)S(t, \Delta t) \geq (N_A \ln N_A)S_A(t, \Delta t) + (N_B \ln N_B)S_B(t, \Delta t)$. Such properties are however not pertinent to the focus of this work.]

In contrast to the above, the practice of aggregating the agents' contact networks over a certain time interval for a temporal network into a static one (Li et al., 2020a; Isella et al., 2011; Riolo et al., 2001; Masuda and Holme, 2013; Stehlé et al., 2011) discards the precise sequences and durations of agent interactions. As for methods that do capture heterogeneities in the temporal domain by means of, e.g., checking how communities evolve in time (Rosvall and Bergstrom, 2010; Peixoto and Rosvall, 2017), constructing temporal counterparts of degree and betweenness (Kim and Anderson, 2012), or treating a temporal network as a multilayer one with layers representing discrete time snapshots (Mucha et al., 2010), we note the following. The emphasis of temporal betweenness on shortest paths makes it ideal to identify topological centrality, but for the application of spreading vulnerability, path *lengths* are only important in terms of which agents are traceable in a fixed time interval. Of critical importance for spreading vulnerability are the exact interaction sequences and frequencies, which are considerably less focused on in the existing literature. Likewise, concerning community detection approaches, under time reversal, the community structure in the multilayer network will remain invariant, while the dynamics of a process taking place on top of the temporal network will be profoundly

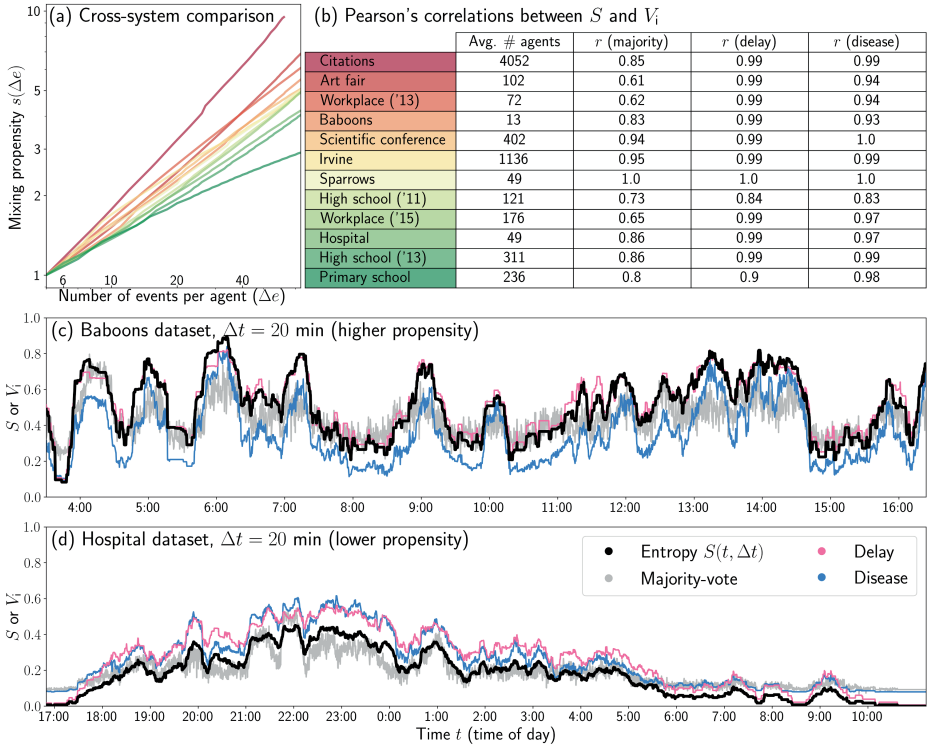


Figure 7.3: Standardised mixing propensity s and the systems' vulnerability V to spreading phenomena (for precise definitions of both, see main text 7.3.3 below). **Panel (a)** Mixing propensity vs the number of events per agent as a system-characterising property, shown for 12 complex systems: from green (slow rate of increase) to red (fast rate of increase). The systems are noted in panel (b) in the same colour coding. **Panel (b)** Pearson's correlation between $S(t, \Delta t)$ and $V_i(t, \Delta t)$, with $i \in \{\text{maj, del, dis}\}$, corresponding to a majority-vote model, a transport delay model and an epidemiological susceptible-infected (SI) model respectively (see text and SI G.3 for details). **Panel (c)** The full time-series of $S(t, \Delta t)$ and $V_i(t, \Delta t)$ on day 2 of the Baboons dataset (Gelardi et al., 2020). Displayed on the x-axis are the times of day. $\Delta t = 20$ minutes (i.e., 60 units of sampling time intervals) is used. **Panel (d)**. Same as in panel (c), but using day 2 of the Hospital dataset (Vanhems et al., 2013).

different.

7.3.2 Representation of mixing propensity

In order to demonstrate that the entropy of temporal entanglement has a clear relationship to the agents' propensity to mixing, we consider two temporal network datasets that track student contacts respectively in a high school and a primary school by wearable sensors, sampled at 20 seconds intervals (Génois and Barrat, 2018; Stehlé et al., 2011; Mastrandrea et al., 2015) (details on the data in SI G.3). Per dataset, in Fig. 7.2(a-b) we plot the time evolution of $S(t, \Delta t)$ for four different Δt -values, varying t over one full school day. Further, delving into the (publicly available) metadata, in Fig. 7.2(c-d) we plot the contact networks of the agents aggregated over two five-minute intervals, one each for

high and low values of $S(t, \Delta t)$. The correspondence between the top and the bottom panels demonstrates that the students' mixing behaviour, adapting to changing school circumstances, is being reflected in the topological complexity of the temporal network, which in turn is being captured by the entropy of temporal entanglement (school breaks invite more mixing in comparison to being separated in classrooms, and correspondingly lead to more topological complexity and higher entropy of temporal entanglement).

Given the above, an unbiased (e.g., from number of agents, sampling time interval) cross-system comparison of topological complexities of temporal networks requires two rescaling operations. The first (natural) one is that Δt must be replaced by Δe , the average number of events per agent in the interval Δt (the assumption here is that the agents' contact networks are sampled frequently enough such that contact sequences are fully captured). The second one concerns normalising the entropy to define a standardised mixing propensity:

$$s(\Delta e) = \frac{1}{T} \sum_{t=0}^T \frac{S(t, \Delta e)}{S(t, \Delta e_0)}, \quad (7.3)$$

where Δe_0 is a constant for standardisation. Note that $s(\Delta e)$ is a monotonically increasing function of Δe since $S(t, \Delta t)$ is a monotonically increasing function of Δt . Plotting $s(\Delta e)$ vs. Δe in a log-log plot in Fig. 7.3(a), using $\Delta e_0 = 5$, we observe the differences in how characteristically fast $s(\Delta e)$ builds up in different real-world systems (see SI G.3 for details on the systems). (Even though Δe_0 is merely a standardisation parameter, it should be chosen not too small to have some stability in the plot. We have experimented with other values such as $\Delta e_0 = 1$ and the results do not substantially differ.) We analyse this further in SI G.2 to demonstrate that the fastest rate of increase in the entropy of temporal entanglement is achieved when the agents' contact sequences, in the temporal domain, only contain trees. Real-world temporal networks of course contain loops (e.g., due to repetitive interactions among agents); note also that social contacts are often structured in 'bubbles', which enhances the chances of having loops.

7.3.3 Relation to spreading vulnerability

We begin by quantifying spreading vulnerability of a complex system, which requires adding dynamical processes on top of the temporal network that the system features. For this purpose, we consider three different types of stochastic processes: (1) a majority-vote model that simulates the opinion on a dilemma that spreads through events by means of the majority votes, (2) a transport delay model, where delay spreads through events due to all the participants copying the delay of the maximally delayed agent, and (3) an epidemiological susceptible-infected (SI) model with infection probability $\beta = 0.8$. Details of the models and their backgrounds can be found in SI G.4. The corresponding process variables are respectively expressed as $V_i(t, \Delta t)$, with $i \in (\text{maj}, \text{del}, \text{dis})$, and just like the entropy of temporal entanglement $S(t, \Delta t)$, they are normalised to the interval $[0, 1]$ (elaborated in SI G.4). Like $S(t, \Delta t)$, for any given t , $V_i(t, \Delta t)$ is a monotonically increasing function of Δt , and the spreading vulnerability is quantified by how fast $V_i(t, \Delta t)$ increases as a function of Δt . [Also, like $S(t, \Delta t)$, $V_i(t, \Delta t)$ exhibits strong heterogeneous behaviour as a function of t .]

We then simulate the above three stochastic processes for all the systems in Fig. 7.3(a). Using $\Delta t = 15$ minutes we plot the process variables together in Figs. 7.3(c-d) for the temporal network dataset for interacting baboons (Gelardi et al., 2020) and hospital patients and workers (Vanheems et al., 2013): visual inspection immediately reveals that for both systems, the variations in S and V_i over time are highly synchronised. For all the complex systems in Fig. 7.3(a), the Pearson's correlation coefficients between $S(t, \Delta t)$ and $V_i(t, \Delta t)$ are tabulated in Fig. 7.3(b) using the same colour scheme [of Fig. 7.3(a)]. The table clearly showcases that the entropy of temporal entanglement is essentially an embodiment of a complex system's vulnerability to spreading phenomena, irrespective of the details of the processes playing out on top of the networks. In other words, the entropy of temporal entanglement secures a clean relation between the spreading vulnerability of a complex system, and the topology of the temporal network featured by the system, and thereby, the current work indicates that computing the entropy of temporal entanglement is sufficient to quantitatively assess a complex system's proneness to facilitate spreading phenomena.

7.4 Discussion and outlook

Summarising, we have developed the entropy of temporal entanglement as a measure of topological complexity of temporal networks. By construction, it is a parameter-free quantity that embodies collective topological property at any given timescale, allowing for comparisons across vastly different complex systems. By simulating three different dynamical processes playing out on the top of real-world temporal networks, we have demonstrated that the entropy of temporal entanglement is a good representation of the systems' spreading vulnerability, irrespective of the details of the processes — less entanglement (lower topological complexity) means lower spreading vulnerability and *vice versa*.

Let us reflect on why this topological entropy measure is such a good predictor of spreading dynamics, irrespective of the details of the dynamical processes put on top on the temporal networks. The reason can be traced back to the fact that the $\varphi_{ij}(t, \Delta t)$ is the probability of a random walker to start at agent i prior to time t , and end up at agent j subsequent to time $t + \Delta t$. With $\sum_j \varphi_{ij}(t, \Delta t) = 1$, the time evolution of $\varphi_{ij}(t, \Delta t)$ as a function of Δt can itself be seen as a spreading process starting at agent i at time t , albeit a *deterministic* one — determined entirely by the topology of the temporal network, free of any parameter. It is therefore logical that the entropy of temporal entanglement serves as a quantitative embodiment of the systems' spreading vulnerability, irrespective of the details of the processes. Nevertheless, the important fact remains that the entropy of temporal entanglement establishes a clean quantitative relation between the topological complexity of a temporal network and the system's proneness to facilitate spreading phenomena.


System vulnerability is receiving progressively more attention in widely diverse areas, especially in the context of large perturbations induced by, e.g., the COVID-19 pandemic, species' habitat loss and fragmentation, disruption of food webs, and robustness of biological systems such as gene regulation, metabolism, neural dynamics, and engineered systems. The common features underlying this diversity are that (a) these are complex

systems, wherein the components (e.g., species, genes) have (developed) time-varying functional dependencies among each other (e.g., sharing resources and by-products, regulating biochemistry) that can be expressed as temporal networks, and (b) the effect of a perturbation, albeit initially localised in space and time, will potentially spread or cascade through the entire system. Predicting the spreading vulnerability of perturbations directly from topological complexity would be an asset — especially considering that the availability of real-world data on large perturbations are typically limited — for example, for mapping their tipping points (Dakos et al., 2019), or alternatively, for tracing the boundaries of their safe operating spaces (Rockström et al., 2009) that separate successful recovery from irreversible degradation. Albeit the dynamical processes running on top of the (temporal) networks will undoubtedly possess a diverse range of time-scales (and interact with the time-scales of the changes in the network topology), the topological complexity of the networks will be a crucial factor in dictating the spreading dynamics. The general nature of our methodological approach sets the stage for these applications in such wide and far-reaching areas.

CHAPTER 8

QUANTIFYING AGENT IMPACTS ON CONTACT SEQUENCES IN SOCIAL INTERACTIONS

This chapter can be found as a preprint: Dekker, M. M., Blanken, T. F., Dablander, F., Ou, J., Borsboom, D., and Panja, D. (2021a). Quantifying agent impacts on contact sequences in social interactions. *arXiv:2107.01443*, accepted in *Scientific Reports*. Supplementary material to this chapter can be found in SI H.



ABSTRACT

Human social behaviour plays a crucial role in how pathogens like SARS-CoV-2 or fake news spread in a population. Social interactions determine the contact network among individuals, while spreading, requiring individual-to-individual transmission, takes place on top of the network. Studying the topological aspects of a contact network, therefore, not only has the potential of leading to valuable insights into how the behaviour of individuals impacts spreading phenomena, but it may also open up possibilities for devising effective behavioural interventions. Because of the *temporal* nature of interactions — since the topology of the network, containing who is in contact with whom, when, for how long, and in which precise sequence, varies (rapidly) in time — analysing them requires developing network methods and metrics that respect temporal variability, in contrast to those developed for static (i.e., time-invariant) networks. Here, by means of event mapping, we propose a method to quantify how quickly agents mingle by transforming temporal network data of agent contacts. We define a novel measure called *contact sequence centrality*, which quantifies the impact of an individual on the contact sequences, reflecting the individual's behavioural potential for spreading. Comparing *contact sequence centrality* across agents allows for ranking the impact of agents and identifying potential 'behavioural super-spreaders'. The method is applied to social interaction data collected at an art fair in Amsterdam. We relate the measure to the existing network metrics, both temporal and static, and find that (mostly at longer time scales) traditional metrics lose their resemblance to *contact sequence centrality*. Our work highlights the importance of accounting for the sequential nature of contacts when analysing social interactions.



8.1 Introduction

Human behaviour plays a central role in generating patterns of interaction that allow for the spreading of a great variety of entities — from rumours (Choi et al., 2020) to viruses (Stehlé et al., 2011) to memes (Kotsakos et al., 2015). Most recently, the importance of such patterns of interaction has been borne out by the COVID-19 crisis. A virus like SARS-CoV-2, which caused the COVID-19 pandemic, spreads through physical contact, or through aerosols that have a finite range; for the virus to be transmitted, people will have to have make close contacts with each other. It is for this reason that behavioural interventions, aiming to break the chains of close human contacts, were our only weapon against COVID-19 for the better part of 2020 and remain so until vaccine coverage is sufficient.

As such, the pandemic underscored the need for better models and measures of human behaviour and the contact patterns it generates. On the technological side, recording individual-level contacts through wearable sensor technology has recently seen considerable development that render the study of contact patterns scientifically feasible (Génois and Barrat, 2018; Cattuto et al., 2010; Stehlé et al., 2011). For example, albeit still imperfect, sensor accuracy is beginning to approach the level and scale needed to inform epidemiology and to model interventions (Masuda and Holme, 2013; Zhang et al., 2021; Liu et al., 2018). In contrast, on the methodological side, we still lack adequate tools to assess, represent, and model the myriad ways in which human behaviour generates patterns of individual-level contacts. This limits progress in many disciplines that require such information, but especially hampers progress along the behaviour-epidemiology interface that is so important in improving our preparedness to deal with the current and future pandemics. This chapter develops methodologies that can begin to fill this gap, by exploiting connections among improved empirical assessments of contact patterns and novel quantitative metrics that are sufficiently advanced to analyse behavioural contact patterns. In particular, it focuses on quantifying the role of each individual in the overall contact patterns, reflecting the agent's impact on the potential of (e.g., epidemiological) spreading in a system.

To achieve these goals, we utilise a network approach in which patterns of contacts which are generated by individual behaviour are mapped on to a contact network. In such a network, each individual corresponds to a node and a link between two nodes represents a contact between two individuals. The advantage of using a network approach is that a general spreading process of some entity (of which pathogen spreading in epidemiology is one, where the concerned pathogen is the entity) can be analysed by means of a two aspects: (a) a contact network, borne out of people's behaviour in social interactions, and (b) a spreading process occurring within a population, which can then be modelled as a dynamical process that uses the network connections as pathways. From a network science perspective, it then becomes natural to expect that the topological properties of interaction networks can be used to analyse the potential of spreading dynamics. In addition, this approach allows us to analyse the relative importance of different individuals in this dynamical process by analysing the effect that removing the individual would have on the network structure. Individuals who contribute greatly to the spread of entities on the network may then represent potential targets of intervention.

It is important to note, however, that spreading dynamics can take on many forms:

infectious diseases will spread differently from rumours or political opinions, which are yet different from spreading of delays in network transport (Dekker and Panja, 2021). Nevertheless, the common element they all share is (deterministic or stochastic) transmission upon contact, and spreading is therefore dependent on the topology of the contact network: who interacts with whom, for how long and in which precise sequence. For example, if a person *A* spreads the rumour, a person *B* can only spread the rumour to person *C* if (s)he has been in contact with *A* before the contact with *C*. Tracking these sequences of interactions, therefore, provides the basis of any spreading phenomenon. While the work presented here was developed in the context of modelling behaviour during the COVID-19 pandemic, the importance of better models for human behaviour in generating contact networks should therefore be seen as generic.

Before a satisfactory analysis of this process can be found, however, several problems that arise in the currently used network analytic techniques (for evaluating the structure of contact networks and the role of different individuals in generating them) must be addressed. First and foremost, in reality, a contact network is dynamic, i.e., the topology of a contact network is changes over time. In Network Science such networks are called *temporal networks*. Many existing methods that analyse contact networks, empirically obtained through wearable sensors (Blanken et al., 2020), treat contact networks as static, since, for instance, recorded contacts may not be adequately logged or the analysis needs to be simplified by collapsing networks over the time domain. From a Network Science perspective, this is in fact reminiscent of one way to reduce the complexity of a temporal network by downscaling it to a static one: a network with a time-invariant topology, obtained through aggregation of all edges over a certain time interval (Li et al., 2020a; Isella et al., 2011; Riolo et al., 2001; Masuda and Holme, 2013; Stehlé et al., 2011). However, information on who interacts with whom, for how long and in which precise sequence gets lost in such aggregations. Static representations therefore do not reflect the dynamic aspects of human behavior: changing interactions, structures and sequences (Li et al., 2017, 2020a; Mucha et al., 2010; Génois and Barrat, 2018; Centola, 2010; Schläpfer et al., 2014; Kiti et al., 2019; Cattuto et al., 2010) have been shown to have large impacts on various dynamical processes (Mucha et al., 2010; Perra et al., 2012; Scholtes et al., 2014; Peixoto and Rosvall, 2017). Static representations miss a crucial opportunity to intervene at the contact network level, e.g., by targeting interventions at specific points in the time domain. Secondly, while there are *static* topological metrics available to quantify the impact of agents on spreading mechanisms, static network techniques cannot differentiate between agents whose contacts are structured in the same way but located at different time points. Examples of such topological metrics are degree, closeness or betweenness centrality (Dezsó and Barabási, 2002; Pastor-Satorras and Vespignani, 2002), *k*-shell decomposition algorithm (Kitsak et al., 2010), methods based on clustering and the topology potentials (Wang and Zhao, 2015), improved coreness centrality and eigenvector centrality (Ahajjam and Badir, 2018), influential individuals (Zhang et al., 2019).

More prudent (existing) approaches to deal with the complexity of temporal contact networks entail setting a threshold to filter out the non-essential edges or edges that exist only by chance: for example, Grabowicz et al. (2014) used a simple threshold based on the number of events (i.e., count of edges across time) between two nodes; Kobayashi et al. (2019) defined a temporal null model to identify pairs of nodes having more interactions than expected given their activities; and in yet another thread of studies, Mellor

(2018) and Mellor (2019) introduced the temporal event graph (TEG), which uses event (interaction between two individuals) as nodes and shared event attendees as edges in a directed graph. These efforts may focus on only part of the complexity of temporal networks and provide possible ways to identify structures (Grabowicz et al., 2014; Kobayashi et al., 2019), communities (Génois and Barrat, 2018; Peixoto and Rosvall, 2017; Peel et al., 2017) and quantify connectivity (Mellor, 2019; Badie-Modiri et al., 2020; Torricelli et al., 2020). However, in the context of spreading phenomena, these metrics are not meant for quantifying agent impacts on contact chains in human interactions (and thereby the spread of a concerned entity).

A more principled approach to unravelling behavioural heterogeneity and the impact of individual actors, which we develop in this chapter, is to consider the full dynamics of people's contacts in the form of a temporal network. Specifically, we provide a way forward by developing (a) a systematic methodology for representing people's (dynamical) contact networks as event graphs, and (b) techniques to analyse the role that individuals play in generating these network structures. The latter information can be used to assess individuals' impacts on the contact sequence *while respecting (temporal) sequence of contacts*.

The chapter is structured as follows. We start by introducing the methodology to extract an individual's impact on contact sequences in a system in section 8.2, for which we need to introduce a reconstruction of the temporal graph using 'event mapping'. In section 8.3, we apply the methodology to data collected during an art fair in Amsterdam in August 2020, where we also provide experimental details and discuss the evolution of the most important metrics over time, as well as relate them to the existing network metrics like contact degree. In section 8.4, we discuss these findings and relate them to the broader discussion on spreading phenomena in networked systems.

8.2 Methods

In this section, we introduce the mapping of temporal network data into an event graph in order to facilitate the identification of agent impacts on a sequence of contacts.

8.2.1 Event mapping of contact sequences

We track the contact sequences of the agents in the following manner. In Fig. 8.1(a) we show an example of a (randomly generated) temporal system with eight agents and four time steps $t \in \{0, 1, 2, 3\}$ in three types of graphs. In the centre, the temporal graph is shown where agents are denoted in squares. The contacts among the agents at every snapshot are denoted by black links, while the agents' movements across snapshots are shown by grey links. To the left of the panel, we show the corresponding aggregated *contact network* over the entire time range, wherein two agents are connected by an edge if they have been in contact at any of the four time steps: the aggregated contact network therefore does not distinguish any sequence between the interactions.

To the right of Fig. 8.1(a), we combine connected components within time snapshots into single circular nodes, which we refer to as *events* (Dekker et al., 2021e), resulting in yet another graph: the *event graph*. Events may consist of either single or more agents,

as marked by the numbers indicated inside each node. The events are linked by the ‘re-usage’ of agents; this is exactly how every agent can act like a spreader between consecutive events, which makes event mapping a natural choice for describing the dynamical evolution of a spreading phenomenon. For example, at $t = 0$, agent 5 is interacting with agent 3. From there, they go their separate ways: at $t = 1$ agent 5 interacts with agent 8, and agent 3 interacts (in a larger group) with agents 1, 2 and 7. Indeed, the event (3, 5) at $t = 0$ links to both (5, 8) and (1, 2, 3, 7) at $t = 1$. Using these events, one can convert the (temporal) network snapshots to an event graph (directed in time), shown at the right of Fig. 8.1(a). Not only does this *event mapping* collapse complex temporal network data into a single directed graph, but most importantly, it also preserves the sequences of contacts. [An important question is to what extent the information for spreading phenomena is already embedded in the static contact graph, and when it is important to do event mapping. This is a question we address multiple times in this chapter, by comparing the results of our metrics (defined below) to properties of the static contact network.]

The event graph allows for an intuitive analysis of which (agent contacts in) events in later time steps are eventually traceable back to the an event in earlier time steps via both direct and indirect connections. An example of such a contact sequence is given in Fig. 8.1(b). There, we track the events that are linked to agent 5 at $t = 0$ (coloured red), over four sequential time snapshots. At $t = 0$, only the event where agent 5 itself is involved is coloured red [i.e., agents (3,5)]. Over time, ‘secondary’ events become coloured red in the event graph, meaning that in due course of time, progressively more events, and the agents contained therein, can be traced backwards to agent 5 at $t = 0$. In the end, for most applications, how many agents can be traced backwards to agent 5 at $t = 0$ is the most relevant quantity (rather than the number of events, i.e., the number of red dots). This prompts us to define $P(j, t, \Delta t)$:

$$P(j, t, \Delta t) = \frac{\text{Number of agents traceable to agent } j \text{ at time } t \text{ within time interval } [t, t + \Delta t]}{N}, \quad (8.1)$$

where N is the total number of agents. In other words, $P(j, t, \Delta t)$ is the fraction of *unique* agents at time $t + \Delta t$ that can ultimately be linked backwards to agent j at time t — clearly, $P(j, t, \Delta t)$ is a function of the starting time t and elapsed time Δt . In Fig. 8.1(b), we show $P(5, 0, \Delta t)$ for different values of Δt , using agent 5 as ‘agent zero’: at $\Delta t = 0$, $P(5, 0, 1) = 2/8 = 25\%$ since at $\Delta t = 0$ two (3 and 5 itself) out of eight agents are linked to 5 at $t = 0$. Similarly, $P(5, 0, 2) = 6/8 = 75\%$ since there are six agents (1, 2, 3, 5, 7 and 8) that can be traced backwards to agent 5 at $t = 0$, and so on. Clearly, a high value of $P(j, t, \Delta t)$ for low Δt would indicate fast mingling of the agents in the time interval $(t, t + \Delta t)$, making the system highly prone to spreading phenomena from the perspective of starting with agent j at $t = 0$. Conversely, a low value of $P(j, t, \Delta t)$ for high Δt would imply the opposite. It is obvious that for fixed t , $P(j, t, \Delta t)$ is a monotonically increasing function of Δt , while for fixed Δt and a different initialisation time t , $P(j, t, \Delta t)$ may vary and therefore possibly result in lower values with increasing t .

In panel (b), we use agent 5 as a mere example to illustrate the calculation of $P(j = 5, t, \Delta t)$. A more representative quantity is obtained by taking the average over all agents j :

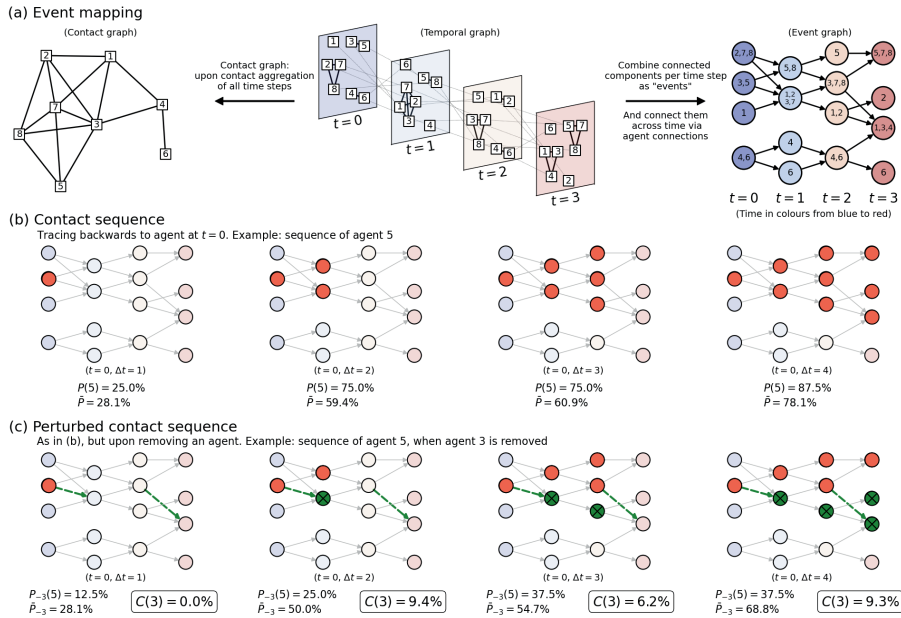


Figure 8.1: Illustration of the event mapping and an individual’s contact sequence centrality using a toy example. All throughout this figure, individual agents are in squares (and numbers), events are in circles. **Panel (a):** In a temporal network, the time evolution of the contact graph (middle) captured in a sequence of snapshots. Contact data over the four snapshots can be aggregated to obtain a static (and aggregated) contact network (left), but in doing so, the sequential information is lost. We term the connected network components within each snapshot ‘events’, they are seen as nodes in an event graph (right). In an event graph the event-to-event trajectories of agents are the links. Colours of circles and snapshots indicate time from blue to red. **Panel (b):** Example of a contact sequence when tracing all events and subsequent agents back to agent 5 at $t = 0$. Starting time t , time elapsed Δt , the percentage unique agents $[P(5)]$, traceable to 5 and the average percentage of agents in the contact sequence \bar{P} are indicated. **Panel (c):** Same as (b), but after removal of agent 3, resulting in a ‘perturbed’ contact sequence of agent 5. The (now absent) links that were associated with agent 3 are in dashed green arrows, and the events that are consequentially not traceable to 5 anymore are coloured green. The associated (lowered value of) $P_{-3}(5)$ is shown. Averaging over all agents provides \bar{P}_{-3} , and allows us to calculate the *Contact Sequence Centrality* of agent 3: $C(3) = \bar{P} - \bar{P}_{-3}$, which is clearly marked in the bottom panel. For abbreviation, we excluded the t and Δt between brackets when showing values of P , \bar{P} and C .

$$\bar{P}(t, \Delta t) = \frac{1}{N} \sum_j P(j, t, \Delta t) \tag{8.2}$$

the value of which is also indicated in Fig. 8.1b. The value of \bar{P} says something about the (unperturbed) contact sequences: high values indicate that spreading to many agents within $(t, t + \Delta t)$ is likely, and vice versa. We aim to identify the impact of each agent on the overall contact sequences. We measure this by looking at the \bar{P} value when we remove an agent — the intuition being that the resulting change in \bar{P} reflects the impact of the removed agent. Again, we start with the example of the contact sequence of

agent 5 (i.e., $P(5)$ instead of \bar{P}). Now, to assess the impact of *another* agent — agent 3 — on agent 5's contact sequence, we *remove* this agent from the system and show what it does to agent 5's contact sequence in Fig. 8.1(c). The removal results in the absence of important links (denoted in green dashed lines) and, consequently, in limiting the contact sequence's reach (i.e., excluding the green nodes). The proportion of agents in agent 5's contact sequence is now written as $P_{-3}(5, t, \Delta t)$, where the subscript $-i$ refers to the removal of agent i . In particular, we have $P_{-3}(5, 0, 2) = 2/8 = 25\%$, which is lower than $P(5, 0, 2)$, since the large event comprised of agents 1, 2, 7 and the removed 3 cannot be traced back to agent 5 at $t = 0$ anymore. More generally, for any agents i, j the relation $P_{-i}(j, t, \Delta t) \leq P(j, t, \Delta t)$ holds, because the contact sequence is either equal or shrunk by the removal of agent i .

The impact of an agent on the contact sequence is naturally quantified by the difference in \bar{P} between the unperturbed [panel (b)] and the perturbed event graph. For the impact of agent 3, the latter is shown in panel (c). Averaging over the contact sequences of each agent j , we obtain an expression for the impact of an individual agent on the contact sequence which we refer to as *contact sequence centrality*:

$$C(i, t, \Delta t) = \frac{1}{N} \left[\sum_j P(j, t, \Delta t) - \sum_j P_{-i}(j, t, \Delta t) \right] = \bar{P}(t, \Delta t) - \bar{P}_{-i}(t, \Delta t). \quad (8.3)$$

A high $C(i, t, \Delta t)$ indicates that the removal of agent i decreases the fraction of agents in an average contact sequence sharply, which reflects an important role of i in connecting contact sequences. This way, $C(i, t, \Delta t)$ becomes an attribute of agent i . It might just reflect that agent i itself has had a lot of contacts in the interval $(t, t + \Delta t)$ (i.e., the degree of agent i in the aggregated contact graph is high), but that is not necessarily the case. It may also be that agent i serves as the intermediary between two larger 'bubbles' or temporal communities, which will become disjointed if agent i is removed from the system. Example values of $C(3, t, \Delta t)$ are shown in Fig. 8.1(c) for the case of agent 3.

8.2.2 Other agent-based topological network metrics

It is important to assess whether the contact sequence centrality C merely reflects static properties of the network topology that could have been found more easily using traditional methods, or whether it provides us with new information regarding an agent's role in the system. This prompts us to compare C values to various other agent-based metrics that have been around in the literature. Some of them are derived from the contact network like the one shown Fig. 8.1(a) on the left, obtained by aggregating all the temporal network snapshots into one single, unweighted, undirected contact network within a certain (or the whole) time interval. The *degree* of an agent i equals the number of links this agents has, i.e. the number of unique agents i had been interacting with. Similarly, the *contact betweenness* is the betweenness centrality in the aggregated contact network; betweenness centrality is defined as the fraction of all shortest paths (i.e., between all unique pairs of agents in the aggregated network) that passes through an agent. As for event maps of temporal networks, the *number of events* at any time snapshot is the num-

ber of events of sizes larger than 1 (i.e., ‘non-individual events’) the agent attends. The *average event size* determines the average number of attendants of events in general.

Finally, for temporal networks, many non-static quantities have been developed, as variants of their static counterparts, in order to study the temporal structural properties of the system (Hafiene et al., 2020; Masuda and Lambiotte, 2020). While they have been developed for degree and closeness as well (Kim and Anderson, 2012), we choose to compare $\Delta\bar{P}$ to two types of betweenness other than the aforementioned contact betweenness. To explain the difference, we refer to the graphs shown in Fig. 8.1(a). On the left, we see the contact graph, in which the contact betweenness (already mentioned above) is defined. On the right, we see an event graph, where the nodes are events. The betweenness — i.e. fraction of shortest paths going through them — of nodes in this graph is computed, and when taking the average of all events where any agent i is participating, we obtain the value of what we refer to as the *event betweenness* of this agent. Likewise, we can focus on the network in the middle: the temporal graph. Nodes here are agents at a specific time snapshot. The betweenness in this graph is computed, and then we again average over all nodes in this graph belonging to each agent individually, which results in what we refer to as the *temporal betweenness* of an agent. The event betweenness and temporal betweenness are related, but not necessarily the same, as finer network structures within events are not incorporated in the event betweenness (while they are in temporal betweenness), and the event betweenness calculates shortest paths between pairs of events rather than pairs of agents.

Because of computational limitations, we choose to calculate temporal betweenness and event betweenness in time intervals smaller than the full datasets (i.e., we take ten subsequent 4-min time intervals). This results in multiple values per agent (each belonging to one of these intervals), of which the average is reported in later usage of these metrics.

8.3 Application to human interaction data

In this section, we perform event mapping and contact sequence calculations on human interaction data measured at an art fair in Amsterdam. After describing the experiment details in section 8.3.1, we discuss the resulting \bar{P} and C values in section 8.3.2, and compare the latter to other metrics in section 8.3.3.

8.3.1 Experiment and sensor data

In August 2020, after the first peak of COVID-19 infections was decaying in the Netherlands, right before the start of the second wave of infections, an art fair was organised in Amsterdam with the goal of assessing the effectiveness of physical distancing interventions. For an illustration of the art fair, see Fig. 8.2. During this three-day art fair, different interventions to promote physical distancing were implemented: walking directions, face masks, and a buzzer-notification (a buzzing sound when coming within 1.5m of another visitor). The three day art fair was split into eleven time slots and the aforementioned interventions varied across these time slots. Each visitor was asked to wear a sensor that recorded the distance to the sensors of other visitors within line of

sight (max 30m) using ultra wide band technology. Whenever two sensors were within 1.5m of one another the opposing tag ID was registered and the contact was logged locally (at a 1-second resolution), which was sent to a database via access points, placed near the entrance of the art fair. Researchers from the University of Amsterdam collected the questionnaire and sensor data and have published this in previous work (Tanis et al., 2021; Blanken et al., 2020). The ethics review board of the University of Amsterdam (2020-CP-12488) approved data collection, and all participants provided informed consent before participating. All personally identifiable information used to link the questionnaire and sensor data has been destroyed. All methods were performed in accordance with the relevant guidelines and regulations.

From time to time, in most experiments, the interaction registration was hampered by not synchronising to the access points well, altering some of the time stamps. Because the time stamps are crucial to the analysis in this chapter, we chose to limit the analysis to parts of the experiments that did not contain such gaps (see SI H for a supplementary figure on choosing the correct parts of the data).

Table 8.1 shows an overview of the experiments, their (filtered) duration, conditions and the resulting number of agents. More details can be found in Tanis et al. (2021).

Exp	Duration [min]	Walking direction	Supplementary intervention	SDS setting	# Agents	Avg. degree	Avg. degree per agent
Exp 2	19 min	Bidirectional	Facemask	No feedback	98	10.2	0.10
Exp 3	25 min	Bidirectional	None	No feedback	95	5.1	0.05
Exp 4	50 min	Bidirectional	Buzzer	Buzzer after 2 sec	88	18.1	0.21
Exp 6	55 min	Unidirectional	None	No feedback	130	19.6	0.15
Exp 7	38 min	Unidirectional	Buzzer	Buzzer immediately stops after 2 sec	88	7.9	0.09
Exp 9	85 min	No direction	Buzzer	Buzzer immediately stops after 2 sec	109	17.3	0.16
Exp 10	42 min	No direction	Buzzer	Buzzer immediately persists after 2 sec	120	11.7	0.10
Exp 11	44 min	No direction	None	No feedback	89	17.4	0.20

Table 8.1: Details of the conditions and attributes of the experiments. In the second column ('Duration') we note the filtered window durations, rather than the full experiment times. Event counts and link frequencies are determined using only events with sizes larger than 1. Number of agents may differ from that in Tanis et al. (2021) because we focus on only a part of each experiment (see Fig. H.1). The average degree and the average degree per agent have been calculated from the aggregated contact network, aggregated over the full valid duration of the experiments.

8.3.2 Evolution of \bar{P} and C

In Fig. 8.3, in order to illustrate differences among the experiments, we show the time evolution of the system property \bar{P} and the contact sequence centrality C (for specified agents) in three of these experiments: 7, 9 and 11. The conditions behind these experiments are quite different as shown in Tab. 8.1: experiment 11 contained no interventions at all, while direction and buzzer feedback interventions are differently applied in experiments 7 and 9. Importantly, we use the experiments to highlight several aspects of \bar{P} and C .

First, in the top row, we show the evolution of $\bar{P}(t, \Delta t)$ — see definition in Eq. (8.3) — as a function of (the starting time) t . A parameter that needs to be chosen is Δt , modu-



Figure 8.2: The art fair on Amsterdam on the first day with bidirectional walking directions (top) and schematic layout of the location (bottom). The art fair consisted of 28 stands, spanning $1,080\text{m}^2$. Visitors entered on the left of the plan in the bottom panel, where also the access point was placed. Figure adapted and reprinted with permission from Tanis et al. (2021).

lating the time interval across which the tracing between agents is done (see Fig. 8.1(b) for an example). This choice requires special attention, because while some agents may have most impact in the short run, others may have more impact on the long run. In other words: the individual differences in impact on contact sequences, the main topic of this chapter, are dependent on the ‘time scale’ Δt we are interested in (we will come back to this later). In Fig. 8.3(a)-(c), we choose $\Delta t = 300$ and 600 seconds. For example, the black curve represents the average fraction of unique agents in the time interval $(t, t + 300$ seconds) in the contact sequence of any agent at time t . High values at time t (may) indicate that large and frequent events are taking place in the interval $(t, t + 300$ sec), involving many unique mingling people, as opposed to mingling only within specific bubbles or clusters. Clearly, the grey curve ($\Delta t = 600$) is always above the black curve ($\Delta t = 300$), which is expected, as over a longer time span \bar{P} has more time to grow. Visual inspection of the panels (a)-(c) immediately reveals a number of differences in the evolution

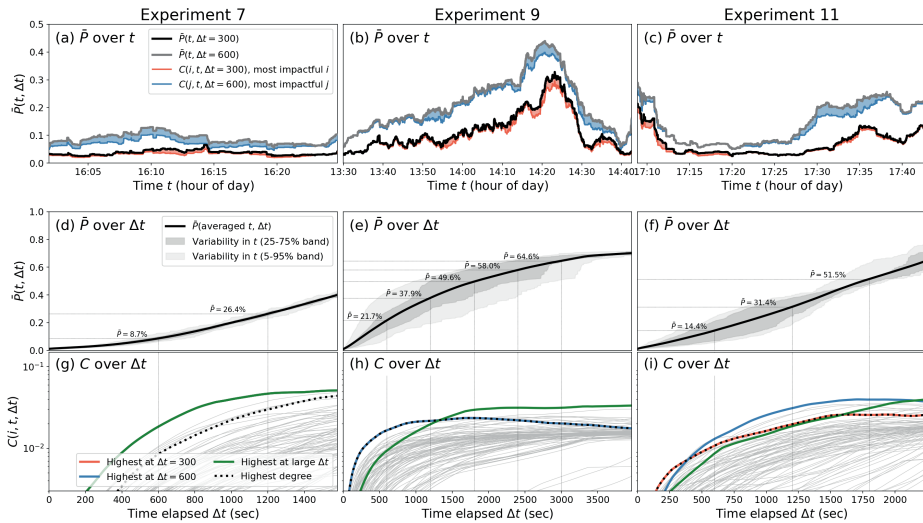


Figure 8.3: Evolutions of \bar{P} and C of experiments 7 (left column), 9 (middle column) and 11 (right column). **Panels (a)-(c):** Evolution of $\bar{P}(t, \Delta t)$ over time t at $\Delta t = 300$ seconds (black/red) and $\Delta t = 600$ seconds (grey/blue). We also show $\bar{P}_{-i}(t, \Delta t)$ (red and blue curve, respectively) and $C(i, t, \Delta t)$ (red and blue area) for the most impactful agents i at these levels of Δt . **Panels (d)-(f):** Evolution of $\bar{P}(t, \Delta t)$ over elapsed time Δt , averaging over t , for the same experiments 7, 9 and 11. For a few values of Δt , we indicate the value of $\bar{P}(t, \Delta t)$. **Panels (g)-(i):** Evolution of $C(i, t, \Delta t)$ for each agent i (in grey). Specific agents are denoted in colours: the agents with the highest C at $\Delta t = 300$ (red) and 600 seconds (blue) as in panels (a)-(c), the agent with the highest C at a near-maximum Δt (i.e., 90% of the maximum time elapsed) in green, and the agent with the highest static contact degree as a black dashed line. Note that in panel (g), the red, blue and green lines overlap (only green shows), and in panel (h), the red and blue overlap (only blue shows).

of $\bar{P}(t, \Delta t)$ among the different experimental settings: while experiment 7 remains rather flat at low values, experiment 9 contains a build-up towards much higher values (up to 0.30 for $\Delta t = 300$ sec), and experiment 11 lies in between the two and seems to be split in alternating phases of highs and lows. Without going into the specifics of these experiments, the shapes of the \bar{P} curves can be interpreted as follows. A constant, flat curve (like in experiment 7) indicates that the interaction configuration is such that the potential of any spreading phenomenon does not vary much in time: if a disease or rumour would start at the start of the time interval, or in the middle, it would spread with roughly equal speed. However, a lot of variation in \bar{P} (like in experiment 9) indicates that the structure of interactions and the temporal graph itself is different when starting at different time points. Sometimes, it promotes tracing many agents through contact sequences (e.g., at 14:20, we have that $\bar{P} = 40\%$ in 10 minutes) by means of very frequent events with randomised attendees, which may imply that the system is vulnerable to spreading dynamics like disease transmissions and rumour spreading. In other cases, the temporal network structure only contains narrow contact sequences, tracing only few agents to each other (e.g., at 13:30, we have that $\bar{P} = 5\%$ in 10 minutes), which is the case in a well-segregated agent population.

In Fig. 8.3(a)-(c), to mark the effect of individual agents in the system, we also plot

C after fictitiously removing the ‘most impactful agent’ at both levels of Δt plotted here: 300 and 600 seconds, in red and blue, respectively. Note that there are different ways to define the ‘most impactful agent’, and the correct choice is depending on the application. In particular, the ranking of agents in terms of their impact on the contact sequences is in reality dependent on both t and Δt , i.e., who is ‘most impactful agent’ varies over time. For illustration purposes we choose to average all C values over t and look at the agent with the highest value of C at fixed, chosen values of Δt (in other panels we use the same procedure). In other words, our definition of ‘most impactful agent’ is the agent that at a chosen value of Δt , has the highest value of C , averaged over the possible time starting points t . In the panels (a)-(c), the reduction in \bar{P} (marked by the red and blue areas) if we would remove these high-impact agents remains relatively invariant over time t in experiment 7, while in experiments 9 and 11, there are clear moments in time where the most impactful agents seem to have their highest impact: for experiment 9, this is near the peak, and for experiment 11, most impact is found either at the start, or around 17:27-17:37. The agents denoted in red and blue are not necessarily the same agent (see below).

While in the upper row, we see how \bar{P} evolves over time, indicating where contact sequences promoted any spreading dynamics most, we now look at how \bar{P} grows with Δt in panels (d)-(f), which reflects how many people you can trace via contact sequences within a given time interval size Δt . These are monotonically increasing curves: taking a larger time window allows the system reach equal or more agents, but never fewer. Cross-experimental differences can be observed in the slope at which \bar{P} is increasing with Δt : at $\Delta t = 10$ minutes, while \bar{P} only reaches 8.7% in experiment 7, it reaches 14.4% in experiment 11, and even 21.7% in experiment 9. These differences based in the topology of the respective temporal networks, but may have implications on potential dynamics on top. For example, the fact that the contact sequences in experiment 9 reach many more agents in the same time interval as experiment 7, implies that spreading phenomena like disease transmission or rumour spreading may be more fast-paced in the setting of experiment 9. There are several considerations to be made when concluding this, as discussed in section 8.4. While the black lines show averages over t , the increase of \bar{P} with Δt varies in time — sometimes this growth is faster than in other moments — which is shown in the grey shaded areas. Note that the variations around the black curve are much larger in experiments 9 and 11 than in experiment 7, also seen in panels (a)-(c).

8.3.3 Relating C to existing network metrics

The bottom panels (g)-(i) in Fig. 8.3 show $C(i, t, \Delta t)$, which is the effect of the removal of agent i on the black curves shown in panels (d)-(f). The curves of all agents i are shown in gray, but we focus on four aspects: the agent with the highest C for $\Delta t = 300$ in red, the same for $\Delta t = 600$ in blue, the same for $\Delta t = 90\%$ of the total time in green, and the agent with the highest degree in the aggregated contact network in the interval $(t, t + \Delta t)$ in black dashed lines. Note that we scanned C based on a few chosen Δt values to highlight particular agents (in red, blue and green), but then plotted these agents over the whole spectrum of Δt values (on the horizontal axis) — ‘their’ Δt value is merely to label them and does not play any further role. We henceforth refer to these agents as the ‘red’, ‘blue’ and ‘green’ agents. The reason for plotting the highest-degree agent (black dashed)

is to assess whether having a high degree in the aggregated contact network is be related to the most impactful agents and the various time scales Δt . The answer clearly depends on the experiments: in experiment 7, there is a single curve clearly above all others on all time scales Δt , which means that the red, blue and green agents overlap. With a clear separation to the second and third places, the highest degree agent is not even close to having most impact on larger time scales (note that the vertical scale is logarithmic, which creates a natural convergence of high-impact curves to the right), albeit that this agent remains in the top 10% of agents with highest C values. In experiment 9, the place for the highest C switches between agents. Up to values of Δt of about 1300 seconds, a particular agent dominates C (overlapping blue and red), which also happens to be the agent with the highest degree. However, at $\Delta t = 1300$ sec, the green agents takes over the dominance in C , who had a modest impact before that. Such separation reveals the dependency on time scale when assessing the impact of a specific agent on contact sequences. In particular, it confirms the hypothesis that aggregated properties do contain most information for contact sequence centrality at small values of Δt , while they lose their relevance at larger values of Δt . This is highly pronounced in experiment 11. At short time scales, the red curve marks the highest- $\Delta \bar{P}_i$ individual, which also happens to be the highest-degree agent. Around $\Delta t \approx 400$ sec, this individual is surpassed by another (marked in blue), which dominates the spectrum of $\Delta \bar{P}_i$ at most time scales, but is caught up by yet another (in green) at the largest time scale, where the red/highest-degree agent is clearly having a lower impact than many other agents.

Further, Fig. 8.4(a)-(f) show the relationships between $C(i, t, \Delta t)$ at a fixed value of Δt (averaged over t) — denoted by $C(i, t_{av}, \Delta t)$ — and six existing network metrics. On the one hand, in panels (a)-(c) we show three such scatterplots: $C(i, t_{av}, \Delta t)$ versus the degree of the agent i in the aggregated network, obtained from the aggregated contact network; the number of (non-individual) events the agent participates in; and the average size of the events (including those including only a single disconnected agent) in the time interval $(t, t + \Delta t)$. Being obtained from the aggregated network, these are all ‘static’ network properties: they do not provide information on the sequences of interactions. For example, the degree is obtained from the static contact network and frequency or sequences are not taken into account. Likewise, the number of non-individual events an agent has been participating in does not necessarily reflect a wide range of agents — all participated events may be containing only the same people, and the average event size reflects how often the people participates in individual events (i.e., isolated from the rest of the agents). All three metrics (especially the [fully aggregated] degree) correlate relatively well with C for smaller values of Δt , across all three experiments (in colours). Still, there is quite some spread: agents with the highest C are not the ones with the highest network metrics. Also, while experiment 7 has generally smaller degree, the scatter in the C values seems to be equal to the other two experiments. On the other hand, panels (d)-(f) describe three forms of betweenness: event, contact and temporal betweenness (see section 8.2.2). A clean relation between C and temporal or event betweenness is barely visible.

Also relevant to the art fair experiment are metadata obtained from questionnaires (for details, see Tanis et al. (2021)). We aggregated questions related to the attitudes of the individuals towards the interventions and dangers of the COVID-19 pandemic that was present during the time of the experiment (August 2020), involving questions like

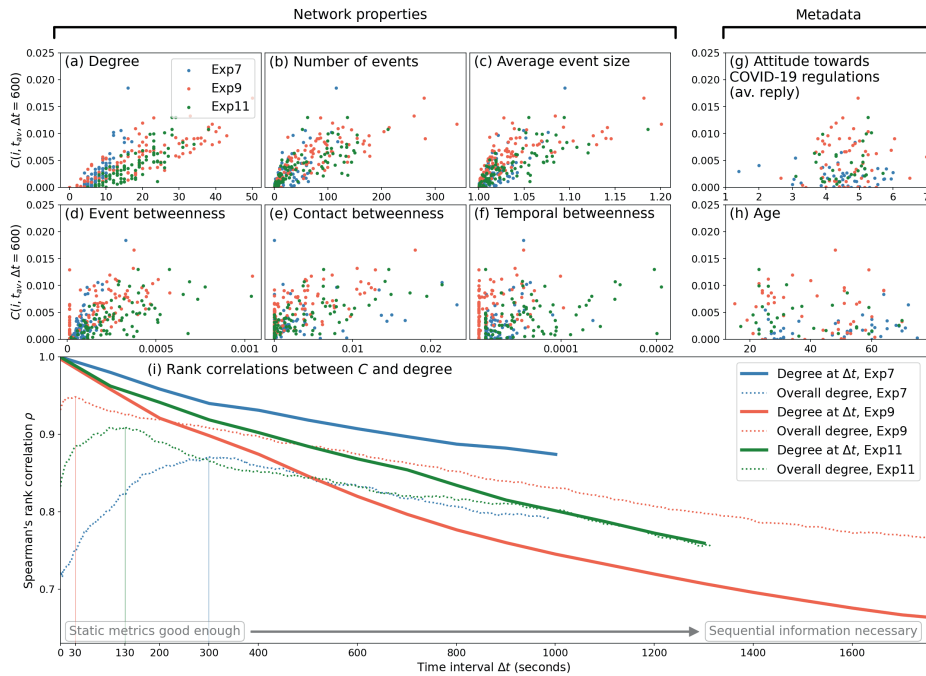


Figure 8.4: Panels (a)-(f): Relation between agent's impact on C and various network properties. Different colours denote different experiments. Panels (g)-(h): Same as in panels (a)-(f), but for two metadata metrics: the inferred attitude of the agent with respect to COVID-19 measures based on questionnaire data, and the agent's age. Panel (i): Spearman's rank correlation coefficients between contact degree and C_i for different values of Δt for each of the three experiments. Thick lines indicate the correlation coefficients when using the average degrees of the agents when looking at interval sizes Δt (as marked by the horizontal axis), while the dotted lines use the overall degree, independent of Δt . Per experiment, Δt values of up to half of the experiment's total time interval are used, to keep enough initialisation times t to average over.

"Do you adhere to the 1.5 meter distance rule as issued by the Dutch government?". Participants' reply on a scale from 1 to 7, where a lower score indicates more scepticism on the pandemic and less strict following of the Government imposed COVID-19 rules, while a higher score means the opposite. The average over 22 of such questions is shown at the horizontal axis in panel (g). Clearly, there are no significant differences between the experiments, and no relation is visible with C . In panel (h), we also plot the age of these participants, also yielding no significant relationship with the impact on spreading capability. Not all agents filled in the questionnaires however, resulting in less dots in the metadata panels than in panels (a)-(f).

Putting together Figs. 8.3(g)-(i) and Fig. 8.4(a)-(f), several relations between network metrics and C can be inferred, but any such relation would be subject to the time scale Δt : for small values of Δt , the static degree may be well suited to describe an agent's impact: having more unique direct contacts means that the agent, potentially, have been in contact with many other agents, which directly affects \bar{P} . In turn, therefore, removing that agent would result in a higher C . At higher time scales, however, the effect of se-

quences becomes more important. Agents with (potentially) smaller initial impact and contact degree may have a larger impact on larger time scales because they serve as conduits of contact for large bubbles or other more subtle temporal structures. This is visible in Fig. 8.3(g)-(i). To explicitly test the effect of time scale on whether traditional network metrics are sufficient, we show the Spearman's rank correlation coefficient between the contact degree and C as a function of Δt in Fig. 8.4(i). The overall degree (independent on Δt is also correlated to C , shown in dotted lines). A clear decay of this correlation with Δt is visible, reflecting that for small Δt , static metrics like the contact degree are sufficient, but for larger time scales, the sequential information becomes important, which motivates a more complex metric like contact sequence centrality C .

8.4 Discussion and Conclusion

In this chapter, we have developed and demonstrated a methodology to quantify agent impact on contact sequences in social interactions from the perspective of spreading phenomena. Spreading phenomena are versatile in their nature and types; however, one common element they share is that they all require individual agents as carriers of the entity, and an agent in possession of the concerned entity can pass it on to others that do not have it, leading to the natural expectation that the temporal topology of agent interactions — who interacts with whom, when and in which sequence — will have a profound influence on the spreading dynamics.

In order to quantify this effect, we have proposed a new measure for individual differences in their impact on contact sequences. The measure is obtained by porting all contact information — while preserving the sequences of contacts — into an event graph, and tagging the connections between events by agents' temporal paths. From the event graph, we determine the metric $\bar{P}(t, \Delta t)$, which is the average fraction of agents at time $(t + \Delta t)$ traceable from any agent at time t . Likewise, contact sequence centrality $C(i, t, \Delta t)$ represents an agent i 's impact on the contact sequence in terms of how strongly \bar{P} changes when all temporal links belonging to agent i are removed.

For event mapping, we have used the convention that at every (integer) snapshot the agents' contact network within an event is a complete one. For example, for the agents' contact network snapshot at time t , agents 1-and-2 and 2-and-3 may have been recorded to directly interact, but not 1-and-3: in such a case event mapping introduces a direct interaction between agents 1-and-3 that are not originally present in the empirical data. While this is a subtle issue discussed elsewhere in detail (Dekker et al., 2021e), we note the value of C is dependent on the choices made in this convention. This is not necessarily a drawback for the concept of contact sequence centrality that we have developed here, but may limit its applicability depending on the time-scale of transmission of a quantity of interest. For example, imagine that the time-scale for transmission is very short relative to the typical duration of an event. Then, in the above example event involving three agents 1, 2 and 3, it will not matter whether 1 and 3 interact directly or not: agent 1 will effectively be in direct contact with agent 3.

Note also that we have calculated both \bar{P} and \bar{P}_{-i} forward in time. This is an obvious choice for spreading dynamics, but for other applications, calculating these quantities backward in time would be more appropriate. For example, when interested in the vul-

nerability of a specific vendor in a supply chain, the more time-backward dependence the vendor has, the more it will be vulnerable to random failure. In other words, the choice of the direction of time for calculating \bar{P} and \bar{P}_{-i} depends on the application. Note also that the removal of a specific agent in our calculations is merely a conceptual construct to assess its impact on the contact sequence — the actual removal of this agent may have unforeseen effects on the topology of the temporal network that cannot be captured in a conceptual framework.

Contact sequence centrality is a purely topological property, reflecting properties of connections. In other words, agents with high levels of C can be treated as what are colloquially referred to as (behavioural) *super-spreaders* (Zhang et al., 2019; Kitsak et al., 2010; Ahajjam and Badir, 2018), being more central to the contact sequences. Our work serves to highlight the fact that super-spreaders are not necessarily those individuals that have most contacts: as shown in Fig. 8.3(g)-(i), we see that the highest-degree agent starts losing its dominance in C when looking at larger time scales. In particular, increasing the time interval Δt makes the degree less and less determinant for the C values, as shown in Fig. 8.4(i). What sets our metric apart from aggregated (and static) and various temporal metrics is that the exact sequence of the interactions matter: in the example of Fig. 8.1(c), removing agent 3 removes an edge early on in the time series, which has cumulative effects later on. Moreover, reversing the direction of time would yield the same values of all static metrics like degree, and also for temporal betweenness for the aggregated networks, even though it would fundamentally change the values C .

Finally, it is important to realise that topology of the contact sequences alone is not enough for the behaviour-epidemiology interface, or for that matter for any dynamical process playing out on top of the temporal network; the process itself may have its own inherent time-scale parameters. Let us elaborate a bit on this point using an example like pathogen spreading, which comes with its own time-scales, such as incubation period, or the time required for an infected agent can itself be infectious. On a given school day, all students in a school may be sequentially traceable to each other as measured by our metrics: $P_i = 1$ for all students i , and $\Delta t = 1$ day. However, the disease may only spread to students directly linked to a patient zero i_0 , simply because secondary contacts do not matter because of the incubation time of COVID-19: the direct contacts were not infectious yet. Given this, it is natural to expect that the eventual dynamics of pathogen spreading will be a combination all time-scales involved, including event frequency in the temporal network.

Indeed, to make the current approach relevant for epidemiology, the observation period of contacts among the agents needs to be sufficiently longer than the incubation period, and the time-lag between being infected and being infectious, bringing us to an issue of practicality. While the event mapping is a relatively quick process to compute, depending on the number of time steps T , the calculation of \bar{P} scales with an extra factor N because the contact sequence tracking [e.g., as in Fig. 8.1(b)] has to be done for every agent. However, when computing the C for every agent i , yet another factor N is brought in, as all calculations have to be done fully over when removing i . Hence, the computational complexity of C is of the order $\mathcal{O}(TN^2)$. This means, given that the contact networks of the agents are sampled at 1-second interval, that for making the current method applicable for COVID-19 epidemiology (with an incubation time of about 5.2 days (Li et al., 2020b) and infectiousness starting time of 12.3 days (He et al., 2020)), fur-


ther course-graining methods to preprocess the network in the the temporal domain will need to be developed.

In this chapter, we have analysed an application inspired by the COVID-19 pandemic, as the structure and dynamics of contact networks have proved pivotal to forge a conceptual link between human behaviour and virus spread. However, our understanding of human behaviour is still limited by a lack of adequate measurement techniques and modelling frameworks. In our view, therefore, further development of technology and methodology to obtain and analyse direct measures of behavioural contact networks is essential to advance our understanding of human behaviour, and to improve the resilience of society in dealing with pandemics.

CHAPTER 9

EVALUATION OF INTERVENTIONS DURING THE COVID-19 FIRST WAVE IN THE NETHERLANDS

This chapter is based on: Dekker, M. M., Coffeng, L., Pijpers, F., Panja, D., and Vlas, S. (2021b). Evaluation of non-pharmaceutical interventions during the COVID-19 first wave in the Netherlands. *Manuscript in preparation*. Supplementary material to this chapter can be found in SI I.



ABSTRACT

Worldwide, the intervention measures to curb the initial spread of SARS-CoV-2 has consisted of a combination of lockdowns, mobility restrictions, social distancing and mask-use. Unravelling their efficacies while simultaneously keeping in perspective their social costs, however, has remained difficult, given that such non-pharmaceutical interventions require high resolution models regarding demography and mobility at a sub-national scale. Here, we achieve such resolution using a country-scale demographically-stratified population and a spatially-explicit individualbased epidemiology, calibrated to hospital admissions data and mobility trends extracted from mobile phone signals and Google. Coupling these datasets allows us to evaluate the intervention measures taken, focusing on the first COVID-19 wave in the Netherlands. We compare a set of national interventions, including travel reductions, behavioural measures such as social distancing and school closure, showing that at least the first two had an important effect on the epidemic spread. Analogously, we test the scenario in which interventions would have been implemented on a local (municipality) level, providing a quantification of the trade-off between case numbers and disruptions due to interventions.



9.1 Introduction

Reduction in mobility, enhanced levels of personal hygiene, social distancing and various combinations thereof — including complete lockdown in case of ‘extreme’ circumstances — are common intervention measures deployed to curb the spread of pathogens at times of epidemics. Every form of intervention measure comes however at the cost of social and economic disruption. For controlling the spread of an epidemic, it is therefore of paramount importance to implement those intervention measures that are not only effective but are also socially minimally disruptive. Well-tuned epidemiology, capable of testing a variety scenarios in terms of measures, timing (Kretzschmar et al., 2020) and targeted sub-national regions, is the standard route for informing the policy base that intervention measures typically rely on.

Precisely herein lies a challenge. The nature of the population and human behaviour is key to epidemiological dynamics, for instance regarding the age-stratification and the demographic composition of a society, how people travel, interact and mingle across different age groups, environmental conditions, demographics and sub-national regions (Lloyd-Smith et al., 2005; Prem et al., 2017; Mossong et al., 2008). Moreover, even though the social perception of any form of intervention measure as a ‘disruption’ from business-as-usual is universal, human behaviour and response to interventions varies strongly across societies and cultures, and is highly heterogeneous even within a population living in a certain geographical region (Marks et al., 2021). It is for this reason that epidemiology meant to inform policy requires local embedding, i.e., incorporation of demographic, mobility and behavioural aspects specific to the concerned population (Vos et al., 2021). When a novel pathogen breaks out, the challenge gets further compounded due to uncertainties in its attributes such as the mode of transmission or the duration of its incubation period (Elias et al., 2021). Building epidemiology taking all this into account is however not only highly complex, but also requires detailed data at high resolutions.

Accounting for the aforementioned heterogeneity and local embedding bridges epidemiology to the study of complex networks. Chapters 7 and 8 in this dissertation analyse human interaction systems to obtain insights in system properties associated with epidemic spreading. The chapters do this from a network point-of-view, contributing to a vast amount of literature concerning a system’s vulnerability to spreading, the concept of superspreading (Lloyd-Smith et al., 2005; Endo et al., 2020) and effective interventions (Brauner et al., 2021; Sharma et al., 2021; Flaxman et al., 2020; Mistry et al., 2021). When using human interaction networks in the context of epidemiology, depending on the scale and scope of the analysis, and the resolution of available data, an important consideration concerns the temporal variation of the network structure. To fully resolve epidemic spreading, high resolution data on temporal variations of the contact network is arguably required, due to the importance of sequences in spreading processes (as discussed in chapters 7 and 8): a person can only infect a person once he or she has been infected before, not *vice versa*. A practical problem, however, concerns the size of the data, computation power to track infectious sequences, and the near-impossibility of tracking contacts of population sizes at the level of countries. Hence, to work on larger scales, many epidemic models are not individual based, but rather assume full or partial (i.e., through stratification) homogeneity in the population interactions (e.g., see introduction of part I). At smaller population sizes, both static network approaches (Li et al.,

2020a; Isella et al., 2011; Riolo et al., 2001; Masuda and Holme, 2013; Stehlé et al., 2011; Zhang et al., 2021), where the contact networks are aggregated or assumed constant, and temporal network approaches (Li et al., 2017, 2020a; Mucha et al., 2010; Génois and Barrat, 2018; Centola, 2010; Schläpfer et al., 2014; Kiti et al., 2019; Cattuto et al., 2010) aim to characterise systems to obtain useful insights for the larger system, or focus on outbreaks in (small) systems assumed to be closed, for example in the context of the spreading of sexually transmitted diseases in small communities (Liljeros et al., 2003; Helleringer and Kohler, 2007). This chapter relates to this literature by accounting for the temporal and geographic nature of contacts on an individual basis, but at the same time avoiding high-resolution contact data requirements. We do this by working on a semi-aggregated spatial resolution (municipalities) and calculating transmission probabilities via a (demographically and geographically) stratified force-of-infection instead of individual contacts.

This chapter focuses on the effectiveness of implemented interventions. By and large, intervention measures deployed during the first COVID-19 wave in spring 2020 were not enough to spatially constrain the disease: the worldwide spread of SARS-CoV-2 along the pathways of globalised travel was too fast to allow for it to happen. Reliable data (specifically, near-real time data needed for policy-informing epidemiology) on community-transmission were not readily available to researchers during most part of the first wave. Neither were there sufficient number of test kits to track infections. In such a situation, for setting intervention policies, large parts of the world used epidemiological insights that were emerging from China, where the outbreak came a couple of months earlier, even though it meant that local contexts specific to regions in China were being used elsewhere. By the time test kits and reliable data started to become available as the first wave wore off, the pandemic had already taken firm roots worldwide. In other words, in real-time, it was difficult to assess the effectiveness of intervention measures, and questions regarding the necessity of targeting full countries rather than sub-national regions remained unchecked.

The aim of this chapter is to make these assessments in hindsight. We evaluate the intervention measures taken during the first COVID-19 wave in the Netherlands using a large individual-based model, which accounts for heterogeneities in demography, mobility and behaviour and is validated by a combination of multiple data sources (on demography, mixing, hospitalisation, mobility and serology). Because of the large combination of data sources used and model components, we move many of the methodological details to SI I.1. Hence, section 9.2 comprises the main steps and dynamical aspects of the model. The results of the model are threefold. First, in section 9.3, we calibrate the model to the first wave and distinguish how different demographic and geographic entities played a role in the first wave. In the second part of the results (section 9.4), we experiment with removing some of the national interventions that were issued by the Dutch government, which reflect their effectiveness in counteracting the spreading. In the final part of this chapter, we discuss the effectiveness of regional interventions (at the level of municipalities), addressing whether national interventions were really necessary.

9.2 Methodology

To start with, the first COVID-19 wave in Netherlands covered the period February 27, 2020 (the first tested case of COVID-19 in the Netherlands) till June 1, 2020 (lifting of most intervention measures). In this timeline, there are four distinguishable periods in terms of the national interventions, which we refer to as *phases*: (i) Phase 1 (February 27 - March 11) when transmission of the pathogen progressed unchecked, (ii) Phase 2 (March 12 - March 22) with minor interventions involving a working-from-home policy, cancellation of large events and some social distancing and face mask advice, (iii) Phase 3 (March 23 - May 11) involving a lockdown with closed schools, event centres, mandated social distancing and working-from-home policies, and (iv) Phase 4 (May 11 - May 31) involving a gradual lifting of all measures.

The model steps are subdivided into eight steps, illustrated in Fig. 9.1. We divide our (simulation) analysis into two parts: (i) demography, mobility and mixing considerations, and (ii) epidemiology and first-wave phenomenology, both consisting of four steps (1-4 and 5-8 respectively in Fig. 9.1). The key steps of the dynamics are summarised below, supplemented by methodology related to the mobility, transmission and interventions later in this section. SI I.1 contains additional details on the methods.

9.2.1 Model steps

Step 1, demographic categories. We resolve the Dutch population into 11 demographic categories. Similarly, we resolve the Netherlands geographically into its 380 official municipalities. From the registry data made available by Statistics Netherlands (CBS, the national statistics agency), we obtain the composition of each municipality in terms of these demographic groups as per their residency. We then distribute 170,721 *agents* — each representing approximately 100 Dutch residents since the population size of the Netherlands is about 17 million — with their homes located proportionately in the 380 municipalities.

Steps 2 and 3, mobility. Using (a Dirichlet distribution constructed from) hourly mobility data based on mobile phone signals, we make the agents move stochastically across municipalities at hourly time resolution, including a differentiation between regular, frequent and incidental travels. In this way, every agent on average conforms to the mobility pattern that is typical for the demographic category and the home-municipality it belongs to (e.g., work-related, day-night and weekday-weekend travel rhythms).

Step 4, mixing. Steps 2-3 also yield us the exact composition of the agents that are found in any municipality at any given hour, including details on their demographics and residency. Agents that are present in the same municipality in the same hour have the possibility to mix, i.e., to interact with each other at close proximity. In our approach, we distinguish four mixing situations: ‘home’, ‘school’, ‘work’ and ‘other’, corresponding to those distinguished in survey data that define our mixing rates, known as the POLY-MOD study (Prem et al., 2017). Every mixing situation distinguishes the involved agents’ demographic categories, the hour of day, and whether the agent, at the concerned hour, is found to be present in its home municipality.

Steps 5-6, pathogen transmission. We use SEIR dynamics for agent-to-agent pathogen transmission, which means that every agent at any time has one of the following four la-

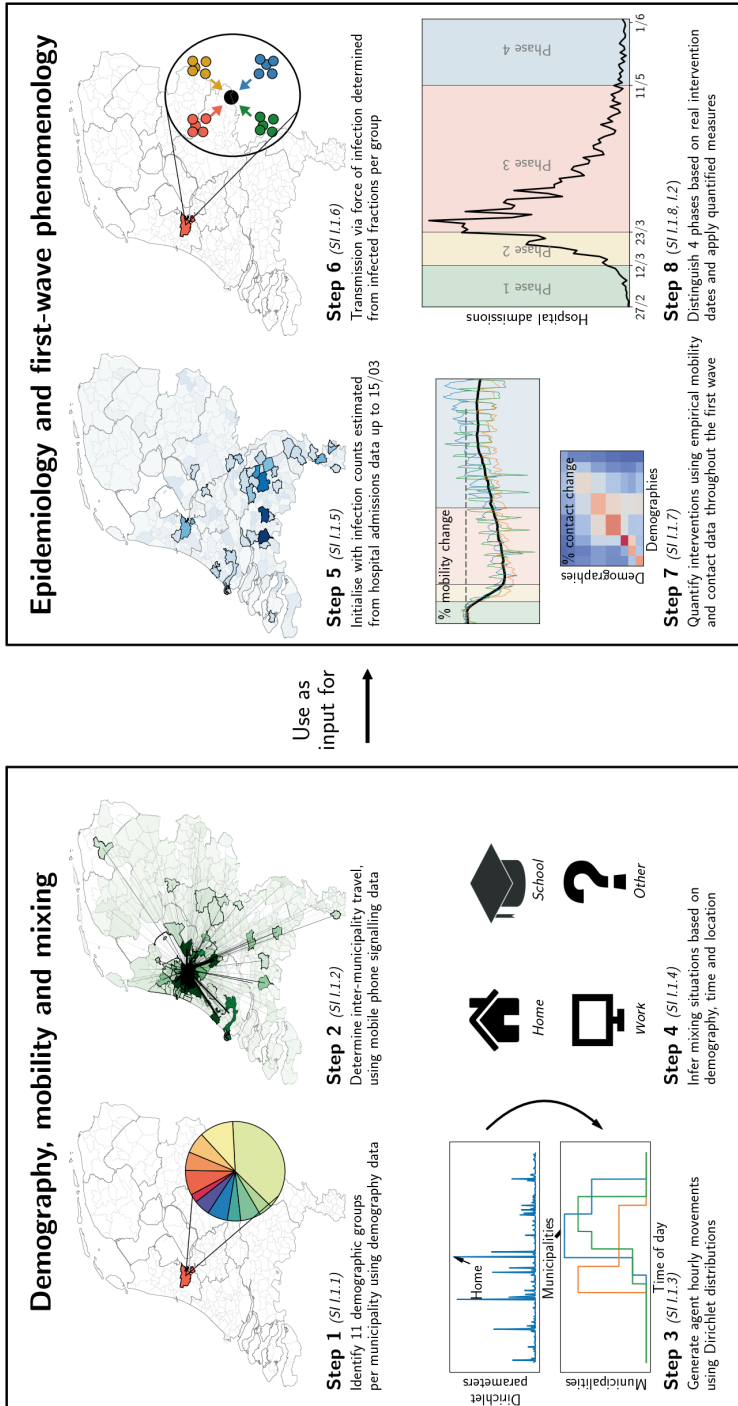


Figure 9.1: Our analysis consists of two parts: Demography, mobility and mixing (left panel), and Epidemiology and first wave phenomenology (right panel); each consisting of four steps (see text for details). See SI 1.2.1 for a detailed description of some of the data used in the figure. A number of processes in these steps are stochastic, especially steps 2, 3 and 6.

bels: susceptible (S), exposed (E), infected (I) and recovered (R). In this process, we assume a closed population during the wave, i.e., we ignore births/deaths as well as migration across the border. Every susceptible agent can be potentially exposed: a susceptible agent belonging to a demographic category g , present in a municipality m and (hourly) time t experiences a *force of infection* that is a function of the overall susceptibility of demographic category g , the phase of the first wave (phases 1 through 4, described earlier), mixing with agents present in municipality m at time t , as per step 4, and their respective infectiousness. Because testing capacity during the first wave was limited, the initial group of agents infected (I) is inferred from hospital admission data, which were mainly confined in south: the Noord-Brabant and Limburg provinces.

Steps 7-8, syncing the epidemiological analysis to national-level first wave phenomenology. Different types of intervention measures put in place in different phases of the first wave contributed to reductions in mobility, mixing and thereby, pathogen transmission. We extract the reduction in mobility from observed Google mobility data during the first wave, and apply this to the working demographics. Similarly, for epidemiological dynamics following the SEIR-model, we account for the reduction in pathogen transmission due to changes in human behaviour (such as mask-wearing, enhanced level of hygiene and social distancing) by means of varying a country-wide parameter — resulting in β_1 through β_4 , one each for each phase of the wave, calibrated to observed hospital admission data. Changes in mixing are incorporated by applying relative mixing changes as found by survey studies in the Netherlands (Backer et al., 2021).

9.2.2 Mobility patterns

The basis for the mobility patterns is mobile phone signal data gathered by the company Mezuro, resulting in numbers of daily travels by people living in any municipality i to municipality j , split into frequent, regular and incidental movements. The demographic data provided by Statistics Netherlands (CBS) allowed us to distinguish 170,721 agents with demographic details, which we link to movement data by assuming working and school-going agents have frequent movements, while the other demographics have regular or incidental movements.

The construction of the actual mobility pattern of an agent in the model is done by creating a Dirichlet distribution, using the movements data (normalised) as shape parameters, from which we draw fractions of the day spent in each municipality (i.e., resulting in 380 fractions for each of the 380 municipalities). These fractions are rounded to hours and concatenated to form a day's schedule, such that for every hour, the agent's position is specified. This is done for 7 days, after which the agent's movements are repeated.

9.2.3 Transmission mechanisms

Transmission from susceptible (S) to exposed (E) in this SEIR-based model is stochastic and based on a force of infection λ , which is used as an hourly infection probability. The intuition behind λ exerted on a susceptible agent is that each demographic category contributes to additional chance of transmission, weighted by the expected mixing between the agent and this group and the fraction infected in this group. The full equation for λ

for people from demographic group g in municipality m at time t is as follows, involving a summation over all demographic groups g' adding to the force of infection:

$$\lambda(g, m, t) = \underbrace{h(g)}_{\text{Susceptibility of } g} \cdot \underbrace{\beta(t) \cdot \bar{s}(t)}_{\text{Phase \& daily cycle}} \cdot \underbrace{\sum_{\text{Group } g'} n_{g,g'} \cdot \frac{I(g', m, t)}{N(g', m, t)}}_{\text{Mixing with groups } g'} \quad (9.1)$$

The first part of the equation involves a parameter $h(g)$ that reflects the susceptibility of agents from demographic group g to the disease (see SI I.1.5). The second part ($\beta(t) \cdot \bar{s}(t)$) contains the aforementioned phase parameter β that involves behavioural aspects like wearing face masks and keeping social distance and a daily cycle parameter $\bar{s}(t)$ (see SI I.1.6), specifying that people barely have any contacts in the middle of the night, for example. The third part involves the mixing with the eleven different demographic groups: $n_{g,g'}$ is the expected number of contacts that group g has with group g' , based on the mixing matrix that reflects the situation (i.e., 'home', 'school', 'work' or 'other'). The fraction $\frac{I(g', m, t)}{N(g', m, t)}$ is the fraction of the total number (N) of people of group g' in municipality m that are infectious (I).

Transmission from exposed (E) to infectious (I) and from infectious (I) to recovered (R) are not probabilistically implemented, but involve a time lag: an incubation and an infectious time scale, respectively. The exact duration of these lags differ per case and are drawn from Weibull distributions: $W(4.6, 20)$, and $W(5, 1)$, respectively, based on de Vlas and Coffeng (2021).

9.2.4 Interventions

The first wave of COVID-19 cases in the Netherlands happened across the period of February 27 (first tested case) to June 1, 2020. We split this period into four phases based on the interventions taking place, in which we mimic changes in mobility, mixing, behaviour and school closure. Details about these phases are shown in Tab. 9.1.

First, we address changes in mobility due to a working-from-home policy by reducing inter-municipality travel in the model using Google Mobility data (Google, 2021), which reports percentual mobility changes across this period. We implement this by placing the reported percentage of people, randomly drawn from the working categories, at home. Second, changes in mixing patterns are addressed by determining percentual changes in the mixing among different age groups from Dutch survey data (Backer et al., 2021) in the months February, April and June 2020, and applying these changes element-wise to the mixing matrices used in our model. Third, behavioural changes are represented by varying $\beta(t)$ in Eq. (9.1) across the phases. Fourth and final, schools are closed by placing school-going agents, i.e., *Primary school children*, *Secondary school children* and *Students*, and parents of primary school children (see SI I.1.7 for the derivation of these parents) at home, both in terms of location and mixing.

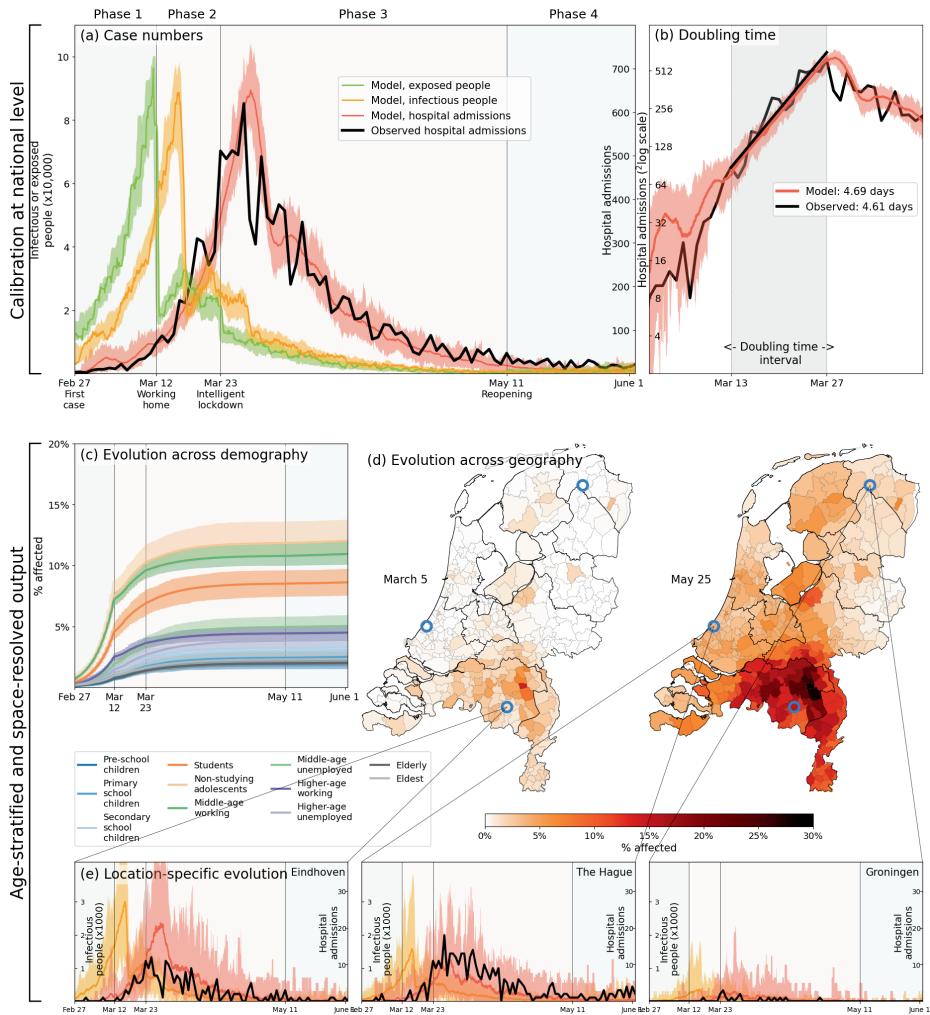


Figure 9.2: Calibration (a-b), and demography- and geography-resolved results from our analysis (c-e). **Panel (a)**, left axis: the daily number of new infections and exposures in yellow and green, respectively. Right axis: daily hospital admissions from analysis output (red) and observed data (black). Background colours and vertical black lines denote the four phases (arbitrary colouring). Uncertainty intervals mark minima and maxima in the ensemble of realisations used in the analysis; the same holds for panels (b), (c) and (e). **Panel (b)**: Hospitalisation doubling time over the period March 13 - March 27, 2020 (shaded grey shaded time domain) in analysis (red, 4.69 days) and observed data (black, 4.61 days). **Panel (c)**: % affected agents (i.e., E , I or R) per demographic group over time. **Panel (d)**: % affected agents per municipality on two days (March 5, May 25). Blue circles indicate the geographical locations of the three example municipalities shown in panel (e). **Panel (e)**: Infected agents (yellow) and hospital-admitted agents (analysis in red, and observed data in black) in three municipalities in different parts of the country: Eindhoven, The Hague and Groningen. Analysis data correspond to an ensemble of 10 independent realisations.

Phase	Start	End	Travel	Mixing	Behaviour	Schools
1	Feb 27	Mar 11	-	-	$\beta = 0.135$	Open
2	Mar 12	Mar 22	-31.7%	Reduced as per Apr 2020	$\beta = 0.11$	Closed halfway*
3	Mar 23	May 10	-42.4%	Reduced as per Apr 2020	$\beta = 0.09$	Closed*
4	May 11	Jun 1	-20.1%	Reduced as per Jun 2020	$\beta = 0.11$	Open

Table 9.1: Overview of the four phases distinguished in the first wave of COVID-19 in the Netherlands, and how they are expressed in the model. Note that the first phase does not have an explicit starting date in the model, because the model is temporally calibrated to the start of the second phase based on the moment at which a threshold value of $\theta = 1.8\%$ of the population has been affected. * Schools were closed during Mar 16 - May 10, which is also what we use in our model, resulting in a slightly asynchronous intervention from the four phases.

9.3 Reproducing the first wave

Having initialised the analysis by the cumulative infection numbers throughout the Netherlands per March 1, 2020, we simulate the epidemiological dynamics following steps 1-8 described above. The direct model output is in terms of infections, which is converted into hospital admissions following a similar procedure as used for the initialisation, because of the limited testing capacity (i.e., making it impossible to compare simulated infection numbers to accurate numbers of tested cases). We further forward-interpret the (simulated) hospitalisation data for the entire first wave. Country-wide hospitalisation data have been made available by the Nationale Intensive Care Evaluatie registratie (NICE) throughout the first wave [thick black line, Fig. 9.2(a)]. The simulated hospitalisations are clearly functions of the four β -parameters, whose numerical values we fix by matching to the NICE-data [Fig. 9.2(a)]. Specifically, β_1 is determined by matching the simulated (initial) doubling time for hospital intake to that obtained from the NICE-data [Fig. 9.2(b)]. The corresponding values of the β -parameters can be found in Tab. 9.1. Overall, this allowed us to match the hospital admission data fairly well, both in terms of absolute numbers as well as in the doubling time — with a model doubling time of 4.69 days, and an observed value of 4.61 days.

Importantly, the stratification embedded in our simulations in terms of 11 demographic groups and 380 municipalities allows us to reveal how the first wave played out demographically and geographically. We show these, averaged over 40 runs that constitute the ensemble of simulation realisations, in Figs. 9.2(c) and (d) respectively. The variations between individual runs in how the demographic categories are affected, visualised in Fig. 9.2(c), are found to be relatively small due to the relative well-mixed spread of the demographic groups across the various municipalities: precisely where the epidemic exactly is does not seem to matter. In terms of seropositivity, adolescents and middle-age working agents dominate the signal, while the youngest and oldest groups are least affected.

The limited across-runs variation however does not hold for the geographical spread. Fig. 9.2(d) shows the average in terms of geographical spread — for example, we see the relatively high prevalence of the disease in the south, which is a consequence of the fact that the epidemic started mainly in the provinces of Noord-Brabant and Limburg in the south of the country. Individual runs (see SI I.2.2), however, distinguish epidemic foci in different places of the country, resulting in a relatively wider spread in municipalities, as

shown in Fig. 9.2(e) for three municipalities across the country.

Still, from these results we conclude that — notwithstanding only four epidemiological parameters for the entire country, one for each phase of the wave — our simulations capture the observed epidemiological dynamics reasonably well. Importantly, the demographic and geographic-explicit stratification in the analysis has the important attribute that it allows us to explicitly test the epidemiological impact of specific types of intervention, or intervention measures implemented specific regions or specific demographic groups. For these, we use the ensemble of runs in Fig. 9.1, using the β -values as per Table 9.1 in the Methods section, as ‘the reference’.

9.4 Evaluation of national-level intervention measures

Like many other countries, the Dutch government implemented mainly nationwide intervention measures during the first wave rather than regional-specific ones. We focus on three categories of interventions: (i) *behavioural* measures, in the form of wearing masks, enhanced hygiene levels and social distancing, (ii) *mobility reduction*, associated with encouraging people to work from home and closing large events, and (iii) *school closure*. In order to unravel the impact of any (single) one of these on the epidemiological dynamics, we alter the simulations in the following manner, and compare the corresponding hospitalisation curves to the reference, which is shown in green in Fig. 9.3(a) (it essentially coincides with the observed data because of calibration as per Fig. 9.2).

Testing the impact of the *behavioural* intervention measures is straightforward: we simply use the same β -value in our simulations in the last three phases of the wave as in the first one, i.e., $\beta_2 = \beta_3 = \beta_4 = \beta_1 = 0.135$, while keeping everything else the same as in the reference. The results of leaving out behavioural measures are shown by the blue curves in Fig. 9.3. Checking the epidemiological impact of limiting *mobility* is somewhat more cumbersome, even though using the same β -values as in the reference ensures that limiting mobility does not interfere with the behavioural intervention measures. During the first wave, there were never any fully closed borders between sub-national regions (such as, e.g., in Italy (Berardi et al., 2020)). However, a working-from-home policy was implemented which most people adhered to, as visible in public transport usage (van der Drift et al., 2021) as well as in Google mobility data (Google, 2021). For testing the epidemiological impact of the mobility-limiting intervention measure, we simply remove the mobility decrease percentages that were found in the Google Mobility data (Tab. 9.1). The corresponding time evolution of the epidemic is shown in red curves in Fig. 9.2. As for the impact of *closing schools*, we note that the Dutch schools were closed relatively late (March 16, while the first confirmed case was on Feb 27) following a long political debate, which exact dates are currently followed in the reference simulation. To evaluate the epidemiological impact of closing schools, we let the school-going children to continue to go to school in the simulations (keeping β -values the same as in the reference). The resulting curves for this intervention measure are shown in yellow in Fig. 9.2.

We evaluate the impact of these intervention measures in terms of three indicators in Fig. 9.2. Note that none of the intervention measures applied to phase 1, such that approximately identical infection numbers are reached in each of the scenarios up to March 12, which with a lag of 14 days between infection and hospitalisation means that

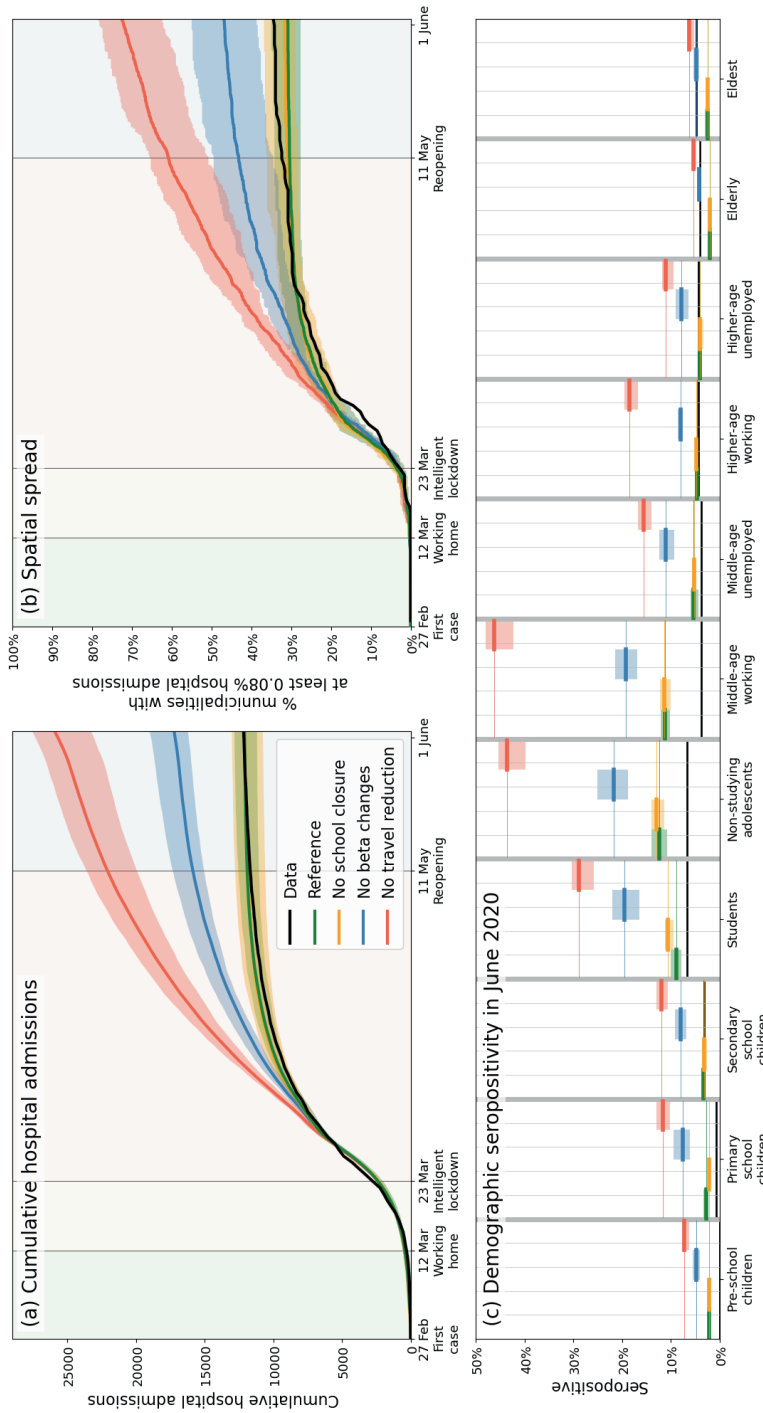


Figure 9.3: Comparison of impacts of nationally administered intervention measures. In all panels, observed data are shown in black, the reference in green, and the impacts of (i) no behavioural changes like wearing masks, enhanced hygiene and social distancing in blue, (ii) no mobility reduction in red and (iii) no closing of schools in yellow. Bandwidths indicate the minima and maxima around the mean of a simulation ensemble of 40 realisations. **Panel (a):** Cumulative national hospital admissions. **Panel (b):** Geographical spread of hospital admissions, measured by the fraction of municipalities that have at least 0.08% of the population admitted to the hospital. **Panel (c):** Comparison of disease prevalence across the eleven demographic groups (horizontal), expressed in % seropositivity (vertical).

all hospitalisation curves are well-aligned up to approximately March 27, 2020. In panel (a), we show the cumulative hospital admissions, which allows for easy comparison of the severity of each scenario. A clear hierarchy is visible: leaving out travel reductions (red) supersedes the reference (green) by more than 100%, from which we conclude that the travel reductions that we observed from the Google Mobility data prevented a doubling of the hospitalisation cases. Note that these are only calculations up to June 1. This is followed by the blue curve; behavioural changes such as wearing face masks and social distancing prevented another 40% of hospital admissions. Then, strikingly, the closure of schools seems to have no significant effect. There are a few important discussion points related to the closure of schools, which are mentioned in the discussion section.

The second indicator, displayed in panel (b), concerns the geographic spread of the disease, measured by the fraction of municipalities in which the daily hospital admission rate has reached at least 0.08% of its population, or had reached this value before (i.e., this metric never decreases). The same hierarchy is visible as in panel (a). Expectedly, omitting travel reductions results in a more than doubling of the spatial spread, reaching (with significant hospital admissions in) more than 70% of all municipalities in June, while the other scenarios are more close together. Finally, in panel (c), we show national seropositivity levels across the 11 demographic categories. Although we add observed seropositivity (in black) to this panel, those values are highly uncertain because of a variety of biases involved in the data collection (Vos et al., 2021). We add them to do a comparison of the general tendency across the demographic groups, which is higher seropositivity for adolescents, lower for older people and very low for the youngest — a tendency also found in the model output of the reference (green). Comparing the four scenarios reveals the same hierarchy as in panel (a), with a disproportionately high level of seropositivity for non-studying adolescents and middle-age working agents when travel reductions are omitted (red).

9.5 Evaluation of regional intervention measures

Next, we evaluate the potential of subnational interventions, implemented at the municipality level. The Dutch government did not implement subnational intervention measures, aside from bringing out an early advice to work from home in the south of the country, the epicenter for the first wave. The reasoning for this was as follows: once COVID-19 cases were discovered locally, likely lagged temporally by incubation and infectious periods, the pathogen would have spread throughout the entire country anyway. This is in line with observations that the Netherlands is spatially well-connected in terms of people's mobility patterns, facilitated by a robust public transport system and a high population density. Nevertheless, given that intervention measures do have a social cost for the populace, we test the epidemiological impacts of two types of subnational intervention measures. The timing and nature of municipal interventions is naturally dependent on (a) mandates from the national government, and (b) actual number of cases in the municipality. We simulate (a) by restricting the nature of municipal interventions to what has been implemented in reality on a national level at the same dates — i.e., if in our simulations, a municipality decides to issue interventions at March 29, it will implement the interventions that in reality the national government had in place

on March 29. The number of cases (I) in real-time is tracked, and the local timing and decision making to intervene is determined when the fraction of the population that is infected exceeds a certain threshold. Choosing the value of this threshold poses a trade-off: a lower threshold ensures implementation of local interventions in an early stage of the COVID-19 wave which would suppress hospital admission counts, but could unnecessarily shut down economic and social activity in some parts of the country that are less affected by the disease. Vice versa, a higher threshold would target areas where the epidemic has progressed most, but could pose the risk of starting control too late, resulting in more hospital admissions. To show the effect of different thresholds for prevalence of infectious cases, we chose a wide range of 3%, 1% and 0.33%.

The results are shown in Fig. 9.4. On the left side, the epidemiological impact of subnational interventions is quantified in terms of the number of hospital admissions (graphs). The lockdown as implemented in the Netherlands is represented by the black (observed) and green lines (prediction), which resulted in approximately 13 thousand hospital admissions up to 1 June 2020. Higher thresholds for deciding to implement a local lockdown clearly result in higher numbers of cumulative hospital admissions (top left) and a lower percentage of municipalities affected (top right). Here, a decision threshold of 3% (dark red) is too high; although it allowed for 185 million additional person-days spent without interventions, it resulted in a 157% increase in number of admissions (approximately 19 thousand). The more stringent thresholds of 1.0% and 0.33% result in numbers of hospital admissions closer to a national lockdown (4,670 and 355 additional admissions, respectively), but at a more modest social benefit (103 million and 35 million additional person-days free from interventions, respectively). Interestingly, at the lower threshold of 0.33% (yellow), approximately 10% of the municipalities remain without interventions for the full duration of the first wave. The maps show the corresponding geographical distribution of percentages of affected people, i.e., people who have experienced infection (left) and the social benefits of subnational interventions in terms of the fraction of repeated simulations in which a municipality is without interventions (right). Municipalities that remain free from interventions are mainly located in the north and east of the country, as can be most clearly seen for the scenario with the 0.33% threshold. From a mobility perspective, these are also the more rural, isolated, and less densely populated subnational regions of the country. We also test a scenario in which the national policies are implemented, but where, additionally, the borders of the municipalities are closed when confronted with 1.8% of the local population being infectious (I). Panel (a) shows that this scenario (grey) decreases the number of hospital admissions by approximately a third, for this particular threshold chosen.

It is important to note that because of the high computational demand of these local scenarios, only a small amount of runs are shown (ensemble sizes of five members for 3% and 1%, and two members for 0.33%) — larger ensemble sizes will provide more accurate results. Still, as we also saw in Fig. 9.2 and 9.3, we mainly expect geographical differences, i.e., which municipalities are affected, rather than strong differences in aggregated metrics such as cumulative hospital admissions and spatial spread.

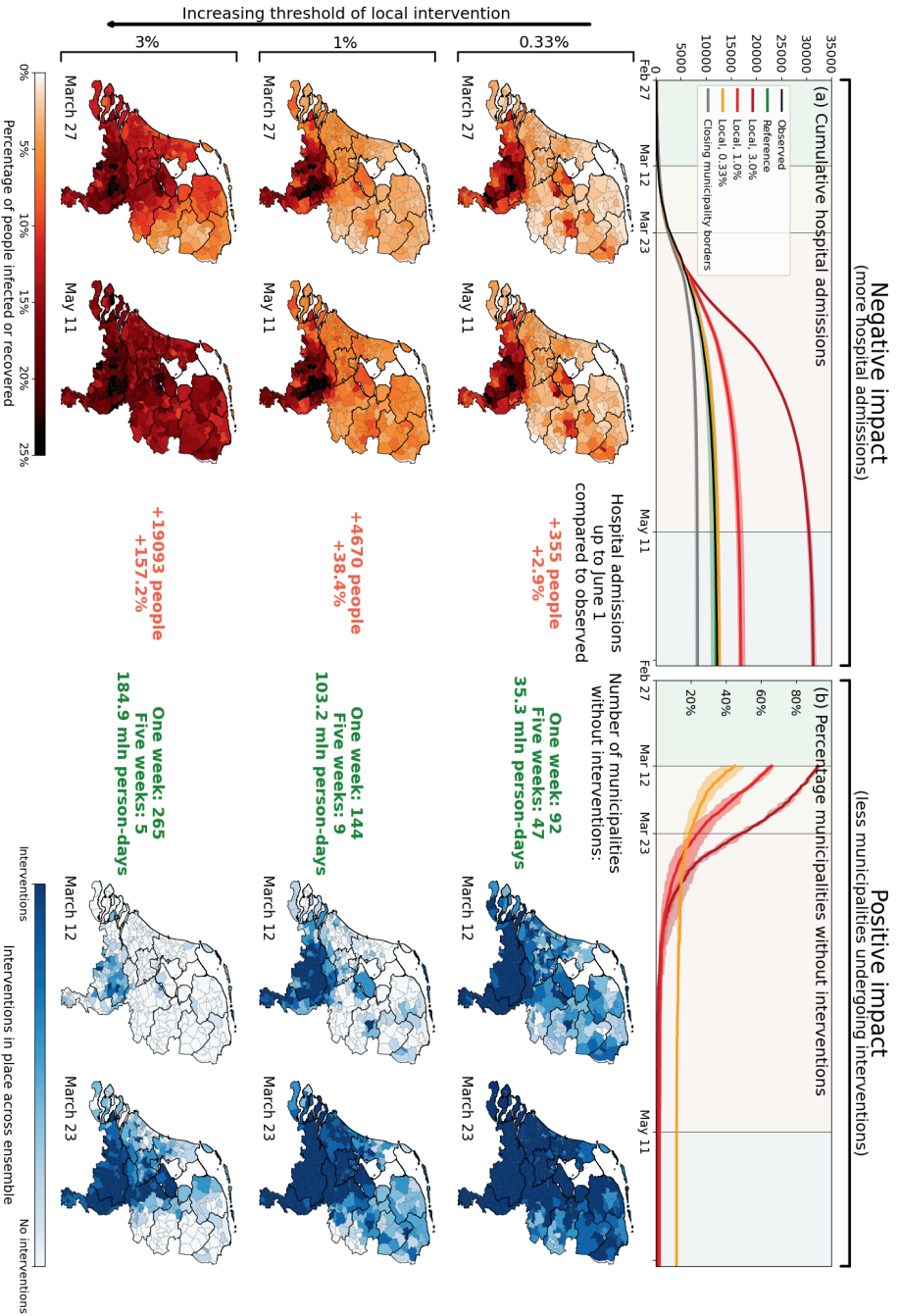


Figure 9.4: Quantification of the trade-off between negative impacts (left) and positive impacts (right) of locally-adjusted interventions at three levels of local intervention thresholds (0.33%, 1% and 3% of population simultaneously infected). **Panel (a):** Cumulative hospital admissions for different scenarios. **Panel (b):** Fraction of municipalities that do not have any interventions in place. **Panel (c):** Cumulative fraction of infection cases per municipality for the three local intervention thresholds. The additional number and percentual growth in hospital admissions as compared to the observed national interventions is indicated. **Panel (d):** Geographical indication of which municipalities have interventions and which not. The amount of municipalities that do not is indicated.

9.6 Discussion and Outlook

In summary, the model indicated national travel reduction and national behavioural change as effective measures in the Dutch first wave of COVID-19. At the same time, we assessed the effectiveness of local interventions, providing a quantification of the trade-off of being affected by interventions and increasing case numbers. In current literature, there is a variety of approaches used to investigate similar questions. Using a hierarchical Bayesian transmission model, Brauner et al. (2021) also evaluated first-wave intervention measures using a statistical approach, estimating that group size limitations and school closures were the most important measures. Analogously, also with a Bayesian hierarchical model, Sharma et al. (2021) analysed the resurgence of COVID-19 in Europe and found a much more limited effect of school closures, and a more pronounced effect for the closure of non-essential businesses and prevention of gatherings. Flaxman et al. (2020), also following Bayesian hierarchy, indicated that in the European first wave, (full) lockdowns were the main reducer of further spread, far exceeding other interventions in effectiveness. Mistry et al. (2021) proposes a method to infer sub-national mixing matrices from key socio-demographic features, and indicate how these can be used for population-wide epidemic modelling, while still accounting for a number of sub-national heterogeneities. The above papers indicate that there are different methodological approaches to this question — in fact, all of these works are different than the individual-based approach proposed in this chapter, with associated advantages and disadvantages.

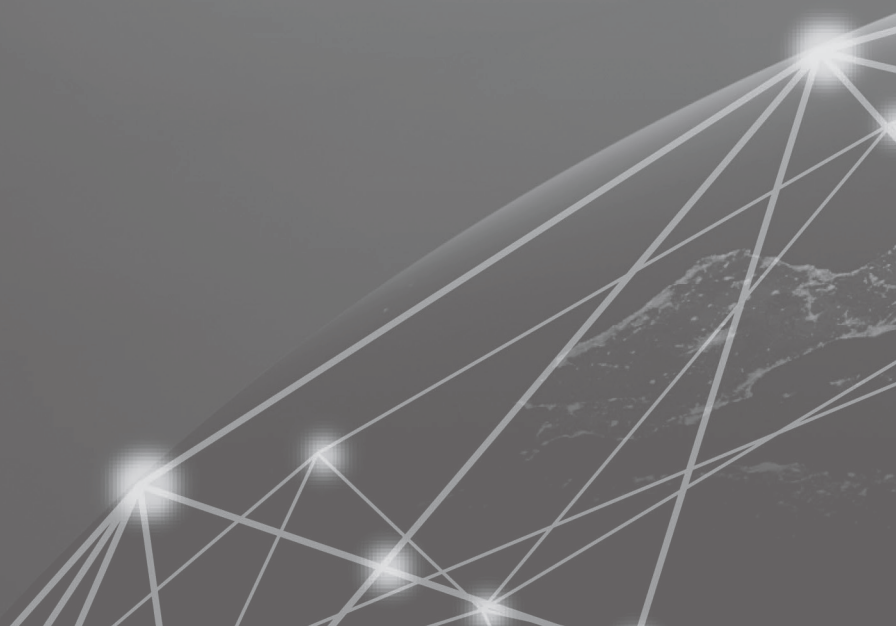
The methodology proposed in this chapter comprises demographic and geographic stratification, and distinguishes multiple circumstances of mixing. Still, there are still forms of granularity that limits our ability to precisely omit (and therefore: evaluate) specific interventions. This is of particular importance when discussing the effect of the closure of schools. In practise, school-going children would, instead of meeting many other like-aged people, be placed at home in near-isolation because of their indoors studying activities. In the model, this is simulated as a combination of factors: we place them at home, the mixing matrix changes from ‘School’, to ‘Home’, it is reduced based on survey studies in the Netherlands throughout this wave, their mobility is reduced and the β value changes. In the ‘no school closure’ scenario (yellow curves in Fig. 9.3), however, we only omit their placing at home and the mixing matrix changes, which are the only measures we could explicitly link to the closure of schools. The result is that we cannot omit all changes that happened due to school closures, which may lead to an underrepresentation of the impact of school closure. Still, it is expected that school closure in the Netherlands specifically had a minor effect in the first wave, because it was relatively late (16 March, which was after the peak; see the amount of exposed people, in green, in Fig. 9.2). Note that conclusions from previous work (Brauner et al., 2021) that indicated a large effects of school closures were based on groups of countries, not solely the Netherlands.

Other limitations to the methods in this chapter concern the generation of mobility patterns of agents. We use the Meuzuro mobility data to infer weights of where people are likely to go (used as input for the scale parameters of the Dirichlet distributions from which we draw their locations, see SI I.1.3). In practise, however, there is more structure in these movements: we cannot distil sequential of visits by the same person in the

Mezuro data. Also, if people work in a particular other municipality, they are likely to go there many times, instead of redrawing their locations each day, resulting in a wide variety of municipalities visited — as it is now in the model. The reduction of mobility is, as mentioned, based on Google Mobility data, which is chosen because they kept track of mobility changes throughout the first wave of COVID-19 in the Netherlands. However, their data is relatively aggregated: it is nation-wide and split into a small number of categories, making it difficult to specify mobility reductions per demographic group or geographic part of the Netherlands.

A main contribution of this work is in found in the geographic explicit nature of the transmission model, which allows for experimentation with locally-adjusted measures rather than only national ones. For now, we implemented this by allowing each municipality to intervene separately, based on a local hospitalisation threshold. When inferring implications from these results, ultimately, they should be translated onto the (local) level at which such decision making takes place. In the Netherlands, there are different levels of governance in this respect, such as 12 provinces and 25 *Veiligheidsregio's*. In particular the latter could be considered for a more tailored advice of how locally-adjusted measures may be more efficient (from a cost-benefit point-of-view) to limit epidemic spreading. Still, the chapter showcases how heterogeneity can be addressed in highly complex situations, how (implicit) network approaches can help understanding these patterns and how one can intervene to limit macroscopic spreading phenomena happening on top of these networks — which adds to the position of this chapter as the conclusion of this dissertation's Part III.

GENERAL DISCUSSION





In short, the three parts in this dissertation covered research questions concerning (I) macroscopic states, (II) the impact of network architecture and (III) spreading vulnerability in complex systems. In all chapters, these concepts are related to a form of collective change in the respective complex system — be it behavioural states expressed in the brain, climate tipping points, large-scale railway disruptions or epidemic spreading. The dissertation is not a comprehensive overview of all possible approaches to these questions. Rather, the chapters aim to contribute to both methodological advancement and the understanding of the example systems studied here, and, throughout this variety, provide an overarching perspective on methodology to macroscopic dynamics in complex systems in general. This section contains a reflection and outlook from this perspective.

An overarching perspective

A main challenge in this dissertation is the necessary level of detail in any of the analyses or models: while the phenomenology may be macroscopic, information on the driving mechanisms may be found in microscopic details, e.g., local disruptions leading to nation-wide railway delays (chapter 6), or a confined group of infectious individuals spreading a virus throughout a country (chapter 9). By nature, complex systems consist of large amounts of components, and in the real world, there are many sources of heterogeneity and external influences that may not be possible to fully account for. As the reader might have noticed, throughout the three parts of this dissertation, a growing level of detail in the methodology can be recognised.

In part I, the effort was to express three unique complex systems — the Dutch railway system, the brain of mice and coupled climate sub-systems — into low-dimensional state variables. In other words, either statistically or from physical reasoning, the aim was to fuse the large amount of microscopic heterogeneity and stochasticity in these systems together into a few macroscopic variables. The resulting analyses were on a macroscopic level: formulating and analysing the phase-spaces between system-wide principal components (chapters 1 and 2), and investigating transitions across macroscopic states (chapters 1 and 3). Even though part I does not ignore microscopic interactions in the system — e.g., chapter 1 implicitly incorporates these by using observed data — the chapters in part I do not explicitly account for them. This is in contrast with Part II, where the effect of network architecture and component interactions are explicitly addressed. For example, in chapter 4, the role of spatial non-uniformities in developing geographical patterns in railway dynamics is researched, and the main question of chapter 6 is how microscopic interactions between railway resources may lead to wide-spread railway disruptions. The role of detailed interactions is pursued even further in part III, where interaction sequences of individual agents are related to a system-wide vulnerability to spreading (chapter 7) or their impact on spreading is quantified (chapter 8). The ultimate product of this trend is the final chapter (chapter 9), which accounts for many sorts of heterogeneity, such as demography, mobility and government policy, in its simulation of SARS-CoV-2 spreading — in sharp contrast to the low-dimensional state variables in chapters 1-3.

How microscopic interactions lead to macroscopic behaviour and which level of de-

tail to include in the analysis are fundamental problems in complex systems literature. From an overarching perspective, I state three considerations concerning the development of methods in this dissertation.

Consideration #1: Concepts versus predictions

The first considers the aim of the research: whether it is conceptual understanding, explaining the macroscopic phenomenology, or prediction, concerning the accurate simulation of a particular phenomenon to be able to predict or hindcast it in scenarios or a predefined state. Even though conceptual understanding and accurate prediction are not mutually exclusive, the aim of the study implicates the level of detail to include. For example, in chapter 3, the low-dimensional climate model connecting the stability of the Atlantic Meridional Overturning Circulation (AMOC) to potential oscillations in the equatorial Pacific is greatly simplified, but its aim was to illustrate conceptually the potential of such events and the respective role of different parameters or physical drivers. In fact, earliest findings of multiple stable equilibria of the AMOC were also derived from such a simple model (Stommel, 1961). An accurate simulation of such an event, involving timing, the exact magnitudes and their implications, is a different question and requires advanced and high-resolution climate models. Also when researching railway dynamics, this contrast is visible. In chapter 4, averaged statistics and schedule information was used to obtain more understanding in geographical structures, the proposed model is not suitable for accurate prediction purposes. (In fact, the spectral clustering method in chapter 4 was made possible because of particular simplifying aspects of the model.) Chapters 1 and 5 focused more on prediction of railway delays, for example, by using statistical analysis on observed data and including many more variables explicitly (such as start and ending stations, train turnover rate, edge-specific running times, etc.), respectively. One would argue that a perfect model provides both conceptual understanding and predictive power, but that is not always possible.

Consideration #2: Level of relevant heterogeneity

In complex systems, agents and their interactions are hardly ever fully homogeneous. An important consideration also in this dissertation concerns the level of heterogeneity that is *relevant* to add in the analysis. In chapter 3, the temperature in the North Atlantic in reality varies enormously due to geographic differences, differences in water masses and many more factors. Still, for the purpose of identifying equilibrium states, it sufficed to use a single temperature difference between the North and Equatorial Atlantic. In contrast, chapter 9 aims to test the effectiveness of governmental interventions in the Dutch first wave of COVID-19, which in particular implied the closure of schools and regional strategies. The latter two particularly affect certain demographic groups (parents and children) or have a geographic component, respectively, which require the model to include demographic and geographic stratification explicitly. In railways, the choice of relevant details to distinguish is not trivial, illustrated by the variety of approaches across the four railway chapters (1, 4, 5 and 6). In other railway literature, highly microscopic models are commonly used to test timetables (e.g. in Middelkoop and Loeve (2006)), which do require details such as types of tracks, trains and infrastructure connections superseding the level as included here. On the other hand, when studying mechanisms

driving large-scale spread of delay (e.g., in chapter 6), such details may not always be necessary and excluding them allowed us to abstract the network into single types of nodes and edges. Another form of heterogeneity is introduced in part III, differentiating it from part II. In part III, we account for temporal changes in the network architecture. Whether it is important to account for such ‘temporal heterogeneity’ is partially matter of time scales, as discussed in the introduction to parts II and III.

Consideration #3: Application perspective

A third consideration is the mere practical side of including additional levels of detail in any analysis, which can be constraint by aspects of data collection and computational limits. When studying spreading phenomena in systems where the spreading pathways change over time, ideally, detailed information on these temporal variations is used, such as the temporal networks in chapters 7 and 8. Moreover, the time scale of this interaction data should be equal or lower than the actual dynamics to have enough temporal resolution (as discussed in the introduction of part III on p. 148), and the length of the dataset should be long enough to be able to capture any dynamical evolution at all — which is a problem mentioned at the end of chapter 8 (p. 177). Such data requirements make doing the same analysis on a large scale unfeasible, which advocates for techniques such as coarse graining. In chapter 9, we did not have enough information to infer mobility, demography and mixing in more refined geographical locations than municipalities, or at more refined temporal resolutions than hours. At this municipality scale, we had to work without explicit interactions between couples of agents, but instead used a demographic mixing-weighted ‘force of infection’ (p. 185) exerted on each person within the same municipality, probabilistically accounting for interactions he or she might have during the specified hour. Another example was the impact of external factors affected railway dynamics, which are known to be important (Wei et al., 2015; Ludvigsen and Klæboe, 2014; Dekker et al., 2021f), but difficult to quantify dynamically: it relates to details such as individual trees falling on tracks and weathering of the tracks and electric lines, of which data is not available. Even though the role of the availability of microscopic information may sound trivial, it is an important consideration in the approach. It also may define an upper limit of the accuracy of prediction in general: given that no information at the detailed level of trees near railway tracks (chapters 1 and 6), or the exact neurological differences between one mouse brain and another (chapter 2), is available, can we even simulate these systems with full accuracy at all? This practical problem relates to the question of Laplace’s demon (p. 8) and the numerical (practical) argument against its ability to predict the future from the current state. For the railway case, I elaborate on this further on below.

Consideration #4: Interdisciplinary

Under the umbrella of complex systems methods, this dissertation contains a variety of example systems, each with their own intricate details, requiring input from experts of the respective fields — neuroscientists (chapter 2), epidemiologists (chapter 9), transportation researchers (chapter 6) and behavioural scientists (chapter 8). Still, working on dynamics in each of these systems reveals that many methodological principles are

interdisciplinary. A clear example is the re-usage of the same methodology from an application in railways (chapter 1) to an application in neuroscience (chapter 2): the condensing of high-dimensional data into a few principal components that span a low-dimensional phase-space in which clusters have a macroscopic meaning. Other examples can be found in the spreading vulnerability metrics (chapters 7 and 8), applied to a variety of systems such as human interaction at schools, natural systems, citation networks and social media.

Application to railways

The reader might have distinguished a second narrative in this dissertation, one besides the more general relation between microscopic elements and macroscopic behaviour in dynamical phenomena. It concerns application of railways and a particular associated emergent phenomenon: large-scale railway disruptions. Railway dynamics are explicitly addressed in chapters 1, 4, 5 and 6, and, arguably, the spreading vulnerability research in chapters 7 and 8 also connects to railways via the principles of the delay cascades in chapter 6. These events typically involve initially local perturbations. Common causes of these perturbations are infrastructure problems due to weather, frost, or trees falling on tracks and overhead lines. Still, such local problems may spread delays across large geographical areas, such as in the case of February 2, 2012 shown in Fig. 1.1 on p. 25 — making railway disruptions a perfect example in which microscopic interactions result in macroscopic phenomena. More details on a few case studies of these events together with a small review on relevant problems and opportunities can be found in Dekker et al. (2021f) (not included in this dissertation). An important factor in the most extreme cases of these events turns out to be the lack of synchronous information, which severely limits the ability for dispatchers to remain in control of the traffic flow, which is why these events are also colloquially referred to as ‘out-of-control situations’.

The results of chapters 1, 4, 5 and 6 are from a dynamical and predictive nature, whereas the direct operational implications were not the focus. In fact, embedded in a project framework as shown in Fig. vi, the research in these chapters was conducted in close collaboration with researchers from other disciplines that did focus on this practical side of these disruption events. The project is split in three parts: subproject A concerns the understanding, anticipation and characterisation of these events, subproject B addresses how, during such events, the logistics should be handled, and subproject C involved microscopic simulations to test the findings of subprojects A and B. The interplay between microscopic interactions and macroscopic phenomena is especially true for large-scale disruption events, making complexity science a suitable start to work on the topics in subproject A (left side), in which the anticipation and the geographic identification of such events are separated into two steps. The right side of the framework concerns subproject B, in which four steps were identified in how to effectively manage such events. Subproject B was led by researcher Rolf van Lieshout. In his research, he produced a range of alternative line plans that were self-organised, i.e., without the need of a functioning central dispatching (Van Lieshout et al., 2020; Van Lieshout et al., 2021). This is especially useful in the case of large-scale disruptions, characterised by asynchronous information flow and limitations to central dispatching.

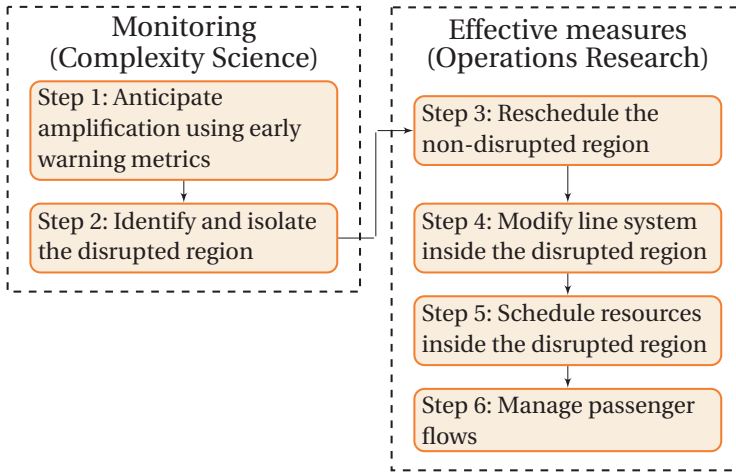


Figure vi: The project framework from Dekker et al. (2021f), describing how the characterisation of large-scale disruptions and the subsequent adaptation to such events are connected, by linking research in chapters 1, 4, 5 and 6 (left side) to that of another PhD project (right side).

The chapters in this dissertation aim to either explicitly (chapters 1, 4, 5, 6) or implicitly (chapters 7 and 8) add to our understanding of these events. Step 1 in the framework of Fig. vi is mainly addressed in chapters 1 and 6. The first chapter was also chronologically the first effort to characterise such disruptive events. Using a principal component analysis, we identified covarying patterns in the delay dynamics (Fig. 1.2 on p. 32), from which a low-dimensional phase-space was found. Expressed in these macroscopic variables, large-scale disruptions could be defined and anticipated using an early warning metric. Given the abstract nature of this phase-space, an understanding of more detailed dynamics driving such (pseudo-)transitions became evidently beneficial, and this resulted in the discrete-event model in chapter 6, which distanced from macroscopic statistics, but focused on the mechanics and drivers of delay dynamics. In particular, we found and simulated potential delay cascades across railway resources. In cases of severe delays, rolling stock and crew members associated with the delayed train lines are also affected. When these resources are transferred to other train lines, these train lines may become additionally delayed, as well (this is explained more elaborately in Fig. 6.2 on p. 138). We used a model for such cascades to diagnostically identify when this effect played a role, and when a lot of pressure was exerted on dispatchers to prevent this effect ('delay mitigation') — see Fig. 6.4 on p. 142.

Chapter 4 concerns step 2 in the framework of Fig. vi, even though it focuses on a general geographical characterisation of delay dynamics in four European countries. The results are a first step towards identifying dynamically sensible areas to isolate (i.e., the 'disrupted region' in step 2) in case of severe disruption. It also involves an international perspective on the role of various regions in the total delay and their exchange with neighbouring regions (Fig. 4.5 on p. 104). Chapter 5 is somewhat in between steps 1 and

2: it describes a model in which we aim to predict and simulate delays at various geographical scales and discuss advantages of defining delays as variables on larger scales in general. Chapter 7 formulates the obtained understanding of delay cascades in chapter 6 into a theoretical framework. In particular, the proposed ‘entropy of temporal entanglement’ can be applied to compare and assess the vulnerability against such cascades in railway schedules.

Even though these chapters each analyse these railway events from a different perspective, a practically useful early warning metric was not found — one that provides railway practitioners with explicit tools to implement, beyond mere statistical alarms (chapter 1) or simulation models focusing on particular aspects of such events (e.g., chapter 6). A reason for this is that the dynamical models in chapters 1, 4, 5 and 6 (along with many other railway models) miss two key ingredients: human control (‘real-time dispatching’) and external or incidental influences such as weather, accidents and infrastructure malfunctioning. Both ingredients are inherently difficult to model and important topics for future research. A possibility would have been to add agent-based control based on protocols, but not all dispatching decisions are fully captured in protocols due to the wide range of dispatching possibilities. Dynamical modelling of human dispatching and conflict resolution would contribute significantly in this respect, which requires highly case-specific behavioural descriptions of agents and their adaptive decision-making. Such directions could be beneficial to pursue for railway research.

The other missing ingredient — external and incidental influences — is arguably even more complex. For example, in Dekker et al. (2021f), we discuss three cases of severe disruptions that were caused by extreme weather (February 3, 2012 and January 18, 2018) and an electric outage (January 17, 2017), respectively. In theory, such influences could be dynamically implemented in simulation models in a probabilistic way, resulting in low but nonzero probabilities for such events. Quantifying such probabilities, and their dynamical impact may, because of their rarity, is notoriously difficult to do accurately, especially in real-time. Consequentially, even if we could build a perfect model of how delays spread and how dispatching reacts (i.e., using the first aforementioned ingredient), if the cause of the disruption related to an electric outage, we would have no prediction of the event prior to its initialisation. Hence, throughout the railway chapters in this dissertation, I focused on the dynamical aspects of these events *after* a potentially externally-caused initial perturbation has taken place.

Also note that even if a perfect early warning could be found (including both ingredients), it is of low practical use when it is a boolean alarm on its own. Sufficient details need to be provided to do something with that information, such as the type of problem, the identification of certain trains that play an important role or a particular disrupted area. Anecdotally, the importance of de-synchronisation of information, loss of oversight and IT problems turned out to be very important in large-scale disruptions. Such problems are less related to dynamics and predictions, but more about information and (IT-)infrastructure. Therefore, part of the solution to such events lies in analysing the information systems, rather than the dynamical aspects of delay evolution.

Summarised, the railway chapters in this dissertation, together with the concept of spreading vulnerability in chapters 7 and 8, aim to contribute to understanding dynamical phenomena in railway systems: the evolution, cause and geographical aspects of large-scale disruptions. A practically useful early warning — i.e., the step ‘before’ this

understanding — could not be found, but, given that such a feat would be possible at all, the work here made a first step. Future research directions concern the aforementioned missing ingredients, and more general ones as discussed on p. 203.

Network architecture and dynamical phenomena

The relation between network architecture and how determining it is for the dynamics happening on top of the network is one of the key research questions in this dissertation. In parts II and III, these networks are included in the analyses. In most cases (e.g., chapters 4, 5, 7 and 8), the networks were incorporated explicitly. However, networks also have played implicit roles in some of the analyses: in chapter 6, the network architecture is merely used to form pathways on which the proposed delay cascades can evolve, and in chapter 9, the network of human interaction (to simulate SARS-CoV-2 transmission) is only implicitly formed by means of large clusters of people in each municipality — agent-to-agent interaction is explicitly not modelled there. Searching for the connection between network architecture and dynamics on top of the network throughout the applications in this dissertation, revealed a few critical remarks on this interplay.

First, many topological metrics are useful to study the structure of network architectures, but their relation to dynamical phenomena happening on top of the network architectures is not trivial, and potentially even very weak. For example, dynamical phenomena may be associated with highly discrete and anisotropic evolution — such as in railway systems, where delays follow trains that are prescheduled to follow particular lines instead of evolving in every direction (which one might expect from a network point-of-view). In chapter 5 we tried to model railways from a diffusion-like perspective by using edge and node weights, but in the end this approach has limited predictive power for cases in which the delays are relatively weak, or when individual trains drive the delay evolution — the latter is not unusual as highlighted in the delay cascades in chapter 6. Following the example of railways further, two well-known network metrics are displayed in Fig. vii for the Dutch railways: the node degree and betweenness centrality of stations. While Utrecht central station (Ut) is clearly visible in these metrics, the station with most passengers per day (Amsterdam central, Asd) has a node degree of 3 and a mediocre betweenness centrality. Dynamically less significant service control points can also be distinguished to have high values in these metrics, such as one near Zwolle (Zlgea). Note that the degree scale only runs up to 6 and many nodes have degree 2, which is common in transportation networks due to their 2D, planar nature. These observations illustrate that these network metrics do not provide much insight in the delay dynamics happening on top of the network — which can be a problem for many more complex systems: standard network metrics do not necessarily provide much insights in the dynamics, and the relationship might be obscured by a discrete, probabilistic or anisotropic nature of the dynamics. This also implicates the limited practical use of percolation theory and node failure techniques to railway dynamics: in many disrupted railway situations, stations are still operating (i.e., not ‘removed’ from the network), and rerouting barely happens in the Dutch railway case.

Second, there is growing criticism on the *direct* applicability of standard network models to real-world systems in general. As mentioned in the general introduction, the

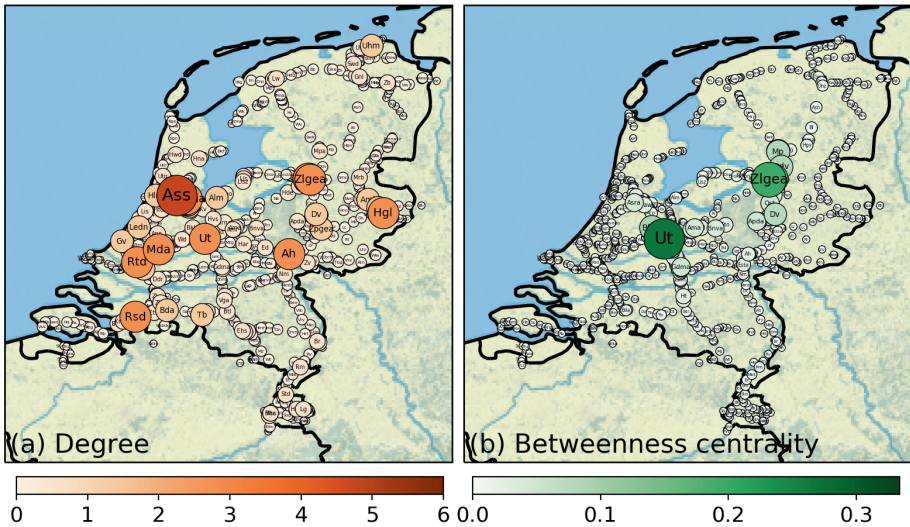


Figure vii: Node degree (left) and betweenness centrality (right) of all nodes in the Dutch railway network. Station abbreviations are used. A few notable ones are Utrecht (Ut), Amsterdam (Asd), Rotterdam (Rtd) and Arnhem (Ah) central stations, a (non-passenger) service control point near Zwolle (Zlgea) and an important transfer station near Amsterdam, called Amsterdam Sloterdijk (Ass). Node degree is defined as the amount of links per node, and betweenness centrality is defined as the fraction of shortest paths (of all node pairs) that includes the node.

discovery of scale-free networks (Barabási and Albert, 1999) was significant, but the universality of it as it was proposed in the real-world is questionable (Mitchell, 2009). The proposed network growth mechanism of preferential attachment (or Yule process (Yule, 1925)), which implies that nodes with many links are more likely to obtain more, indeed results in network structures that are scale-free. That is, the degree distribution follows Zipf's law or the related Pareto's principle (Newman, 2005). However, there are many other mechanisms that drive such network structures, such as combining exponentials, inverse quantities and random walks (Newman, 2005; Mitzenmacher, 2003). Additionally, many examples of real-world systems can be found where such power law distributions are *not* present, resulting in (stretched-)exponential or log-normal distributions, even in systems that were previously found to follow power laws (Clauset et al., 2009). Besides limitations of generative network models, standard networks principally include nodes and links that (albeit weighted) involve some form of comparability or homogeneity, which may not reflect several heterogeneities that are crucial for the system dynamics. In railway networks, for example, all stations are homogenised as 'nodes', while stations in bigger cities are logistically very different from rural ones — not only in terms of network topology. In the case of human interaction, a scalar link weight might not be enough to capture all complexities regarding disease transmission: it assumes that with a single scalar you can compare the link's strength with another link's strength. However, external conditions such as temperature and humidity may alter the transmissivity over time, which would theoretically be possible to include in the weight of links by making

their weights a function of time or climate, but is often not done. For example, in the primary school network (Fig. iib on p. 14), some links between the children are made during a break outside, which is for epidemiological purposes very different from those in a small, confined room. Or, in the citation network in Fig. iid on p. 14, some citations may be in a negative context, when counteracting an argument made in an earlier paper, which may be very different from a positive context when studying how information spreads through scientific literature. Pursuing this problem, the network science community has made significant advancements in the direction of multiplex networks (e.g., in Mucha et al. (2010)) and sophisticated ways of weighing or stratifying links and nodes, to account for such qualitative differences.

A third remark concerns temporal networks and echoes the importance of the network sampling time scale and its relation to the time scale of the targeted dynamics, as mentioned in multiple instances across this dissertation, such as the introduction to part III (p. 148) and chapter 8. In light of limited availability of data and potentially not having a choice in the sampling time scale of empirical temporal networks, this particularly means that temporal networks are not automatically the better choice to describe spreading phenomena.

A final note concerns the adaptivity of network architecture. In part III, many examples of temporal networks are discussed: networks that change over time. In those cases, the temporal networks were prescribed and treated as non-reactive to the dynamics or environment. In other words, while the dynamics on top was affected (bound) by these architectures, it was not the other way around. However, the architecture response is an important principle in many complex systems: cases in which the network architecture adapts to either the environment or the dynamical state on top of the graph. Such systems are referred to as *complex adaptive systems* and are an active research topic of today. In fact, many systems in nature and society, even some in this dissertation, may be regarded as complex adaptive systems. In epidemiology, people adapt to the state of the disease: in the early stage of COVID-19 in the Netherlands, for example, people did not regard it as a major threat and only started reducing their contacts (altering the typical network patterns) when the disease was more widespread. The effect of the adaptivity of the system's structure to environmental conditions may therefore be important to consider when utilising network architectures to describe network phenomena.

Future directions

Although in each of the chapters, case-specific follow-ups are possible and mentioned in each of the chapter discussion sections, three general future directions can be distinguished and are listed below.

Direction #1: Prediction of cascading dynamics

Cascading dynamics are conceptually addressed in multiple chapters of this dissertation and are roughly described as causal pathways in the dynamics across multiple agents or subsystems. In chapter 3, cascades across coupled climate subsystems are found in the form of interacting critical transitions. In chapter 6, we find cascades on a more microscopic detail: delayed train lines that automatically delay associated rolling stock and

crew, which may pass on their delays onto other train lines upon transfer, which in turn also affect associated rolling stock and crew, leading to a potential cascade of delays towards large geographic areas. In chapters 7 and 8, cascades are explicitly found in the interaction sequences, where changing network architectures act as pathways for cascading spreading phenomena. Especially when it concerns cross-system interactions, it is of high societal relevance to predict cascading dynamics — e.g., as in chapter 3 and 6, or in a famous electric outage in Italy, associated with cascading interactions between communication networks and their dependence on electricity (Buldyrev et al., 2010).

In the case of critical transitions that set in motion other transitions, such as discussed in chapter 3, prediction of such cascades relates to the vast literature on early warnings of tipping points. The theory of ‘critical slowing down’ is well established, associated with indicators such as increased autocorrelation, variance, skewness and flickering (Dakos et al., 2008; Kuehn, 2011; Scheffer et al., 2009). However, many of these metrics concern individual transitions rather than cascading transitions, which may involve more complex approaches such as detrended fluctuation analysis (used in chapter 3 and its accompanying SI C). This is an active branch of research and contributes to our understanding of cascading dynamics of critical transitions. However, there are also other forms of cascading dynamics, that do not involve critical transitions. Such pathways may be difficult to recognise and call for studying individual mechanisms. The inference of causal links in such complex systems requires more sophisticated tools such as tools proposed in Runge et al. (2019), or studies focusing on the detection of specific cascades, such as in chapter 6.

In general, chained causal links that form cascading dynamics are vital to many dynamical phenomena discussed in this dissertation. Predicting their onset can be of vital importance for intervention and system design purposes.

Direction #2: Spreading vulnerability and system design

Several chapters in this dissertation address spreading dynamics. Chapter 7 defines an entropy-based metric that specifies a system’s vulnerability to spreading, and chapter 8 quantifies the impact of individual agents in this vulnerability using ‘contact sequence centrality’. Both link spreading dynamics to aspects in a system’s network architecture. A natural question is what exactly makes a system vulnerable to spreading. Knowing the answer to this questions allows for more robust system design and cross-comparison of systems.

Fig. 7.3 on p. 158 in chapter 7 already makes first steps in this direction by comparing the vulnerability of a variety of systems using the entropy metric. Ultimately, these resulting spreading vulnerabilities are dependent on the systems’ *designs*: particular configurations of the sequences of agent interactions favour, and others limit spreading in the system. This is of particular interest in the cases of epidemiology and railways. In chapters 6, 7 and 8, several suggestions of important factors have been made, concerning system modularity, entanglement of interactions and sequence centrality. In epidemiology, an obvious direction is the creation of ‘bubble’ structures in the interaction networks where people are allowed to interact within, but not across. More subtle would be to investigate and target those people that connect those bubbles, arguably the ones with high contact sequence centrality (see chapter 8). In the railways example, this would

imply paying extra attention to spreading connections made via the utilised resources; allocating crew and rolling stock to service lines in the same confined area. Such modular timetable designs are already naturally in place in other countries such as Italy and Denmark, where a larger variety of railway companies handles each their own region, with associated advantages and disadvantages. In the Netherlands, a large fraction of the train lines are managed by a single railway company.

It is clear from these examples that a proper quantification of spreading vulnerability and its drivers allows for the designing of systems that are robust to spreading phenomena. While chapter 7, we already proposed a pure network-based metric to quantify this, more research is needed to characterise spreading vulnerability in cases where dynamical time scales play an important role, or where such high-resolution temporal network data is not available.

Direction #3: Spreading vulnerability and system interventions

While system design is focused on spreading prevention, information on spreading vulnerability can also be used on-the-fly spreading limitation, which has to do with system control or *interventions*. On a more general and theoretical level, on-the-fly limiting of spreading is highly relevant to societal problems such as the aforementioned spreading of infectious diseases and railway delays. Chapter 9 already discusses this from an applied perspective in the case of SARS-CoV-2 spreading in the Netherlands and the effectiveness of governmental interventions. It involves reacting to spreading rather than preventing it, and has to do with keeping up the system's performance as much as possible, while preventing further spread — such as isolating a disrupted region as discussed in the above railway application section, or implementing regional COVID-19 measures instead of nationally. In the railways, the degrees of freedom are partially confined to protocols which can be studied to include a form of control module in railway modelling. Analogous to the second direction mentioned above, a general approach to spreading vulnerability and interventions may be of high societal value.

Final remarks

This dissertation contains efforts to contribute to our understanding of emergent phenomena in complex systems, both from a methodological and an applied perspective. The work contains a variety of applications and their associated dynamical phenomena. We have defined states in railway, neurological and climate systems, found geographic patterns in nation-wide transport systems, and discussed the vulnerability of various human interaction and mobility structures to spreading dynamics. While answering some of the questions in these systems, as research goes, many more appeared — none of the proposed methods and models have 'fully solved' the observed macroscopic dynamical phenomena. Ideally, one might argue, we eventually reach the ability to simulate nation-wide mobility, human behaviour and infectious disease spreading to the highest resolution and accuracy. But is this realistic? The perfect, full-resolution simulation of reality will always, or at least in the near future, be a fantasy. (E.g., see Scarpino and Petri (2019) for fundamental limits to the predictability of outbreaks.) This relates back to Laplace's demon, mentioned in the general introduction (p. 8). The implicit consequence of the

demon's inability to predict the past and future is that methods linking the effect of microscopic features to collective change in a system will, although flawed, always remain important. This realisation lies at the core of this dissertation and motivates the efforts presented. I hope to have contributed to solutions for building that bridge. Regarding the paradox of the simultaneous imperfection and usefulness of models, the British statistician George E.P. Box once wrote in the context of the ideal gas law:

“Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. (...) For such a model there is no need to ask the question ‘Is the model true?’. If ‘truth’ is to be the ‘whole truth’ the answer must be ‘No’. The only question of interest is ‘Is the model illuminating and useful?’.” (Box, 1976)

Or, reformulated in the shorter, well-known aphorism:

“All models are wrong, but some are useful”

SUPPLEMENTARY INFORMATION

Appendix A

Supplementary information to Chapter 1

A.1 Data processing

This section concerns processing of the discrete logging data from the Dutch railways to continuous time series.

A.1.1 Data aggregation

We start by defining ‘nodes’ as the points at which logging takes place, and ‘segments’ as the tracks between them. In the Dutch railway system, not only the number of delays are logged, but also whether they concern departures, passings or arrivals. To make the data uniform over the entire network, at any given time, we aggregate all logging activities on to the segments that the trains should be travelling on. Precisely: we aggregate the departure and arrival activities respectively on the segment that the designated train moves to and where it comes from, in the following manner (see Fig. A.1).

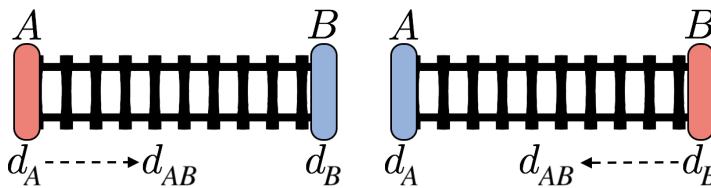


Figure A.1: Transformation of logged delay at nodes to segmental delay. Consider a train travelling from node A and to node B, where delays d_A and d_B are logged. **Left:** departure logging at A, meaning that we ‘forward’ this logged delay in space towards segment AB. **Right:** arrival logging at B, we aggregate this event also to segment AB.

The choice of dealing with delays on segments (instead of nodes) is motivated in threefold. First and foremost, to improve on spatial precision during aggregation: at larger stations, where many different train lines cross, we would lose information about the direction of the delay propagation if we summed up all delays for the incoming and/or the

outgoing trains into one node. Secondly, delays mostly change on the segments themselves. Only in specific cases (e.g., crew transfers, blocking of parts of a station), changes in delay actually take place at a node. (Consequently, the logged activities almost always relate to movements of trains, except for departures at large stations where there is a planned gap between arrival and departure.) The third reason for the choice of segments to aggregate the time series on is that it facilitates easy visualisation.

The logs are discrete events. To get a real-time progression of delays on the network, we need to process the logs to generate continuous time series data. At the logging events (e.g., a departure event at node A in Fig. A.1), both the planned (t_p) and realised (t_r) times are recorded. At the moment of logging, the delay carried by this train is obviously $t_r - t_p$. But between t_r and t_p , the train is already delayed and consequently affecting the segment, although the train is yet to enter the segment. This leaves us with two choices to deal with the delay: (a) keep the delay with the train, i.e., keep it on the segment where the train presently is, or (b) put the delay on the segment where it is planned to be. We choose the latter, since it more accurately indicates the effect of the delay on the network: e.g., when a train is an hour late at a given segment, it already causes problems before it actually gets there (an hour later than scheduled).

Mathematically, we define delay as explained in Eq. (1.6) and (1.6) in chapter 1. In this form, the equation for $d_i(t)$ contains a discontinuity when t reaches t_r . This is unwanted, since for prediction purposes, we need to treat the system as a continuous dynamical system. We therefore introduce a Gaussian weighted running window with two parameters: the Gaussian variance σ and a parameter tuning the width of the window p :

$$\bar{d}_i(t) = \frac{\sum_{t'=t-p\sigma}^{t+p\sigma} d_i(t') \cdot e^{-\frac{1}{2}\left(\frac{t'-t}{\sigma}\right)^2}}{2p\sigma + 1} \quad (\text{A.1})$$

We use $p = 5$ and $\sigma = 3$ seconds, tuned to give a relatively smooth curve without losing too much detailed information. This means that the aggregation is done on a 5-second time resolution, which, for computational reasons, is further reduced to 1-minute time resolution, implying that the precise values of σ and p do not affect the results. In Fig. A.2 we show a smoothed sample time series from February 4, 2018, wherein the distinctive (smoothed) saw-tooth pattern is clearly visible.

A.1.2 Problems with data during large disruptions

In case of very large disruptions, errors in the data logging occurs due to information loss on the exact locations of trains. Another source of errors is the large-scale stand-still of trains, while not all trains are registered as ‘cancelled’. A consequence of these errors is that trains that are restarted after a network-shutdown are sometimes treated as if they are merely delayed (rather than being cancelled and restarted). This results in unrealistically high delay values (sometimes over 13 hours). An instance of this situation occurred on January 18, 2018, also considered as one of the case studies in Dekker et al. (2021f). On this particular day, trains stood still for a full afternoon without being cancelled, and were simply restarted.

Even for a dataset consisting of 365 days, these strong but erroneous delay sources have a major impact on the PCA due to their large impact on the variance, while they

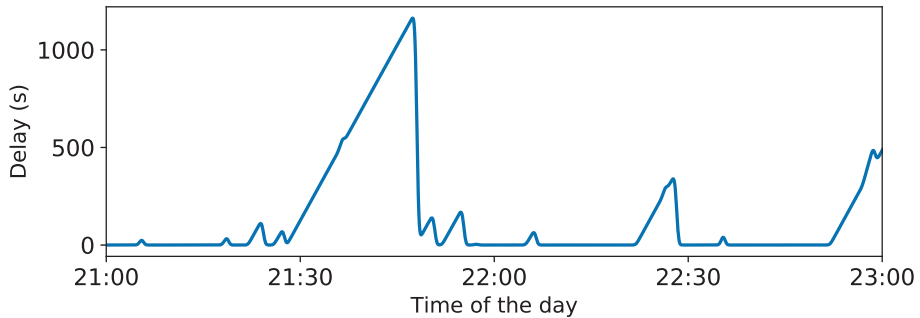


Figure A.2: Delay for the segment from station Utrecht Vaartsche Rijn towards station Utrecht Central Station, between 21:00 and 23:00. This particular data is from 1 July 2017.

clearly do not play any dynamical role whatsoever. In order to remove these spurious events, each train activity that is registered to move with a delay larger than three hours are considered first cancelled, and then replanned as a 'new' train such that it loses its delay signature.

A.1.3 Classification of days

To get a rough indication of the performance of the system, the manager of the main railway network in the Netherlands (ProRail) uses a day-to-day labelling of each day. The labelling is based on two variables: punctuality and cancellations. These metrics are calculated on a fraction of all train activities. First, the focus is only on the arrival activities at the final destinations, making sure that each train activity is counted only once. Secondly, only passenger (and no freight) trains are considered. Thirdly, a specific subset of service control points (SCPs) are evaluated, but not all of them. This subset is chosen by expert judgement and in communication with the Dutch ministry of Infrastructure, based on impact on the infrastructure, personnel stress level and passenger flow, and is based on the coverage of lines such that only important stations (in terms of where delay can easily arise, spread or change) are considered. Again, this is motivated to not count a delayed train passing through many SCPs more times than trains that are travelling through less SCPs. Finally, The delay data at the specified service control points is only calculated a few times per day per train line.

Table A.1 shows the specified values per label. Formally, the black days are a subset of red days, but we separate these for analysis purposes.

The SCPs considered are the following (Dutch) passenger stations: Rode School, Groningen, Leeuwarden, Delfzijl, Bad Nieuweschans, Winschoten, Veendam, Harlingen Haven, Sneek, Stavoren, Emmen, Coevorden, Hardenberg, Almelo, Oldenzaal, Hengelo, Enschede, Zwolle, Raalte, Kampen, Deventer, Apeldoorn, Zutphen, Winterswijk, Doetinchem, Nijmegen, Arnhem, Venray, Venlo, Roermond, Sittard, Heerlen, Maastricht, Maastricht Randwyck, Kerkrade Centrum, Eindhoven, Tilburg, 's-Hertogenbosch, Geldermalsen, Tiel, Ede, Barneveld Zuid, Amersfoort, Almere, Utrecht, Gorinchem, Dordrecht, Roosendaal, Vlissingen, Rotterdam, Gouda, Amsterdam Zuid, Amsterdam, Hoorn, Alkmaar, Haarlem, Schiphol

Airport, Alphen aan den Rijn, Leiden, Den Haag Centraal and Den Haag HS.

Label	Punctuality	Cancellations	Number of days
Green	>92.5%	<1%	46
Neutral	-	-	292
Red	<85%	>5%	21
Black	<75%	>10%	6

Table A.1: Day-to-day severity labels by Dutch asset manager ProRail. The last column shows the number of days in the dataset used in this study (July 1, 2017 - June 30, 2018). Note that the label ‘Neutral’ applies when none of the other labels does.

A.1.4 Choosing the relevant principal components

Figure A.3 shows the distribution of the explained variance and autocorrelation decay time scale for the first ten PCs, explaining our choice for using only the first two components for the reduced phase-space. As mentioned in Sec. 1.3, we perform the PCA on the data from the ‘disrupted days’, i.e., the red and the black days combined together.

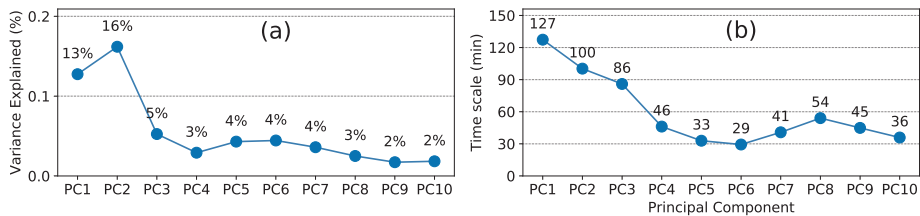


Figure A.3: Variance explained (a) and time scale of autocorrelation decay rate (b) for the first ten principal components (PCs).

The explained variance (upper panel) by the first two PCs is distinctively more than for the other components, while the persistence (lower panel, defined as the time scale for the decay of the autocorrelation function) of the first three PCs stands out. Because the third PC contains so little variance, we retain only on the first two PCs to construct the reduced phase-space.

A.2 Sensitivity analyses

The choices we have made for p_c , ϵ and t_{\max} in chapter 1 are 0.08, 30 min and 90 min, respectively. The choices should not be based on optimising the PSS alone, as they are system-dependent: some systems require more time accuracy in practise (thus forcing ϵ to be small), while other require a strong reduction in false alarms (lowering p_c). This section is devoted to analysing the results for various choices made in chapter 1 (among which the values of the mentioned parameters). Summarised, it follows that the results are robust and that an intermediate value of the parameters often suffices.

A.2.1 Principal components

Since the PCs — calculated from the data on the ‘red’ and the ‘black’ days — form the backbone of our analysis, their choice must be robust. In order to ensure this within the scope of the full dataset, with respect to various time resolutions, periods and delay severity labels (see SI A.1.3), we compare the results of the first 10 PCs to those calculated from the full dataset (365 days, 1-min time resolution) as a reference, using Pearson’s correlation coefficient. Note that in this comparison between the two sets of PCs, their order (decreasing in variance within each set) may not be the same; hence, the correlations shown are obtained without maintaining the order of the PCs, but maximising these correlations within the first 25 reference PCs. The results are shown in Fig. A.4.

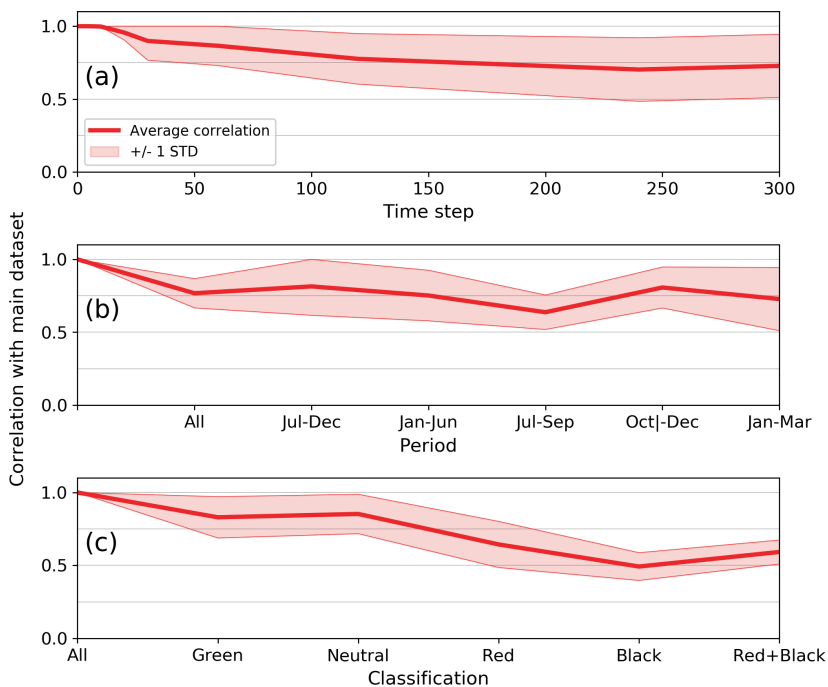


Figure A.4: (Absolute) correlation of the coefficients of the first 10 PCs with the main dataset over various (a) time steps, (b) periods and (c) day-classifications. Average is shown in thick red line, shaded areas reflect +/- one standard deviation. The order of the eigenvectors is optimised in terms of correlation with the main dataset.

Figure A.4a concerns robustness of the PCs with respect to time resolution. For time steps between 1 and 60 minutes, the correlation with the main dataset remains above 0.8 on average, presenting strong evidence of time step robustness. The high robustness up to 60 minutes can be explained by the fact that large amplification (implying most variance) happens on timescales of about an hour. Decreasing the time resolution also increases the amount of variance explained by the first four PCs (not shown), as a result of the fact that long-(time-)scale variance is more structural than short-scale fluctuations.

A second PC robustness check is performed with respect to period within the one year of data; the results of 2 half-year sets and 4 three-months sets are compared to the whole year, providing insight in non-stationarities related to seasonal variation and changes in the timetable. The resulting correlations are shown in Fig. A.4b. The average correlations remain above 0.7, reflecting sufficient robustness. The period October to December seems to correlate the weakest, possibly related to the weather.

The third robustness check focuses on the severity labels (see SI A.1.3), with results shown in Fig. A.4c. The PCs of the green and neutral sets compare best with the main dataset, which makes sense as these are the largest sets. However, this might also point to a bias (of the main set) to describe common small fluctuations rather than large disruptions. Overall, the green, neutral and red sets perform reasonably well, with averages of 0.83, 0.85 and 0.64, respectively. The black days correlate worst, with an average of 0.49, indicating that these black days (few as they are) are only weakly represented by the current principal components.

A.2.2 Clustering

The (sub)clusters calculated in Sec. 1.3.2 need to be checked for robustness against gridding (resolution, grid size and symmetric-logarithmic gridding) and the time lag chosen for the transition matrix that formed the basis for clustering.

The results of the clustering with respect to various choices of the grid resolution is shown in Fig. A.5. The similarities across different resolutions are self-evident, as long as one does not choose too low or too high resolutions. We therefore decided to stick to the middle range, i.e., 123×123 .

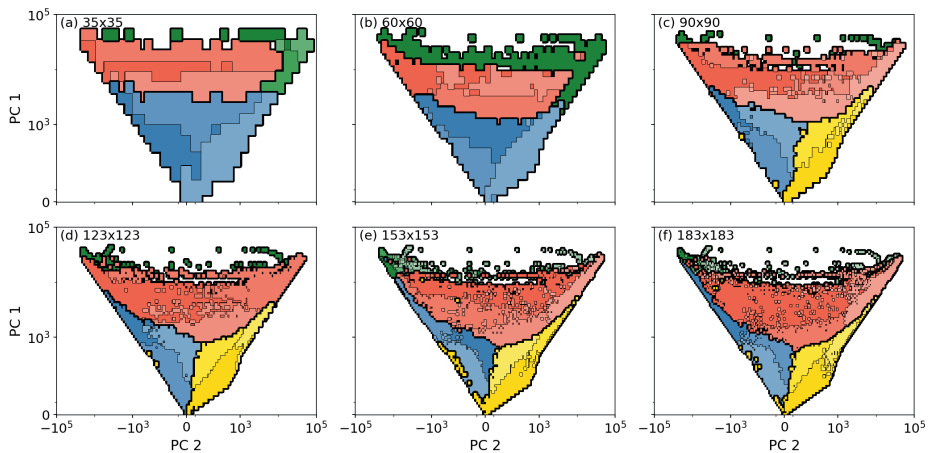


Figure A.5: Phase-space, containing clusters in colours, and subclusters in various shades of the same (arbitrary) colour using transition matrices of various grid resolutions. Areas where there are realisations have been coloured. The used value of time lag τ is 30 min.

Figure A.6 shows the clustering for various time lags used in calculating the transition matrix. Here too the similarities across different time lags are self-evident, as long as one

does not choose too small or too large time lags. We therefore decided to stick to the middle range, i.e., 30 min.

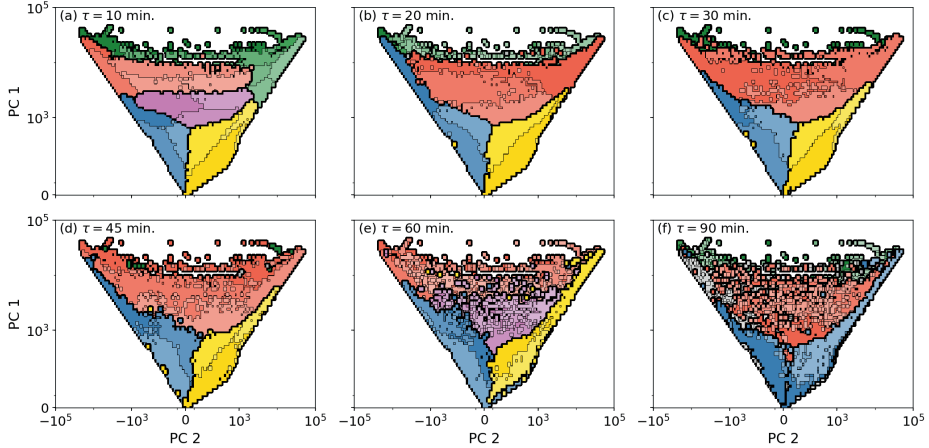


Figure A.6: Phase-space, containing clusters in colours, and subclusters in various shades of the same (arbitrary) colour using transition matrices of various time lags. Areas where there are realisations have been coloured. Used grid resolution is 123x123.

A.2.3 Alarm lags

We continue with checking how various choices of the critical probability affect the time lag τ_{alarm} at which an alarm is given, and in turn, whether an alarm is given at all within the time horizon t_{max} .

Figure A.7 shows the τ_{alarm} for $p_c = 0.01$ and $p_c = 0.25$, to compare with Fig. 1.6 of chapter 1 (where $p_c = 0.08$). For $p_c = 0.01$, almost the whole phase-space is coloured, indicating that, with high enough t_{max} , alarms are practically always given, which is not useful. When p_c increases to 0.25, we see that only a very small part of the phase-space is coloured, limiting predictions only to short times. Intuitively, all of this makes sense: long-term predictions are simply more uncertain. We therefore need to choose p_c somewhere in the intermediate range.

A.2.4 Parameter sensitivity of Peirce skill score

The parameters ϵ , p_c and t_{max} have a great modulating impact on the skill score.

Figure A.8a-c show the dependence of the skill score on ϵ and p_c . The lower ϵ is, the more precise the predictions are in time, but at the cost of skill (i.e., less alarms when there should be alarms). The score also decreases for higher values of p_c (this can be understood as: there will be a lot more missed alarms, since only a few grid cells will give an alarm at high p_c ; see SI A.2.3). In other words, there is an optimum for the skill with respect to p_c .

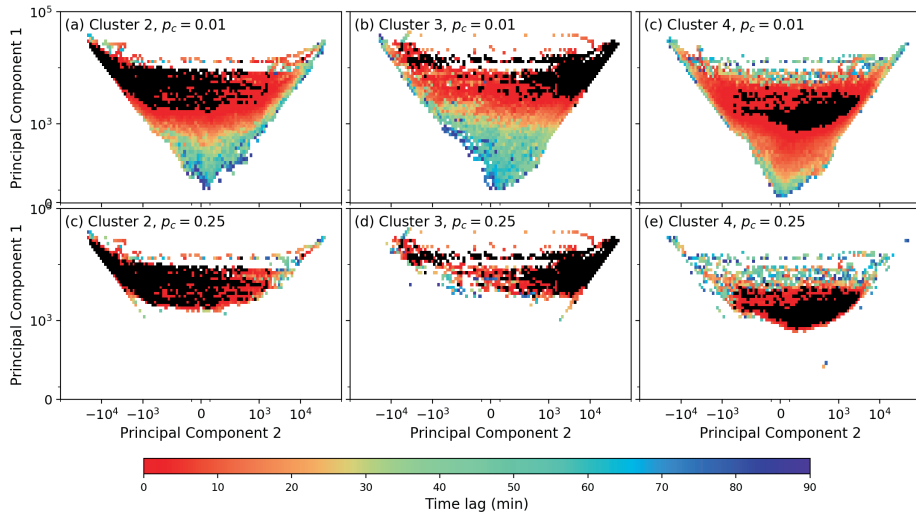


Figure A.7: Same plots as in Fig. 1.6 of chapter 1, but for (a-c) $p_c = 0.01$ and (d-f) $p_c = 0.25$.

Figure A.8d-f reveal the dependence of the skill on the time horizon t_{\max} and p_c . A small time horizon means that one only needs to predict instances of disruption shortly beforehand, which is intuitively a more easily achieved. This is visible in the much higher skill scores at low values of t_{\max} . Again, the feature of increasing, then decreasing skill with respect to p_c is visible in these figures.

Another feature visible in Fig. A.8 is that the skills score curves of subclusters 2 and 3 do not differ much with respect to each other, while the curves of subcluster 4 are different: primarily in the form of the shift of the optimum toward higher p_c values. This can be explained by the fact that false alarms are important in the signal of this subcluster: for a p_c value of 0.08, alarm is given already at 79% of the phase-space. An increase in the value of p_c reduces the part of the phase-space where an alarm is given, which also reduces the false alarms.

A final feature in Fig. A.8a-c is the median τ_{alarm} . Up to $p_c \approx 0.15$, more than half of all alarms given are above 15 min. This is connected to the choice of ϵ , which should be smaller than the alarm lag given. If this is not the case, then the alarm's precision in time becomes arbitrary.

A.3 Case study of a regular day

Here, we discuss the performance of the prediction scheme if there are no instances when the system reaches the disrupted state. Figure A.9 shows the system's trajectory in the phase-space on April 27, 2018, along with all prediction outcomes towards subclusters 2 and 3.

The system does not attain high PC1 values and remains within the lower subclusters, mainly in the positive PC2 area (i.e., delays mostly on L1 and L3). It does not reach sub-

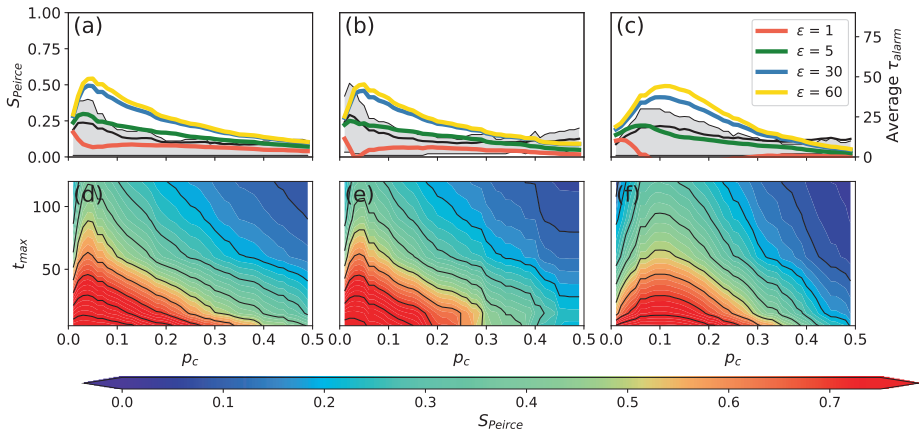


Figure A.8: Peirce Skill Score for the predictions done towards entering subclusters (a,d) 2, (b,e) 3 and (c,f) 4. Top: Peirce Skill Score for different values of p_c (x-axis) and ϵ (in coloured solid lines, legend). Black line (right axis) indicates the median lag at which an alarm is given (i.e., average τ_{alarm} , right axis), including a band width of 25% and 75% percentiles, shown in grey shading. A time horizon of $t_{\text{max}} = 90$ minutes is chosen. Bottom: Peirce Skill Score for different values of p_c and t_{max} . A band width of $\epsilon = 30$ minutes is chosen.

clusters 2 and 3 at all, but does enter subcluster 4 twice (as noted earlier, this is common for regular days). In the right panels, we see that in the night or early morning, the system stays ‘quiet’: correctly, almost no alarms are given, resulting in a lot of correct rejections (blue dots). Later in the day, delay builds up and after a lot of false alarms, the metric correct predicts (‘hits’) the entrance to subcluster 4.

Correct predictions to enter subcluster 4 are made up to 56 minutes in advance, with a band width of $\epsilon = 30$ minutes. The time horizon τ_{max} is 90 minutes, so there are no correct predictions between 56 and 90 minutes prior to the entrance of the subcluster. This actually reflects the situation on regular days: the system may touch upon, or even reach, subcluster 4, but it remains difficult to predict. As the choice of $p_c = 0.05$ is relatively low, a lot of false alarms (58%) are given, while no missed alarms are registered. The absence of missed alarms combined with the occurrences of hits maximises the ratio $\#H/\#O$ (i.e., it becomes 1), increasing the Peirce skill to a positive score despite the number of false alarms. However, while predicting the entrance in subclusters 2 and 3, we obtain a negative PSS due to the fact that no Hits are predicted (it never reaches these subclusters).

The right panels of Fig. A.9 show the evolution of the total delay and cancellations. For cancellations we use the number of train activities (departure, arrival, short stops etc.) that were scheduled, but cancelled, per minute. This is a measure of the reduction of ‘stress’ on the system by human decision. Note that there is a strong increase in total delay roughly from 19:00 onward, coinciding with the entrance to subcluster 4. The signal in cancellations per minute is however difficult to interpret. It seems that during two events, roughly at 06:00 and 13:00, there were a lot of cancellations; we could not retrieve the reasons behind these cancellations. It could be that these cancellations avoided the entrance to subcluster 4, by limited the system’s overall delay.

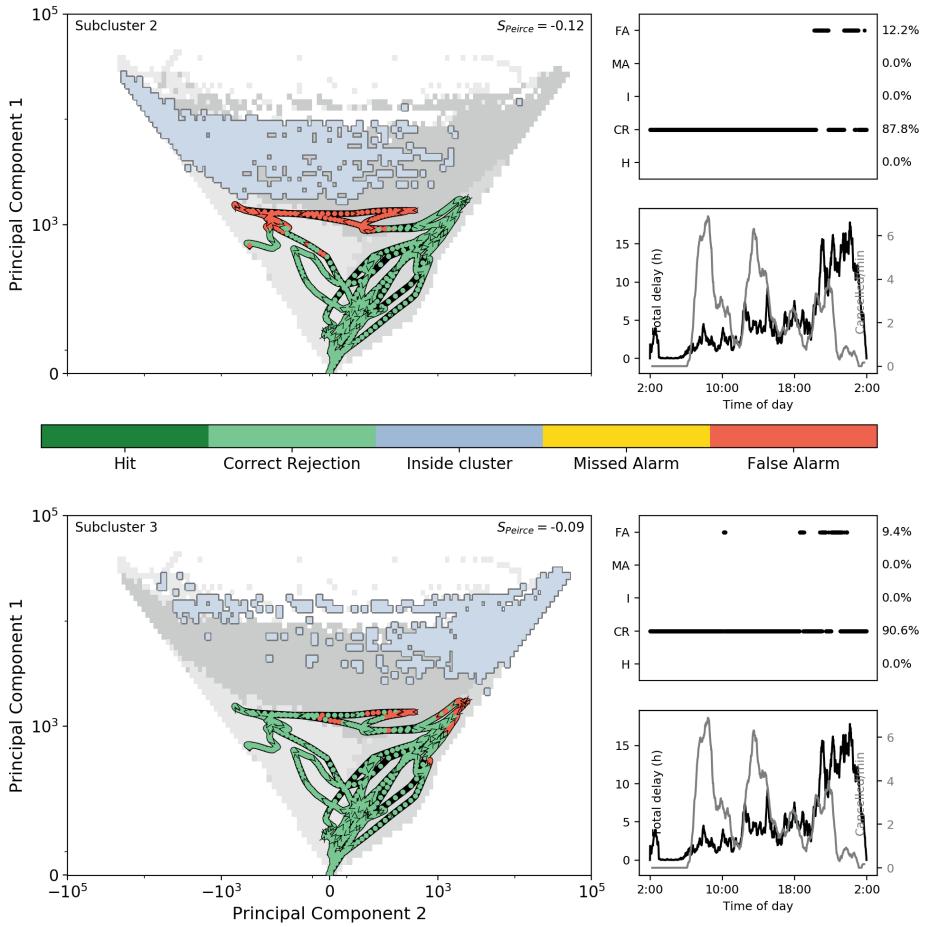


Figure A.9: Same as in Fig. 1.7 of chapter 1, but for the case of April 27, 2018.

Appendix B

Supplementary information to Chapter 2

B.1 Example time series

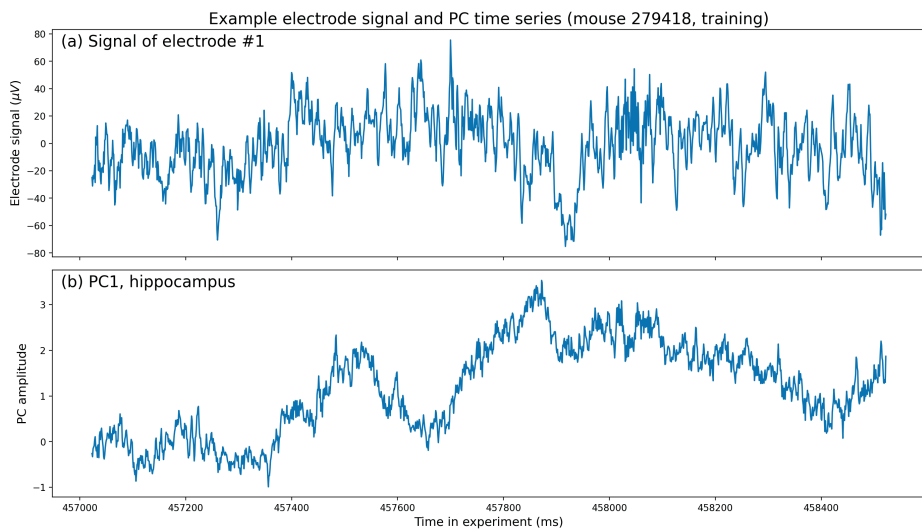


Figure B.1: Example time series of (a) the raw electrode signal, and (b) a principal component (PC1 Hippocampus) of mouse 278418, the training dataset. Only 1500 ms is shown out of approximately 600,000 ms in total.

Supplementary information to Chapter 3

C.1 Parameter and simulation settings

	Double fold - Eq. (3.3)	Fold-Hopf - Eq. (3.4)	Hopf-fold - Eq. (3.5)	Double Hopf - Eq. (3.6)
Leading system	$\phi_c = \pm 0.19$ (Fold) Bistable for $ \phi < \sqrt{\frac{-4a_1^3 a_2^3}{27a_1^4}}$ (if $a_1 < 0, a_2 > 0$)	$\phi_c = \pm 0.38$ (Fold) Bistable for $ \phi < \sqrt{\frac{-4a_1^3 a_2^3}{27a_1^4}}$ (if $a_1 < 0, a_2 > 0$)	$\phi_c = 0$ (Hopf) Oscillatory for $\phi > 0$ (if $a_1 b_1 < 0$)	$\phi_c = 0$ (Hopf) Oscillatory for $\phi > 0$ (if $a_1 b_1 < 0$)
Coupling	$\gamma = 0.48x$	$\gamma = -0.1 + 0.12x$	$\gamma = 0.05 + 0.5x$	$\gamma = -0.05 + 2x$
Following system	$\gamma_c = \pm 0.54$ (Fold) Bistable for $ \gamma < \sqrt{\frac{-4b_1^3 b_2^3}{27b_1^4}}$ (if $b_1 < 0, b_2 > 0$)	$\gamma_c = 0$ (Hopf) Oscillatory for $\gamma > 0$ (if $b_1 c_1 < 0$)	$\gamma_c = \pm 0.38$ (Fold) Bistable for $ \gamma < \sqrt{\frac{-4c_1^3 c_2^3}{27c_1^4}}$ (if $c_1 < 0, c_2 > 0$)	$\gamma_c = 0$ (Hopf) Oscillatory for $\gamma > 0$ (if $c_1 d_1 < 0$)
Parameters	$a_1 = -0.5$ $a_2 = 0.5$ $b_1 = -0.5$ $b_2 = 1.0$	$a_1 = -1$ $a_2 = 1$ $b_1 = b_2 = 1$ $c_1 = -1; c_2 = 1$	$a_1 = 0.05; a_2 = 1$ $b_1 = -0.05; b_2 = 1$ $c_1 = -1$ $c_2 = 1$	$a_1 = 0.04; a_2 = 2$ $b_1 = -0.04; b_2 = 2$ $c_1 = 0.4; c_2 = 1$ $d_1 = -0.4; d_2 = 1$

Table C.1: Parameter values and coupling for the four types of cascading tipping as shown in Figs. 3.1 and 3.2.

C.2 Simulations

In this section we discuss the previously described metrics applied in ensemble simulations of cascading transition events. This provides insight into the statistical characteristics of these events, the causal relation between tipping of both systems and the potential prediction of these events. We focus on the double-fold and the fold–Hopf cascading tipping cases for multiple reasons. First, these cases are most illustrative in terms of relation to physical systems. Second, in these cases the leading system starts with a fold bifurcation, which creates a clear threshold for the start of the event (which is convenient for analysis purposes).

	Double fold - Eq. (C.1) (following system tips)	Double fold - Eq. (C.1) (following system does not tip)	Fold-Hopf - Eq. (C.2)
Forcing and coupling	$\phi(t) = 0.0012t$ $\gamma(x) = 0.05 + 0.37x$	$\phi(t) = 0.0012t$ $\gamma(x) = 0.05 + 0.37x$	$\phi(t) = 0.002t$ $\gamma(x) = -0.2 + 0.3x$
Parameters	$a_1 = -0.5$ $a_2 = 0.5$ $b_1 = -0.5$ $b_2 = 1.0$	$a_1 = -0.5$ $a_2 = 0.5$ $b_1 = -0.25$ $b_2 = 1$	$a_1 = -1$ $a_2 = 1$ $b_1 = 0.1; b_2 = 1$ $c_1 = -0.5; c_2 = 1$
Integration time	$t_{\max} = 500$ $\Delta T = 0.5$	$t_{\max} = 500$ $\Delta T = 0.5$	$t_{\max} = 500$ $\Delta T = 0.5$
Noise	Noise mean = 0 Noise variance = 0.1	Noise mean = 0 Noise variance = 0.1	Noise mean = 0 Noise variance = 0.1
Initial conditions	$(x_0, y_0) = (-0.8, -1)$	$(x_0, y_0) = (-0.8, -1)$	$(x_0, y_0, z_0) = (-0.5, 1, -1)$

Table C.2: Parameter values, coupling and initial conditions for the ensemble simulations of the Double Fold and Fold-Hopf systems as shown in Figs. C.1, C.2 and C.3.

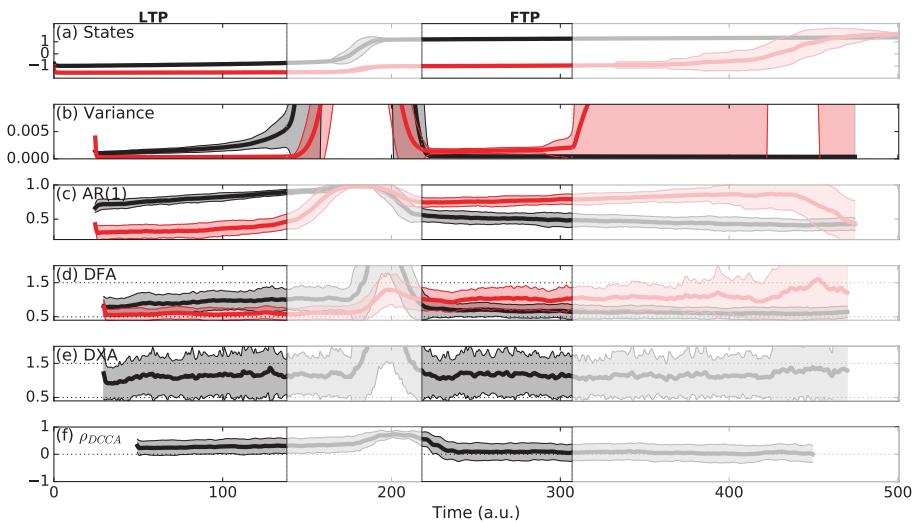


Figure C.1: Ensemble (100-member) simulations of a dynamical system undergoing a double-fold cascade (Eq. (C.1)) where both systems undergo a transition, parameter values as in Tab. C.2. **Panel (a):** states of x (black) and y (red). **Panel (b):** variance of x (black) and y (red). **Panel (c):** autoregressive coefficient at lag 1 of x (black) and y (red). **Panel (d):** detrended fluctuation analysis scaling exponent of x (black) and y (red). **Panel (e):** detrended cross-correlation analysis scaling exponent. **Panel (f):** detrended cross-correlation coefficient by Zebende (2011). White-shaded areas indicate windows containing the actual transitions. The increased variance, AR(1) and DFA scaling exponent prior to transition in the leading system and following system, respectively confirms the predicted increased memory through critical slowing down.

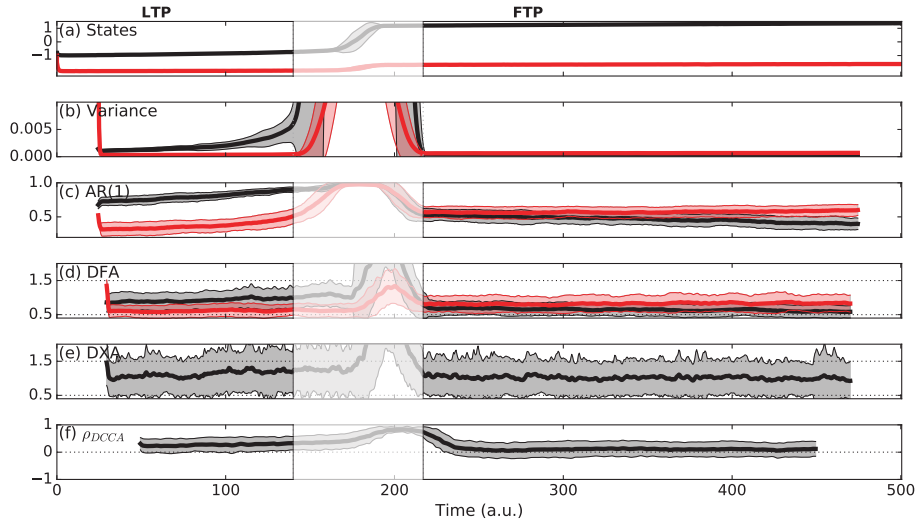


Figure C.2: As in Fig. C.1, but without any transition in the following system. In this case, only the leading system has a transition. Parameter values are given in Tab. C.2.

Double-fold cascading tipping

To simulate these events and use statistical indicators, noise has to be included. The system of equations used here is

$$\begin{cases} \frac{dx}{dt} = a_1 x^3 + a_2 x + \phi + \zeta_x \\ \frac{dy}{dt} = b_1 y^3 + b_2 y + \gamma(x) + \zeta_y \end{cases} \quad (\text{C.1})$$

where in addition to Eq. (3.3), ζ_x, ζ_y are Gaussian white noise terms. We simulate an ensemble of 100 members with the parameter settings and initial conditions as displayed in Tab. C.2. The results of this ensemble are displayed in Fig. C.1. Running windows containing the transition are shaded white as these data are misleading when one wants to know what happens before the bifurcation points. We make the distinction between the leading-transitional period (LTP), which is the time series before the tipping point in the leading system, and the following-transitional period (FTP), which is the time series between the first tipping point and the tipping point in the following system. The FTP can be seen as an in-between state that is stable for both systems, although the following system is close to bifurcation, meaning it is rather prone to noise. Therefore, the duration of this state is highly unpredictable. However, as we will see in this section, the FTP does provide information regarding the diagnosis of a potential second transition.

In the LTP, we clearly see the gradual increasing leading system's variance, the AR(1) coefficient and DFA scaling coefficient. These are all evidence of the leading system slowly approaching a bifurcation point, according to the change in the parameter. There is not much evidence of long-range auto-correlations in the time series of the following system, as its variance is low and the DFA scaling exponent remains below 0.5, pointing towards

the fact that the detrended fluctuations are statistically white noise. The AR(1) coefficient of the following system does increase just prior to the first tipping, but it also stays small (compared to unity).

The detrended cross-correlation scaling exponent (abbreviated here as DXA) does give > 0.5 values, but the range throughout the ensemble members is too large most of the time to see any structural development when approaching the bifurcation point. During the leading system's transition, a strong increase is visible, pointing towards rather strong cross-correlated behaviour during this period (as the following system also shifts its equilibrium a little).

The quantity ρ_{DCCA} seems to attain a small positive value (around 0.3) and stays relatively constant throughout the whole time interval. One important aspect of the calculation of ρ_{DCCA} , as we found by experimentation, is that the values are very sensitive to the segment size s and the moving window size. The moving window determines the amount of data that is available to find long-range correlations, and the segment size has a strong impact on the accuracy of the fits and on the related segmented fluctuations. Due to the fact that we need a temporal evolution of the statistical indicators in this type of problem, we need moving windows and, thus, encounter a problem. As these indicators (DXA and ρ_{DCCA}) have been successfully applied in simpler systems (Zebende, 2011; Podnobik and Stanley, 2008; Zhang et al., 2001; Zhou, 2008), more research on the sensitivity of the indicators with respect to the segment size and moving window size may lead to more robust results.

During the FTP, the variance, AR(1) and DFA of the leading system are strongly reduced. However, the gradual increase of the following system's variance, AR(1) coefficient and DFA scaling coefficient are definitely visible, pointing towards the approach of a bifurcation in the following system. Also notable is the contrast in the DFA of the following system before and after the tipping of x . The DFA of y goes from a white-noise regime (around 0.5) before the tipping of x towards a regime where the detrended fluctuations point to long-range auto-correlations after the tipping of x (1–1.5). This illustrates the relation between the leading system's state and the following system's DFA scaling exponent. The DXA remains relatively high, but overall no structural development is seen in this graph. The quantity ρ_{DCCA} exhibits the same behaviour as in the LTP, most likely due to the previously mentioned reasons. To assess the consequences of the cascading effect on the statistics mentioned, we compare the results with a case where the system (Eq. (C.1)) does not undergo a tipping in the following system (so only one transition remains). The resulting ensemble results are shown in Fig. C.2. The most important differences between a regular cascading event and a single tipping event can be found when comparing the variance, AR(1) and DFA scaling coefficient changes between LTP and FTP (or the period after the first transition). During the LTP, the leading system is close to transition and has strong auto-regressive behaviour, whilst the following system is far from its bifurcation point. During the FTP, the following system generally gains memory because it is brought closer to its transition point, and the leading system loses memory because it just arrived at a new state. Therefore, we expect that from the LTP towards the FTP, the variance, AR(1) and DFA will decrease in the leading system, and increase in the following system. To quantify this effect, Tab. C.3 shows the ratios of the different quantities, indicated by Q , during the FTP and LTP phases ($\bar{Q}_{FTP}/\bar{Q}_{LTP}$), for the cases *with* a second tipping (corresponding to runs shown in Fig. C.1) and *without* second tipping

Ratios	Mean	Standard deviation
With second tipping		
Leading variance	0.24	0.44
Leading AR(1)	0.62	0.09
Leading DFA	0.79	0.27
Following variance	3.95	1.53
Following AR(1)	1.92	0.39
Following DFA	1.70	0.49
Without second tipping		
Leading variance	0.15	0.06
Leading AR(1)	0.60	0.08
Leading DFA	0.74	0.23
Following variance	1.63	0.42
Following AR(1)	1.34	0.31
Following DFA	1.40	0.41

Table C.3: Comparison of the ratios of autoregressive variables prior to and after the first transition, using the ensembles shown in Fig. C.1 and Fig. C.2.

t-test variable	T-statistic	dF	p-value*
Leading variance	1.62	101.67	0.11027
Leading AR	1.32	197.07	0.18904
Leading DFA	1.23	190.60	0.22264
Following variance	13.73	112.38	0.0
Following AR	9.93	184.12	0.0
Following DFA	4.13	190.17	0.00006

Table C.4: Results of Student's t test on the differences between the ratios (in Tab. C.3) of the cases with and without second tipping. * p -values calculated using the `scipy.stats` Python package.

(Fig. C.2). All ensemble members are included in these values, accounting for a mean and standard deviation of the ratios. As expected, the leading system's auto-regressive metrics decrease in both cases, which is visible in the fact that the mean values of the ratios of the leading system's auto-regressive variables are lower than 1. Furthermore, as expected, the following system's auto-regressive behaviour increases (ratios > 1) in both cases; however, it is striking that in the case of a cascading tipping event (with second tipping), the following system's ratios are much higher than those in the case of a single tipping event (without second tipping). To investigate whether the difference in the ratios between single or double tipping is indeed significant, a Student's t test was carried out. The results are shown in Tab. C.4. The high p -values for the leading system's ratios indicate no significant difference between single or double tipping, but the low p -values for the following system's ratios indicate a significant difference. This shows the potential of using the ratio of auto-regressive variables before and after a transition to assess whether a cascading transition may follow. Further research is needed to quantify this

expectation and to assess the sensitivity of these ratios to the system's parameters.

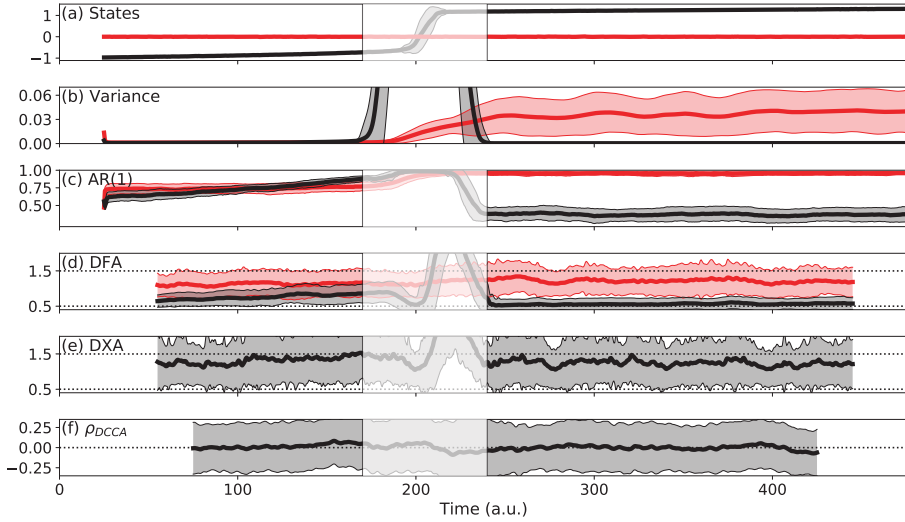


Figure C.3: As in Fig. C.1, but for the fold-Hopf cascade (Eq. (C.2)). Parameter values are given in Tab. C.2.

Fold-Hopf cascading tipping

Many statistical indicators have been specifically applied on fold bifurcations, because these transitions show a clear sign of critical slowing down and increased auto-correlation; this is due to their irreversibility and the process of going from one equilibrium towards another. A state approaching a Hopf bifurcation reacts differently to perturbations than a state approaching a saddle-node bifurcation. Therefore, we will consider the fold-Hopf cascade in the light of the previously described statistical indicators. For this, we use the following stochastic dynamical system:

$$\begin{cases} \frac{dx}{dt} = a_1 x^3 + a_2 x + \phi + \zeta_x \\ \frac{dy}{dt} = b_1 z + b_2 (\kappa(x) - (y^2 + z^2))y + \zeta_y \\ \frac{dz}{dt} = c_1 y + c_2 (\kappa(x) - (y^2 + z^2))z + \zeta_z \end{cases} \quad (\text{C.2})$$

This system is similar to Eq. (3.4), but white noise is added through the terms ζ_x, ζ_y and ζ_z . We used an ensemble of 100 simulations with the parameter settings and initial conditions as displayed in Tab. C.2. The results of the ensemble are shown in Fig. C.3. Here, we do not make the distinction between the LTP and the FTP, because in contrast to the double-fold cascade, the following system undergoes a transition that is easily reversed and the system is either stationary or oscillating. We subtract a running average from the states and calculate the statistics from those series to prevent the oscillation from dominating the signal in the auto-correlation. The following system (red) is quickly drawn towards the equilibrium state $(y, z) = (0, 0)$, and the leading system (black) is in a steady

state. During the time leading up to the bifurcation point, the variance, AR(1) coefficient and DFA of the leading system x gradually increase, as is expected as we force this system towards its bifurcation point. It also seems that the AR(1) of the following system slightly increases during this period.

After the transition of the leading system, the oscillation of the following system immediately starts due to noise. The variance and AR(1) of the following system after the bifurcation are strongly increased with respect to before the bifurcation, despite the removal of the running average to eliminate the oscillation's signal. On average, the DFA scaling exponent of the following system also increases, which relates it to the leading system's state. The DXA sharply increases just prior to the critical transition, although it retains relatively high values throughout the whole time series. The reason behind this might be the low level of noise that is chosen, or other simulation-specific parameters. It could also be due to the fact the following system has a high, weakly varying DFA scaling exponent on average, which in turn might affect the height and variability of the cross-correlation. The ρ_{DCCA} coefficient remains positive and small, as in the double-fold case. Again, this may have to do with the choice of window and segment sizes.

Appendix D

Supplementary information to Chapter 4

D.1 Data sources and processing

D.1.1 Data sources

Operational data for the Netherlands is obtained from the Dutch infrastructure manager ProRail and covers the period July 2018 to June 2019. A selection of 16 days is made, based on delay severity (4 days in each of four categories; Green, Neutral, Red and Black, for more details see the SI F and chapter 1), taken to statistically cover multiple situations. Network data on locations and links of the nodes and edges of the railway network, including delay time series can be found in an Open Science Framework related to previous work (Dekker et al., 2019): <https://osf.io/tps4r/>. Data for the German and Italian railways are obtained from the supplementary material of Monechi et al. (2018), who received this data from the OpenDataCity (<http://www.opendatacity.de/>) and the ViaggiaTreno (<http://www.viaggiatreno.it/>) websites. The German data spans across March 2015, and the Italian data across March 2015 and April 2015. The Swiss railway data is obtained from the OpenTransportData website (<https://opentransportdata.swiss/>) over January 2018. The German data consists of several stations across the border in mainly the south and west, which are strongly connected to its lines.

D.1.2 Data cleaning

From the infrastructure data, the network topologies of these systems is determined. In all but the Dutch case, the data had to be cleaned for unrealistically long edges. These edges arise from trains travelling long distances between major cities without stopping in subsequent smaller stations, effectively adding direct edges between major cities that in practice are not direct edges. This cleaning is done by looking at the distribution of edge lengths. Also, because the timetable may vary from day-to-day, I set the minimum train frequency on any to be 1 per hour, and scheduled running times to be at least 1 minute, to avoid values of α to become zero or infinity.

Furthermore, it is important to note that there are disconnected (e.g. Sardinia in the

Italian case) and non-operational components in some of the railway systems analysed in chapter 4. These parts are omitted - we focus on the giant component and where necessary, make all links bidirectional. Note that these problems only concern a small fraction of the links in these systems.

D.1.3 Train types

We focus on passenger trains only, and exclude freight trains in the analysis. One reason for this is that freight trains are (economically) privacy sensitive, meaning that it is difficult to get a complete dataset. Another reason is that delay is not always well-defined for such trains, as their routes and schedules are mostly separated from the main schedule. It should be noted that their contribution to delay is usually only minor, as, for example in the Netherlands, they cover only 5.7% of all Dutch train kilometres in 2017 (numbers courtesy of the Dutch infrastructure manager ProRail). More details can be found in Dekker et al. (2019), section 3.1.

Appendix E

Supplementary information to Chapter 5

E.1 Model derivations

In this section we explain how we arrive at an equation for the evolution of the time delays D_i , defined on the stations of the network. Recall that the evolution over a time Δt is written in general as

$$D_i(t + \Delta t) - D_i(t) = F_{1,i}(\vec{D}) - F_{2,i}(\vec{D}),$$

where \vec{D} is the vector with all delays in the network. The two terms above describe how the delay on a station can increase and decrease respectively. We will describe how we obtain an explicit formula for these terms below.

E.1.1 Expressing the delay of a train as function of the delay on a station

Since we aggregate delay on stations but have data on train flows, we need a way of expressing a train's delay d as function of the station's delay D it is contributing to. In other words, we need to find a function f such that:

$$d_T(t) = f(D_{S(T,t)}(t)),$$

where we have introduced the new notation $S(T, t)$, which is the station *towards which* train T is moving at time t . By definition, the delay of a station is the sum of the delays of trains moving to that station. Under assumption that all trains currently moving to i *equally contribute* to the delay D_i , this leads to:

$$d_T(t) = \frac{D_{S(T,t)}(t)}{\# \text{ trains moving to } S(T, t) \text{ at time } t} = \frac{D_{S(T,t)}(t)}{\sum_{\text{edges } e \text{ to } S(T,t)} f_e t_e}, \quad (\text{E.1})$$

where to compute the number of trains moving towards a station i , we use the formula

$$\# \text{ trains moving to } i \text{ at time } t = \sum_{\text{edges } e \text{ to } i} f_e t_e = \sum_{\ell \in \mathcal{N}_{\text{in}}(i)} f_{\ell i} t_{\ell i}. \quad (\text{E.2})$$

Here $e = (\ell, i)$ refer to edges, f_e denotes the frequency of trains on edge e and t_e average time it takes to travel edge e . The product of f_e and t_e is then the average number of trains on this edge. By $\mathcal{N}_{\text{in}}(i)$ we denote the in-neighbour of station i , i.e., the stations which have edges directed towards i (similarly we denote the out-neighbours by $\mathcal{N}_{\text{out}}(i)$). For example, if a station has only one edge outgoing towards it with 2 trains per hour and it takes 15 minutes to cross the edge, on average there is 1/2 train on this edge all the time.

E.1.2 First term: New incoming trains

First we compute $F_{1,i}(\vec{D})$, the contribution of new trains on their way to i . These are trains that, in the time interval $[t, \Delta t]$, passed a neighbour $\mathcal{N}_{\text{in}}(i)$ of i and are now moving towards i . At time t , these train's delay was counting towards one of the in-neighbours of i , but at time $t + \Delta t$ it counts towards D_i . The total contribution is

$$\begin{aligned} F_{1,i}(\vec{D}) &= \sum_{\substack{\text{Trains } T \text{ that started on a track to } i \\ \text{in the time } \tau \in [t, t + \Delta t]}} d_T(\tau) = \\ &= \sum_{j \in \mathcal{N}_{\text{in}}(i)} \sum_{\substack{\text{Trains } T \text{ to } j \text{ that hit } j \\ \text{and continue to } i \\ \text{in } \tau \in [t, t + \Delta t]}} d_T(\tau) \\ \text{Eq. (E.1)} \quad &= \sum_{j \in \mathcal{N}_{\text{in}}(i)} \sum_{\substack{\text{Trains } T \text{ to } j \text{ that hit } j \\ \text{and continue to } i \\ \text{in } \tau \in [t, t + \Delta t]}} \frac{D_j(t)}{\sum_{\ell \in \mathcal{N}_{\text{in}}(j)} f_{\ell j} t_{\ell j}} = \\ &= \sum_{j \in \mathcal{N}_{\text{in}}(i)} \# \text{Trains that come into } j \text{ and continue to } i \frac{D_j(t)}{\sum_{\ell \in \mathcal{N}_{\text{in}}(j)} f_{\ell j} t_{\ell j}} = \\ &= \sum_{j \in \mathcal{N}_{\text{in}}(i)} \# \text{Trains that hit } j \times \frac{\text{Fraction of trains that come into } j \text{ and continue to } i}{\sum_{\ell \in \mathcal{N}_{\text{in}}(j)} f_{\ell j} t_{\ell j}} \frac{D_j(t)}{\sum_{\ell \in \mathcal{N}_{\text{in}}(j)} f_{\ell j} t_{\ell j}} \quad (\star) \end{aligned}$$

We now make an approximation: we assume that the fraction of trains that continue from j to i can be computed as follows:

$$\text{Fraction of trains that come into } j \text{ and continue to } i = \frac{f_{ji}}{\sum_{\ell \in \mathcal{N}_{\text{out}}(j)} f_{j\ell}} \times (1 - \text{Probability that train has end station at } j) =: r_{ji} \times (1 - s_j) =: p_{ji}.$$

This is a strong assumption, as also discussed in chapter 5: here, we assume that trains, when arriving at any station, pick a random direction on which to continue next, and the probability is weighted by the overall frequency of trains on the edges. In reality, where the train comes from matters for where it will go to. Note that if $f_{ji} = \text{constant}(j)$ inde-

pendent on i , i.e. all tracks starting from j have the same frequency of trains, then

$$\frac{f_{ji}}{\sum_{\ell \in \mathcal{N}_{\text{out}}(j)} f_{j\ell}}$$

reduces to $(k_{j,\text{out}})^{-1}$, the out-degree of station j .

We also approximate the number of trains that reach station j in the time interval $[t, t + \Delta t]$:

$$\# \text{Trains that hit } j = \Delta t \times \sum_{\ell \in \mathcal{N}_{\text{in}}(j)} f_{\ell j}.$$

In other words, we go over each edge towards j and assume that the number of trains on each edge e that reaches j in a time interval Δt is given by $f_e \Delta t$. For example, if there are 6 trains per hour on a given edge, we expect that in an interval of 20 minutes, 2 trains reach the end of the edge.

So now we have

$$\star = \sum_{j \in \mathcal{N}_{\text{in}}(i)} \left(\Delta t \sum_{\ell \in \mathcal{N}_{\text{in}}(j)} f_{\ell j} \right) \times p_{ji} \times \left(\frac{D_j(t)}{\sum_{\ell \in \mathcal{N}_{\text{in}}(j)} f_{\ell j} t_{\ell j}} \right) = \Delta t \sum_{j \in \mathcal{N}_{\text{in}}(i)} p_{ji} B_j D_j(t),$$

where we set

$$B_j := \frac{\sum_{\text{edges } e \text{ to } j} f_e}{\sum_{\text{edges } e \text{ to } j} f_e t_e} = \frac{\sum_{\ell \in \mathcal{N}_{\text{in}}(j)} f_{\ell j}}{\sum_{\ell \in \mathcal{N}_{\text{in}}(j)} f_{\ell j} t_{\ell j}}.$$

E.1.3 Term 2: Removal of trains

The delay of trains that arrive to station i in the interval $[t, t + \Delta t]$ has to be subtracted from D_i (since now this train's delay is not counting towards D_i , but towards the next station on this train's route). We have

$$F_{2,i}(\vec{D}) = \sum_{\substack{\text{Trains } T \text{ that reached } i \\ \text{in the time } \tau \in [t, t + \Delta t]}} d_T(\tau) \stackrel{\text{Eq. (E.1)}}{=} \frac{D_i(t)}{\sum_{\ell \in \mathcal{N}_{\text{in}}(i)} f_{\ell i} t_{\ell i}} \times \# \text{trains that reach } i \text{ in interval } [t, t + \Delta t].$$

The number of trains that arrive to i in the interval $[t, t + \Delta t]$ is

$$\sum_{\text{edges } e \text{ to } i} f_e \Delta t = \Delta t \sum_{\ell \in \mathcal{N}_{\text{in}}(i)} f_{\ell i},$$

such that we now can state that

$$F_{2,i}(\vec{D}) = \frac{\sum_{\ell \in \mathcal{N}_{\text{in}}(i)} f_{\ell i}}{\sum_{\ell \in \mathcal{N}_{\text{in}}(i)} f_{\ell i} t_{\ell i}} D_i(t) \Delta t \stackrel{\text{Eq. (5.5)}}{=} \Delta t B_i D_i(t).$$

E.2 Disrupted days

We apply our model to highly disrupted day. They were identified as follows: for each day between January 2019 and May 2020, we calculated the peak total amount of delay in the

network. We then took the 50 most delayed days. Most days are from 2019: Jan 15, Jan 18, Jan 21, Jan, 22, Jan 23, Jan 30, Jan 31, Feb 4, Feb 25, Mar 10, Apr 2, Apr 29, May 2, May 3, May 14, May 16, May 27, Jun 5, Jun 19, Jun 24, Jun 25, Jun 27, Jul 23, Jul 24, Jul 25, Oct 2, Oct 6, Oct 15, Oct 28, Nov 4, Nov 8, Nov 13, Nov 14, Nov 18, Nov 19, Nov 20, Nov 21, Nov 26, Nov 28, Dec 2, Dec 4, Dec 5, Dec 11 and Dec 27. Some of them were from 2020: Feb 9, Feb 10, Feb 15, Feb 25, Mar 17, May 4.

Figs. E.1, E.2 and E.3 show the situation for these highly disrupted days, at the moment of peak delay. These images were obtained as follows. From the data, we have a list (s, d) of stations and the associated delay at the peak moment. We also have the longitude x and latitude y of each station. We assume that the dataset consisting of (x, y, d) is a sample of a geographical distribution with weights given by the delay. We then use kernel density estimation with a Gaussian kernel with bandwidth 0.1 to obtain a smooth approximation to this distribution. We use the class `KernelDensity` from the Python package `scikit-learn` for this. Note that the weights must be positive. We therefore set all delays that are negative (when trains are ahead of schedule) or zero to a small value (0.001). Note that the colour on these maps does not hold absolute quantitative information. It only shows the relative magnitude of the delay in different locations. Since the colours are normalised for each day separately, colours between different days cannot be directly compared.

The maps indicate the time of peak delay on top (up to 30 second resolution) and the total delay in the network at the bottom. The cross marks the location of the station Brussel-Centraal. Each map also mentions in which group we classified the disrupted situation. BXL means that high delays only appear around Brussels, the capital and central hub of the railway network. OP means that there is a single localised peak of delay at another location than Brussels. Situations classified as MP have multiple distinct peaks and finally S stands for ‘spread out’: in these situations the delay is present in large regions.

E.3 Non-trivial optimal level of K

Figure 5.6 indicates that on a subset of the 50 disrupted days, performance of the model is optimal for values of K around 10. In this section we discuss possible explanations for this. From a theoretical perspective, we expect the model performance to increase with decreasing K , as on a more coarse-grained level, the mean-field approximations in the model are more valid. For example, delay propagation in all possible directions weighted by schedule-based parameters like train frequencies mainly works when there are a lot of trains going in many different directions — the train frequency then becomes a good weighting factor of where delay really goes. The increasing performance with decreasing K well visible in Fig. 5.6.

Still, on some days, the performance drops when K becomes too high. To investigate possible explanations, we show the clustering configuration at $K = 10$ in Fig. E.4. The clusters are approximately equal in size and well distributed across the country. The major cities of Brussels, Antwerp, Liège, Ghent and Charleroi are all located in separate clusters (i.e., 3, 9, 2, 1 and 6, respectively). These cities are also roughly in the centre of each of these clusters. Notably, all of the disrupted situations with an optimal performance around $K = 10$ fall in the group of situations with delay localised around Brussels

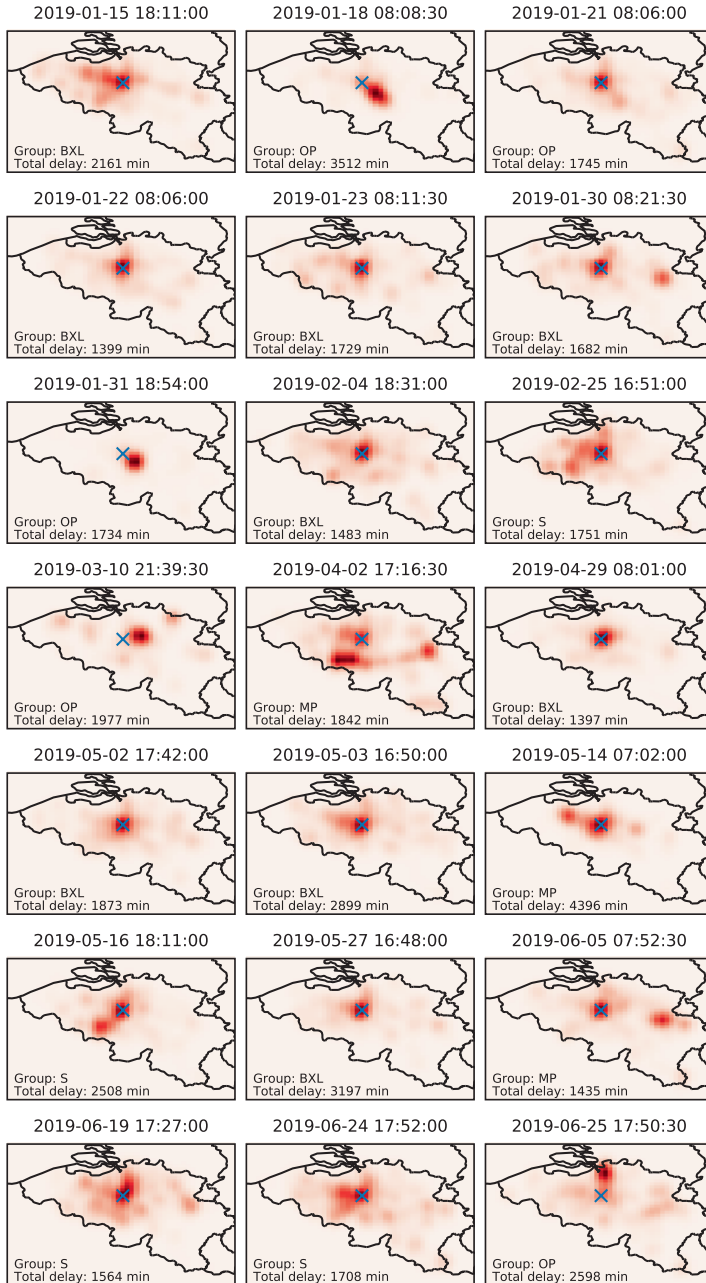


Figure E.1: Spatial distribution of time delays on disrupted days.

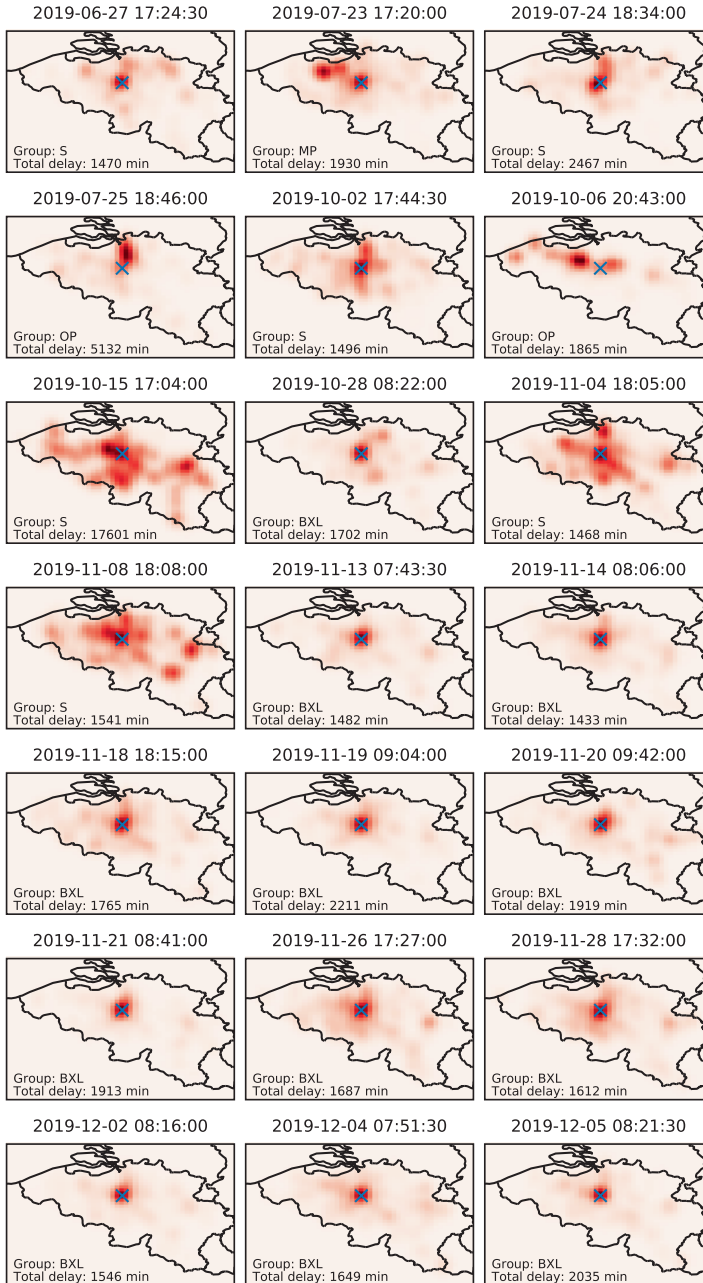


Figure E.2: Spatial distribution of time delays on disrupted days.

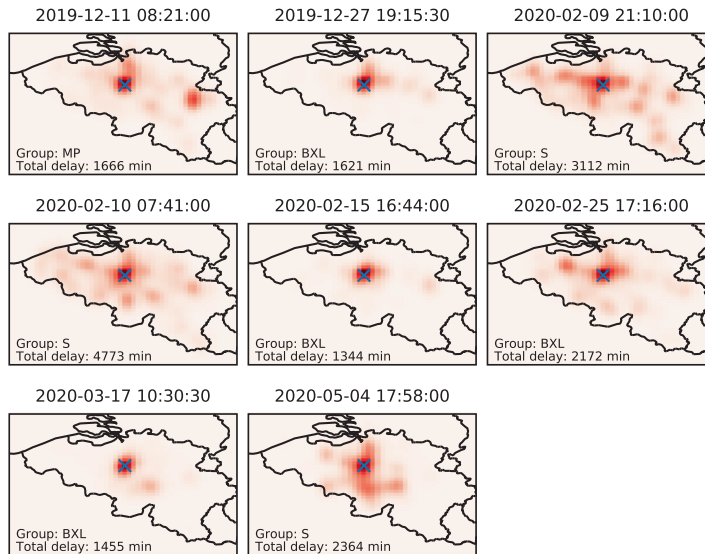


Figure E.3: Spatial distribution of time delays on disrupted days.

(see SI E.1).

One possible explanation for the good performance of this number of clusters is as follows. In Fig. E.4, the major Belgian cities are centred in different clusters. Such hubs are main routes of delay spreading in severely disrupted situations, but can exhibit complicated dynamics in the real system, since they bring together many trains, from different directions, can have complicated track layouts and scheduling to ensure connections. In our clustered version of the model, having these big hubs centred or well enclosed within clusters entails that the more complicated delay spreading happens within the cluster, compressed into a single diagonal term in the reduced G -matrix. The inter-cluster spread mostly contains straight lines between smaller stations and relatively weak connections (often just a few edges between clusters). So, one factor which may explain the $K = 10$ optimum is the clustering configuration.

Another factor is of a more theoretical nature. When K becomes too small and the peak delays are at a boundary of a cluster, inter-cluster delay spreading becomes very important and difficult to capture. For example, in Fig. 5.3a we see that for $K = 5$, Brussels — a location with most delays on disrupted days — is situated at the boundary of the red cluster. The exact direction spreading just after the peak delay is of high importance because it determines the delays of the yellow and blue cluster to a large extent. Calculating a Spearman's correlation of only five clusters will make it prone to a slight deviation from our model's expected delay spread, even leading up to negative Spearman's correlations for the lowest values of K , which is visible at the bottom of Fig. 5.6. Whether these observations are robust will require more case study-specific research.

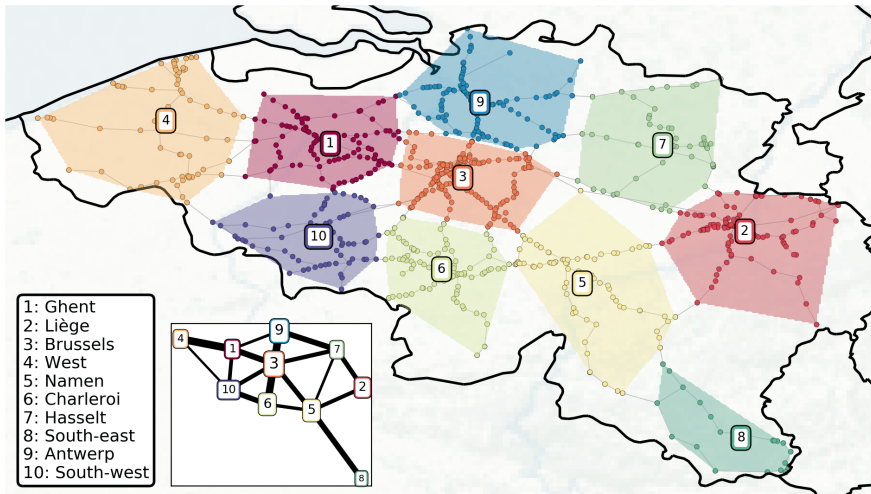


Figure E.4: Clusters at optimal aggregation level of 10 clusters. Indicative names of the clusters are shown in a legend. The elements of the resulting 10 by 10 G -matrix are reflected in the subpanel at the bottom left: the edge widths show the off-diagonal elements (i.e., dynamical ‘connectedness’ between clusters), node size indicates diagonal elements (i.e., smaller nodes mean stronger delay sinks).

E.4 Toy models

E.4.1 Toy model generation

We aim to test the model on smaller and simple networks. For this, we generate ‘toy models’, involving both an synthetic topology of the infrastructure, as well as an synthetic line schedule on top of this network. We use the following procedure to generate such toy models:

- we generate the graph \mathcal{G} ,
- for the fixed visualisation layout of the graph \mathcal{G} we determine the Euclidean coordinates of each of the vertices for this layout - these are the geographic coordinates of the station in our example,
- we draw P pairs of graph’s vertices - they will be the beginnings and the ends of the lines (note: we take each line to be bidirectional in terms of placing trains),
- at the beginning of each line we place a train with a random delay, the trains then move at a constant (Euclidean) speed, going towards their destination, along the shortest path in the graph,
- at each moment of time we count the delay of each station and compare it through the Spearman’s correlation coefficient with our model predictions, averaging over 200 runs.

E.4.2 Toy model examples

While there are endless possibilities of generating the topology of the infrastructure networks for the above toy models, we focus only on two types, of which examples are shown in Fig. E.5:

- random graphs with 15 vertices and 20 edges. If such drawn graph was not connected, we consider its largest connected component.
- star graph with $n + 1$ vertices, where the vertices $1, \dots, n$ are connected only to the vertex 0.

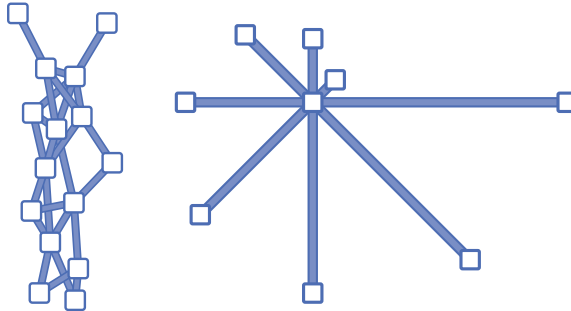


Figure E.5: Considered toy-network topologies: random graph with 15 nodes and 25 edges (left) and star graph with 9 nodes (right).

Supplementary information to Chapter 6

F.1 Data

F.1.1 Data source for the Dutch railways

Realised data (i.e., delay information) and train line schedules were obtained from ProRail, and rolling stock and crew schedules were obtained from the main Dutch railway company Nederlandse Spoorwegen (NS). Although the latter is not a complete set of all rolling stock and crew schedules, NS covers about 90% of all train activities in the Netherlands, allowing for relative completeness in most of the country. All assets are anonymised for privacy reasons and can be tracked within each individual day. For the rolling stock schedules on any given day, only the realised schedules are recorded (due to the nature of how these assets are planned), while for the crew data, both realised and planned schedules are fully known, including all intermediate mutations to the schedule. We focus on passenger trains only, and exclude freight trains in the analysis. The main reason for this is that freight trains are (economically) privacy sensitive, meaning that it is difficult to get a complete dataset. It should be noted that their contribution to delay is usually only minor, as they cover only 5.7% of all Dutch train kilometres in 2017 (numbers courtesy of ProRail), and having schedules and routes that are often partly separated from the passenger train dynamics. More details can be found in Dekker et al. (2019), section 3.1. Results in chapter 6 are obtained from a dataset ranging from 1 July, 2017 to 30 June, 2018. The Dutch data can be found in an Open Science Framework repository (<https://osf.io/tps4r/>).

F.1.2 Data source for the other railway systems

Data for the German and Italian railways in Fig. 6.1 is gained from the supplementary material of Monechi et al. (2018), which got this data from the OpenDataCity (<http://www.opendatacity.de/>) and the ViaggiaTreno (<http://www.viaggiatreno.it/>) websites, respectively for March and April 2015. The Swiss railway data is obtained from the OpenTransportData website (<https://opentransportdata.swiss/>) over January

2018. The delay data and network topologies are separate files, such that the topology is not inferred from the delay data files. No information on rolling stock and crew resources used in delay activities is known.

F.1.3 Infrastructure and service details of the Dutch railways.

The Dutch railways consist of 402 passenger stations, linked together in a dense network with the highest line utilisation rate and busiest tracks in Europe according to statistics from the European Union (European Commission, 2016) and the Dutch Statistics Agency (CBS) (Ramaekers et al., 2016). Delay, location, time, train line number, activity nature, and more details are logged each time a train passes by a so-called ‘service-control-point’ (SCP) — which are 801 measure stations on the network, of which the passenger stations are a subset. The SCPs are used as nodes in this analysis, interlinked by 1438 edges (tracks between the SCPs). Network data on locations and links of the nodes and edges of the railway network, including delay time series can be found in an Open Science Framework related to previous work (Dekker et al., 2019): <https://osf.io/tps4r/>. The three layers contain, on average (December 2017, only NS), ~6500 train service lines, ~600 rolling stock units and ~2300 crew members per day, respectively. The train lines are subdivided into 500 unique train *series* — groups of train lines that are relatively similar in their trajectories and type of rolling stock.

ProRail and railway operators in the Netherlands classify each individual (whole) day as one out of four performance categories; ‘Green’, ‘Neutral’, ‘Red’ and ‘Black’. The classification is based on train punctuality and cancellations — while Green days involve high punctuality and a low percentage of cancelled trains, Black days involve low punctuality and a high percentage of cancelled trains. The exact definitions are displayed in Tab. F.1. Punctuality and cancellation percentages are calculated with a number of constraints, leading to an ‘effective punctuality’ and ‘effective cancellation’ percentage. First, only *arrival* activities are used. Secondly, only passenger trains are considered (no freight trains). Thirdly, only activities on a subset of stations are considered. These stations are chosen based on impact on not only future delay, but also passengers, pressure on personnel and infrastructure. This choice is the result of a joint concession with the Dutch Ministry of infrastructure (Ministerie van Infrastructuur en Milieu, 2014). The fourth and final constraint to calculate the effective punctuality and cancellation metrics is that these metrics are calculated on a few specified points on the day, and averaged. Note that the above is used to determine the classification of the days, while the results in chapter 6 are not constrained by the above — in particular, our results are based on all stations.

Label	Punctuality	Cancellations	Number of days
Green	> 92.5%	< 1%	46
Neutral	-	-	292
Red	< 85%	> 5%	21
Black	< 75%	> 10%	6

Table F.1: Day-to-day severity labels by Dutch asset manager ProRail. The last column shows the number of days in the dataset (July 1, 2017 - June 30, 2018). The unofficial label ‘Neutral’ applies when none of the other labels does. Table re-used from Dekker et al. (2019).

F.1.4 Crew rescheduling data

The data for on-the-fly mutations to the crew schedules have been obtained from Nederlandse Spoorwegen (NS). For every crew member, it contains not only the final, realised schedule, but also earlier stages of their schedules — with versions marked by an integer and accompanying *mutation time stamp* t_s . We need to use the latter to determine the version of the global crew schedule at any time: for example in the simulation of Fig. 6.3, or when calculating P in Fig. 6.5. To be precise, at any time t_0 , all crew mutations with $t_s < t_0$ are known and used in the model starting at t_0 . One difficulty we face for the mutation data is that every now and then, there are system-wide updates of crew schedules: many mutations appear with identical time stamps as ‘hindsight’ updates at unrealistic times, e.g., at 02:00h on the day after. While these are clearly distinct from on-the-fly mutations carried out by traffic controllers, we cannot neglect them. In our simulations, we account for these by resetting the mutation times towards those at which the crew activities in question took place.

F.2 Processing delay data

F.2.1 Delay calculation

As mentioned in the previous section, delays are discretely logged at the service-control-points, using indices to refer to certain activities: departure, arrival or passing-through. We define delay d of such an activity a as the difference between realised time $t_{\text{real}}(a)$ and the planned time $t_{\text{planned}}(a)$ of any activity:

$$d(a) = t_{\text{real}}(a) - t_{\text{planned}}(a) \quad (\text{F.1})$$

This means that $d(a)$ is usually positive (when the train is late), but can be negative (when the train is early); the latter is rare due to the mandate that trains should not depart early. The transformation of discrete delay logs of individual activities to a continuous time series of delay may be useful for analysing the data itself (e.g. in principal component analysis as in previous literature (Dekker et al., 2019)) or for visualisation purposes. Note that this is not done in the model itself — in which we have discrete delay values for each activity (rather than a time series of delay in space). We therefore aggregate the discrete delay values per activity $d(a)$ to a time-dependent delay activity per edge e : $D(t, e)$. This is done by interpolating between the delay of departure $d(a_{\text{dep}})$ and arrival $d(a_{\text{arr}})$ activities of the same service, where the departure is being done on one side of edge e , and the arrival at the other side of e :

$$D(t, e) = \begin{cases} t - t_{\text{planned}}(a_{\text{dep}}) & \text{if } t_{\text{planned}}(a_{\text{dep}}) < t < t_{\text{real}}(a_{\text{dep}}) \\ d(a_{\text{dep}}) + \frac{d(a_{\text{arr}}) - d(a_{\text{dep}})}{t_{\text{real}}(a_{\text{arr}}) - t_{\text{real}}(a_{\text{dep}})} \cdot (t - t_{\text{real}}(a_{\text{dep}})) & \text{if } t_{\text{real}}(a_{\text{dep}}) < t < t_{\text{real}}(a_{\text{arr}}) \\ 0 & \text{otherwise} \end{cases} \quad (\text{F.2})$$

Here, we use 1 second resolution, but in principle the aggregation can be done at any resolution. Depending on the purpose, Eq. (F.2) is done either per train, or aggregated over all trains that travel across the same edge at a particular time. The resulting time

series are shaped in a sawtooth-like manner in which individual trains can be re-traced (Dekker and Panja, 2019).

F.2.2 Delay jumps

A delay jump $\delta(a)$ is defined as the change in delay of activity a compared to its previous activity a_{previous} . Formally:

$$\delta(a) = d(a) - d(a_{\text{previous}}), \quad (\text{E3})$$

Note that if a is the train line's starting activity (i.e., there is no a_{previous}), then the delay jump is not defined. Activities a and a_{previous} can be of various types. For a delay jump at a station, a is the arrival of a certain train at the station, and a_{previous} its re-departure. Delay jumps at stations can arise due to, e.g., resource allocation problems and late transfers. Delay jumps can also occur on tracks, when a is the departure of a train onto a certain track, and a_{previous} is its arrival at the end of the track. In Fig. 6.4 we consider *all* delay jumps of all possible types. In practise, railway companies often consider delay jumps to be only those values of δ that are larger than 3 minutes ('significant delay jumps'). We do not make this distinction and consider all values of δ to be 'delay jumps' — δ can even be negative, as can be seen in Fig. 6.4.

F.2.3 Classification of delay jumps

In Fig. 6.4m we label all delay jumps found in the data; while delay jumps can be found in the data, their causes are not recorded. Here, we describe each of these categories individually, using the (trilayer) model. The core of this exercise is that we compare the observed delay jump $\delta_{\text{obs}}(a)$ of every activity a with its simulated counterpart $\delta_{\text{sim}}(a)$ when initialising the model closely before the scheduled time of activity a , i.e. $t_0 < t_{\text{planned}}(a)$. In Fig. 6.4, model runs are reinitialised every 15 minutes, such that the maximum time between t_0 and $t_{\text{planned}}(a)$ is (less than) 15 minutes. Focusing on labelling delay jumps found in data, four categories can be found:

- I Cascading due to a delayed crew transfer. A delay jump is labelled (I) if $\delta_{\text{sim}} > 0$ because of a delayed crew member in the simulation output, and $\delta_{\text{obs}} > 0$.
- II Cascading due to a delayed rolling stock transfer. Analogous to label (I), delay jump is labelled (II) if $\delta_{\text{sim}} > 0$ because of a delayed rolling stock unit in the simulation output, and $\delta_{\text{obs}} > 0$.
- III Delay jumps larger than 10 minutes, but unaccounted for in the model.
- IV Noise, i.e. all other delay jumps. Individual units in this category, as well as the net sum within a time window can be both positive and negative (as seen in Fig. 6.4).

However, when for any train activity a , a (nonzero and positive) delay jump is found in the model output, but not in the observed data, we distinguish two reasons describing this discrepancy:

- V Mitigation of the delay jump (δ_{sim}) due to rescheduling of the activity: if the simulation initialised closest to $t_{\text{planned}}(a)$ did not predict a positive delay jump, but

other simulations (initialised earlier) did, we assume that changes in the schedules were made to mitigate this delay jump.

VI Mitigation of the delay jump (δ_{sim}) due to cancellation of the activity. Whether a train activity is cancelled can be found directly in the raw data, so no heuristic is needed here.

To calculate the magnitude of mitigation labels (V and VI), we use the simulated delay jumps δ_{sim} rather than the observed ones δ_{obs} as used in labels I-IV. In Fig. 6.4, all delay jumps (δ_{obs} for labels I-IV, δ_{sim} for labels V-VI) are accumulated per label within a certain time window. The relative importance $v(l)$ of each label l can then be calculated as:

$$v(l) = \frac{\sum_a \delta(a) \text{ has label } l}{\sum_a \delta(a)} \quad (\text{F.4})$$

where a denotes train activities.

F.2.4 Plotting details

For plotting delay data in Figs. 6.1 and 6.3, a smoothening procedure is applied to the spatial representation of delay, by applying a weighted averaging of neighbouring edges delay values. The weights are 1 for the edge considered, and 0.25 for neighbours. This is only done for visualisation purposes, and analyses include only raw values. In Fig. 6.4, a running window of 5-minutes length (i.e., +/- 2.5 minutes) is used at each (1 minute-) time step, followed by a 30-minute running mean smoothening over all the percentage curves. In Fig. 6.4b, four days per class are used from the period July 1, 2017 to 30 June, 2018. In particular, these included Dec 11, 2017, Nov 23, 2017, May 1, 2018 and Apr 30, 2018 for Black days; Nov 18, 2017, Nov 20, 2017, Nov 22, 2017 and Dec 13, 2017 for Red days; Nov 21, 2017, Dec 3, 2017, Nov 28, 2017 and Dec 16, 2017 for Neutral days; and July 28, 2017, Apr 15, 2018, July 22, 2017 and Sep 17, 2017 for Green days. The days were selected randomly in the dataset. To calculate the averages in Fig. 6.4b, only time windows around points between 6:00 AM and 1:00 AM (next day) are taken, to avoid confusing biases in the middle of the night (when only very few trains are travelling).

F.3 Model details

F.3.1 Model basics

The basic principle of the mono-, bi- and trilayer models used in this work is the explicit simulation of the planned schedules. Starting the the model at any chosen point in time t_0 , it iterates over all subsequent planned activities $a \in \{a(t) | t > t_0\}$, sorted by scheduled activity time (a list of roughly 150,000 activities if initialising at the start of day) — that is, when an activity $a(t_i)$ is concerned, the delay of all activities before that ($\{a(t) | t < t_i\}$) are already determined. For each iteration (activity), the delay is calculated and attributed if any of the following conditions are met:

- The train line involved in a had delay in the previous activity (sometimes referred to as ‘advection’).

- A rolling stock unit used in this activity was transferred from another train line, and had delay.
- A crew unit used in this activity was transferred from another train line, and had delay.

This is captured in Eq. (6.1) in chapter 6. Although possible, in chapter 6 we did not include any noise, such that delay passed on without any (stochastic) reduction (see below under ‘Noise’), but buffers are always taken into account.

F3.2 Formal model

Given the set of all train service activities $\{a(t)\}$ with t referring to their planned time, we initialise the model at time t_0 with knowledge about all activities $\{a(t)|t \leq t_0\}$. The aim of the model is to simulate how delay spreads towards other activities later than t_0 : $\{a(t)|t > t_0\}$. In the monolayer model, we propagate delay of activity $a_0 \in \{a(t)|t \leq t_0\}$ to future activity $a_1 \in \{a(t)|t > t_0\}$ if a_1 is the subsequent activity of the train service performing a_0 , minus a scheduled buffer $\beta(a_0, a_1) = t_{\text{plan}}(a_1) - t_{\text{plan}}(a_0)$ for each activity i . Similarly, in the bilayer model, we not only propagate delay from a_0 to a_1 when they are performed by the same train service, but also when the same rolling stock unit is used. In the trilayer model, delay is propagated via also the re-usage of crew members. In general, we can express the potential delay propagation that may be propagated from activity a' to subsequent activity a as $\mathcal{H}[d(a') - \beta(a', a)]$. The Heaviside function $\mathcal{H}(x)$ is used to prevent trains from departing early (which in practise rarely happens, and if so, with small delay jumps). For any activity $a \in \{a(t)|t > t_0\}$, then, we compute its delay from delays of previous activities $a' \in \{a(t)|t \leq t_0\}$ as in Eq. (6.1), using the maximum of all potential propagations of delay — of all unique resources used in activity a .

Buffers

Concerning the matter of buffers, we include up to 5 minutes as a necessary transfer time for resources. For example, if a crew member has 14 minutes to transfer towards another train, we assume that it needs at least 5 minutes to get to that train, leaving an effective buffer of only 9 minutes (which is then used in the calculation). If the planned buffer would be 3 minutes, this results in 0 minutes buffer — instead of a negative buffer of -2 minutes, we assume that a 3-minute planned buffer must be associated to an easy transfer that does not require 5 minutes. The role of buffers in the in delay propagation has been investigated in-depth previously (Zieger et al., 2018), but it is not the goal of the research in chapter 6.

F3.3 Initialisation.

Initialising the model from a real snapshot in the past (e.g., as in Fig. 6.3) requires extra attention. Trivially, the train activities that were realised before the initialisation time (19:00) need to be put in by the data (as ‘initialisation’), and those far past 19:00 need to be fully calculated by the model. However, the simulation of delays of train activities realised just after 19:00 are less trivial — as some information is already available about

these delays at 19:00 meaning that the delay can be more accurately estimated. We treat these boundary cases in the following manner (assuming positive delay):

- If $t_{\text{plan}}(a_{\text{dep}}) < t < t_{\text{real}}(a_{\text{dep}})$: only $d(a_{\text{dep}})$ is initialised from the data, and $d(a_{\text{arr}})$ is extrapolated using only real-time information: equal to $d(a_{\text{dep}})$.
- If $t_{\text{real}}(a_{\text{dep}}) < t < t_{\text{real}}(a_{\text{arr}})$: both $d(a_{\text{dep}})$ and $d(a_{\text{arr}})$ are initialised from the data.

Negative delays in a similar manner, but the case of $t_{\text{real}}(a_{\text{dep}}) < t < t_{\text{plan}}(a_{\text{dep}})$ arises, where also only $d(a_{\text{dep}})$ is initialised from the data.

E.3.4 Noise

As shown in Eq. (6.1), noise can be put in the model to simulate small (new) delay jumps, as in the real data: delay changes due to numerous small events like slow driving due to difficult weather conditions or a large number of passengers being slow to get into the train. For the model results in chapter 6, we set noise ζ to zero, in order to avoid confusion between model noise and noise in the real data; inclusion of model noise would have made interpretation of delay jumps in Fig. 6.4 much more difficult.

In the past, we have used a variation of this model that included explicit noise, by estimating noise from the data of all delay jumps per aggregated edge e and train series s over 100 days (Dekker et al., 2019) — e.g., by considering all delay jumps of the 3000 train service series between Utrecht and Amsterdam. From this, we assembled a probability density function (PDF) per such pair (e, s) . In that earlier work, for calculating the delay of activity a as in Eq. (6.1), we drew the (uncorrelated) noise from these PDFs. As we ran a similar procedure for chapter 6; we found that cascading effects are much stronger than the effect of noise for large-scale disruptions, leading us to the conclusion that stochasticity does not play a major role in their development. For this reason, as well as to improve interpretability of the real data as in fig. 4, we have excluded explicit noise from the model.

E.3.5 Mitigation elements

The model describes delay propagation, given that the schedules of train lines and assets remain the same, i.e., *if dispatchers take no mitigation actions* (although we do quantify mitigation in hindsight in Fig. 6.4). This assumption reduces the model's predictability horizon. In practise, the Dutch railway organisations have various more detailed contingency plans (*Versperringsmaatregelen*) in place, partially predetermining action in given disruption events. These involve protocols on highly specific situations which are beyond the scope of this research, but may in the future be used to predict delay development even further (e.g., specifically for the Dutch railways). For the purpose of model performance metrics, we exclude trains that end up being cancelled (as we have no prediction means for that in the model). Mutations on the crew schedules are not implemented throughout a single model simulation, but re-initialising multiple runs with different version of the schedules can be utilised to have updated information on what happens in the real data (see Figs. 6.4 and 6.5).

Appendix G

Supplementary information to Chapter 7

G.1 Formal derivations

In this section, we derive the building blocks $Q^{(\text{in})}$ and $Q^{(\text{out})}$ matrices for calculating the entropy of temporal entanglement. In Sec. G.1.1, we do this for connected components of the network. In Sec. G.1.2, we follow this up with the derivation of $Q^{(\text{in})}$ and $Q^{(\text{out})}$ matrices for the agents' precise contact networks at the snapshots as recorded in the data [e.g., as can be seen in Fig. 7.1(a) at $t = 0$], which we henceforth refer to as 'fine structures' for brevity. We will witness that the Q -matrices for both cases are determined from topological symmetry considerations, and are free of fitting parameters. The contribution of individual agents to entanglement is defined in Sec. G.1.3.

G.1.1 Entropy of temporal entanglement for connected components

Let us start with a single connected component at time t , e.g., as shown in Fig. 7.1(b). For this, it means that there is a single event α containing n agents at time t . Following our convention regarding the agent strings as explained in chapter 7 surrounding Fig. 7.1, such a situation is shown in Fig. G.1(a). Therein, we denote the (topological) weight of a link from agent i coming into event α by $w_{i\alpha}$, and similarly, the topological weight of a link exiting event α and reaching agent j by $w'_{\alpha j}$. These weights are the elements of the $Q^{(\text{in})}$ and $Q^{(\text{out})}$ matrices, respectively. In this section, we determine these weights from the following considerations, and subsequently construct the corresponding Q and P matrices.

- A1. The first observation is that the entanglement topology is invariant under time reversal. This implies

$$\frac{w_{1\alpha}}{w'_{\alpha 1}} = \frac{w_{2\alpha}}{w'_{\alpha 2}} = \dots = \frac{w_{n\alpha}}{w'_{\alpha n}} = c_\alpha, \quad (\text{G.1})$$

for some (yet unknown) c_α .

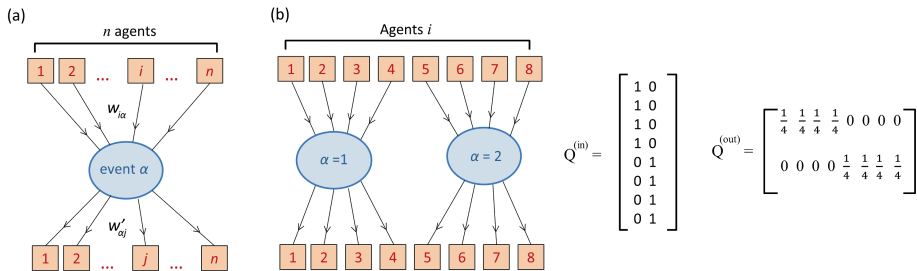


Figure G.1: Panel (a) Event α taking place at time t for one of the connected components with n agents, labeled by Roman indices. Panel (b) As a working example we consider the same event structure for eight agents at $t = 0$ as in Fig. 7.1(b) in chapter 7.

- A2. The second observation is that the entanglement topology is invariant under an exchange operation $i \leftrightarrow j$ for all pairs of the agent identities i and j . This further implies $w_{1\alpha} = w_{2\alpha} = \dots = w_{n\alpha} = w$, for some (yet unknown) w .
- A3. Finally, w and c are determined by considering random walker hops, as is standard in Network Science, in the following manner. (a) First, $w_{i\alpha}$ is the probability for a random walker to start at agent i prior to time t and end at event α at time t in one hop, meaning that $w = 1$. (b) Similarly, $w'_{\alpha j}$ is the probability of a random walker to start at event α at time t and end at agent j subsequent to time t , yielding $c_\alpha = n$. Returning to the original notation from Fig. G.1(a), this means that $w_{i\alpha} = 1 \forall i$ and $w'_{\alpha j} = 1/n \forall j$.

Important to note here is that the random walkers only move forward in time; i.e., the link weights respect the direction of time.

Upon putting A1-A3 together, we can construct the relevant $Q^{(in)}$, $Q^{(out)}$ and P matrices for this connected component. First, $Q^{(in)}$, whose $i\alpha$ -th element equals $w_{i\alpha}$, becomes an $n \times 1$ matrix with all entries unity, and $Q^{(out)}$, whose αj -th element equals $w'_{\alpha j}$, becomes a $1 \times n$ matrix with all elements equalling $1/n$. Next, $P = Q^{(in)} \cdot Q^{(out)}$ becomes an $n \times n$ matrix, for which all elements are equal to $1/n$, meaning that at time t all involved ($= n$) agents contribute $n \ln n$ to the entropy of temporal entanglement. [Note here that both $Q^{(in)}$ and $Q^{(out)}$ are row-normalised, which makes P both row- and column-normalised. These normalisations stem from the conservation of random walkers — any random walker starting from any agent prior to time t must end up at some agent subsequent to time t .] In other words, the agents participating in the event are *maximally entangled* at time t : indeed, if at time $(t + 1)$ the same n agents also participate in an event together that involves no new agents, then ensured by the relation $P^2 = P$, the contribution of these n agents to the entropy of temporal entanglement remains $(n \ln n)/(N \ln N)$ also at time $(t + 1)$.

The method is trivially extended when there are multiple events taking place at time t , as shown for an example in Fig. G.1(b) for 8 agents, which, along with the corresponding P in Fig. G.2, is then copied in Fig. 7.1(c) in chapter 7.

$$P = \begin{bmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{bmatrix}$$

Figure G.2: The 8×8 P -matrix constructed as $P = Q^{(\text{in})} \cdot Q^{(\text{out})}$ from Fig. G.1(b).

G.1.2 Entropy of temporal entanglement for the fine structures

Calculation of the entropy of temporal entanglement involving the fine structures in the agents' contact networks at integer times can also be performed by constructing the corresponding $Q^{(\text{in})}$, $Q^{(\text{out})}$ and $P = Q^{(\text{in})} \cdot Q^{(\text{out})}$ matrices — along very similar lines as in Sec. G.1.1. The procedure is in fact best illustrated by taking another working example, which we do in Fig. G.3: note that the fine structures correspond to those shown in Fig. 7.1(a) at $t = 0$ (in chapter 7). The key difference between the two working examples in Fig. G.1 and in Fig. G.3 is that in the latter the fine structures of the agents' contact network at time t is possible to capture only by binary events (i.e., each event involves two agents, corresponding to the link between them). Nevertheless, we follow the same convention for the denoting the agents, the events, and the weights of the agent-to-event and event-to-agent links as we did in Sec. G.1.1.

First, the topology, as seen in Fig. G.3, is again time-reversal invariant, which allows us to write, analogous to Eq. (G.1),

$$\frac{w_{i\alpha}}{w'_{\alpha i}} = \frac{w_{j\alpha}}{w'_{\alpha j}} = c_{\alpha} \quad (\text{G.2})$$

where i and j are the two agents participating in (binary) event α , for some (yet unknown) c_{α} . [That said, the topology is no longer invariant under an exchange operation w.r.t. the agent identities $i \leftrightarrow j$ as was the case in Fig. G.1(a)]. Next, we assign $w_{i\alpha}$ to be the probability for a random walker to start at agent i prior to time t and end at event α at time t . Using the principle that, starting at agent i , the random walker chooses any of the connecting events with equal probability, we obtain $Q_c^{(\text{in})}$, whose $i\alpha$ -th element equals $w_{i\alpha}$ (Fig. G.3). Finally, $Q_c^{(\text{out})}$, whose αj -th element equals $w'_{\alpha j}$, and obeys Eq. (G.2), is obtained by transposing $Q_c^{(\text{in})}$ and subsequently row-normalising it. The matrix $Q_c^{(\text{out})}$, obtained by means of this procedure, is also shown in Fig. G.3.

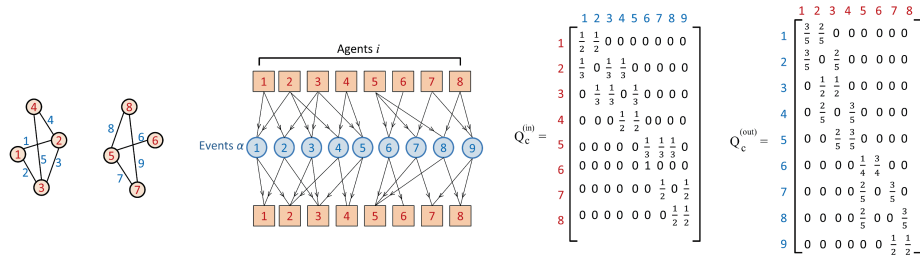


Figure G.3: Constructing the $Q_c^{(in)}$ and the $Q_c^{(out)}$ for the working example, containing the agents' actual contact network at $t = 0$ in Fig. 7.1(a).

The corresponding P matrix, obtained as $P^{(c)} = Q_c^{(in)} \cdot Q_c^{(out)}$ is shown in Fig. G.4. It can be shown, using the above procedure to determine $Q_c^{(in)}$ and $Q_c^{(out)}$, that

$$P_{ij}^{(c)} = \frac{1}{\mathcal{K}_i + \mathcal{K}_j} \quad \text{for } i \neq j \quad \text{and} \quad P_{ii}^{(c)} = 1 - \sum_{j \neq i} P_{ij}^{(c)}, \quad (G.3)$$

where \mathcal{K}_i is the degree of agent i in the contact network at time t . Clearly, $P_{ij}^{(c)} = 0$ if agents i and j are not in contact at time t . Note here that $P^{(c)}$, just as in P , is row- as well as column-normalised (here too, the normalisations stem from conservation of random walkers — any random walker starting from any agent prior to time t must end up at some agent subsequent to time t). Once $P^{(c)}$ is calculated in this way, the corresponding entropy of temporal entanglement S_c can be calculated following Eq. (7.2).

$$P^{(c)} = \begin{bmatrix} \frac{3}{5} & \frac{1}{5} & \frac{1}{5} & 0 & 0 & 0 & 0 & 0 \\ \frac{1}{5} & \frac{13}{30} & \frac{1}{6} & \frac{1}{5} & 0 & 0 & 0 & 0 \\ \frac{1}{5} & \frac{1}{6} & \frac{13}{30} & \frac{1}{5} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{5} & \frac{1}{5} & \frac{3}{5} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{7}{20} & \frac{1}{4} & \frac{1}{5} & \frac{1}{5} \\ 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{5} & 0 & \frac{11}{20} & \frac{1}{4} \\ 0 & 0 & 0 & 0 & \frac{1}{5} & 0 & \frac{1}{4} & \frac{11}{20} \end{bmatrix}$$

Figure G.4: The 8×8 $P^{(c)}$ -matrix constructed as $P^{(c)} = Q_c^{(in)} \cdot Q_c^{(out)}$ from Fig. G.3.

A discussion on the differences between the P (Fig. G.2) and the $P^{(c)}$ (Fig. G.4) matrices is in order. The difference between the two stems from the fact that the agents

within the two events in Fig. G.1(b) are maximally entangled, which is not the case for the agents in Fig. G.3. Mathematically, the difference lies in the fact that $P^2 = P$ in contrast to $[P^{(c)}]^2 \neq P^{(c)}$.

In a more precise formulation, we note that while constructing $P^{(c)}$ by means of combining the probabilities of random walkers to start at agent i prior to time t and end at event α at time t , or start at event α at time t and end at agent i subsequent to time t , we have allowed the random walker to take one combined agent-to-agent hop across time t . From this, one would expect that if a large number of such sequential agent-to-agent random walker hops would be allowed while the agents' contact network remains the same over time [i.e., the same agent contact network holds at times $t, (t+1), \dots, (t+k)$ for large k] then the random walker starting prior to time t at agent i to reach all agents within the corresponding connected component subsequent to time $(t+k)$ with equal probability. In fact, this is indeed the case, since

$$\lim_{k \rightarrow \infty} [P^{(c)}]^k = P. \quad (\text{G.4})$$

This limiting case is made possible by the condition that $P^{(c)}$ matrices are row- as well as column-normalised (which in essence is the detailed balance condition in statistical physics, effected by the conservation of random walkers as stated above (van Kampen, 2007)).

Further reflection reveals that the subtleties regarding the differences between P and $P^{(c)}$ arise due to the discreteness of the time snapshots. In chapter 7 we noted that real-world temporal network data are often sampled at some fixed interval τ_s , and also that the random walker hops are coupled to τ_s . Given that in theory, τ_s can be taken to be infinitesimally small (in comparison to the time-scales of change in the topology of the agents' contact network), and that in real world, every interaction lasts for a finite amount of time, the limit in Eq. (G.4) ensures that P is the correct descriptor for measuring entanglement.

G.1.3 Contribution of a specific agent to entanglement

The contribution S_i of agent i to the entropy of temporal entanglement is the difference between the entropy of temporal entanglement when string i is embedded in the temporal network, and that when string i is removed from of the temporal network. Note that two entanglements of agent i at two different snapshots, with agents j and k respectively, makes agent i an 'in-between temporal stop' for agents j and k . With this in mind, S_i can be seen as *a measure of betweenness of agent i in a temporal network*.

The quantity S_i can be calculated as follows. For every time snapshot we construct the P matrices, and the corresponding $P^{(-i)}$ matrices by replacing all the i -th row and column elements of the P matrices by zeros, except the diagonal element P_{ii} . We then row normalise both of them separately to obtain the P and the $P^{(-i)}$ matrices, and construct the product matrices

$$\wp(t, \Delta t) \equiv P(t)P(t+1) \dots P(t+\Delta t). \quad (\text{G.5})$$

and

$$\varphi^{(-i)}(t, t + \Delta t) \equiv P^{(-i)}(t)P^{(-i)}(t + 1) \dots P^{(-i)}(t + \Delta t). \quad (\text{G.6})$$

This is followed by correspondingly calculating $S(t, \Delta t)$ and $S^{(-i)}(t, \Delta t)$ using Eq. 7.2). This yields us

$$S_i(t, \Delta t) = S(t, \Delta t) - S^{(-i)}(t, \Delta t). \quad (\text{G.7})$$

G.2 Entropy of temporal entanglement as a function of Δt

In chapter 7, especially in Fig. 7.3, we have analysed the evolution of the entropy of temporal entanglement. In this section, we investigate the highest rate at which entropy can increase with the time interval size Δt , which we in the end extrapolate to the special case of a perfect tree — i.e., a system we argue has the maximum possible entropy growth with time.

Although in real-world networks, event sizes vary considerably, for the sake of an illustration, given a certain initialisation time t , let us consider the case where the event sizes are constant ($= n$) for $\Delta t = 1$. Following Fig. 7.1, all rows of φ will have exactly n entries containing $\frac{1}{n}$, and 0 elsewhere, leading to:

$$\begin{aligned} S_{\text{tree}}(t, \Delta t = 1) &= -\frac{1}{N \ln N} \cdot N \cdot n \cdot \left(\frac{1}{n} \ln \frac{1}{n} \right) \\ &= \frac{\ln n}{\ln N}. \end{aligned}$$

Extending this to higher values of Δt requires more formality. To this end, we consider an event \mathcal{E} of size $n_{\mathcal{E}}$ at some time $t' = t + \Delta t$ where $\Delta t > 1$, and conceptually imagine that agents from different ‘groups’, labelled by colours such as red, blue etc. participate. We define a group \mathcal{G} as follows: in between times t and t' , $\forall i$ belonging to one group, and $\forall j$ belonging to another, there is no event where both i and j are temporally connected to. Prior to time t' , let us denote the number of red agents (i.e., belonging to the red group \mathcal{G}_R) by \mathcal{N}_R , the number of blue agents (i.e., belonging to the blue group \mathcal{G}_B) by \mathcal{N}_B and so on. Similarly, the number of red agents participating in event \mathcal{E} at time t' is given by n_R (forming the corresponding subgroup g_R), the number of blue agents participating in event \mathcal{E} at time t' is given by n_B (forming the corresponding subgroup g_B), and so on. Clearly,

$$n_R + n_B + \dots = n. \quad (\text{G.8})$$

We also label the elements of $\varphi(t, \Delta t - 1)$ belonging to a given coloured group by indices of the same colour; e.g., we re-index the red agents by red indices $1_R, 2_R, \dots, \mathcal{N}_R$, of which the agents $1_R, 2_R, \dots, n_R$ participate in the event. Then the elements of the $\varphi(t, \Delta t - 1)$ matrix corresponding to the red group of agents can be extracted to form an $\mathcal{N}_R \times \mathcal{N}_R$ matrix. Let us denote these elements by the notation p_{ij} , with $1_R \leq (i, j) \leq \mathcal{N}_R$. Then the

following relations will hold:

$$\sum_{i=1_R}^{\mathcal{N}_R} p_{ij}(t-1) = 1, \quad i \in \mathcal{G}_R, j \in \mathcal{g}_R;$$

$$\sum_{i=1_B}^{\mathcal{N}_B} p_{ij}(t-1) = \dots = 1, \quad i \in \mathcal{G}_B, j \in \mathcal{g}_B; \dots$$

These relations simply follow from the fact that $\varphi(t, \Delta t - 1)$ is column-normalised, having noted that $\varphi_{ij}(t, \Delta t - 1) = 0$ when agents i and j belong to two different (ly coloured) groups.

Subsequent to the event at time t , for an agent $j \in \mathcal{E}$ the elements of $P(t)$ will be given by

$$p_{ij}(t) = \frac{1}{n_{\mathcal{E}}} \begin{cases} \sum_{j'=1_R}^{n_R} p_{ij'}(t-1) & i \in \mathcal{G}_R, j' \in \mathcal{g}_R \\ \sum_{j'=1_B}^{n_B} p_{ij'}(t-1) & i \in \mathcal{G}_B, j' \in \mathcal{g}_B \\ \dots \end{cases} \quad (\text{G.9})$$

and is independent of j [i.e., $p_{ij}(t) \equiv p_i(t)$]. This allows us to express the corresponding change in entropy due to the event \mathcal{E} , by separating the agents in different (ly coloured) subgroups, as

$$\Delta S = \frac{1}{N \ln N} \left\{ \left[n_{\mathcal{E}} \sum_{i=1_R}^{\mathcal{N}_R} p_i(t) \ln p_i(t) - \sum_{i=1_R}^{\mathcal{N}_R} \sum_{j=1_R}^{n_R} p_{ij}(t-1) \ln p_{ij}(t-1) \right] \right. \\ \left. + \left[n_{\mathcal{E}} \sum_{i=1_B}^{\mathcal{N}_B} p_i(t) \ln p_i(t) - \sum_{i=1_B}^{\mathcal{N}_B} \sum_{j=1_B}^{n_B} p_{ij}(t-1) \ln p_{ij}(t-1) \right] + \dots \right\}. \quad (\text{G.10})$$

Using Eq. (G.9) to replace $p_{ij}(t)$ in Eq. (G.10), and thereafter having dropped the $(t-1)$ argument for the matrix elements, we have

$$\Delta S = \frac{n_{\mathcal{E}} \ln n_{\mathcal{E}}}{N \ln N} + \frac{1}{N \ln N} \left\{ \underbrace{\left[\sum_{i=1_R}^{\mathcal{N}_R} \sum_{j=1_R}^{n_R} p_{ij} \ln p_{ij} - \sum_{i=1_R}^{\mathcal{N}_R} \left(\sum_{j=1_R}^{n_R} p_{ij} \right) \ln \left(\sum_{j=1_R}^{n_R} p_{ij} \right) \right]}_{C_R} \right. \\ \left. + \underbrace{\left[\sum_{i=1_B}^{\mathcal{N}_B} \sum_{j=1_B}^{n_B} p_{ij} \ln p_{ij} - \sum_{i=1_B}^{\mathcal{N}_B} \left(\sum_{j=1_B}^{n_B} p_{ij} \right) \ln \left(\sum_{j=1_B}^{n_B} p_{ij} \right) \right]}_{C_B} + \dots \right\}. \quad (\text{G.11})$$

Given that $p_{ij} \leq 1$, it is easily argued from Eq. (G.11) that C_R, C_B etc. terms are ≤ 0 :

the only case when they can be zero is when n_R, n_B etc. either 0 or 1. This in turn means that an event can maximally contribute to the entropy of temporal entanglement only when, for every agent pairs (i, j) participating in the event, between t and t' there exists no event that both i and j are temporally connected to.

Note however that this condition is not possible to maintain for an arbitrary temporal depth of between t and $t + \Delta t$. A toy case where this is possible — returning to the one we started this section with — is with fixed event size n and also fixed event frequency n per snapshot, requiring the condition $N = n^k$ to be satisfied for some integer k . We illustrate the procedure with an example below. When increasing Δt to 2 in that toy example, we add an extra time step. For the system that is a perfect tree in the temporal domain, any agents that have participated in events at $\Delta t = 1$, do not participate in an event at $\Delta t = 2$; instead, all n agents from any event in the first time step participate in n separate events. This ensures that the all rows of the φ matrix have exactly n^2 elements with entries $\frac{1}{n^2}$, resulting in $S(t, \Delta t = 2) = 2 \frac{\ln n}{\ln N}$. Continuing this for any value of Δt , this can be generalised to:

$$\begin{aligned} S_{\text{tree}}(t, \Delta t) &= -\frac{1}{N \ln N} \cdot N \cdot n^{\Delta t} \cdot \left(\frac{1}{n^{\Delta t}} \ln \frac{1}{n^{\Delta t}} \right) \\ &= \Delta t \frac{\ln n}{\ln N}, \end{aligned}$$

leading to a linear relationship between S and Δt for a perfect tree in the temporal domain.

From a point of view of analysing spreading dynamics, the perfect tree is the system with the highest spreading vulnerability. For a perfect tree in the temporal domain, agents only meet agents that do not have a (higher-order) connection in common in the past. For the application of rumour spreading, for example, in a perfect tree, people that *know* the rumour will therefore only meet people that *do not know* the rumour — obtaining the fastest possible spread of the rumour.

G.3 Data from real-world temporal network datasets

In Fig. 7.1 in chapter 7, we use a network constructed by hand for illustration purposes. In Figs. 7.2 and 7.3, we use temporal network data from real-world systems. In this section we discuss the preprocessing and availability of these datasets.

G.3.1 High school and Primary school

In Fig. 7.2, we use data on the interactions among students of a high school and a primary school. Both datasets are frequently used in network science papers. The high school data is referred to as the ‘Thiers13’ dataset (Mastrandrea et al., 2015), from which we use two days in the analysis, and the primary school data is the ‘LyonSchool’ dataset (Gemmetto et al., 2014; Stehlé et al., 2011), from which we also use two days of data in the table in Fig. 7.3(b). Both are freely accessible at the Sociopatterns project website (<http://www.sociopatterns.org/datasets/>). Agent interactions in these datasets

are defined as close face-to-face proximity of students and teachers using wearable sensors and proximity-sensing infrastructure based on radio frequency identification devices. In the table in Fig. 7.3(b), we also use a second high school dataset (also from the Sociopatterns project) from 2011, referred to as ‘Thiers11’, from which we use four days in the analysis. The sampling time interval τ_s for these datasets is 20 seconds.

G.3.2 Other Sociopatterns datasets

Data from the scientific conference, hospital and workplaces (2013) and (2015) are also accessible via the Sociopatterns project and are often referred to as the SFHH conference, LH10, InVs13 and InVs15 datasets. The scientific conference refers to the 2009 SFHH conference in Nice, France (June 4-5, 2009) (Cattuto et al., 2010; Stehlé et al., 2011). The workplace datasets InVs13 and InVs15 were experiments conducted in French office buildings in 2013 and 2015 respectively (Génois and Barrat, 2018). The hospital dataset contains the temporal network of contacts among patients and health-care workers (HCWs) and among HCWs in a hospital ward in Lyon, France, from Monday, December 6, 2010 at 1:00 pm to Friday, December 10, 2010 at 2:00 pm — we use four separated days in our analysis from this time interval. The study included 46 HCWs and 29 patients (Vanhems et al., 2013). The sampling time interval τ_s for these datasets is 20 seconds.

G.3.3 High-energy physics citations

The citations network we analyse in chapter 7 is from the e-print arXiv server HEP-PH (high energy physics phenomenology) and covers all the citations within a dataset of 34,546 papers with 421,578 edges (Gehrke et al., 2003). Because of the enormous number of agents (i.e., papers) and (temporal) links, we only use a subset of this dataset. The dataset comprises papers over approximately 10 years of data. We only use papers between March 11, 1996 and August 3, 1998. These dates are obtained by taking 10,000 interactions in approximately the middle of the time series (to best relate to the temporal dynamics of this system). This results in a much smaller number of 4,052 papers (approximately 10% of the total). If a paper i cites paper j , we treat this as an ‘interaction’ between these two papers. The sampling time interval τ_s for these datasets is in days.

G.3.4 Baboons

The baboons temporal network we analysed is from an experiment on a group of 20 Guinea baboons living in an enclosure of a primate centre in France, between June 13, 2019 and July 10, 2019 (Gelardi et al., 2020). Only 13 out of the 20 baboons wore proximity sensors — which is the data we use for this analysis. In our analysis, we use the first five days of this dataset. The sampling time interval τ_s for these datasets is 20 seconds.

G.3.5 Sparrows

The sparrows dataset is from a study at the University of California, Santa Cruz Arboretum (Shizuka et al., 2014). The sparrows arrive there in October-November and depart for their breeding grounds in March-April each year. The study spanned three non-breeding

seasons: January-March 2010 (Season 1), October 2010-February 2011 (Season 2) and October 2011-April 2012 (Season 3). Each season, the authors captured birds using baited traps and attached individually unique combinations of colour bands. In Season 2, they did not band any birds between October-December 2010. Network edges are found based on co-memberships between flocks (defined as a group of individuals found within a single 5m radius) by identifying the colour-banded individuals in each flock. Data are available at $\tau_s = 1$ day.

G.3.6 Irvine social app

This dataset is comprised of private messages sent on an online social network at the University of California, Irvine. Users could search the network for others and then initiate conversation based on profile information. An edge (i, j, t) means that user i sent a private message to user j at time t (Panzarasa et al., 2009). The data as a whole spans of approximately 195 days. As with the aforementioned Citations dataset, for this analysis, we focused on only part of the dataset to improve computational speed. In particular, we used the first 34 days, resulting in a set of 25,000 interactions and 1136 unique agents (users). The time is aggregated to $\tau_s = 6$ hrs.

G.3.7 Art fair

The art fair data is gathered during ‘Smart Distance Lab: The Art Fair’ between August 28 and 30, 2020 in Amsterdam, the Netherlands (Tanis et al., 2021). It consists of eight experiments, where various conditions such as walking direction, face masks and proximity alerting systems were varied. We use the data from the wearable sensors, that defined an interaction by two people coming in the proximity of less than 1.5 m. The dataset is available on Figshare and in a MySQL database. The sampling time interval τ_s for this dataset is one second.

G.4 Spreading processes on top of temporal networks

In order to reveal the relationship between the entropy of temporal entanglement and the dynamical processes playing out on top of the network, we simulate three models describing three different (stochastic) processes on the real-world systems in Fig. 7.3. The results of the simulations, reported in chapter 7, have been obtained by performing ensembles of 25 re-initialised model runs (per initialisation time t) over 80% of the full time window per system as starting points t , using $\Delta t = 12$ time steps for the correlations in Fig. 7.3b, and 60 time steps for the visualisations in panels (c) and (d). Only ensemble averages are reported in chapter 7. For most datasets, this procedure is repeated over multiple subsets. For example, the high school data Thiers13 contains five days of data. Those days are individually run using this ensemble procedure, and the average across those days is shown in the table in Fig. 7.3(b). Below, we describe the stochastic models. For all models, the number of agents is denoted by N .

G.4.1 Majority-vote

For the first model we simulate the propagation of opinions on a dilemma, which is associated to spreading of information, such as (fake) news, memes or political opinions. The opinion of agent i at time t can have values 0 or 1. We denote the overall fraction of agents having opinion 0 with f_0 , and analogously we define f_1 . At initialisation (starting time t , interval size $\Delta t = 0$), we start with half of the population having opinion 0, and the other half opinion 1. The agent interactions are specified as they are defined by the (real-world) temporal network's topology. At each event, all agent's change their opinion to the opinion of the majority of the participants. For example, if in an event with three agents, the opinions are 001, then, the third agent will change its opinion to 0. If the opinions are tied, all participants will swap to a single opinion, chosen at random. If there is only one single agent in an event, its opinion will remain the same. No noise is introduced. We define a single state variable that is dependent on the starting time of the model (t) and the time interval (Δt) that is used to progress it from t onward:

$$V_{\text{maj}}(t, \Delta t) = 2 \cdot [\max(f_0(t, \Delta t), f_1(t, \Delta t)) - 0.5].$$

The subtraction of 0.5 and subsequent multiplication with 2 make sure that $V_{\text{maj}} \in [0, 1]$, which is a standardised form for easy comparison to the entropy of temporal entanglement.

G.4.2 Transport delay

The second model concerns the propagation of a continuous variable through interactions (events), all event participants will attain the highest value among them. This relates to various real-world phenomena, such as transportation delays (hence the title of this model), in which transport assets (such as a physical train and a driver to drive it; assets may need to come from different physical locations (Dekker and Panja, 2021), and have to wait for the maximally-delayed one — assigning all assets in an event with a newly generated delay that equals the highest of all the delays in that event. (Another example is the spread of dominant genes in phylogenetics.) At the start of the model, all agents are given a random number between 0 and 1, we refer to as the 'delay' d_i of agent i , which propagates as per the above rule. Over time, the average delay will increase up to the maximum existing delay. We define a state variable, again dependent on t and Δt :

$$V_{\text{del}}(t, \Delta t) = 2 \cdot \left[\frac{1}{N} \sum_{i=1}^N d_{i=1}(t, \Delta t) - 0.5 \right],$$

which has the same standardisation as the state variable of the majority-vote system, to bring V_{del} between 0 and 1.

G.4.3 Infectious disease

While numerous infectious disease models for networks exist, we chose the simple susceptible - infected (SI) model. Each of the model's agents i can be in either one of two states q_i : susceptible ($q_i = 0$) or infected ($q_i = 1$). There is no exposed or recovered state,

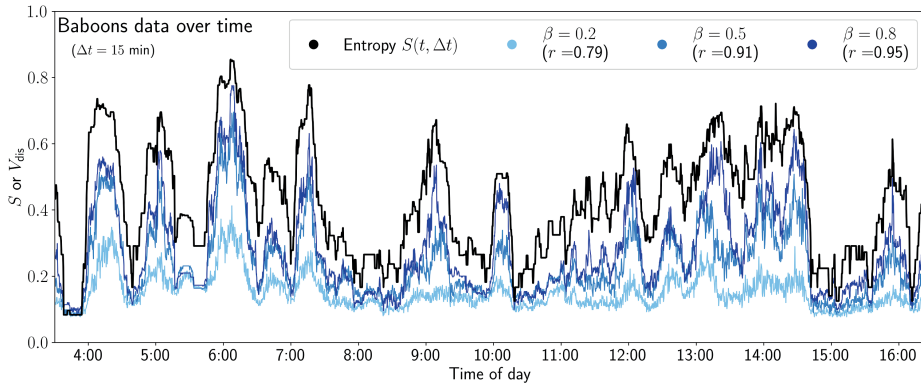


Figure G.5: Evolution of the entropy of temporal entanglement S , and the state variable V_{dis} , for three values of β , using $\Delta t = 15$ minutes, for day 2 in the Baboons dataset. Pearson's correlation coefficients (r) are also noted in the figure (which are the values only for this particular day). Note that Fig. 7.3(b) uses five days of the Baboons dataset, leading to a slightly smaller Pearson's correlation coefficient for $\beta = 0.8$, namely 0.93.

as commonly used in epidemiology. Each model run is initialised with 10% (randomly chosen) agents being infected. Upon interaction in an event, if one of the participants is infected, there is a probability β of other (susceptible) participants becoming infected as well. We define a state variable dependent on t and Δt , which is the average status q over all agents:

$$V_{\text{dis}}(t, \Delta t) = \frac{1}{N} \sum_{i=1}^N q_i(t, \Delta t),$$

which is a number between 0 (no infected) and 1 (all agents infected).

Clearly β is a parameter for the spreading of the pathogen. For chapter 7, we have chosen $\beta = 0.8$. For the Baboons dataset, below we present the data for two additional values of β . While the Pearson's correlation between S and V_{dis} decreases with decreasing β , the correlation still remains substantial even for β as low as 0.2.

Appendix H

Supplementary information to Chapter 8

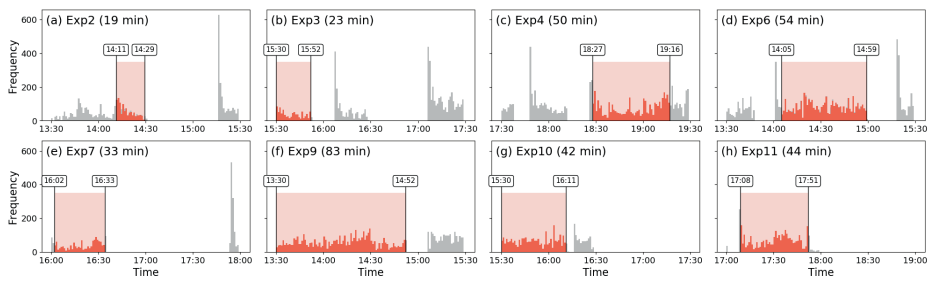


Figure H.1: Histograms of the time stamps of events across the eight experiments. In red, the filtered windows used in chapter 8 are highlighted.

Supplementary information to Chapter 9

I.1 Model components

This SI is devoted to explaining the eight steps in Fig. 9.1 more elaborately. The general intuition of the model is as follows. We start with a population of 170,721 agents divided into 11 demographic groups that live throughout the 380 municipalities as depicted by demographics data. Using mobile-phone signalling data, we stochastically infer movements of these agents throughout the country, based on their demography and home locations. This is done in the ‘mobility part’ (steps 1-4). Then, knowing all locations and movements, we simulate an individual-based SEIR model on top of this, described in the ‘transmission part’ (steps 5-8). The model is on an hourly basis and has a spatial resolution of the 380 municipalities.

I.1.1 Agents and demographic groups [Step 1]

The first step concerns the definition of the agents and their attributes. We simulate the Dutch population on a scale 1:100, meaning that each agent in the model represents 100 people in the population. The total size of the Dutch population in 2019 was 17,256,870. All demographics and population data is obtained from Statistics Netherlands (Centraal Bureau voor de Statistiek, CBS, in Dutch). Because we need an integer number of agents in each municipality, we end up with a total of 170,721 agents by rounding down, which results in a slightly lower number of agents than exactly the total divided by 100 (which would be 172,569).

Only data aggregated on the municipality level is used, in avoidance of the violation of privacy. The exact municipality division changes a little bit every year. Because the mobility data (see SI I.1.2) was using the municipality division of 2018, we projected the demographic data from 2019 onto the municipalities as they were in 2018. In turn, for plotting purposes (e.g. for the maps in Fig. 9.1), the municipality borders and shapefiles of 2020 are taken, requiring an additional projection for plotting. Note that year-to-year changes of municipality divisions are minor and only affect small municipalities.

We split the Dutch population (per municipality) into eleven demographic categories,

based on age, whether they go to school or not, and employment status. Data on employment status and whether people go to school are based on tax records and education institutional registers, also from Statistics Netherlands. The resulting categories are displayed in Tab. I.1. Distinguishing other criteria other than age provides additional information on how these people move (being frequent movements or not), specifies mixing situations (Mossong et al., 2008) and provides additional information on demographic distribution across the different municipalities. Importantly, it allows us to more explicitly target the right agents when implementing interventions (see SI I.1.7-I.1.8).

The overall fractions of each demographic group is displayed in Tab. I.1. The *Middle-age working* category is by far the largest. Note that these fractions are not constant across municipalities: some Dutch municipalities, especially those in the north and east of the country, have a higher-than-average number of elderly and fewer children, and the opposite is usually true for larger cities in the west.

Group	Criteria			Aspects			
	Age (y)	Work	School	Amount (fraction)	Av. time in home municipality	Daytime mixing home (not home)	Nighttime mixing home (not home)
Pre-school children	0-4	-	-	851880 (4.9%)	6%	Home (Other)	Home (Other)
Primary school children	5-11	-	Yes	1295380 (7.5%)	6%	School	Home (Other)
Secondary school children	12-16	-	Yes	991290 (5.7%)	6%	School	Home (Other)
Students	17-24	-	Yes	1086240 (6.2%)	26%	School+Work	Home (Other)
Non-studying adolescents	17-24	-	-	632530 (3.6%)	26%	Work	Home (Other)
Middle-age working	25-54	Yes	-	5530360 (31.8%)	26%	Work	Home (Other)
Middle-age unemployed	25-54	-	-	1231780 (7.1%)	6%	Home (Other)	Home (Other)
Higher-age working	55-67	Yes	-	1623040 (9.3%)	26%	Work	Home (Other)
Higher-age unemployed	55-67	-	-	1109170 (6.4%)	6%	Home (Other)	Home (Other)
Elderly	68-80	-	-	2102530 (12.2%)	6%	Home (Other)	Home (Other)
Eldest	80+	-	-	802670 (4.6%)	6%	Home (Other)	Home (Other)

Table I.1: The eleven demographic groups along with the criteria used to define them, and aspects of how they are implemented in the model.

I.1.2 Inter-municipality travel [Step 2]

Given the distribution of the agents and the demographic categories across the Netherlands, we require an empirical basis for the inter-municipality movements of the agents. This is provided in the form of the average Dutch municipality-to-municipality movements over the period of March 1, 2019 up to and including March 14, 2019, by the data company Mezuro. They directly measure these movements using cellular mobile phone data and this has been verified using GPS data. For more details, see <https://www.mezuro.com/>. Specifically, the data comprises a matrix M_{ij} showing how many (daily average) visits there were from people living in municipality i to municipality j . The data does not contain information about movements within a given municipality. Neither does it provide information about sequenced movements — for example when a person living in Amsterdam goes to Rotterdam and later on the day goes to Utrecht, the data indicates that there is a movement of an Amsterdam inhabitant to Rotterdam, and a movement of an Amsterdam inhabitant to Utrecht (not from Rotterdam to Utrecht). There is also no information on the duration of the visit nor about demographic aspects of the people moving.

The daily average visits in M_{ij} are split into three categories: ‘frequent’, ‘regular’ and ‘incidental’, which is used to crudely infer which movement data to use for which demo-

graphic category: frequent and regular movements are applied to working and school-going agents, while incidental movements are applied to all other agents. So in fact, two mobility matrices M are determined: one (M_{freq}) with regular and frequent movements, and one (M_{inc}) with incidental movements, both averaged across the 14 days that this data comprises.

1.1.3 Agent movements [Step 3]

The third step of the model concerns linking the mobility data to the movements of individual agents in the model. The movements are determined per agent. Each agent belongs to a particular demographic group g and lives in a municipality m . Given demographic group g , we use either M_{freq} or M_{inc} , as mentioned above. In the following, we use M to denote the choice of one of these matrices. The municipality m points us to the particular row M_m to use from the mobility matrix, containing the number of movements from m to other municipalities. We use this information to create the scale parameters for the Dirichlet distribution used to draw the fractions of the day from, that the agent (living in m , belonging to g), spent in each municipality. We do this as follows.

First, we have to normalise these movements in a proper manner. To account for the fact that people living in some municipalities have above-average number of movements, we do not normalise the movements M_m by the total (i.e., resulting in a row-sum of 1), but by the amount of people living in m , effectively obtaining the number of movements to each other municipality *per inhabitant of m* . However, because we aim to use these elements to eventually draw fractions of the day spent in each municipality, we also need to set how much time is spent in the home municipality itself (because M_m is not obtained from the data — it only contains movements between municipalities). We assume that working people and students spend approximately 25% of their time, and not-working people spend approximately 5% of their time in other municipalities. Using the average row sums divided by the populations as mentioned above, this results in values of 1 and 1.5, respectively, for weighing the time spent in the home municipality. Summarised, this boils down to the following expression of the m' th scale parameters δ of the Dirichlet distribution, for people living in municipality m , belonging to working people and students:

$$\delta(m, m') = \begin{cases} \frac{M_{\text{freq}, mm'}}{P(m)} \cdot \frac{2.5}{\sum_i \delta(m, i)} & \text{if } m \neq m' \\ 1 \cdot \frac{2.5}{\sum_i \delta(m, i)} & \text{if } m = m' \end{cases} \quad (I.1)$$

where $P(m)$ is the population of municipality m . And analogously for other demographic groups:

$$\delta(m, m') = \begin{cases} \frac{M_{\text{inc}, mm'}}{P(m)} \cdot \frac{2.5}{\sum_i \delta(m, i)} & \text{if } m \neq m' \\ 1.5 \cdot \frac{2.5}{\sum_i \delta(m, i)} & \text{if } m = m' \end{cases} \quad (I.2)$$

The factor $\frac{2.5}{\sum_i \delta(m, i)}$ sets the total sum of the Dirichlet parameters to 2.5, to set the variability between draws from the resulting distribution to be constant. Histograms of

the row sums of the mobility matrices, divided by the municipality population sizes are shown in Fig. I.1. The resulting Dirichlet distribution is used to draw fractions of the day that a person belonging to the respective demographic group and living in the respective municipality spends in each municipality — meaning that, per person, 380 fractions are drawn, one for each municipality. Because of the definition of the shape parameters, the largest fraction is usually his/her home municipality. These fractions are converted to integer-hours by multiplying with 24 and rounding down (leftover hours are spent in home municipalities).

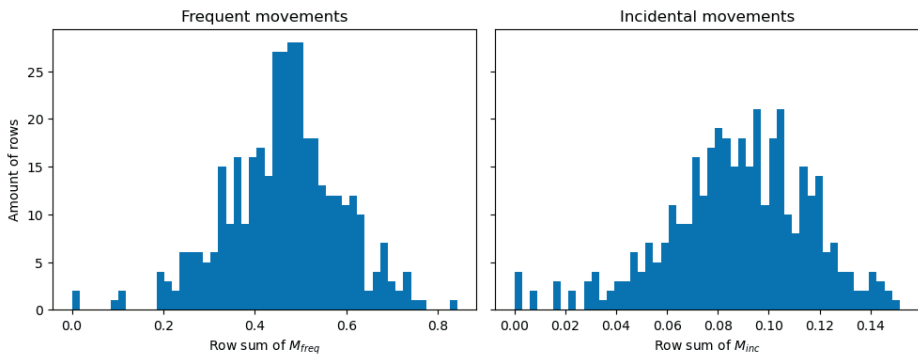


Figure I.1: Histograms showing the row sums of (a) matrix M_{freq} and (b) matrix M_{inc} divided by the population sizes belonging to those rows.

To convert this list into an actual schedule of this person on this particular day, the order of these visits should be decided. The time spent in the home municipality is cut in two and the halves is placed at the beginning and ending of the day, marking staying at home overnight. For example, given that p spends 14 hours of the day in their home municipality, then p is assumed to spend the hours 0:00-7:00AM and 17:00-24:00PM in this municipality. Duplicate hours spent in other municipalities are concatenated and these concatenated periods are placed in the leftover hours in a random order. This leads to daily ‘schedules’ such as visualised in the bottom panel of Step 3 in Fig. 9.1.

We repeat this procedure seven times to end up with a weekly schedule. This ultimately results in a movement pattern that varies from day to day, but is the same for each day of the week (e.g., Monday in week 1 are equal to Mondays in other weeks). The resulting fraction of people in other (not-home) municipalities is shown in Fig. I.2 for different moments of the day, having an average of 16.6% of the time spent in other municipalities.

We check the approximate validity of this pattern by comparing these results to survey data from Dutch governmental research agencies. In particular, from a regular survey done by the Sociaal Cultureel Planbureau (SCP), we know that people (12 years old and older) in 2016 spent on average 20.5 hours on paid work, 3.3 on schooling and 42.1 on recreation (Social Cultureel Planbureau, 2016). Also, 38% of the people lives in the municipality that they work in (Centraal Bureau voor de Statistiek, 2021), i.e., 62% has to travel to another municipality for work. This means that $\frac{20.5}{7 \cdot 24} \cdot 62\% = 7.6\%$ of the total time of a day is on average spent in other municipalities because of work. Furthermore, we know

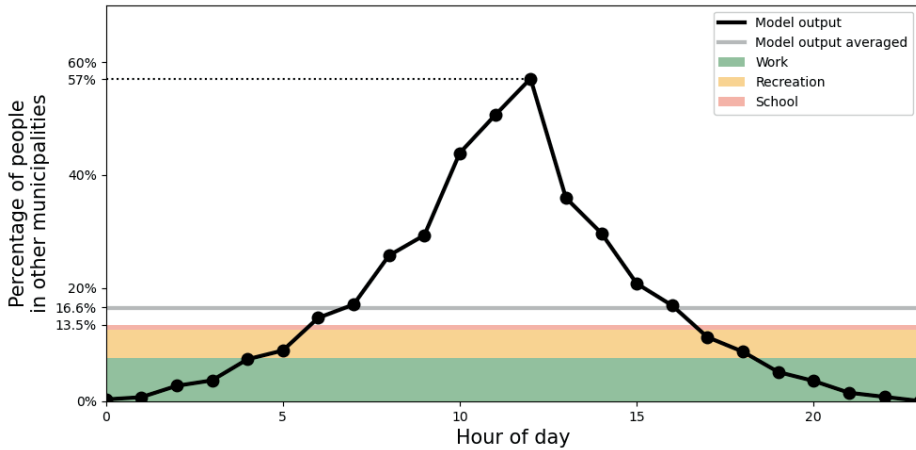


Figure I.2: Percentage of people outside of their home municipality (black). At the bottom, in colours, daily-averaged estimates of time spent outside the home municipality are displayed, due to the activities of work (green), recreation (yellow) and school (red).

that 48% of students do not live in dorms (Kenniscentrum Studentenhuisvesting, 2020), which we can use to approximate how many students have to travel between municipalities for their schooling. Analogous to working time, we reason that an additional fraction of time spent outside of the home municipality due to schooling is $\frac{3.3}{7.24} \cdot 48\% = 0.94\%$, assuming that students living in dorms outside of their school municipality and the fact that youth between 12 and 18 do not live in dorms partially counterbalance. Concerning recreation, we assume 10% of recreation being outside of the home municipality, which results in yet another additional fraction of $\frac{42.1}{7.24} \cdot 20\% = 5.0\%$. Summing them results in 13.5%. This is less than the observed 16.6% (grey line in Fig. I.2), but there are many large uncertainties in these calculations (e.g., the time spending survey (Social Cultureel Planbureau, 2016) is only based on people of 12 years and older), but we use them to have an approximate validation.

Another validation to be made concerns the 2.5 in Eqs. (I.1) and (I.2), which is the total sum of the parameters of the Dirichlet distributions. Statistically, the total sum of the scale parameters in a Dirichlet distribution indicates the variability across draws. Therefore, we need to make sure that the variability we set here makes sense. We do this by calculating the fraction of the day spent in the home municipality of people, and visualise the between-people variability in this metric. The results are shown in Fig. I.3. In the left panel, the near-separation is noticeable between groups that had increased cross-municipality movement (i.e., home-scale parameter of 1 in the Dirichlet distribution; students and working people) and those that did not (i.e., parameter of 1.5; other groups). Also, we see that in many groups (also in the ‘All’ category), the distributions of how long people are in their own municipality varies, and the tails overlap. The differences across the municipalities in the right column are explained by differences in their mobility-to-population ratio (illustrated in the row sums in Fig. I.1 and demographic differences. Even though we do not have observed data to compare these numbers, they

do not seem to be unrealistic: large cities such as Utrecht, Amsterdam and Rotterdam involve people that are probably working there, and may indeed therefore have a higher fraction of time spent in their home municipality.

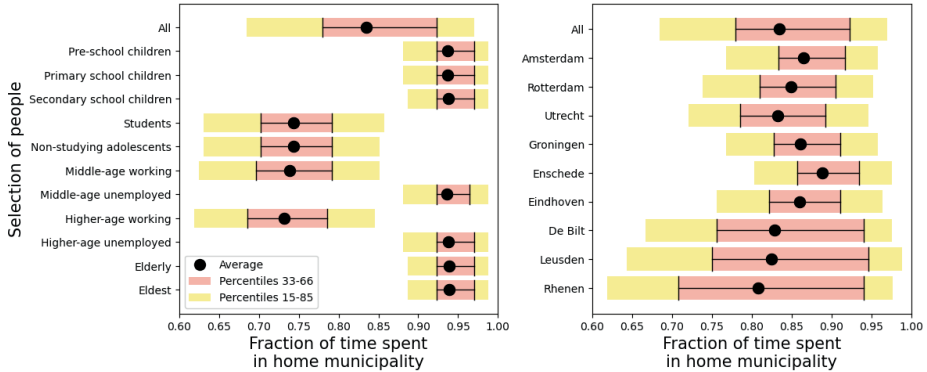


Figure I.3: Variability in the fraction of time spent in the home municipality. **Panel (a):** across different demographic groups. **Panel (b):** across a selection of municipalities.

I.1.4 Mixing [Step 4]

From steps 1-3, we know which people are in the same municipality at which moment of time. In step 4, we determine the mixing of different demographic groups *within* a municipality, assuming proportionate mixing. We base this on two factors: a demographic stratification in the mixing, and a distinction of four different mixing situations. We follow the POLYMOD study (Mossong et al., 2008; Prem et al., 2017) by distinguishing four unique situations: ‘home’, ‘school’, ‘work’ and ‘other’. From Prem et al. (2017), we downloaded the average contact rate matrices, and converted them from their 16 demographic groups to our 11 demographic groups.

Which of the four situations applies to a person, is determined by time of day, whether the person is in his/her home municipality, and the demographic category the person belongs to. The exact mixing matrices used is shown in Tab. I.1, under ‘Daytime mixing’ (corresponding to time between 8 and 18) and ‘Nighttime mixing’ (other times of day). For the Students group, we make an exception by not taking one specific mixing matrix, but averaging the ‘work’ and ‘school’ mixing matrices during daytime.

I.1.5 Initialisation and hospital admissions [Step 5]

Hospital admissions

Let us start with the hospital admission data itself. This is obtained from the Nationale Intensive Care Evaluatie (NICE) registration, which is the official institution for hospital reports. In particular, the data can be found under <https://data.rivm.nl/meta/srv/dut/catalog.search#/metadata/4f4ad069-8f24-4fe8-b2a7-533ef27a899f?tab=relations>, which are daily numbers per municipality. In particular, this means that all

model results had to be translated into daily number (summing over a running window of 24 time steps).

We proceed by discussing the translation between infection cases and hospital admissions, which is required for both the initialisation and calibration. Following de Vlas and Coffeng (2021) we use a time lag between becoming symptomatic (in model terms: infectious I) and a potential hospital admission of being Weibull-distributed with mean 14 and scale parameter 10. The probability of hospital admission p_{hos} , given that a person becomes infectious is not equal for all people: elderly are more susceptible to being hospital admitted than young children, for example. In Tab. I.2, we show p_{hos} across the various demographic groups. These probabilities are determined as follows. Using the seropositivity data during the Dutch first wave (Vos et al., 2021) and knowing how many people there are in each age group (see SI I.1.1), we can estimate the cumulative number of infection cases per group. Combining this with cumulative hospital admission data per age group, we can divide the two to get the hospitalisation probability per age group, which can be translated to the 11 demographic categories we use in chapter 9. These probabilities p_{hos} per demographic category are shown in Tab. I.2. The Weibull-distributed temporal translation and demography-stratified probability of hospitalisation are done when translating the model results into hospital admissions, for example in the red curve in Fig. 9.1a.

Demographic group	p_{hos}	$h(g)$
Pre-school children	0	1.0
Primary school children	0	2.0
Secondary school children	0.0018	3.051
Students	0.0006	5.751
Non-studying adolescents	0.0006	5.751
Middle-age working	0.0081	3.6
Middle-age unemployed	0.0081	3.6
Higher-age working	0.0276	5.0
Higher-age unemployed	0.0276	5.0
Elderly	0.0494	5.3
Eldest	0.0641	7.2

Table I.2: Susceptibility parameter $h(g)$ and probability of hospital admission p_{hos} .

Initialisation

The model is initialised with an estimated number of infectious cases in the period up to March 1, 2020. Initialising with fewer cases (i.e., up to before March 1) would increase the sparsity, and taking a longer initialisation input (i.e., up to later than March 1) would limit our ability to test intervention measures in Phase 1. This had to be derived from hospital admissions, because the testing capacity was so low in this period that the tested infection cases cannot be used to estimate the real number of infections.

While the conversion of model output to hospital admissions is done using a Weibull distributed time lag, we do the initialisation simpler. We start with hospital admission

data in the observed data (i.e., it is the other way around), which are specified per municipality. We use a flat percentage $p_{\text{hos},0}$, which is the weighted average of all hospital admission probabilities in Tab. I.2, weighted by the respective category size, and we translate these numbers back 14 days into the past (instead of fully Weibull-distributed). The respective affected agents are randomly drawn (within each specified municipality) based on the resulting numbers. These were assigned the disease stage ‘infectious’ I , while the rest of the population was assigned ‘susceptible’ S . We aim to approximate in the initialisation the number of infection cases up to March 1, which means we have to take hospital admission data up to March 15.

I.1.6 Disease transmission and λ [Step 6]

The transmission dynamics and force of infection λ are already discussed in the Methods section of the chapter 9. This section focuses on the susceptibility parameter and the daily cycle, both part of the equation for λ . The first part of the equation involves a parameter $h(g)$ that reflects the susceptibility to the disease, based on the demographic group g . The age-specific susceptibility parameter shown in Tab. I.2 is based on estimations used in Rijksinstituut voor Volksgezondheid en Milieu (RIVM) (2021), table A3. The second part ($\beta(t) \cdot \bar{s}(t)$) contains a parameter β that involves behavioural aspects like wearing face masks and keeping social distance — this parameter is used for distinguishing four phases in the epidemic as described later. The parameter $\bar{s}(t) = \frac{s(t)}{\sum_t s(t)}$ (where $t \in [0, 24]$ in hours) is what we refer to as the ‘daily cycle parameter’, reflecting the fact that people hardly mix during the night, and more throughout the day (see for values of $s(t)$ per hour in Tab. I.3).

Hour of day (t)	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
$s(t)$	1	1	1	1	1	1	0.75	0.5	0.25	0	0	0	0	0	0	0	0	0	0	0.2	0.4	0.6	0.8	1

Table I.3: Daily cycle parameter $s(t)$, from which we obtain $\bar{s}(t)$.

I.1.7 Intervention data sources [Step 7]

There are four factors that we apply to mimic intervention measures. The first are behavioural changes such as wearing face masks and social distancing, represented by the time-varying values of β . We use the β 's as calibration parameters and do not use any external data source for their determination, other than the qualitative fact that governmental policy on this was issued from phase 2 and on, being more strictly adhered to in phase 3, and loosened in phase 4. In other words: β should decrease a bit in phase 2, even further in phase 3, and increase again in phase 4. The phases are summarised in Tab. 9.1.

Other simulated intervention measures were related to mobility, restricting various events and working branches, and applying a working-from-home policy, which were all informed by data. The reduction in inter-municipality travel was quantified using Google Mobility data, which describe how mobility changed across this period across six categories (shown in Fig. I.4). Using the average of three of these categories — transit stations, workplaces and retail & recreation, chosen because these reflect inter-municipality travel best — and averaging the mobility changes within each phase, we end up with three

scalar percentages for phases 2, 3 and 4, representing a mobility reduction. These percentages are implemented by randomly selecting the respective percentage of the population among the employed demographic categories, and placing them at home.

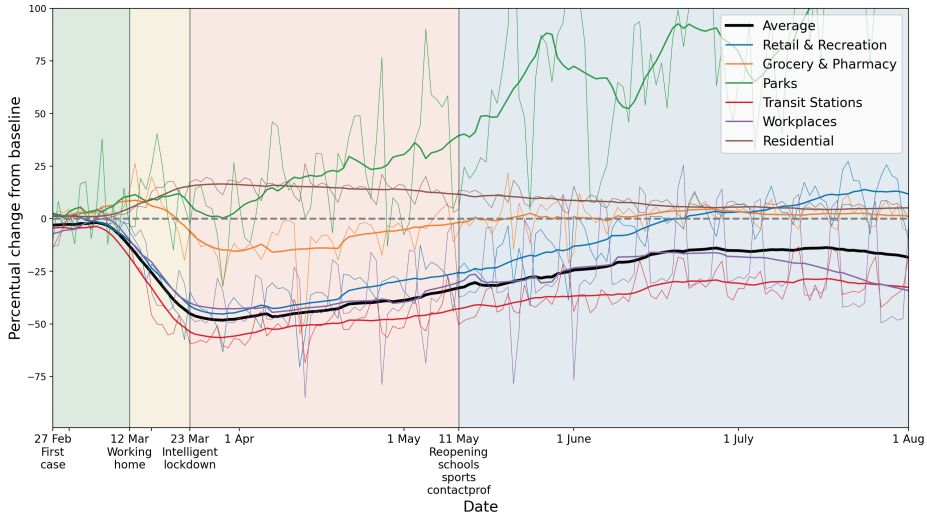


Figure I.4: Percentual mobility changes with respect to the baseline, provided by Google Mobility data. Six categories are distinguished, in different colours. For the purpose of measuring how inter-municipality travel reduced, we use the average of the categories 'Transit Stations', 'Workplaces' and 'Retail and Recreation'. For each colour, the raw data (thin lines) and 14-day running averaged data (thicker lines) are shown.

Across the first wave of COVID-19, people also started mixing differently. This is explicitly measured in the Netherlands using surveys in the PIENTER study (Vos et al., 2021). There, three survey studies are done: one in February 2020, one in April 2020 and one in June 2020. We only use this data in terms of their percentual changes: that of April with respect to February, and that of June with respect to February. The authors did not distinguish the same four unique mixing situations (home, work, school and other) as we use here, so we translate their (age-stratified) mixing changes into a single 11-by-11 matrix representing percentual changes in the contact rates between the demographic categories, and apply these percentual changes to all four mixing matrices in the same manner. In particular, the mixing changes in phase 2 and 3 are determined by the percentual mixing changes from the April surveys with respect to February surveys, and analogously, the changes in phase 4 are determined from the June surveys. See Fig. I.5.

A final important relevant factor to address among the changes and interventions during the first wave, was the closing of schools. The specific school closure dates were March 16, 2020 - May 10, 2020, which are directly used in the model, which in particular means that the schools get closed halfway phase 2. School closure is implemented by two aspects. First, during daytime, all agents belonging to the demographic categories *Primary school children*, *Secondary school children* and *Students* are placed at their home municipality, and we now utilise the 'home' mixing matrix instead of the 'school' mixing matrix to determine their mixing. Second we incorporate the effect of parents of pri-

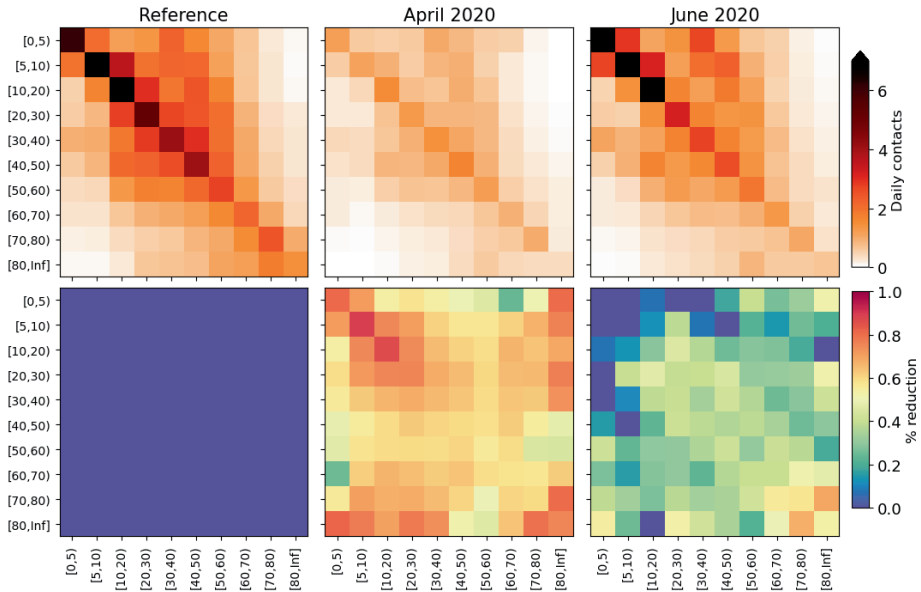


Figure I.5: Changes applied to the model mixing matrices, calculated from survey mixing matrices in the PIENTER study (Vos et al., 2021). **Top panels:** daily contacts in absolute numbers, stratified by age as depicted in the reference study (left), April 2020 surveys (middle) and June 2020 surveys (right). **Bottom panels:** element-wise reductions in contact numbers relative to the reference. These reduction percentages are converted to the demographic groups we use in this study and applied throughout the phases.

mary school children being forced to stay at home because their children are not going to school anymore. This means that, also during daytime, we place 12% of the *Middle-age working* people in their home municipality and set their mixing to ‘home’. These are chosen separate from people working at home due to the mobility changes, to prevent double-counting.

The 12% is calculated as follows. In the Netherlands, 84% of people around 45 have children (CBS, 2004). Applying the crude assumption that agents in the Middle-age working group (25-54 years) have equal number of children, we deduce that the ages of those children are uniformly distributed between -6.6 and 22.4 years old (from which you clearly see why we only use the Middle-age working category), using the average age of mothers, which is 31.6 (CBS, 2021). Using the Dutch primary-school ages of 4-12 years, this means that $8/(22.4-6.6) = 28\%$ of these children are at primary schools. Assuming a rough estimate of 50% of those parents actually staying home (the rest having babysitters, family members or other means of taking care of their children), we end up with $0.84 \cdot 0.28 \cdot 0.50 = 0.12$, which is 12%.

I.1.8 Connecting model runs to real dates and calibrate [Step 8]

Even though the model is initialised with an approximation of the infection cases up to March 1, each simulation requires a spinup to start mimicking the observed data well.

This initial evolution is different for each simulation, and therefore we have to link each simulation separately to real dates. In other words, the model output has to be calibrated to actual dates. For this, we use the onset of phase 2 as a reference point. In particular, in each time step of the model simulation, we calculate the total number of I and R agents. If this crosses the threshold of 1.8% of the population, phase 2 starts, meaning that this is March 12. Once this is done, the aforementioned calibration using the β values could be done.

I.2 Additional model results

I.2.1 Explanation colours in Fig. 9.1

In Fig. 9.1, step 2, the green colours in the map and widths of the lines are both showing the same data, i.e., the total number of visitors from Amsterdam to other municipalities on March 1, 2019, as marked by the Mezuro data (see SI I.1.2). Several high-visited municipalities are highlighted by using black contours. In Fig. 9.1, step 5, the total number of hospital admissions are shown between February 27 and March 15 in blue shades. Several municipalities with high hospital admission counts are highlighted using black contours. This information is used for the initialisation, as mentioned in SI I.1.5.

I.2.2 Variability in geographic evolution within the ensembles

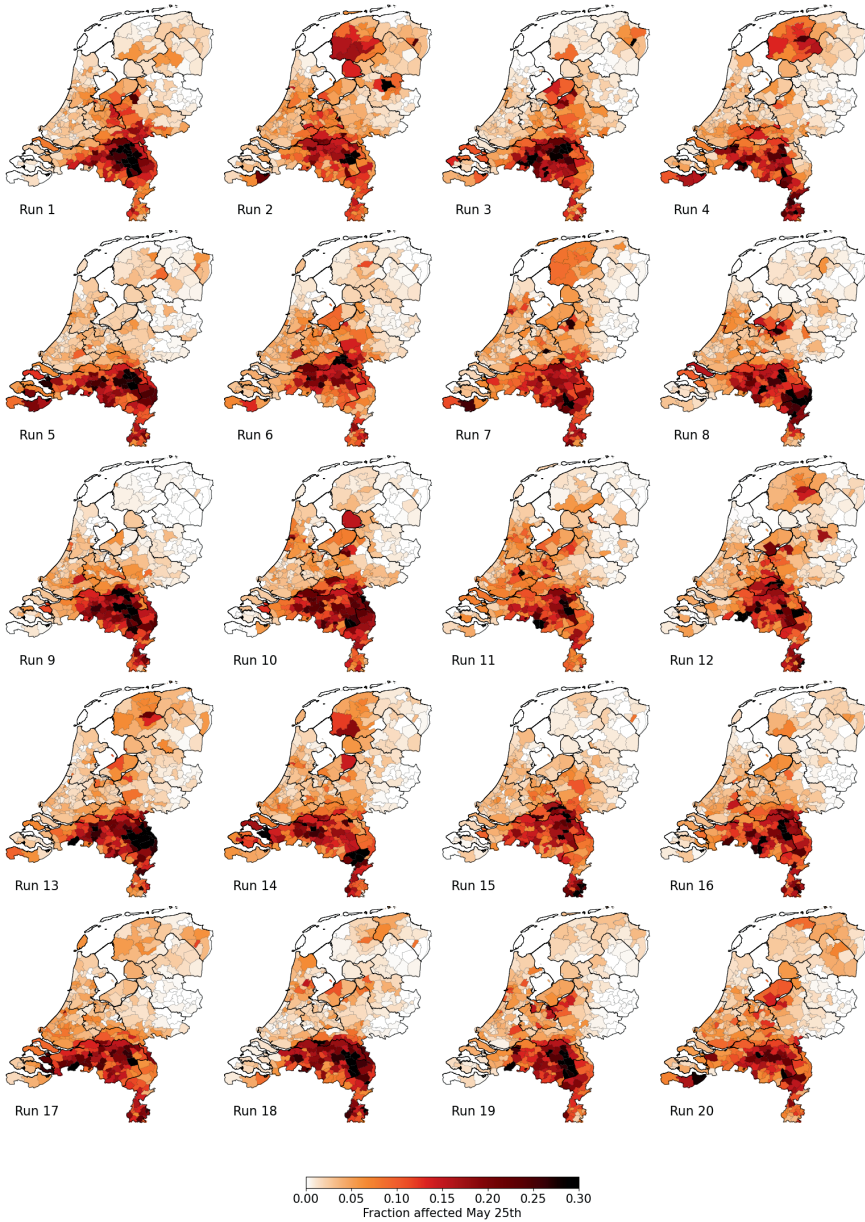


Figure I.6: Percentages affected on May 25 using the Reference scenario, for 20 unique runs (out of an ensemble of 40). Each column represents identical mobility seeds.

Bibliography

- Ahajjam, S. and Badir, H. (2018). Identification of influential spreaders in complex networks using HybridRank algorithm. *Scientific Reports*, 8(1):11932.
- Albert, R., Jeong, H., and Barabási, A. L. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382.
- Aleina, F. C., Baudena, M., D’Andrea, F., and Provenzale, A. (2013). Multiple equilibria on planet dune: Climate-vegetation dynamics on a sandy planet. *Tellus B*, 65(0).
- Allouche, O., Tsoar, A., and Kadmon, R. (2006). Assessing the accuracy of species distribution models: Prevalence, kappa and the true skill statistic (TSS). *Journal of Applied Ecology*, 43(6):1223–1232.
- Almaas, E., Kovács, B., Vicsek, T., Oltvai, Z. N., and Barabási, A. L. (2004). Global organization of metabolic fluxes in the bacterium *Escherichia coli*. *Nature*, 427(6977):839–843.
- Anderson, P. W. (1972). More is different. *Science*, 177(4047):393–396.
- Aristotle (350 BC). *Metaphysics*, volume 2. Translated in 1924 in Harvard University Press Cambridge, MA.
- Ashwin, P., Creaser, J., and Tsaneva-Atanasova, K. (2017). Fast and slow domino effects in transient network dynamics. *Phys. Rev. E*, 96(5):459.
- Ashwin, P., Wieczorek, S., Vitolo, R., and Cox, P. (2012). Tipping points in open systems: Bifurcation, noise-induced and rate-dependent examples in the climate system. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 370(1962):1166–1184.
- Backer, J. A., Mollema, L., Vos, E. R., Klinkenberg, D., van der Klis, F. R., de Melker, H. E., van den Hof, S., and Wallinga, J. (2021). Impact of physical distancing measures against COVID-19 on contacts and mixing patterns: repeated cross-sectional surveys, the Netherlands, 2016–17, April 2020 and June 2020. *Eurosurveillance*, 26(8).
- Badie-Modiri, A., Karsai, M., and Kivela, M. (2020). Efficient limited-time reachability estimation in temporal networks. *Phys. Rev. E*, 101:052303.
- Balcan, D., Colizza, V., Gonçalves, B., Hud, H., Ramasco, J. J., and Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences of the United States of America*, 106(51):21484–21489.
- Ball, R., Panja, D., and Barkema, G. (2016). A two component railway network model exhibiting service collapse. *Warwick Research Archive eprint 81367*.
- Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286(5439):509–512.
- Barriopedro, D., García-Herrera, R., Lupo, A. R., and Hernández, E. (2006). A climatology of Northern Hemisphere blocking. *Journal of Climate*, 19(6):1042–1063.

- Bartolo, R., Saunders, R. C., Mitz, A. R., and Averbeck, B. B. (2020). Dimensionality, information and learning in prefrontal cortex. *PLoS Computational Biology*, 16(4):e1007514.
- Bassett, D. S. and Sporns, O. (2017). Network neuroscience. *Nature Neuroscience*, 20(3):353–364.
- Bathiany, S., Claussen, M., and Fraedrich, K. (2012). Implications of climate variability for the detection of multiple equilibria and for rapid transitions in the atmosphere-vegetation system. *Climate Dynamics*, 38(9):1775–1790.
- Bathiany, S., van der Bolt, B., Williamson, M., Lenton, T., Scheffer, M., van Nes, E., and Notz, D. (2016). Statistical indicators of Arctic sea-ice stability - prospects and limitations. *The Cryosphere*, 10:1631–1645. DOI: 10.5194/tc-10-1631-2016.
- Batra, S. W. T. (1968). Behavior of some social and solitary halictine bees within their nests: A comparative study (hymenoptera: Halictidae). *Journal of the Kansas Entomological Society*, 41(1):120–133.
- Berardi, C., Antonini, M., Genie, M. G., Cotugno, G., Lanteri, A., Melia, A., and Paolucci, F. (2020). The COVID-19 pandemic in Italy: Policy and technology impact on health and non-health outcomes. *Health policy and technology*, 9(4):454–487.
- Berger, A., Gebhardt, A., Müller-Hannemann, M., and Ostrowski, M. (2011). Stochastic delay prediction in large train networks. *OpenAccess Series in Informatics*, 20:100–111.
- Bhatia, U., Kumar, D., Kodra, E., and Ganguly, A. R. (2015). Network science based quantification of resilience demonstrated on the indian railways network. *PLOS ONE*, 10(11):e0141890.
- Blanken, T., Tanis, C. C., Nauta, E., Dablander, F., Zijlstra, B., Bouten, R., Oostvogel, Q., Boersma, M., van der Steenhoven, M., van Harreveld, F., and et al. (2020). Smart distance lab: A new methodology for assessing social distancing interventions.
- Blondel, V. D., Guillaume, J.-L., Lambiotte, R., and Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10):P10008.
- Box, G. E. P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356):791–799.
- Brauner, J. M., Mindermann, S., Sharma, M., Johnston, D., Salvatier, J., Gavenčič, T., Stephenson, A. B., Leech, G., Altman, G., Mikulik, V., Norman, A. J., Monrad, J. T., Besiroglu, T., Ge, H., Hartwick, M. A., Teh, Y. W., Chindelevitch, L., Gal, Y., and Kulveit, J. (2021). Inferring the effectiveness of government interventions against covid-19. *Science*, 371(6531):eabd9338.
- Brier, G. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.
- Brockmann, D. and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *Science*, 342(6164):1337–1342.
- Brummitt, C. D., Barnett, G., and D'Souza, R. M. (2015). Coupled catastrophes: sudden shifts cascade and hop among interdependent systems. *Journal of The Royal Society Interface*, 12(112):20150712–12.
- Büchel, B., Spaninger, T., and Corman, F. (2020). Empirical dynamics of railway delay propagation identified during the large-scale Rastatt disruption. *Scientific Reports*, 10(1):18584.
- Büker, T. and Seybold, B. (2012). Stochastic modelling of delay propagation in large networks. *Journal of Rail Transport Planning and Management*, 2(1-2):34–50.
- Buldirev, S. V., Parshani, R., Paul, G., Stanley, H. E., and Havlin, S. (2010). Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028.
- Buzsáki, G. (2010). Neural syntax: cell assemblies, synapsembles and readers. *Neuron*, 68(3):362.
- Buzsáki, G. and Llinás, R. (2017). Space and time in the brain. *Science*, 358(6362):482.
- Calhoun, V. D., Liu, J., and Adali, T. (2009). A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *NeuroImage*, 45(1, Supplement 1):S163–S172.
- Campanelli, B., Fleurquin, P., Eguiluz, V. M., Ramasco, J. J., Arranz, A., Etxebarria, I., and Ciruelos, C. (2014). Modeling Reactionary Delays in the European Air Transport Network. *Fourth SESAR Innovation Days*, 1:1–6.

Bibliography

- Cardoso, J.-F. (1999). High-order contrasts for independent component analysis. *Neural Computation*, 11(1):157–192.
- Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J.-E., and Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed rfid sensor networks. *PLOS ONE*, 5(7):1–9.
- CBS (2004). Kinderloosheid en opleidingsniveau. <https://www.cbs.nl/nl-nl/nieuws/2004/22/kinderloosheid-en-opleidingsniveau>. Online; accessed 13 November 2021.
- CBS (2021). Kinderen krijgen. <https://www.cbs.nl/nl-nl/visualisaties/dashboard-bevolking/levensloop/kinderen-krijgen>. Online; accessed 13 November 2021.
- Centola, D. (2010). The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197.
- Centola, D., Eguíluz, V. M., and Macy, M. W. (2007). Cascade dynamics of complex propagation. *Physica A*, 374(1):449–456.
- Centraal Bureau voor de Statistiek (2021). Bijna 4 op de 10 werkt en woont in dezelfde gemeente. <https://www.cbs.nl/nl-nl/nieuws/2017/32/bijna-4-op-de-10-werkt-en-woont-in-dezelfde-gemeente>. Online; accessed 6 November 2021.
- Cerreto, F., Nielsen, B. F., Nielsen, O. A., and Harrod, S. S. (2018). Application of Data Clustering to Railway Delay Pattern Recognition. *Journal of Advanced Transportation*, 2018.
- Cessi, P. (1994). A simple box model of stochastically forced thermohaline flow. *Journal of Physical Oceanography*, 24:1911–1920.
- Chen, C. (2014). Dynamic network zoning method based on community detection. In *2014 17th IEEE International Conference on Intelligent Transportation Systems, ITSC 2014*, pages 2749–2755. Institute of Electrical and Electronics Engineers Inc.
- Chen, V. and Ruan, S. (2009). Graph Cut Based Segmentation of Brain Tumor From MRI Images. *International Journal on Sciences and Techniques of Automatic control & computer engineering*, 3(2):1054–1063.
- Choi, D., Chun, S., Oh, H., Han, J., and Kwon, T. T. (2020). Rumor propagation is amplified by echo chambers in social media. *Scientific Reports*, 10(1):310.
- Chung, S. H., Ma, H. L., and Chan, H. K. (2017). Cascading Delay Risk of Airline Workforce Deployments with Crew Pairing and Schedule Optimization. *Risk Analysis*, 37(8):1443–1458.
- Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4):661–703.
- Cohen, M. X. (2021a). A tutorial on generalized eigendecomposition for source separation in multichannel electrophysiology. *arXiv*.
- Cohen, M. X. (2021b). A data-driven method to identify frequency boundaries in multichannel electrophysiology data. *Journal of Neuroscience Methods*, 347:108949.
- Cohen, M. X., Englitz, B., and França, A. S. C. (2021). Large-Scale and Multiscale Networks in the Rodent Brain during Novelty Exploration. *eNeuro*, 8(3).
- Corman, F., D’Ariano, A., and Hansen, I. A. (2014). Evaluating disturbance robustness of railway schedules. *Journal of Intelligent Transportation Systems*, 18(1):106–120.
- Corman, F. and Kecman, P. (2018). Stochastic prediction of train delays in real-time using Bayesian networks. *Transportation Research Part C: Emerging Technologies*, 95:599–615.
- Creaser, J., Tsaneva-Atanasova, K., and Ashwin, P. (2018). Sequential noise-induced escapes for oscillatory network dynamics. *SIAM Journal on Applied Dynamical Systems* *SIAM Journal on Applied Dynamical Systems*, 17(1):500–525.
- Şahin, I. (2017). Markov chain model for delay distribution in train schedules: Assessing the effectiveness of time allowances. *Journal of Rail Transport Planning and Management*, 7(3):101–113.

- Cuadra, L., Salcedo-Sanz, S., Del Ser, J., Jiménez-Fernández, S., and Geem, Z. W. (2015). A critical review of robustness in power grids using complex networks concepts.
- Cunningham, J. P. and Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.*, 17(11):1500–1509.
- Dakos, V., Matthews, B., Hendry, A. P., Levine, J., Loeuille, N., Norberg, J., Nosil, P., Scheffer, M., and De Meester, L. (2019). Ecosystem tipping points in an evolving world. *Nature Ecology & Evolution*, 3(3):355–362.
- Dakos, V., Scheffer, M., van Nes, E. H., Brovkin, V., Petoukhov, V., and Held, H. (2008). Slowing down as an early warning signal for abrupt climate change. *Proceedings of the National Academy of Sciences*, 105(38):14308–14312.
- Dam, A. van, Dekker, M. M., Morales-Castilla, I., Rodríguez, M. Á., Wichmann, D., and Baudena, M. (2021). Correspondence analysis, spectral clustering and graph embedding: applications to ecology and economic complexity. *Scientific reports*, 11(1):1–14.
- Dan, J. M., Mateus, J., Kato, Y., Hastie, K. M., Yu, E. D., Faliti, C. E., Grifoni, A., Ramirez, S. I., Haupt, S., Frazier, A., Nakao, C., Rayaprolu, V., Rawlings, S. A., Peters, B., Krammer, F., Simon, V., Saphire, E. O., Smith, D. M., Weiskopf, D., Sette, A., and Crotty, S. (2021). Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection. *Science*, 371(6529).
- Dasgupta, A., Leskovec, J., Mahoney, M. W., and Lang, K. J. (2009). Community structure in large networks: Natural cluster sizes and the absence of large well-defined clusters. *Internet Mathematics*, 6(1).
- de Vlas, S. J. and Coffeng, L. E. (2021). Achieving herd immunity against COVID-19 at the country level by the exit strategy of a phased lift of control. *Scientific Reports*, 11(1):4445.
- DeConto, R. M. and Pollard, D. (2003). Rapid Cenozoic glaciation of Antarctica induced by declining atmospheric CO₂. *Nature*, 421:245–249.
- Dekker, M. M. (2021). Geographic delay characterization of railway systems. *Scientific Reports*, 11(1):1–13.
- Dekker, M. M., Blanken, T. F., Dablander, F., Ou, J., Borsboom, D., and Panja, D. (2021a). Quantifying agent impacts on contact sequences in social interactions. *arXiv:2107.01443*, accepted in *Scientific Reports*.
- Dekker, M. M., Coffeng, L., Pijpers, E., Panja, D., and Vlas, S. (2021b). Evaluation of non-pharmaceutical interventions during the COVID-19 first wave in the Netherlands. *Manuscript in preparation*.
- Dekker, M. M., França, A. S. C., Panja, D., and Cohen, M. X. (2021c). Characterizing neural phase-space trajectories via principal component clustering. *Journal of Neuroscience Methods*, 362:109313.
- Dekker, M. M., Haarsma, R. J., de Vries, H., Baatsen, M., and van Delden, A. J. (2018a). Characteristics and development of European cyclones with tropical origin in reanalysis data. *Climate dynamics*, 50(1):445–455.
- Dekker, M. M., Medvedev, A. N., Rombouts, J., Siudem, G., and Tupikina, L. (2021d). Modelling railway delay propagation as diffusion-like spreading. *arXiv:2105.06111*.
- Dekker, M. M. and Panja, D. (2019). A reduced phase-space approach to analyse railway dynamics. *IFAC-PapersOnLine*, 52(3):1–6.
- Dekker, M. M. and Panja, D. (2021). Cascading dominates large-scale disruptions in transport over complex networks. *PLOS ONE*, 16(1):e0246077.
- Dekker, M. M., Panja, D., Dijkstra, H. A., and Dekker, S. C. (2019). Predicting transitions across macroscopic states for railway systems. *PLOS ONE*, 14(6):e0217710.
- Dekker, M. M., Schram, R., Ou, J., and Panja, D. (2021e). The hidden dependence of spreading vulnerability on topological complexity. *arXiv:2107.01651*.
- Dekker, M. M., van Lieshout, R. N., Ball, R. C., Bouman, P. C., Dekker, S. C., Dijkstra, H. A., Goverde, R. M. P., Huisman, D., Panja, D., Schaafsma, A. A. M., and van den Akker, M. (2021f). A next

- step in disruption management: combining operations research and complexity science. *Public Transport*, pages 1–22.
- Dekker, M. M., Von Der Heydt, A. S., and Dijkstra, H. A. (2018b). Cascading transitions in the climate system. *Earth System Dynamics*, 9(4):1243–1260.
- Dekker, M. M., Walentek, D., Haslbeck, J., and Broere, J. (2020). A general framework for international conflict under incomplete information. *APSA preprints*.
- Delorme, A. and Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134(1):9–21.
- Dermody, B. J., Sivapalan, M., Stehfest, E., Van Vuuren, D. P., Wassen, M. J., Bierkens, M. F., and Dekker, S. C. (2018). A framework for modelling the complexities of food and water security under globalisation. *Earth System Dynamics*, 9(1):103–118.
- Descartes, R. (1637). *A Discourse on the Method*. Translated in 2007 in Oxford: Oxford University Press.
- Dezső, Z. and Barabási, A.-L. (2002). Halting viruses in scale-free networks. *Phys. Rev. E*, 65:055103.
- Diego, X., Marcon, L., Müller, P., and Sharpe, J. (2018). Key Features of Turing Systems are Determined Purely by Network Topology. *Physical Review X*, 8(2):021071.
- Dijkstra, H. A. and Neelin, D. (1995). Ocean-Atmosphere Interaction and the Tropical Climatology. Part II: Why the Pacific Cold Tongue is in the East. *J. Climate*, 8:1343–1359. printed.
- Dong, B. and Sutton, R. (2007). Enhancement of ENSO variability by a weakened Atlantic meridional circulation in a coupled gcm. *Journal of Climate*, 20(19):4920–4939.
- Dong, S., Wang, H., Mostafavi, A., and Gao, J. (2019). Robust component: a robustness measure that incorporates access to critical facilities under disruptions. *Journal of Royal Society Interface*, 16:20190149.
- Elias, C., Sekri, A., Leblanc, P., Cucherat, M., and Vanhems, P. (2021). The incubation period of COVID-19: A meta-analysis. *International Journal of Infectious Diseases*, 104:708–710.
- Elsworth, G., Galbraith, E., Halverson, G., and Yang, S. (2017). Enhanced weathering and CO₂ draw-down caused by latest Eocene strengthening of the Atlantic meridional overturning circulation. *Nature Geoscience*.
- Endo, A. et al. (2020). Estimating the overdispersion in COVID-19 transmission using outbreak sizes outside China. *Wellcome open research*, 5.
- Erath, A., Löchl, M., and Axhausen, K. W. (2009). Graph-theoretical analysis of the swiss road and railway networks over time. *Networks and Spatial Economics*, 9(3):379–400.
- Erdos, P., Rényi, A., et al. (1960). On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci.*, 5(1):17–60.
- European Commission (2016). Statistical Pocketbook EU Transport. *European Union*.
- Firth, J. A., Hellewell, J., Klepac, P., Kissler, S., Jit, M., Atkins, K. E., Clifford, S., Villabona-Arenas, C. J., Meakin, S. R., Diamond, C., Bosse, N. I., Munday, J. D., Prem, K., Foss, A. M., Nightingale, E. S., Zandvoort, K. v., Davies, N. G., Gibbs, H. P., Medley, G., Gimma, A., Flasche, S., Simons, D., Auzenbergs, M., Russell, T. W., Quilty, B. J., Rees, E. M., Leclerc, Q. J., Edmunds, W. J., Funk, S., Houben, R. M. G. J., Knight, G. M., Abbott, S., Sun, F. Y., Lowe, R., Tully, D. C., Procter, S. R., Jarvis, C. I., Endo, A., O'Reilly, K., Emery, J. C., Jombart, T., Rosello, A., Deol, A. K., Quaife, M., Hué, S., Liu, Y., Eggo, R. M., Pearson, C. A. B., Kucharski, A. J., Spurgin, L. G., and Group, C. C.-. W. (2020). Using a real-world network to model localized COVID-19 control strategies. *Nature Medicine*, 26(10):1616–1622.
- Flaxman, S., Mishra, S., Gandy, A., Unwin, H. J. T., Mellan, T. A., Coupland, H., Whittaker, C., Zhu, H., Berah, T., Eaton, J. W., et al. (2020). Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe. *Nature*, 584(7820):257–261.
- Fleurquin, P., Ramasco, J. J., and Eguiluz, V. M. (2013). Systemic delay propagation in the US airport network. *Scientific Reports*, 3(1):1159.

- Fortunato, S. and Hric, D. (2016). Community detection in networks: A user guide. *Physics Reports*, 659:1 – 44. Community detection in networks: A user guide.
- França, A. S. C., van Hulst, J. A., and Cohen, M. X. (2020). Low-cost and versatile electrodes for extracellular chronic recordings in rodents. *Heliyon*, 6(9).
- Franks, N. R. (1989). Army ants : a collective intelligence. *Am. Scientist*, 77:139–145.
- França, A. S. C., Borgesius, N. Z., Souza, B. C., and Cohen, M. X. (2021). Beta2 Oscillations in Hippocampal-Cortical Circuits During Novelty Detection. *Frontiers in Systems Neuroscience*, 15:8.
- Gambardella, L. M., Rizzoli, A. E., and Funk, P. (2002). Agent-based Planning and Simulation of Combined Rail/Road Transport. *SIMULATION*, 78(5):293–303.
- Gao, J., Buldyrev, S. V., Havlin, S., and Stanley, H. E. (2011). Robustness of a network of networks. *Phys. Rev. Lett.*, 107:195701.
- Gasson, E., Lunt, D. J., DeConto, R. M., Goldner, A., Heinemann, M., Huber, M., LeGrande, A. N., Pollard, D., Sagoo, N., Siddall, M., Winguth, a., and Valdes, P. J. (2014). Uncertainties in the modelled CO₂ threshold for Antarctic glaciation. *Climate of the Past*, 10(2):451–466.
- Geels, F. W. (2004). From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. *Research Policy*, 33(6-7):897–920.
- Gehrke, J., Ginsparg, P., and Kleinberg, J. (2003). Overview of the 2003 kdd cup. *SIGKDD Explor. Newsl.*, 5(2):149–151.
- Gelardi, V., Godard, J., Paleressompoulle, D., Claidiere, N., and Barrat, A. (2020). Measuring social networks in primates: wearable sensors versus direct observations. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 476(2236):20190737.
- Gemmetto, V., Barrat, A., and Cattuto, C. (2014). Mitigation of infectious disease at school: targeted class closure vs school closure. *BMC infectious diseases*, 14(1):695.
- Génois, M. and Barrat, A. (2018). Can co-location be used as a proxy for face-to-face contacts? *EPJ Data Science*, 7(1):11.
- Ghaemi, N. (2018). *Short-turning Trains during Full Blockages in Railway Disruption Management*. PhD thesis, TU Delft.
- Ghaemi, N., Cats, O., and Goverde, R. M. (2018). Macroscopic multiple-station short-turning model in case of complete railway blockages. *Transportation Research Part C: Emerging Technologies*, 89:113–132.
- Gildor, H. and Tziperman, E. (2000). Sea ice as the glacial cycles' climate switch: role of seasonal and orbital forcing. *Paleoceanography*, 15(6):605–615.
- Girvan, M. and Newman, M. E. J. (2002). Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, 99(12):7821–7826.
- Glaser, J. I., Whiteway, M., Cunningham, J. P., Paninski, L., and Linderman, S. W. (2020). Recurrent Switching Dynamical Systems Models for Multiple Interacting Neural Populations. *bioRxiv*, page 2020.10.21.349282.
- Google (2021). Google Community Mobility Report 2020. <https://www.google.com/covid19/mobility/>. Online; accessed 31 October 2021.
- Goverde, R. M. P. (2010). A delay propagation algorithm for large-scale railway traffic networks. *Transportation Research Part C: Emerging Technologies*, 18(3):269–287.
- Grabowicz, P. A., Aiello, L. M., and Menczer, F. (2014). Fast filtering and animation of large dynamic networks. *EPJ Data Science*, 3(1):27.
- Gu, Y., Fu, X., Liu, Z., Xu, X., and Chen, A. (2020). Performance of transportation network under perturbations: Reliability, vulnerability, and resilience. *Transportation Research Part E: Logistics and Transportation Review*, 133:101809.
- Guo, W., Toader, B., Feier, R., Mosquera, G., Ying, F., Oh, S.-W., Price-Williams, M., and Krupp, A. (2019). Global air transport complex network: multi-scale analysis. *SN Applied Sciences*, 1(7):680.

- Hafiene, N., Karoui, W., and Romdhane, L. B. (2020). Influential nodes detection in dynamic social networks: A Survey. *Expert Systems with Applications*, page 113642.
- Harrod, S., Cerreto, E., and Nielsen, O. A. (2019). A closed form railway line delay propagation model. *Transportation Research Part C: Emerging Technologies*, 102:189–209.
- He, X., Lau, E. H. Y., Wu, P., Deng, X., Wang, J., Hao, X., Lau, Y. C., Wong, J. Y., Guan, Y., Tan, X., Mo, X., Chen, Y., Liao, B., Chen, W., Hu, F., Zhang, Q., Zhong, M., Wu, Y., Zhao, L., Zhang, F., Cowling, B. J., Li, F., and Leung, G. M. (2020). Temporal dynamics in viral shedding and transmissibility of COVID-19. *Nature Medicine*, 26(5):672–675.
- Hechmann, I., Comes, T., and Nickel, S. (2015). A critical review of supply chain risk — Definition, measure and modeling. *Omega*, 52:119–132.
- Hekkert, M. P., Suurs, R. A., Negro, S. O., Kuhlmann, S., and Smits, R. E. (2007). Functions of innovation systems: A new approach for analysing technological change. *Technological Forecasting and Social Change*, 74(4):413–432.
- Helbing, D. (2013). Globally networked risks and how to respond. *Nature*, 497(7447):51–59.
- Held, H. and Kleinen, T. (2004). Detection of climate system bifurcations by degenerate fingerprinting. *Geophysical Research Letters*, 31(23):n/a–n/a. L23207.
- Helleringer, S. and Kohler, H.-P. (2007). Sexual network structure and the spread of HIV in Africa: evidence from Likoma Island, Malawi. *Aids*, 21(17):2323–2332.
- Hirota, M., Holmgren, M., Van Nes, E. H., and Scheffer, M. (2011). Global resilience of tropical forest and savanna to critical transitions. *Science*, 334(6053):232–235.
- Holme, P., Kim, B. J., Yoon, C. N., and Han, S. K. (2002). Attack vulnerability of complex networks. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 65(5):14.
- Holme, P. and Saramäki, J. (2012). Temporal networks.
- Huang, P., Wen, C., Peng, Q., Jiang, C., Yang, Y., and Fu, Z. (2019). Modeling the Influence of Disturbances in High-Speed Railway Systems. *Journal of Advanced Transportation*, 2019.
- Infrabel (2021). Open data.
- Inoue, H. and Todo, Y. (2019). Firm-level propagation of shocks through supply-chain networks. *Nature Sustainability*, 2(9):841–847.
- International Transport Forum (ITF) (2009). *Competitive Interaction Between Airports, Airlines and High-Speed Rail*. ITF Round Tables. OECD Publishing.
- International Union of railways (2020). Railway statistics - synopsis 2020.
- Isella, L., Stehlé, J., Barrat, A., Cattuto, C., Pinton, J.-F., and Van den Broeck, W. (2011). What's in a crowd? Analysis of face-to-face behavioral networks. *Journal of Theoretical Biology*, 271(1):166–180.
- Ivanov, D. (2020). Predicting the impacts of epidemic outbreaks on global supply chains: A simulation-based analysis on the coronavirus outbreak (COVID-19/SARS-CoV-2) case. *Transportation Research Part E: Logistics and Transportation Review*, 136:101922.
- Jaffino, B. A., Kwakkel, J., and Verbraeck, A. (2020). Transport network criticality metrics: a comprehensive analysis and a guideline for selection. *Transport Reviews*, 40(2):241–264.
- Janić, M. (2018). Modelling the resilience of rail passenger transport networks affected by large-scale disruptive events: the case of HSR (high speed rail). *Transportation*, 45(4):1101–1137.
- Jespersen-Groth, J., Potthoff, D., Clausen, J., Huisman, D., Kroon, L., Maróti, G., and Nielsen, M. N. (2009). Disruption management in passenger railway transportation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5868 LNCS, pages 399–421. Springer, Berlin, Heidelberg.
- Jungclauss, J. H., Haak, H., Esch, M., Roeckner, E., and Marotzke, J. (2006). Will Greenland melting halt the thermohaline circulation? *Geophysical Research Letters*, 33(17):n/a–n/a. L17708.

- Kadir, R. A., Shima, Y., Sulaiman, R., and Ali, F. (2018). Clustering of public transport operation using K-means. In *2018 IEEE 3rd International Conference on Big Data Analysis, ICBDA 2018*, pages 427–432. Institute of Electrical and Electronics Engineers Inc.
- Kaluza, P., Kölzsch, A., Gastner, M. T., and Blasius, B. (2010). The complex network of global cargo ship movements. *Journal of the Royal Society Interface*, 7(48):1093–1103.
- Kantelhardt, J., Koscielny-Bunde, E., Rego, H., Havlin, S., and Bunde, A. (2001). Detecting long-range correlations with detrended fluctuation analysis. *Physica A*, 295:441–454.
- Kecman, P. and Goverde, R. M. P. (2015a). Online data-driven adaptive prediction of train event times. *IEEE Transactions on Intelligent Transportation Systems*, 16(1):465–474.
- Kecman, P. and Goverde, R. M. P. (2015b). Predictive modelling of running and dwell times in railway traffic. *Public Transport*, 7(3):295–319.
- Kendall, D. G. (1956). *Deterministic and stochastic epidemics in closed populations*, pages 149–166. University of California Press.
- Kenniscentrum Studentenhuisvesting (2020). Landelijke monitor studentenhuisvesting 2020. <https://www.kences.nl/publicaties/landelijke-monitor-studentenhuisvesting-2020/>.
- Kermack, W. O., McKendrick, A. G., and Walker, G. T. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 115(772):700–721.
- Khorasani, A., Shalchyan, V., and Daliri, M. R. (2019). Adaptive Artifact Removal From Intracortical Channels for Accurate Decoding of a Force Signal in Freely Moving Rats. *Front. Neurosci.*, 13.
- Kim, H. and Anderson, R. (2012). Temporal node centrality in complex networks. *Phys. Rev. E*, 85:026107.
- Kiti, M. C., Melegaro, A., Cattuto, C., and Nokes, D. J. (2019). Study design and protocol for investigating social network patterns in rural and urban schools and households in a coastal setting in Kenya using wearable proximity sensors. *Wellcome Open Res.*, 4(84):17.
- Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., and Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893.
- Kivelä, M., Arenas, A., Barthelemy, M., Gleeson, J. P., Moreno, Y., and Porter, M. A. (2014). Multilayer networks. *Journal of complex networks*, 2(3):203–271.
- Klausberger, T. and Somogyi, P. (2008). Neuronal diversity and temporal dynamics: The unity of hippocampal circuit operations. *Science*, 321(5885):53–57.
- Kobayashi, T., Takaguchi, T., and Barrat, A. (2019). The structured backbone of temporal social ties. *Nature Communications*, 10(1):220.
- Kossinets, G. and Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science*, 311(5757):88–90.
- Kotsakos, D., Sakkos, P., Katakis, I., and Gunopulos, D. (2015). Language agnostic meme-filtering for hashtag-based social network analysis. *Social Network Analysis and Mining*, 5(1):28.
- Kretzschmar, M. E., Rozhnova, G., Bootsma, M., van Boven, M., van de Wijgert, J., and Bonten, M. (2020). Time is of the essence: Impact of delays on effectiveness of contact tracing for COVID-19, a modelling study. *medRxiv*, page 2020.05.09.20096289.
- Kriegeskorte, N. and Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain.
- Kriegler, E., W Hall, J., Held, H., Dawson, R., and Schellnhuber, H. (2009). Imprecise probability assessment of tipping points in the climate system. *Proceedings of the National Academy of Sciences of the United States of America*, 106:5041–6.
- Kuehn, C. (2011). A mathematical framework for critical transitions: Bifurcations, fast-slow systems and stochastic dynamics. *Physica D*, 240(12):1020–1035.

Bibliography

- Kutzbach, J., Bonan, G., Foley, J., and Harrison, S. (1996). Vegetation and soil feedbacks on the response of the African monsoon to orbital forcing in the early to middle Holocene. *Nature*, 384(6610):623–626.
- Ladyman, J., Lambert, J., and Wiesner, K. (2013). What is a complex system? *European Journal for Philosophy of Science*, 3(1):33–67.
- Laplace, P. S. (1814). *Essai philosophique sur les probabilités*. Courcier.
- Le Van Quyen, M. and Bragin, A. (2007). Analysis of dynamic brain oscillations: methodological advances. *Trends in Neurosciences*, 30(7):365–373.
- Lehky, S. R. (2010). Decoding Poisson Spike Trains by Gaussian Filtering. *Neural Computation*, 22(5):1245–1271. PMID: 20028224.
- Lenton, T. M. (2011). Early warning of climate tipping points. *Nature Climate Change*, 1:201–209.
- Lenton, T. M., Held, H., Kriegler, E., Hall, J. W., Lucht, W., Rahmstorf, S., and Schellnhuber, H. J. (2008). Tipping elements in the Earth's climate system. *Proceedings of the National Academy of Sciences*, 105(6):1786–1793.
- Lenton, T. M. and Williams, H. T. (2013). On the origin of planetary-scale tipping points. *Trends in Ecology & Evolution*, 28(7):380–382.
- Leobons, C. M., Gouvêa Campos, V. B., and de Mello Bandeira, R. A. (2019). Assessing Urban Transportation Systems Resilience: A Proposal of Indicators. In *Transportation Research Procedia*, volume 37, pages 322–329. Elsevier.
- Lessan, J., Fu, L., and Wen, C. (2019). A hybrid Bayesian network model for predicting delays in train operations. *Computers and Industrial Engineering*, 127:1214–1222.
- Li, A., Cornelius, S. P., Liu, Y.-Y., Wang, L., and Barabási, A.-L. (2017). The fundamental advantages of temporal networks. *Science*, 358(6366):1042–1046.
- Li, A., Zhou, L., Su, Q., Cornelius, S. P., Liu, Y.-Y., Wang, L., and Levin, S. A. (2020a). Evolution of cooperation on temporal networks. *Nature Communications*, 11(1):2259.
- Li, D., Daamen, W., and Goverde, R. M. (2016). Estimation of train dwell time at short stops based on track occupation event data: A study at a Dutch railway station. *Journal of Advanced Transportation*, 50(5):877–896.
- Li, F., Cheng, D., Hadjieleftheriou, M., Kollios, G., and Teng, S.-H. (2005). On trip planning queries in spatial databases. In Bauzer Medeiros, C., Egenhofer, M. J., and Bertino, E., editors, *Advances in Spatial and Temporal Databases*, pages 273–290, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., Xing, X., Xiang, N., Wu, Y., Li, C., Chen, Q., Li, D., Liu, T., Zhao, J., Liu, M., Tu, W., Chen, C., Jin, L., Yang, R., Wang, Q., Zhou, S., Wang, R., Liu, H., Luo, Y., Liu, Y., Shao, G., Li, H., Tao, Z., Yang, Y., Deng, Z., Liu, B., Ma, Z., Zhang, Y., Shi, G., Lam, T. T., Wu, J. T., Gao, G. F., Cowling, B. J., Yang, B., Leung, G. M., and Feng, Z. (2020b). Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *New England Journal of Medicine*, 382(13):1199–1207. PMID: 31995857.
- Liljeros, F., Edling, C. R., and Amaral, L. A. (2003). Sexual networks: Implications for the transmission of sexually transmitted infections. *Microbes and Infection*, 5(2):189–196.
- Lin, X. (2019). A Road Network Traffic State Identification Method Based on Macroscopic Fundamental Diagram and Spectral Clustering and Support Vector Machine. *Mathematical Problems in Engineering*, 2019.
- Linderman, S., Nichols, A., Blei, D., Zimmer, M., and Paninski, L. (2019). Hierarchical recurrent state space models reveal discrete and continuous dynamics of neural activity in *C. elegans*. *bioRxiv*, page 621540.
- Liu, Q.-H., Ajelli, M., Aleta, A., Merler, S., Moreno, Y., and Vespignani, A. (2018). Measurability of the epidemic reproduction number in data-driven contact networks. *Proceedings of the National Academy of Sciences*, 115(50):12680–12685.

- Livina, V. and Lenton, T. (2007). A modified method for detecting incipient bifurcations in a dynamical system. *Geophysical Research Letters*, 34(3):n/a–n/a. L03712.
- Lloyd, S. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.
- Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., and Getz, W. M. (2005). Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359.
- Lordan, O., Sallan, J. M., Simo, P., and Gonzalez-Prieto, D. (2015). Robustness of airline alliance route networks. *Communications in Nonlinear Science and Numerical Simulation*, 22(1-3):587–595.
- Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of atmospheric sciences*, 20(2):130–141.
- Luczak, A., Barthó, P., and Harris, K. D. (2009). Spontaneous Events Outline the Realm of Possible Sensory Responses in Neocortical Populations. *Neuron*, 62(3):413–425.
- Ludvigsen, J. and Klæboe, R. (2014). Extreme weather impacts on freight railways in Europe. *Natural Hazards*, 70(1):767–787.
- Luxburg, U. V. (2007). A Tutorial on Spectral Clustering A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4):395–416.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–296. The Regents of the University of California.
- Makeig, S., Debener, S., Onton, J., and Delorme, A. (2004). Mining event-related brain dynamics. *Trends Cogn. Sci.*, 8(5):204–210.
- Mancastropa, M., Vezzani, A., Muñoz, M. A., and Burioni, R. (2019). Burstiness in activity-driven networks and the epidemic threshold. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(5):053502.
- Marković, N., Milinković, S., Tikhonov, K. S., and Schonfeld, P. (2015). Analyzing passenger train arrival delays with support vector regression. *Transportation Research Part C: Emerging Technologies*, 56:251–262.
- Marks, M., Millat-Martinez, P., Ouchi, D., Roberts, C., Alemany, A., Corbacho-Monné, M., Ubals, M., Tobias, A., Tebé, C., Ballana, E., Bassat, Q., Baro, B., Vall-Mayans, M., G-Beiras, C., Prat, N., Ara, J., Clotet, B., and Mitjà, O. (2021). Transmission of COVID-19 in 282 clusters in Catalonia, Spain: a cohort study. *The Lancet Infectious Diseases*, 21(5):629–636.
- Mastrandrea, R., Fournet, J., and Barrat, A. (2015). Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys. *PLOS ONE*, 10(9):1–26.
- Masuda, N. and Holme, P. (2013). Predicting and controlling infectious disease epidemics using temporal networks. *F1000Prime Reports*, 5.
- Masuda, N. and Holme, P. (2020). Small inter-event times govern epidemic spreading on networks. *Phys. Rev. Research*, 2:023163.
- Masuda, N. and Lambiotte, R. (2020). *A Guide To Temporal Networks*, volume 6. World Scientific.
- Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., and Bethge, M. (2018). DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289.
- Meester, L. E. and Muns, S. (2007). Stochastic delay propagation in railway networks and phase-type distributions. *Transportation Research Part B: Methodological*, 41(2):218–230.
- Mellor, A. (2018). The temporal event graph. *Journal of Complex Networks*, 6(4):639–659.
- Mellor, A. (2019). Event graphs: Advances and applications of second-order time-unfolded temporal network models. *Advances in Complex Systems*, 22(03):1950006.

Bibliography

- Michel, C. M. and Koenig, T. (2018). EEG microstates as a tool for studying the temporal dynamics of whole-brain neuronal networks: A review. *NeuroImage*, 180:577–593. Brain Connectivity Dynamics.
- Middelkoop, A. D. and Loeve, L. (2006). Simulation of traffic management with FRISO. In *WIT Transactions on the Built Environment*, volume 88 of *WIT Transactions on The Built Environment, Vol 88*, pages 501–509, Southampton, UK. WIT Press.
- Middelkoop, D., Steneker, J., Meijer, S., Sehic, E., and Mazzarello, M. (2012). Simulation backbone for gaming simulation in railways: A case study. In *Proceedings - Winter Simulation Conference*, pages 1–13. IEEE.
- Ministerie van Infrastructuur en Milieu (2014). Concessie voor het hoofdrailnet 2015-2025.
- Mishra, A., Englitz, B., and Cohen, M. X. (2020a). EEG microstates as a continuous phenomenon. *NeuroImage*, 208.
- Mishra, A., Marzban, N., Cohen, M., and Englitz, B. (2020b). Dynamics of neural microstates in the VTA-striatal-prefrontal loop during novelty exploration in the rat. *bioRxiv*.
- Mistry, D., Litvinova, M., y Piontti, A. P., Chinazzi, M., Fumanelli, L., Gomes, M. F., Haque, S. A., Liu, Q.-H., Mu, K., Xiong, X., et al. (2021). Inferring high-resolution human mixing patterns for disease modeling. *Nature communications*, 12(1):1–12.
- Mitchell, M. (2009). *Complexity: A Guided Tour*. Oxford University Press, Inc., USA.
- Mitzenmacher, M. (2003). A Brief History of Generative Models for Power Law and Lognormal Distributions. *Internet Mathematics*, 1(2):226 – 251.
- Monechi, B., Gravino, P., Di Clemente, R., and Servedio, V. D. P. (2018). Complex delay dynamics on railway networks from universal laws to realistic modelling. *EPJ Data Science*, 7(1):35.
- Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G. S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., and Edmunds, W. J. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLOS Medicine*, 5(3):1–1.
- Mucha, P. J., Richardson, T., Macon, K., Porter, M. A., and Onnela, J.-P. (2010). Community structure in time-dependent, multiscale, and multiplex networks. *Science*, 328(5980):876–878.
- Nagel, T. (1998). Reductionism and antireductionism. *The limits of reductionism in biology*, 213:3–14.
- Nakao, H. and Mikhailov, A. S. (2010). Turing patterns in network-organized activator-inhibitor systems. *Nature Physics*, 6(7):544–550.
- Nash, A. and Huerlimann, D. (2004). Railroad simulation using OpenTrack. *WIT Transactions on The Built Environment*, 74:45–54.
- Nederlandse Spoorwegen, ProRail, Ministerie van Infrastructuur en Milieu (2013). Programma winterweer op het spoor. <https://www.tweedekamer.nl/>. Online; accessed 29 January 2014.
- Netea, M. G. and Li, Y. (2021). Immune memory in individuals with COVID-19. *Nature Cell Biology*, 23(6):582–584.
- Newman, M. (2005). Power laws, Pareto distributions and Zipf’s law. *Contemporary Physics*, 46(5):323–351.
- Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, 103(23):8577–82.
- Ng, A. Y., Jordan, M. I., and Weiss, Y. (2002). On spectral clustering: Analysis and an algorithm. In *Advances in Neural Information Processing Systems*.
- Nielsen, L. K., Kroon, L., and Maróti, G. (2012). A rolling horizon approach for disruption management of railway rolling stock. *European Journal of Operational Research*, 220(2):496–509.
- Nogal, M., O’Connor, A., Caulfield, B., and Martinez-Pastor, B. (2016). Resilience of traffic networks: From perturbation to recovery via a dynamic restricted equilibrium model. *Reliability Engineering and System Safety*, 156:84–96.

- Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., and Anguita, D. (2017). Dynamic delay predictions for large-scale railway networks: Deep and shallow extreme learning machines tuned via thresholdout. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 47(10):2754–2767.
- Oneto, L., Fumeo, E., Clerico, G., Canepa, R., Papa, F., Dambra, C., Mazzino, N., and Anguita, D. (2018). Train Delay Prediction Systems: A Big Data Analytics Perspective. *Big Data Research*, 11:54–64.
- Ouyang, M., Dueñas-Osorio, L., and Min, X. (2012). A three-stage resilience analysis framework for urban infrastructure systems. *Structural Safety*, 36-37:23–31.
- Pagani, A., Mosquera, G., Alturki, A., Johnson, S., Jarvis, S., Wilson, A., Guo, W., and Varga, L. (2019). Resilience or robustness: identifying topological vulnerabilities in rail networks. *Royal Society Open Science*, 6(2):181301.
- Pang, R., Lansdell, B. J., and Fairhall, A. L. (2016). Dimensionality reduction in neuroscience. *Curr. Biol.*, 26(14):R656–R660.
- Pant, R., Hall, J. W., and Blainey, S. P. (2016). Vulnerability assessment framework for interdependent critical infrastructures: case-study for Great Britain's rail network. *European Journal of Transportation and Infrastructure Research*, 16(1):174–194.
- Panzarasa, P., Opsahl, T., and Carley, K. M. (2009). Patterns and dynamics of users' behavior and interaction: Network analysis of an online community. *Journal of the American Society for Information Science and Technology*, 60(5):911–932.
- Parra, L. C., Haufe, S., and Dmochowski, J. P. (2019). Correlated components analysis - extracting reliable dimensions in multivariate data. *arXiv:1801.08881*.
- Pastor-Satorras, R. and Vespignani, A. (2002). Immunization of complex networks. *Phys. Rev. E*, 65:036104.
- Pearson, P., Foster, G., and Wade, B. (2009). Atmospheric carbon dioxide through the Eocene–Oligocene climate transition. *Nature*, 461(7267):1110–3.
- Peel, L., Larremore, D. B., and Clauset, A. (2017). The ground truth about metadata and community detection in networks. *Science Advances*, 3(5).
- Peirce, C. (1884). The numerical measure of the success of predictions. *Science*, ns-4(93):453–454.
- Peixoto, T. P. and Rosvall, M. (2017). Modelling sequences and temporal networks with dynamic community structures. *Nature Communications*, 8(1):582.
- Peng, C.-K., Buldyrev, S. V., Havlin, S., Simons, M., Stanley, H. E., and Goldberger, A. L. (1994). Mosaic organization of DNA nucleotides. *Phys. Rev. E*, 49:1685–1689.
- Perra, N., Gonçalves, B., Pastor-Satorras, R., and Vespignani, A. (2012). Activity driven modeling of time varying networks. *Scientific Reports*, 2(1):469.
- Pescaroli, G. and Alexander, D. (2016). Critical infrastructure, panarchies and the vulnerability paths of cascading disasters. *Natural Hazards*, 82(1):175–192.
- Podnobik, B. and Stanley, H. (2008). Detrended cross-correlation analysis: A new method for analyzing two non-stationary time series. *Physical Review Letters*, 100(8):084102.
- Politico (2019). Rail delays prompt German identity crisis. by: *Armando Babani*.
- Potthoff, D., Huisman, D., and Desaulniers, G. (2010). Column generation with dynamic duty selection for railway crew rescheduling. *Transportation Science*, 44(4):493–505.
- Prem, K., Cook, A. R., and Jit, M. (2017). Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLoS Computational Biology*, 13(9):e1005697.
- Pyrgiotis, N., Malone, K. M., and Odoni, A. (2013). Modelling delay propagation within an airport network. *Transportation Research Part C: Emerging Technologies*, 27:60–75.
- Rakshit, S., Bera, B. K., and Ghosh, D. (2018). Synchronization in a temporal multiplex neuronal hypernetwork. *Phys. Rev. E*, 98:032305.
- Ramaekers, P., Wit, T. d., and Pouwels, M. (2016). Hoe druk is het nu werkelijk op het Nederlandse spoor? *Centraal Bureau voor de Statistiek*.

Bibliography

- Rickles, D., Hawe, P., and Shiell, A. (2007). A simple guide to chaos and complexity. *Journal of epidemiology and community health*, 61(11):933–937.
- Righi, A. W., Wachs, P., and Saurin, T. A. (2012). Characterizing complexity in socio-technical systems: a case study of a SAMU Medical Regulation Center. *Work*, 41:1811–1817.
- Rijksinstituut voor Volksgezondheid en Milieu (RIVM) (2021). The expected outcome of COVID-19 vaccination strategies. <https://www.rivm.nl/documenten/expected-outcome-of-covid-19-vaccination-strategies>. Online; accessed 2 November 2021.
- Riolo, C. S., Koopman, J. S., and Chick, S. E. (2001). Methods and measures for the description of epidemiologic contact networks. *Journal of Urban Health*, 78(3):446–457.
- Rockström, J., Steffen, W., Noone, K., Persson, Å., Chapin, F. S., Lambin, E. F., Lenton, T. M., Scheffer, M., Folke, C., Schellnhuber, H. J., Nykvist, B., de Wit, C. A., Hughes, T., van der Leeuw, S., Rodhe, H., Sörlin, S., Snyder, P. K., Costanza, R., Svedin, U., Falkenmark, M., Karlberg, L., Corell, R. W., Fabry, V. J., Hansen, J., Walker, B., Liverman, D., Richardson, K., Crutzen, P., and Foley, J. A. (2009). A safe operating space for humanity. *Nature*, 461(7263):472–475.
- Ross, R. and Hudson, H. P. (1917). An application of the theory of probabilities to the study of a priori pathometry.—Part III. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 93(650):225–240.
- Rosvall, M. and Bergstrom, C. T. (2010). Mapping change in large networks. *PLOS ONE*, 5(1):1–7.
- Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. E., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P., and Vidal, M. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178.
- Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 5(11).
- Scarpino, S. V. and Petri, G. (2019). On the predictability of infectious disease outbreaks. *Nature Communications*, 10(1):898.
- Scheffer, M., Bascompte, J., Brock, W., Brovkin, V., Carpenter, S., Dakos, V., Held, H., van Nes, E., Rietkerk, M., and Sugihara, G. (2009). Early-warning signals for critical transitions. *Nature*, 461:53–59.
- Schipper, D. and Gerrits, L. (2018). Differences and similarities in European railway disruption management practices. *Journal of Rail Transport Planning and Management*, 8(1):42–55.
- Schläpfer, M., Bettencourt, L. M. A., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., West, G. B., and Ratti, C. (2014). The scaling of human interactions with city size. *Journal of The Royal Society Interface*, 11(98):20130789.
- Schöbel, A. (2012). Line planning in public transportation: models and methods. *OR spectrum*, 34(3):491–510.
- Scholtes, I., Wider, N., Pfitzner, R., Garas, A., Tessone, C. J., and Schweitzer, F. (2014). Causality-driven slow-down and speed-up of diffusion in non-markovian temporal networks. *Nature Communications*, 5(1):5024.
- Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, P. A., Mukherjee, G., and Manna, S. S. (2003). Small-world properties of the Indian railway network. *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 67(3):5.
- Sharma, M., Mindermann, S., Rogers-Smith, C., Leech, G., Snodin, B., Ahuja, J., Sandbrink, J. B., Monrad, J. T., Altman, G., Dhaliwal, G., et al. (2021). Understanding the effectiveness of government interventions against the resurgence of COVID-19 in Europe. *Nature communications*, 12(1):1–13.

- Shi, J. and Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.
- Shizuka, D., Chainé, A. S., Anderson, J., Johnson, O., Laursen, I. M., and Lyon, B. E. (2014). Across-year social stability shapes network structure in wintering migrant sparrows. *Ecology Letters*, 17(8):998–1007.
- Shughrue, C., Werner, B., and Seto, K. C. (2020). Global spread of local cyclone damages through urban trade networks. *Nature Sustainability*, pages 1–8.
- Simonsen, I., Buzna, L., Peters, K., Bornholdt, S., and Helbing, D. (2008). Transient dynamics increasing network vulnerability to cascading failures. *Physical Review Letters*, 100(21):218701.
- Siudem, G. and Holyst, J. A. (2019). Diffusion on hierarchical systems of weakly-coupled networks. *Physica A: Statistical Mechanics and its Applications*, 513:675–686.
- Smuts, J. C. (1926). *Holism and evolution*. London: Macmillan.
- Sobkowicz, P., Kaschesky, M., and Bouchard, G. (2012). Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web. *Government Information Quarterly*, 29(4):470–479.
- Social Cultureel Planbureau (2016). Tijdsbestedingsonderzoek 2016. <https://digitaal.scp.nl/eenweekinkaart2/een-week-in-vogelvlucht/>. Online; accessed 6 November 2021.
- Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Colizza, V., Isella, L., Régis, C., Pinton, J.-F., Khanafer, N., Van den Broeck, W., and Vanhems, P. (2011). Simulation of an SEIR infectious disease model on the dynamic contact network of conference attendees. *BMC Medicine*, 9(1):87.
- Stehlé, J., Voirin, N., Barrat, A., Cattuto, C., Isella, L., Pinton, J., Quaghiotto, M., Van den Broeck, W., Régis, C., Lina, B., and Vanhems, P. (2011). High-resolution measurements of face-to-face contact patterns in a primary school. *PLOS ONE*, 6(8):e23176.
- Steinhaus (1956). Sur la division des corps matériels en parties. *Bulletin de l'Académie Polonaise des Sciences, Classe III*, IV(12):801–804.
- Sterl, A., Severijns, C., Dijkstra, H. A., Hazeleger, W., van Oldenborgh, G. J., van den Broeke, M., Burgers, G., van den Hurk, B., van Leeuwen, P. J., and van Velthoven, P. (2008). When can we expect extremely high surface temperatures? *Geophysical Research Letters*, 35:L14703.
- Stevenson, I. H. and Kording, K. P. (2011). How advances in neural recording affect data analysis. *Nature Neuroscience*, 14(2):139–142.
- Stommel, H. (1961). Thermohaline Convection with Two Stable Regimes of Flow. *Tellus*, 13(2):224–230.
- Strogatz, S. H. (1994). *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*. Reading, MA: Perseus Books.
- Tanis, C. C., Leach, N. M., Geiger, S. J., Nauta, F. H., Dablander, F., van Harreveld, F., de Wit, S., Kanters, G., Knoppers, J., Markus, D. A. W., Bouten, R. R. M., Oostvogel, Q. H., Boersma, M. J., van der Steenhoven, M. V., Borsboom, D., and Blanken, T. F. (2021). Smart Distance Lab's art fair, experimental data on social distancing during the COVID-19 pandemic. *Scientific Data*, 8(1):179.
- Tantet, A., Lucarini, V., and Dijkstra, H. A. (2018). Resonances in a chaotic attractor crisis of the Lorenz flow. *Journal of Statistical Physics*, 170(3):584–616.
- Tantet, A., van der Burgt, F. R., and Dijkstra, H. A. (2015). An early warning indicator for atmospheric blocking events using transfer operators. *Chaos*, 25(3):1–48.
- Thompson, J. and Stewart, H. (2002). *Nonlinear Dynamics and Chaos*. Wiley.
- Thompson, J. M. and Sieber, J. (2011). Climate tipping as a noisy bifurcation: A predictive technique. *IMA Journal of Applied Mathematics (Institute of Mathematics and Its Applications)*, 76(1):27–46.
- Tigchelaar, M., von der Heydt, A. S., and Dijkstra, H. A. (2011). A new mechanism for the two-step $\delta^{18}\text{O}$ signal at the eocene-oligocene boundary. *Climate of the Past*, 7(1):235–247.

Bibliography

- Timmermann, A., An, S.-I., Krebs, U., and Goosse, H. (2005). ENSO suppression due to weakening of the North Atlantic thermohaline circulation. *Journal of Climate*, 18(16):3122–3139.
- Timmermann, A., Jin, E., and Abshagen, J. (2003). A nonlinear theory for El Niño bursting. *Journal of the Atmospheric Sciences*, 60(1):152–165.
- Timmermann, A., Okumura, Y., An, S.-I., Clement, A., Dong, B., Guilyardi, E., Hu, A., Jungclaus, J., Renold, M., Stocker, T., R.J., S., Sutton, R., Xie, S.-P., and Yin, J. (2007). The Influence of a Weakening of the Atlantic Meridional Overturning Circulation on ENSO. *Journal of Climate*, 20(19):4899–4919.
- Torricelli, M., Karsai, M., and Gauvin, L. (2020). weg2vec: Event embedding for temporal networks. *Scientific Reports*, 10(1):7164.
- Trautmann, E. M., Stavisky, S. D., Lahiri, S., Ames, K. C., Kaufman, M. T., O’Shea, D. J., Vyas, S., Sun, X., Ryu, S. I., Ganguli, S., and Shenoy, K. V. (2019). Accurate estimation of neural population dynamics without spike sorting. *Neuron*, 103(2):292–308.e4.
- Tsuchiya, S., Tatano, H., and Okada, N. (2007). Economic loss assessment due to railroad and highway disruptions. *Economic Systems Research*, 19(2):147–162.
- Ulam, S. M. (1964). *Problems in modern mathematics*. Science Editions John Wiley & Sons, Inc., New York.
- van der Drift, S., Wismans, L., and Kalter, M.-J. O. (2021). Changing mobility patterns in the Netherlands during COVID-19 outbreak. *Journal of Location Based Services*, 0(0):1–24.
- van Kampen, N. (2007). Stochastic processes in physics and chemistry (third edition). In *Stochastic Processes in Physics and Chemistry (Third Edition)*, pages 193–218. Elsevier, Amsterdam, third edition edition.
- Van Lieshout, R. N., Bouman, P. C., and Huisman, D. (2020). Determining and evaluating alternative line plans in out-of-control situations. *Transportation Science*, 54(3):740–761.
- Van Lieshout, R. N., Bouman, P. C., van den Akker, M., and Huisman, D. (2021). A self-organizing policy for vehicle dispatching in public transit systems with multiple lines. *Transportation Research Part B: Methodological*, 152:46–64.
- van Sebille, E., England, M. H., and Froyland, G. (2012). Origin, dynamics and evolution of ocean garbage patches from observed surface drifters. *Environmental Research Letters*, 7(4):044040.
- Vanhems, P., Barrat, A., Cattuto, C., Pinton, J., Khanafer, N., Régis, C., Kim, B.-a., Comte, B., and Voirin, N. (2013). Estimating potential infection transmission routes in hospital wards using wearable proximity sensors. *PLoS ONE*, 8(9):e73970.
- Varela, F., Lachaux, J.-P., Rodriguez, E., and Martinerie, J. (2001). The brainweb: Phase synchronization and large-scale integration. *Nature Reviews Neuroscience*, 2(4):229–239.
- Venkatesh, M., Jaja, J., and Pessoa, L. (2019). Brain dynamics and temporal trajectories during task and naturalistic processing. *NeuroImage*, 186:410–423.
- Vespignani, A. (2010). Complex networks: The fragility of interdependency. *Nature*, 464(7291):984–985.
- Vespignani, A. (2012). Modelling dynamical processes in complex socio-technical systems. *Nature Physics*, 8(1):32–39.
- Vos, E. R. A., den Hartog, G., Schepp, R. M., Kaaijk, P., van Vliet, J., Helm, K., Smits, G., Wijmenga-Monsuur, A., Verberk, J. D. M., van Boven, M., van Binnendijk, R. S., de Melker, H. E., Mollema, L., and van der Klis, F. R. M. (2021). Nationwide seroprevalence of SARS-CoV-2 and identification of risk factors in the general population of the Netherlands during the first epidemic wave. *Journal of Epidemiology & Community Health*, 75(6):489–495.
- Walentek, D., Broere, J., Cinelli, M., Dekker, M. M., and Haslbeck, J. M. (2021). Success of economic sanctions threats: coercion, information and commitment. *International Interactions*, pages 1–32.
- Wang, S. and Zhao, J. (2015). Multi-attribute integrated measurement of node importance in complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 25(11):113105.

- Wang, W., Ma, Y., Wu, T., Dai, Y., Chen, X., and Braunstein, L. A. (2019). Containing misinformation spreading in temporal social networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 29(12):123131.
- Watts, D. J. (2002). A simple model of global cascades on random networks. *Proc. Nat. Acad. Sci.*, 99(9):5766–5771.
- Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of ‘small-world’ networks. *Nature*, 393(6684):440–442.
- Wbcsd (2009). Mobility for development - Executive summary. *Challenges*.
- Wei, D., Liu, H., and Qin, Y. (2015). Modeling cascade dynamics of railway networks under inclement weather. *Transportation Research Part E: Logistics and Transportation Review*, 80:95–122.
- Wen, C., Li, Z., Huang, P., Lessan, J., Fu, L., and Jiang, C. (2019). Cause-specific investigation of primary delays of Wuhan–Guangzhou HSR. *Transportation Letters*.
- Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., and Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, 338(6104):267–270.
- Wheeler, W. M. (1910). *Ants: their structure, development and behavior*. Number 9 in Columbia University Biological Series. Columbia University Press.
- World Economic Forum (2011). New models for addressing supply chain and transport risk. *Technical report, The World Economic Forum*.
- Xia, J., Huang, W., and Guo, J. (2012). A clustering approach to online freeway traffic state identification using ITS data. *KSCE Journal of Civil Engineering*, 16(3):426–432.
- Xie, S.-P., Miyama, T., Wang, Y., Xu, H., de Szoeko, S. P., Small, R. J. O., Richards, K. J., Mochizuki, T., and Awaji, T. (2007). A regional ocean–atmosphere model for Eastern Pacific climate: Toward reducing tropical biases. *Journal of Climate*, 20(8):1504–1522.
- Yang, S., Wu, J., Qi, G., and Tian, K. (2017). Analysis of traffic state variation patterns for urban road network based on spectral clustering. *Advances in Mechanical Engineering*, 9(9):2017.
- Yosef, N. and Regev, A. (2011). Impulse control: Temporal dynamics in gene transcription. *Cell*, 144(6):886–896.
- Yuan, N., Zuntao, F., Zhang, H., Piao, L., Xoplaki, E., and Luterbacher, J. (2015). Detrended partial-cross-correlation analysis: A new method for analyzing correlations in complex system. *Scientific Reports*, 5(8143).
- Yule, G. U. (1925). II. – a mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F. R. S. *Philosophical Transactions of the Royal Society of London. Series B, Containing Papers of a Biological Character*, 213(402-410):21–87.
- Zebende, G. (2011). DCCA cross-correlation coefficient: Quantifying level of cross-correlation. *Physica A: Statistical Mechanics and its Applications*, 390(4):614 – 618.
- Zebiak, S. E. and Cane, M. A. (1987). A Model El Niño–Southern Oscillation. *Monthly Weather Review*, 115(10):2262–2278.
- Zhang, D., Lee, T. N., Johns, W. E., Liu, C.-T., and Zantopp, R. (2001). The Kuroshio east of Taiwan: Modes of variability and relationship to interior ocean mesoscale eddies. *Journal of Physical Oceanography*, 31(4):1054–1074.
- Zhang, D., Wang, Y., and Zhang, Z. (2019). Identifying and quantifying potential super-spreaders in social networks. *Scientific Reports*, 9(1):14811.
- Zhang, J., Litvinova, M., Liang, Y., Zheng, W., Shi, H., Vespignani, A., Viboud, C., Ajelli, M., and Yu, H. (2021). The impact of relaxing interventions on human contact patterns and SARS-CoV-2 transmission in China. *Science Advances*, 7(19).
- Zhang, R. and Delworth, T. L. (2005). Simulated Tropical Response to a Substantial Weakening of the Atlantic Thermohaline Circulation. *Journal of Climate*, 18(12):1853–1860.
- Zhang, X., Yi, H., Bai, W., and Tian, X. (2015). Dynamic trajectory of multiple single-unit activity during working memory task in rats. *Frontiers in Computational Neuroscience*, 9(SEP).

Bibliography

- Zhou, W. (2008). Multifractal detrended cross-correlation analysis for two nonstationary signals. *Physical Review*, 77(6).
- Zhu, Y. and Goverde, R. M. (2019). Railway timetable rescheduling with flexible stopping and flexible short-turning during disruptions. *Transportation Research Part B: Methodological*, 123:149–181.
- Zieger, S., Weik, N., and Nießen, N. (2018). The influence of buffer time distributions in delay propagation modelling of railway networks. *Journal of Rail Transport Planning and Management*, 8(3-4):220–232.

About the author

Mark Dekker was born on June 7, 1992 in Dordrecht, the Netherlands. He obtained a double Bachelor's degree in Physics and Mathematics from the Utrecht University in 2015, specialising in numerical methods and climate physics. During this time, he spent a semester on geophysics courses at the University Centre in Svalbard, Norway, and did his Bachelor thesis on the structure of extratropical cyclones at the Royal Dutch Meteorological Institute (KNMI) in the Netherlands. In 2017, he obtained a Master's degree in Climate Physics (*cum laude*) at the Institute for Marine and Atmospheric research Utrecht (IMAU) and did his Master thesis about critical transitions in the climate system.

Mark started a PhD candidacy in 2017 at the department of Information and Computing Sciences at Utrecht University, as part of a larger consortium funded by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO). The project was called 'Improving the resilience of railway systems' and was led by dr. Debabrata Panja. Besides Utrecht University, it involved members from TU Delft, Erasmus University Rotterdam, Nederlandse Spoorwegen (NS) and ProRail.

During the PhD candidacy, he was an active member of the Centre for Complex Systems Studies at Utrecht University, where he participated in and organised a number of events as part of the junior advisory board YCRU. He worked at the Statistics Netherlands (CBS) when researching the COVID-19 pandemic as part of the ZonMw consortium project called 'Evidence-based effective monitoring and control of Covid-19 after the initial outbreak', for which he was awarded the NPSO Innovation Prize 2022. He presented his work at a number of conferences such as the Conference on Complex Systems (CCS), Netherlands Platform Complex Systems (NPCS) and the CRITICS workshop.

Mark is currently working as a climate and energy researcher at the Netherlands Environmental Assessment Agency (PBL in Dutch), researching the complexity of our interaction with the Earth's climate.

List of Publications and Working papers

Included in this thesis

All chapters are based on research that is either published, under review or in a near-submission stage:

- Ch. 1 Dekker, M. M., Panja, D., Dijkstra, H. A., and Dekker, S. C. (2019). Predicting transitions across macroscopic states for railway systems. *PLOS ONE*, 14(6):e0217710
- Ch. 2 Dekker, M. M., França, A. S. C., Panja, D., and Cohen, M. X. (2021c). Characterizing neural phase-space trajectories via principal louvain clustering. *Journal of Neuroscience Methods*, 362:109313
- Ch. 3 Dekker, M. M., Von Der Heydt, A. S., and Dijkstra, H. A. (2018b). Cascading transitions in the climate system. *Earth System Dynamics*, 9(4):1243–1260
- Ch. 4 Dekker, M. M. (2021). Geographic delay characterization of railway systems. *Scientific Reports*, 11(1):1–13
- Ch. 5 Dekker, M. M., Medvedev, A. N., Rombouts, J., Siudem, G., and Tupikina, L. (2021d). Modelling railway delay propagation as diffusion-like spreading. *arXiv:2105.06111*
- Ch. 6 Dekker, M. M. and Panja, D. (2021). Cascading dominates large-scale disruptions in transport over complex networks. *PLOS ONE*, 16(1):e0246077
- Ch. 7 Dekker, M. M., Schram, R., Ou, J., and Panja, D. (2021e). The hidden dependence of spreading vulnerability on topological complexity. *arXiv:2107.01651*
- Ch. 8 Dekker, M. M., Blanken, T. F., Dablander, F., Ou, J., Borsboom, D., and Panja, D. (2021a). Quantifying agent impacts on contact sequences in social interactions. *arXiv:2107.01443*, accepted in *Scientific Reports*
- Ch. 9 Dekker, M. M., Coffeng, L., Pijpers, F., Panja, D., and Vlas, S. (2021b). Evaluation of non-pharmaceutical interventions during the COVID-19 first wave in the Netherlands. *Manuscript in preparation*

Other academic work

- Dekker, M. M., van Lieshout, R. N., Ball, R. C., Bouman, P. C., Dekker, S. C., Dijkstra, H. A., Goverde, R. M. P., Huisman, D., Panja, D., Schaafsma, A. A. M., and van den Akker, M. (2021f). A next step in disruption management: combining operations research and complexity science. *Public Transport*, pages 1–22
- Dekker, M. M. and Panja, D. (2019). A reduced phase-space approach to analyse railway dynamics. *IFAC-PapersOnLine*, 52(3):1–6
- Dekker, M. M., Haarsma, R. J., de Vries, H., Baatsen, M., and van Delden, A. J. (2018a). Characteristics and development of European cyclones with tropical origin in reanalysis data. *Climate dynamics*, 50(1):445–455
- Dekker, M. M., Walentek, D., Haslbeck, J., and Broere, J. (2020). A general framework for international conflict under incomplete information. *APSA preprints*
- Walentek, D., Broere, J., Cinelli, M., Dekker, M. M., and Haslbeck, J. M. (2021). Success of economic sanctions threats: coercion, information and commitment. *International Interactions*, pages 1–32
- Dam, A. van, Dekker, M. M., Morales-Castilla, I., Rodríguez, M. Á., Wichmann, D., and Baudena, M. (2021). Correspondence analysis, spectral clustering and graph embedding: applications to ecology and economic complexity. *Scientific reports*, 11(1):1–14

Non-academic publications

- Panja, D., and Dekker, M. M. (2021). Het Modelleren van Infectieziekten. *Nederlands Tijdschrift van Natuurkunde*. Januarinumnummer 2021.
- Dekker, M. M. (2021). Lessons from Complex Systems Science for Railway Practise. *Subreport A of 'Improving the Resilience of Railway systems'*.

Acknowledgements

I suppose this is it. After more than four years, I have wrapped up all lessons and research in the pages before. A final piece is still missing and is written in *this* section, which arguably is one of the most read sections in a dissertation, for not unimportant reasons. The journey towards a PhD's completion is always many-faceted. Of course, part of those facets concern a student's skill, proper guidance, the choice of a solid topic, some perseverance and a bit of luck (a pessimist might say that these are named in increasing order of importance). However, in the end, the success of a PhD candidacy is much broader than that and is crucially dependent on the people surrounding the student. A dissertation is a student's proof of having the ability to be a good researcher — realising that being a good researcher also involves reflecting on how most research is performed (*with people*) illustrates the importance of this section.

While working on this dissertation, I had the pleasure of meeting many people that enriched the journey and contributed to it. Starting with the person that hands-down contributed more to my research than any other person beside myself, I thank my supervisor Debabrata Panja. Deb, I am grateful for your many ideas, time and our frequent brainstorming and discussion sessions. You have a critical and out-of-the-box perspective on pretty much everything. You were always looking for new possibilities to apply our research to, which lead to a variety of interesting opportunities ranging from presentations for diverse audiences to research projects in various areas I was not native to — such as neuroscience (chapter 2), behavioural science (chapter 8) and epidemiology (chapter 9). Thank you for these opportunities. Our discussions have pushed me and my research across several edges, which shaped the research itself and me as a researcher. I am glad for the time we have worked together.

In the department of Information and Computing Sciences, I was part of a small group of physicists working on 'Simulations of Complex Systems', which particularly involved my two colleague PhD-students and friends: Wei and Federico. While PhD candidacy may often feel like a one-person job, these two have been crucial to my time as a PhD student. We have had numerous lunches, coffee breaks and barbecues together. When corona hit the Netherlands, these were mostly replaced by video calls. Thank you, Wei, for our mutual interest in converting our office into an urban jungle, and for carefully cultivating *Euphorbia Leuconeura*, which I still grow at home as part of the legacy of our office. Thank you, Federico, for the many times we worked together at the CCSS,

talking (way too long for a work day?) about politics, the evil nature of certain tech companies and sharing our opinions about corona policy, but also for the honest talks we had about challenges as a PhD student. I am glad to have met you both.

Outside of our department, I had the pleasure of being active member of the Centre for Complex Systems Studies at Utrecht University, involving numerous interesting lectures and events. Thank you Qingyi, for creating and organising this environment. Also during the coronacrisis, I was able to work many hours in the corona-safe open working spaces you created for us, which was the perfect break for the many days working from home. Through the Centre, I have met many interesting and nice people, such as those in the YCRU, some of which I started research projects with together.

This work is part of a larger consortium funded by Nederlandse Organisatie voor Wetenschappelijk Onderzoek (NWO) in a project called 'Improving the resilience of railway systems', which involved people from various universities and organisations. Two people in particular I would like to thank here: Henk and Stefan, who had a crucial role in the first few years of my PhD candidacy and ignited core research ideas that shaped many parts in this dissertation. Further, I thank the other consortium members (in particular Rolf, Marjan, Paul, Rob, Egidio, Dennis, Pieter-Jan, Alfons and Edith) for the many discussions we had, for the patience with me being new to the field of transportation and operation research and for helping to together build the bridge between disciplines that made this project happen. Also within the associated organisations Dutch Railways (NS) and ProRail, apart from the aforementioned consortium members, I have had help from various people in explaining and processing large amounts of data, which I am grateful for. Chapters 8 and 9 were also part of another consortium called 'Evidence-based effective monitoring and control of Covid-19 after the initial outbreak', funded by Nederlandse organisatie voor gezondheidsonderzoek en zorginnovatie (ZonMw). For chapter 9, I was temporarily intern at Statistics Netherlands (CBS), for which I specifically thank Frank.

Beyond the NWO consortium, I have met numerous researchers from a variety of backgrounds, who have contributed to work in this dissertation: Mike, Arthur, Anna, Alexey, Jan, Gregory, Liubov, Raoul, Jiamin, Denny, Fabian, Tessa, Luc, Sake and Frank. I also express my gratitude to Gerard, my promotor, for your more formal role during the PhD candidacy. In other contexts, I am happy to have met Peter, Mara, Alje, Nacho, Miguel, David, Joris, Jonas, Matteo and Dawid. It has been a pleasure to know you and to work with you. Specifically, I thank René for the many coffees we had, for your friendship and the incredibly bad timing of starting a podcast together when I am finishing my PhD.

Needless to say, there are many people outside of academia that contributed informally to this dissertation — it is a long list and there is not enough room to fully thank everyone. Thank you Tycho and Daan, my student roommates and friends, for the many laughs we had and for the home you created for me in the first period of my PhD. Thank you Martijn, Tim, Bram and Isebrand. for the great memories we have as a group, for the conversations we had on subjects that cannot be simulated in mathematical models, and for challenging me in many ways. Thank you Steven and August, for the many lengthy discussions we had on pretty much everything, for really knowing me and for the friendship we have.

I do not have to cite any papers here to say that family is an important ingredient for healthy progress at work, especially when it comes to large projects like this dissertation. Thank you papa, mama, Arno, Eline en Marlieke, for creating a home where disserta-

Acknowledgements

tions and career do not matter for who we are. The many worries that had come from a PhD's meandering trajectory always seemed to dissipate a little bit when I was at home in Veenendaal. Instead of talking about papers and coding problems, I would come there, punch my brother, and enjoy the much better coffee than we have in Utrecht. Thank you for showing interest in the work and helping me with important choices. I also thank my family-in-law in similar ways: in particular, in the last phase of finishing this work, the writing retreat in Otterlo has been instrumental.

There is one person left to mention, whose contribution to this work and me as a person cannot be expressed in words: my wife, the love of my life, whom I in fact married during this PhD candidacy. In a way, she is a co-author of this dissertation. Jolein, thank you for supporting me when I was worrying too much, and for celebrating with me when things were going well. Thank you for always supporting me and being proud of what I do, even when the research is a little technical. Thank you for making me smile when I got cranky when my programming did not work. You make me realise what is really important in life, which ultimately brings me to a few old but unwavering words to conclude with:

Soli Deo Gloria.

