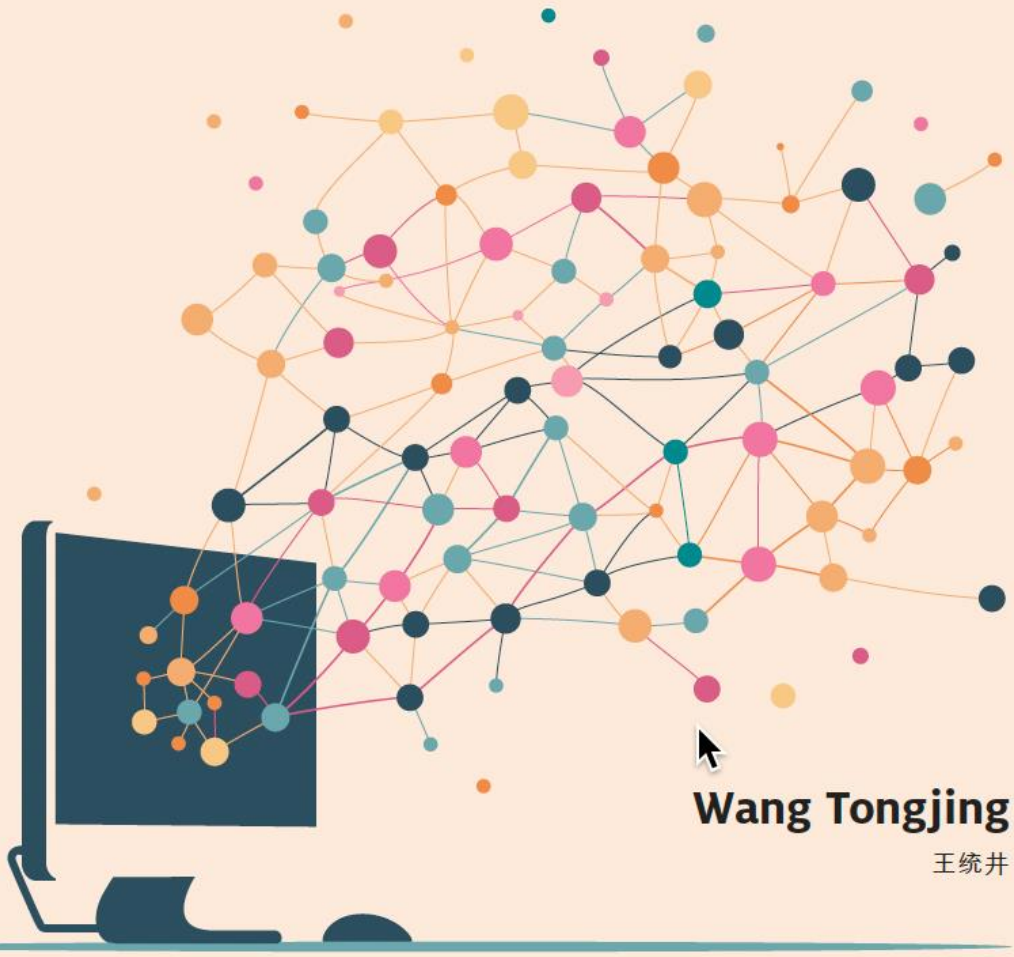
Connected Cities

Deciphering city connections via collocation analysis



Wang Tongjing

王统井

Connected Cities: Deciphering City Relationships Using Collocation Analysis

Wang Tongjing

王统井

Doctoral dissertation

DOI: https://doi.org/10.33540/2500
Cover design: Wang Tongjing, enabled by Midjourney
Design and layout: Wang Tongjing
Print: Ipskamp Printing

Connected Cities: Deciphering City Relationships Using Collocation Analysis

Verbonden Steden:
Stadsrelaties Ontcijferen met Collocatieanalyse

(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. H.R.B.M. Kummeling,
ingevolge het besluit van het college voor promoties
in het openbaar te verdedigen op

dinsdag 24 september 2024 des middags te 4.15 uur

door

Tongjing Wang

geboren op 7 juli 1992
te Hangzhou, China

Promotor:
Dr. E. Meijers

Copromotor:
Dr. H. Wang

Beoordelingscommissie:
Prof. dr. R. Boschma
Prof. dr. E. Buitelaar
Prof. dr. C. Castaldi
Prof. dr. B. Derudder
Prof. dr. B. Sun

至人无己
神人无功
圣人无名

庄子

逍遥游

The perfect man ignores self;
The divine man ignores achievement;
The true Sage ignores reputation.

Zhuang Zi
A Happy Excursion

Contents

Acknowledgments

Little did I expect that a high school English class would foreshadow my eventual research topic. During this class, I was tasked with sharing an article and I selected one called “Illusion of Pastoral Peace” from New Concept English III, an English textbook commonly used in China. This article discussed the advantages and disadvantages of living in cities alongside the idyllic yet often “delusional” life in the countryside. Following my presentation, my English teacher asked me if I preferred to live in the city or the countryside. I said I prefer to live in the city during weekdays and in the countryside on the weekends so I can enjoy the benefits on both sides.

I studied civil engineering during my undergraduate and master’s. After spending four years in undergraduate school and three years in graduate school, I got my master’s degree in transportation engineering at Tongji University, Shanghai, China, specializing on pavement materials. However, I found myself increasingly interested in exploring the broader societal context.

As my background was deeply rooted in engineering, I was aware of the challenges associated with shifting to an entirely different field, so I decided to do a fresh master’s degree before pursuing a Ph.D. I had two options at that time. One was the Metropolitan Analysis Design and Engineering program, offered jointly by Delft University of Technology (TU Delft) and Wageningen University and Research (WUR), conducted at the Amsterdam Institute for Advanced Metropolitan Solutions (the AMS Institute). The other option was a program in Transportation Planning at the University of California, Davis. While the path not taken with UC Davis remains an intriguing “what if”, I am certain that my chosen path has been the right one.

This is because I met my supervisor Evert Meijers in the AMS institute, who was a course teacher there. His lectures and research focus, combined with his amiable personality, fascinated me. So I approached him with the idea of pursuing a Ph.D. with his supervision. Remarkably, despite my lack of background knowledge in this new domain, he agreed. He also introduced me to the use of the toponym co-occurrence method for my

research, sparing me the often daunting task of identifying a viable Ph.D. topic on my own.

I am immensely grateful for Evert's support and supervision. Initially, I doubted the feasibility of pursuing Ph.D. research in human geography, given my background in civil engineering—a field that follows a completely different research paradigm. Yet, Evert's guidance and inspiration made it possible. His optimism and encouragement have been a constant source of strength, instilling a profound confidence in my research pursuits.

I had another unexpected encounter that shaped my academic journey. During my master's studies at the AMS Institute, I took a course on network analysis at TU Delft, led by Huijuan Wang. She introduced me to the network analysis method, which I found instrumental in my research. I asked her if she would be my co-supervisor. She accepted and brought me into the Multimedia group in the computer science field at TU Delft. Huijuan's mentorship introduced me to the value of rigorous discipline in research, a lesson that proved invaluable across all facets of my work.

After studying for one and a half years at TU Delft Architecture School, I followed Evert to the Economic Geography Group at Utrecht University, where I was warmly welcomed. I'm very grateful for Ron Boschma and Carolina Castaldi, who fostered a supportive and encouraging environment there. I also owe a special gratitude to Pierre-Alexandre Balland, who offered me an incredible opportunity to spend a semester abroad at Northeastern University in Boston.

I must also express my appreciation for my colleagues—Martijn, Deyu, Nicola, Kerstin, Jaap, Milene, Sergio, Yuanyuan, Yibo, Dongmiao, Benjamin, Diego, Eduardo, Christian, Adrian, and Zenne. Their companionship has greatly enhanced my Ph.D. journey, making it not only a scholarly pursuit but also many joyful memories. Additionally, I am thankful for the extended family within our larger department, including Hu Yang, Liu Dan, Mengyuan, Hongbo, Wanlin, Yunyao, Li Zhen, Hongmei, Jiakun, Linlin, Zeng Peng, and Zeng Yi.

I reserve a special thanks for Yuanyuan. Conversations with her consistently serve as a source of joy and inspiration.

My appreciation also goes to computer science researchers from the Multimedia group at TU Delft—Zou Li, Shilun, Oma, Alberto, Tianrui, Maosheng, Tianqi, and Xiuxiu. Conversations with them have consistently brought fresh insights and ideas, significantly enriching my research journey.

Additionally, I extend my heartfelt thanks to my friends at the Architecture School of TU Delft—Maarten, Rodrigo, Ali, Song Yan, Liu Mei, Kaiyi, Yizhao, Yinhua, Enshan, Jia Lin, Zhang Gong, Shuyu, and Ana, who has been nothing short of inspiring.

My exchange study in Boston was another highlight of my academic journey, and I owe a debt of gratitude to Amy, Siqi, Guo Zhen, Kaicheng, and Nicholas, who made my stay truly rewarding.

A special acknowledgment must be reserved for Yin Zhao and Bao Ziyu, whom I fortuitously met during a class at TU Delft. Their willingness to assist with the daunting task of processing big data was a turning point in my Ph.D. research. Their expertise and support were crucial in navigating the complexities of big data analysis.

I am fortunate to have met so many wonderful people who have made my Ph.D. journey not only enriching but also joyful.

I extend my appreciation to the Dutch education system where PhD students are treated as equal colleagues, with minimal hierarchy. Importantly, this environment also embraces cultural diversity. Chinese names can be challenging for the Dutch to pronounce or remember, but they make an effort to address Chinese students by their given names.

In the conclusion of my acknowledgments, I extend my gratitude to ChatGPT and Midjourney. This tool was used to polish the English of the later additions to this thesis, revise coding solutions, and create the image on the cover, but these are only some of the many applications possible. It is clear that Artificial Intelligence will, and is reshaping human roles within society.

Wang Tongjing

Chapter 1

Introduction

1. Introduction

Cities benefit from being related (Castells, 1996; Capello, 2000; Gordon and McCann, 2000; Johansson and Quigley, 2004; Meijers, et al., 2018; Taylor and Derudder, 2022; Bathelt and Storper, 2023). These relationships can manifest in various forms. For instance, transportation flows between cities is a type of intercity relationship, and a good transportation flow suggests an efficient transportation system that enables the shared use of large-scale infrastructure for surrounding cities. Collaboration between cities, especially in scientific research, is another type of relationship, that facilitates a city's innovation capabilities (Balland and Boschma, 2022; Boschma et al., 2023). The advancements in information technology have further strengthened these connections, which has been captured by sociologist Manuel Castells (1996), who noted urban systems are shifting from isolated 'spaces of places' to well-networked 'spaces of flows'.

Building on this, a substantial body of literature has introduced myriad concepts to elucidate the increasing impacts of city relationships on urban economic growth and regional development balance. One of the many highlighted concepts is 'borrowed size' (Alonso, 1973; Meijers and Burger, 2017), which suggests that cities can 'borrow' the advantages from other cities through connections, gaining access to broader markets and resources, thereby improving embedded cities' economic performance and innovation capacity. Additionally, the notion of 'urban network externalities' (Capello, 2000) is also important for guiding planning policies, which indicates that cities are not isolated entities but part of a larger system where each city's growth and innovation can be amplified through its connections with others. Expanding on these insights, the idea of 'megaregions' as posited by Florida et al., (2008) shifts the perspective from individual cities to the collective strength and competitive edge of interconnected cities at the regional level, which claims that the competitiveness of a country on a global scale is increasingly determined at the regional level rather than by individual cities.

Besides, in the contemporary era where innovation is a key driver of economic growth, the traditional emphasis on geographical proximity as a determinant of city interactions necessitates a reevaluation. As Boschma (2005) underscores, geographical proximity is merely one of the many types of proximities that shape intercity relationships, and its importance

in innovation is increasingly mediated by cognitive, organizational, social, and institutional proximities. This view aligns with the concept of 'knowledge spillover' (Bathelt et al., 2004), which posits that innovation and economic growth are not exclusive to geographically concentrated locales. Instead, spatially dispersed locations, connected through various kinds of networks, can act as hubs of innovation and growth.

While there is a rich theoretical foundation underpinning the study of city relationships and their impact, early research in this area often leaned more toward theoretical analysis rather than empirical validation. This tendency stemmed primarily from the difficulty of obtaining data on relationships between cities, as traditional approaches to collecting city relational data, such as surveys, are often time-consuming and costly. An illustrative case was the study by Rozelle et al. (1999) who undertook an extensive survey to investigate rural-urban migration networks in China. This research team, composed of two authors and 14 students and fellows, spent two years (1995 and 1996) surveying over 200 villages in China. However, some might still consider this sample size limited given the vastness of the Chinese territory.

The historical dearth of relational data in urban studies was once starkly characterized as the discipline's 'dirty little secret' (Short et al., 1996), a field marked by 'theoretical sophistication and empirical poverty' (Taylor, 2004). This critique was reiterated by Pumain (2003), who highlighted the scarcity of studies genuinely engaging with relational data. She ardently argues for the application of quantitative methods to elucidate the complexities in intercity relationships, thereby enabling more precise policy guidance and empirically validated results. In this context, the availability of robust relational data becomes a prerequisite for conducting meaningful and actionable quantitative research.

The advent of Information and Communication Technologies (ICTs) has ushered in a transformative era for empirical research in urban studies, mitigating earlier challenges of data scarcity, particularly capturing city relationships in spatial mobility. At the city level, researchers can draw from a rich pool of data such as metro card entries and exits (Wang et al., 2020; Kim, 2019), taxi services recording (Huang et al., 2015; Liu et al., 2015), and mobile phone signals (Wang et al., 2013; Cao et al., 2019; Hadachi et al., 2020) and points of interest tracking (POIs) (Li et al., 2018;

Moyo and Musakwa, 2019; Song et al., 2018). These sources present a fruitful ground for exploration.

However, in general, less data is available at larger spatial scales, and cross-scale studies frequently encounter compatibility issues. When broadening the scope to regional and national scales, the frequency of train and bus connections between cities becomes a primary data source (Cao et al., 2013 and 2019; Liu et al., 2016). On an even larger scale, such as intercontinental and global scales, attention often has to turn to airplane routes, as it is one of the few data sources available for indicating city connections on such a large scale (Derudder et al., 2005; Cardillo et al., 2013). However, such data reveals limited information about passengers' actual travel patterns, thereby offering a relatively narrow window into understanding the a city's position on a global network system.

Furthermore, the previously mentioned types of city relational data are location-focused. While such data provide invaluable insights into the geographical proximity between cities, they fall short of capturing other important dimensions of proximity between cities. It is also widely acknowledged that city networks are multilayered (Burger et al., 2014; Hu et al., 2020), encompassing "not only economic but also social, cultural, and environmental activities" (Davoudi, 2008, p.51). Acknowledging and identifying the diversity of city relationships is critical.

In response to this scarcity of city relational data, alternative proxies have been proposed to indicate relationships between cities from other perspectives. These include using the number of scientific papers and patents collaboratively produced by institutes in different cities to represent the scientific collaboration or "shared understandings" between these cities (Nooteboom, 2000; Werker et al., 2019; Capello and Caragliu, 2018; Cao et al., 2019). Additionally, researchers also proxy city relationships based on their institutional and business connections, notable frameworks including the Globalization and World Cities Research Network (GaWC) and the interlocking network model (Taylor et al., 2004 and 2012; Derudder et al., 2003). These models focus on the dispersion of firms' branches, particularly those of advanced producer service firms, as such firms are considered key economic agents in the global economy, and they strategically position their branches to forge connections between cities, thereby establishing them as nodes in a global service network (Taylor, 2014), a notion aligned with Sassen's 'global city' (1991).

Methodologically, these proxies share a common analytical thread: the frequency of city name appearances, albeit focusing on specific textual content for yielding more precise implications into city relationships. Yet, attempts have been made to apply this analytical method to more generalized texts such as Wikipedia, newspapers, and search engine results (Devriendt et al., 2011; Liu et al., 2014; Peris et al., 2020), exploring its potential to obtain a more comprehensive understanding of intercity relationships. This broader method is known as collocation analysis or toponym co-occurrence in geography. This type of method is focused on extracting and analyzing patterns of word co-appearance within text, especially the patterns with a higher-than-usual frequency. City names may coincidentally co-appear in text, but if such co-appearance frequency is significantly higher than usual in a large sample, then a general conclusion can still be confidently drawn that certain pairs of cities are more related than other possible pairs that co-occur less.

While this method is gaining popularity recently, it is not new. This analytical approach is not new; it was first proposed by linguist J.R. Firth in 1957, who discovered the potential of using word co-appearances to evaluate lexical and grammatical relationships between words. One of Firth's often cited quotes is "You shall know a word by the company it keeps (Firth, 1957, p179) ", suggesting that the meaning and usage of a word could be inferred from its most frequent collocates.

This method was later re-envisioned for geographic explorations, pioneered by geographer Waldo Tobler, who reconstructed the urban system of 119 pre-Hittite towns in Capadoccia 4,000 years ago based on toponym co-occurrences on cuneiform tablets (Tobler & Wineburg, 1971). Subsequent research has applied this method to analyze various databases, including newspapers, search engines, Wikipedia, web archives, and historical documents (Wu et al., 2020; Liu et al., 2014; Meijers and Peris, 2018; Salvini and Fabrikant, 2016; Peris et al., 2020). These studies all claim that their results suggest that the frequency of city name co-appearances can reflect city relationships to a certain extent, promising new approaches to identifying city relationships.

While promising, employing collocation analysis as a tool in city network analysis still presents several challenges. On the practical side, the task of data extraction from large databases can be difficult, particularly for

researchers in social science who may not be familiar with data mining techniques. Additionally, the relationships extracted solely from co-appearance frequencies may lack nuance and specificity, thereby raising questions about the depth of the insights generated.

Additionally, translating the collocation patterns into insights for guiding city and regional policy-making poses its own set of challenges. These include assessing the importance of city relationships in enhancing city performances and informing regional planning initiatives effectively.

To address these issues, this Ph.D. thesis has two objectives, a methodological one and an empirical one.

The methodological objective is to explore the potential of this collocation analysis method in extracting relationships between cities. The aim is to construct a more robust analytical framework capable of capturing the multi-dimensional intricacies of city networks. This involves developing practical strategies for extracting collocation patterns of cities from large databases and for classifying these patterns.

The empirical objective is to leverage the collected collocation patterns to deepen our understanding of the role that city relationships play within urban systems. This will be achieved by comparing the relative importance of network externalities with agglomeration externalities and by developing methods for measuring the embeddedness of cities within a planned region to inform megaregion planning strategies.

2. Theoretical background

This section is organized as follows: firstly, it will illustrate the potential benefits a city can receive from being connected with other cities. Following this, the discussion will transition from individual city scale to the regional scale, regarding governmental megaregion plans aimed at improving regional competitiveness. Finally, the section will introduce the method of collocation analysis and highlight both its potential benefits and limitations. By sequentially going through these topics, this section aims to provide a comprehensive foundation for the subsequent studies presented in this thesis.

2.1 Agglomeration vs network externalities

The traditional framework for urban development is rooted in Alfred Marshall's agglomeration theory in the 1920s (Marshall, 1920; Phelps and Ozawa, 2003). This theory advocates for population concentration in central locations to achieve a range of benefits including a more efficient labor market, reduced transportation costs, and facilitating input sharing, labor market pooling, and knowledge spillovers. Later studies have expanded benefits including home market effects, where concentrated demand stimulates agglomeration (Krugman, 1980, 1995), and consumption-related advantages (Glaeser et al., 2001; Waldfogel, 2008).

While agglomeration is still acknowledged as the driving force behind urban development, there is also a growing recognition of the associated costs, such as congestion, pollution, high housing prices, and risks of large-scale social unrest. As cities grow in size, those associated costs, or so-called negative externalities, can potentially outweigh the benefits brought by agglomeration. Therefore, politicians and urban planners are increasingly hesitant to promote traditional agglomeration-oriented developing models (Au and Henderson, 2006; Wei et al., 2015). Instead, strategies that focus on promoting more balanced urban development are gaining more popularity.

One of the common strategies to still capitalize on agglomeration benefits while mitigating the negative externalities is to foster complementary activities that cities can mutually benefit from. For instance, one city might host a major airport that serves the entire region, while another city could feature a large port that handles and redistributes cargo, and yet another could be a regional center for politics and administration. Another strategy is to strengthen collaboration, particularly evident in encouraging manufacturing products that are within a shared supply chain. By specializing in the production of specific components, cities can collectively contribute to assembling complex, high-value products like automobiles or semiconductor devices. The successful realization of these complementary and collaborative activities necessitates cities being well-embedded through various socio-economic relationships, facilitating their integration into a cohesive urban system. In other words, through collaborative endeavors and specialized roles, cities have the potential to realize increasing returns while avoiding traditional agglomeration negativities. This form of benefits that comes from being related with other

cities is conceptualized as ‘network externalities’, distinguishing it from traditional agglomeration externalities.

The concept of ‘network externalities’ is not new; it has historical roots in Howard’s Garden City model (1898) and was later articulated as ‘borrowed size’ by Alonso in 1973. Over the years, this concept has been further nuanced and expanded upon through various theoretical frameworks. The notion of ‘borrowed size,’ for instance, has been further studied by scholars like Phelps et al., (2001 and 2003), Capello and Camagni (2000), Meijers et al. (2016), and Meijers and Burger (2017), who identified that borrowed size is facilitated through networks of interacting cities. Capello (2000) further refined the concept by introducing the term ‘urban network externalities’ to describe how cities can exploit economies of scale through complementary relationships and cooperative activities. In this framework, the economic scale is distributed among embedded cities, despite the individual costs each incurs to participate in the network. Building on this, the discussion of ‘polycentric urban regions’ proposes that networks of cities can lead to more balanced regional development (e.g. Green, 2007; Liu et al., 2016; Meijers et al., 2018; Burger and Meijers, 2012).

Additionally, the concepts of ‘principal of relatedness’ highlight the importance of recognizing non-geographical relatedness between cities, especially in the context of regional innovation (e.g. Boschma, 2005; Boschma and Iammarino, 2009; Neffke et al. 2011). Furthermore, the idea of ‘knowledge spillover’ posits that relationships between cities facilitate the transfer of knowledge, which is crucial for innovation and economic development (e.g. Bathelt et al., 2004; Acs et al., 2009; Rutten, 2017). The concept of ‘megaregion,’ as articulated by Florida et al. (2008), suggests that for optimal economic development, there needs to be a focus on regional cohesion, effectively treating a network of cities as a singular economic entity to exploit economies of scale to a greater extent. Each of these frameworks offers a unique lens through which to understand the complexities of network externalities.

Urban development has undergone significant transformations, particularly in how we understand the role of geographical proximity in agglomeration externalities. Earlier empirical studies often depicted these externalities as being geographically constrained (Rosenthal and Strange, 2004). This was partly due to historical limitations such as poor

infrastructure, lack of efficient communication methods, and the costs associated with competitive trading, which often outweighed the benefits of connectivity. However, technological advancements have revolutionized this scenario. The barriers to connectivity have been substantially reduced. The ubiquity of information and communication technologies means that spatial proximity no longer holds the dominant power it once had in establishing strong ties. This has led to a reevaluation of the role of various alternative types of proximity—organizational, institutional, cognitive, and social—in complementing spatial proximity (Capello, 2020; Johansson and Quigley, 2004; Boschma, 2005).

Moreover, the shift in the urban economy from manufacturing-centric industries to service-oriented sectors has placed a greater emphasis on creativity and innovation (Florida, 2005; Glaeser, 2011). In this evolving context, distant networks have become increasingly important for knowledge spillovers, complementing local interactions (Rutten, 2017). Some recent research even posits that these network externalities may be more effective in promoting innovation than traditional agglomeration externalities (Basile et al., 2012; Galaso and Kovářík, 2021).

Intercity connections are especially important for medium and small cities. Camagni et al. (2015) found that second-tier cities in Europe can overcome the lack of agglomeration through innovation and city networks. While agglomeration economies remain important, many empirical studies show that stronger relationships and connectivity to other regions and (larger) cities can also foster development, and facilitate agglomeration benefits spill over to nearby smaller places, empirical studies including Europe (Camagni et al., 2016; Meijers et al., 2016; Cicerone et al., 2020), USA (Chatman and Noland, 2014), China (Huang et al., 2020) and Japan (Otsuka, 2020). However, regarding cultural amenities, Burger et al. (2015) also warned that the size of a city still matters most—larger cities actually profit more than smaller cities from being embedded in networks, and in fact, often cast an ‘agglomeration shadow’ over smaller cities.

The high-speed railway network has been widely used as a proxy representing the strength of city relationships (Niu and Li, 2018; Jiao et al., 2020; Huang et al., 2020). Generally, these studies show that being connected to a high-speed railway system can have positive effects on the economic growth of cities. However, it must be noted that while there may be a generative effect (a general positive effect of transport infrastructure

improvement) for the economy as a whole, this may hide a distributive effect too in the sense that some cities profit, whereas others lose out due to improved accessibility, and in its wake, increased competition (Meijers et al., 2012). This is particularly salient for peripheral cities that lack targeted development policies. As these cities are driven toward agricultural specialization, they risk losing industrial output (Faber, 2014; Baum-Snow et al., 2020). Here, we may draw an analogy with agglomeration benefits and costs, as there are also network benefits and costs.

However, evaluating the relationship between agglomeration and network externalities remains challenging for several reasons: firstly, agglomeration and network externalities, conceptually, are fuzzy concepts (Van Meeteren et al., 2016), making it difficult to disentangle due to their interconnected nature. Secondly, there are various types of agglomeration and network externalities (Burger et al., 2014; Gross and Ouyang, 2021). Drucker (2012) posits that the rate at which agglomeration externalities diminish with distance depends on the industry, type of agglomeration, and type of externality examined. Thirdly, the significance of agglomeration (size) and network connectivity varies with performance measures (Meijers et al., 2016; Phelps, 2021). From certain perspectives, network connectivity is more critical than size in determining the existence of for instance specific metropolitan functions in cities, implying that networks could effectively substitute for size.

2.2 Megaregion planning

While empirical evidence quantifying the benefits of network externalities remains limited, policymakers and researchers have long envisioned leveraging agglomeration advantages on a regional scale. This vision has led to the proposal of numerous large-scale initiatives, commonly referred to as megaregion plans.

Although there is no single definition or shared standard for determining whether a large-scale sprawling urbanized landscape is a megaregion (Dewar and Epstein, 2007; Fang and Yu, 2017), the aim of employing the megaregion concept is quite clear, namely to leverage the possibility of expanding agglomeration, enhancing the competitiveness, cohesiveness, and sustainability of a given regional territory (CEC, 2011), as megaregion

development plans are associated with promises that global economic growth is and will continue to be concentrated in a few regions.

Researchers and planners in China particularly embrace this promise, as numerous megaregion proposals have been incorporated into various types of central and local planning documents. This enthusiasm even extends from the government to the general public, who would view inclusion in a megaregion as a precondition for the city's prosperity, whereas exclusion would lead to stagnation (Fang, 2015). This perspective reflects the high stakes involved in the conceptualization and operationalization of megaregions, which needs particular caution (Wu and Zhang, 2007; Li and Wu, 2012; Wu, 2018).

The early approaches to defining megaregions often departed from a morphological perspective (Harrison and Hoyler, 2015), marking out space by observing the physical landscape, selecting a small number of proximate large cities, and then including the surrounding cities (Fang, 2015; Florida et al., 2008). This early morphological perspective focused on stock indicators such as the number of cities, the total population within the delineated area, the whole built-up area, and the total economic size. (Fang and Yu, 2017). However, later research shows that just being close to each other is not enough to guarantee a strong metropolitan economy (Ahrend et al., 2017)—summing small cities does not necessarily make a successful megaregion, as fragmentation looms: "institutional and spatial fragmentation, functional redundancies, uncoordinated transport planning, disconnected housing, and labor markets, imbalanced distribution of investment, unwillingness to cooperate by local authorities in the absence of a metropolitan government, and a lack of common historical, cultural or political references able to shape a joint identity and shared strategic priorities" (Cardoso and Meijers, 2020, p362). Fragmentation, except physical ones, is not visible on the map and therefore, renders cartographic/morphology-based approaches a rather limited perspective.

It is now widely acknowledged that to achieve the agglomeration economies associated with a large megaregional critical mass, the cities constituting megaregions should be integrated well in multiple dimensions (Ahrend et al., 2017; Meijers et al., 2018). Such integration processes take a long time, and are characterized by periods of rapidly increasing integration when the interests of economic and governmental actors as well

as the public at large align, but can also be characterized by throw-backs if this is not the case (Cardoso and Meijers, 2021).

A megaregion is also composed of a wide variety of places whose governments may have different interests that do not necessarily align. The interest of smaller cities can, for instance, be to obtain better access to urban functions in large cities, and/or to acquire a higher political status for negotiating with higher-level government by being part of a larger consortium (Meijers and Burger, 2017; Cardoso and Meijers, 2017; Derudder et al, 2022). Large cities in megaregions may want to extend the support base for their metropolitan functions, or use their relationships with other cities to mitigate negative returns (Camagni et al, 2017). Beyond the local government level, national governments see the development of megaregions as a way for certain regions to obtain a favorable investment and business environment and consequently improve their global competitiveness (Harrison and Hoyler, 2014; Wu, 2020a). In addition, and beyond economic imperatives, a series of urban challenges also calls for recognizing using the megaregion scale to facilitate collaboration and overcome trans-territorial issues, such as economic inequalities (Farole et al., 2011), competition between adjacent cities (Pan et al., 2017), inter-regional infrastructure services (Wang et al., 2021), and environmental hazards and pollution (Chen et al., 2023).

Megaregion plans also comprise more than what common metropolitan plans at the smaller scale focus on in functional terms (e.g. strengthening accessibility for daily commuting), as the megaregion plans are often envisioned as 'associational repertoire' (Cooke and Morgan, 1994), a framework that can facilitate effective interactions. As indicated by Cardoso and Meijers (2020), urban planning should contribute to a process of metropolisation to make "institutionally, functionally, and spatially fragmented urbanized regions become integrated along various dimensions and emerge as connected systems at a higher spatial scale". In other words, actively developing a coherent region is what most megaregion plans have to strive for to reap the benefits associated with megaregions (Meijers et al., 2018).

In conclusion, the success of a megaregion depends on the multi-dimensional relationships among cities within the planned megaregion. To foster successful megaregion development, it is necessary to first understand how the individual cities within the planned megaregion are

actually related to each other. This necessitates a comprehensive assessment to identify the existing level of integration, serving as a prerequisite for any future planning and implementation strategies.

2.3 Collocation analysis method

Collocation analysis measures word association degrees through the word co-appearances in text. When this method is applied in geographic entities, this method is also referred to as toponym co-occurrence. This approach has shown the potential to capture the strength of the relationship between cities through the co-occurrence frequency of the city names in the text.

Collocation analysis was initially introduced by linguist J.R. Firth (1957) as a lexical analysis tool. As was observed already in the introduction, his phrase "You shall know a word by the company it keeps" (Firth 1957, p179) captures the essence of collocation analysis well. Building on Firth's contributions, collocation is defined as "a combination of two words that exhibit a tendency to occur near each other in natural language" (Evert, 2008, p4). This notion is further underscored by Hunston (2011, p14) who notes, "the meaning of any word cannot be identified reliably if the word is encountered in isolation."

One of Firth's contributions to collocation analysis is that Firth set the empirical-focused analysis framework (Evert, 2008), emphasizing that the relationship between words should be determined by their actual co-appearance frequency derived from empirical studies rather than theoretical conjectures. This aspect resonates with the principles of contemporary big data analysis methodologies that prioritize deriving insights directly from empirical evidence.

Over the years, the scope of collocation analysis has expanded from measuring word relationships to analyzing the topic of the text. This extension has been informed by the realization that keywords can serve as indicators of a text's thematic focus if these keywords are more frequently associated with specific contexts (Stubbs, 2001). Thus, collocation analysis has evolved from its lexical origins to become a versatile tool for examining the content and structure of the text.

A distinguishing feature of collocation analysis, as compared to frequency analysis, is its focus on words that are more associated with each other than

would be expected by chance. In other words, a high co-occurrence frequency between two words does not necessarily indicate a high association between the two words (Evert, 2008). Collocation analysis goes a step further by comparing the observed frequency of word pairs with what would be expected under normal circumstances, thereby providing a more nuanced understanding of word associations.

Collocation analysis is also a summative-based distant-reading method, distinguishing it from close-reading techniques such as Critical Discourse Analysis (CDA) and New Criticism. Its strength lies in its ability to conduct large-scale, systematic analyses, making it particularly valuable for extensive examination of digital text. When contrasted with other distant-reading methods like machine learning-based methods, collocation analysis holds distinct advantages (Baker, 2006; Baeza-Yates and Castellanos, 2019). Despite machine learning methods having the capability to analyze and compare enormous textual features, they often have difficulties in quantifying relationships with social relevance, providing meaningful interpretations, and ensuring reliability and controllability (Lazer, 2009; Kelling et al., 2009; Loukides, 2010; Miller, 2010). Such limitations can make relationships that are quantified by machine learning methods less effective in social studies, where nuanced interpretation and contextual understanding are paramount. In contrast, collocation analysis offers unique benefits, as it focuses on relationships between keywords, which are easier for calibration and interpretation compared to machine learning methods.

Collocation analysis can be conducted in two approaches: a ‘driven’ approach and a ‘based’ approach. In the driven approach, a specific keyword is selected, and all words that frequently co-occur with this given keyword are identified. This method allows for a deeper understanding of the contextual usage of the chosen keyword. Typically, the next step involves using Critical Discourse Analysis (CDA) to explore the reasons behind the frequent co-occurrence of certain words with the keyword. The research conducted by Baker et al. (2008) is an illustrative example of how collocation analysis can be effectively combined with other methodologies like Critical Discourse Analysis (CDA) for a more nuanced understanding. In their study, they analyzed a 140-million-word corpus of British news articles focusing on refugees, asylum seekers, immigrants, and migrants (collectively referred to as RASIM). Initially, they employed collocation analysis to identify words that frequently co-occurred with the terms under

the RASIM umbrella. This quantitative approach allowed them to systematically sift through an extensive dataset to pinpoint recurring lexical patterns. Following this, they utilized CDA for a deep qualitative analysis to explore the reasons behind the frequent co-occurrence of these words with the different RASIM terms. This enabled them to delve into the underlying meanings, implications, and contextual factors that influenced these lexical choices. By doing so, they could compare and contrast the differences in discourse surrounding the various groups categorized under RASIM.

In contrast, the 'based' approach begins with a predetermined set of keywords and aims to assess the degree of collocation among them. Following this initial assessment, the analysis can take one of two directions. It may adopt elements of the 'driven' approach by delving into the reasons why certain word combinations exhibit a higher degree of collocation. The study by Porter et al. (2018) is an example of how this 'based' approach can be effectively applied to yield insights in other research fields. They extracted co-occurrence frequency between predetermined keywords related to place names and diseases from a 100 million words corpus of UK newspapers. The co-occurrence frequency was then used to map the temporal evolution of mortality patterns, providing insights into the spatial and temporal distribution of mortality rates and disease prevalence in the UK during the 1900s.

Alternatively, the findings from the collocation analysis can also serve as a foundation to support other study fields. For instance, Nyman et al. (2021) used words and their collocated sentiment words to predict stock market trends. In the tourism sector, Camprubí and Coromina (2016) employed place names and their collocated words on tourism websites to assess the appeal of various tourism destinations. Overell and Rüger (2008) utilized place name co-occurrence on Wikipedia to enhance the accuracy of toponym disambiguation. Similarly, Wu et al. (2019) incorporated toponym co-occurrence results to refine the field strength model, thereby improving the identification of a metropolitan area's hinterland.

In geography, collocation analysis—specifically focusing on toponym co-occurrence, has demonstrated considerable promise in uncovering the complex relationships between cities. Its adaptability and versatility have allowed it to be applied across a diverse range of textual data sources, including online newspapers, search engines, Wikipedia, and web archives

(Devriendt et al., 2011; Liu et al., 2014; Peris et al., 2020). Furthermore, these applications have shown the method's effectiveness, with findings that align with other types of city relationships.

Regarding the application of toponym co-occurrence in China, some pioneering studies have shown the potential this method has for analyzing intercity relationships. Liu et al. (2014) applied this method to Baidu, a Chinese Internet search engine, to investigate the relationship between geographical entities. They found relationship patterns reflect similarities between neighboring provinces and indicate the spatial organization of China. Zhong et al. (2017) further developed this method by applying complex network theory to evaluate the topological structures of the toponym occurrence network which was extracted daily from a newspaper over the course of a year. They found that the network showed strong cluster characteristics, and the frequency of toponym co-occurrence was negatively correlated with the administrative hierarchy, but less so with geographic distance. Guo et al. (2022) calculated a city's total appearance with other Chinese cities on the search engine Baidu and examined the factors that can contribute to the frequency of this appearance, finding that factors such as GDP, administration level, tourism, and the number of enterprises all significantly increase a city's appearance on the Internet.

By focusing only on keywords, collocation analysis often faces criticism for its apparent simplicity, arguing that collocation may overlook the actual text content of each word co-appearance. For instance, relying on immediate co-text can risk a skewed or superficial understanding of the broader context. The patterns detected through collocation analysis may not inherently carry meaningful significance, and without appropriate contextualization, the analysis could yield surface-level interpretations. Another concern is the method's emphasis on higher-than-usual patterns of word co-occurrence, which could potentially overshadow more subtle nuances. As Baker et al. (2013) caution, there is a risk of merely 'uncovering the obvious', thereby missing out on less frequent but equally significant patterns. Additionally, collocation analysis tends to prioritize inter-textual patterns over intra-textual nuances. This focus can result in the overlooking of significant meanings or patterns that are specific to individual texts (Tognini-Bonelli, 2001).

Indeed, these concerns are reasonable, and collocation analysis is not a perfect method. However, as Baker et al. (2008) mentioned, it is necessary

to analyze word patterns in an interpretative rather than purely descriptive nature. A key justification for this method is the principle of local interpretation, which suggests that effective communication doesn't necessitate a context more complex than what's required for interpretation (Brown and Yule, 1982; Tognini-Bonelli, 2001). As such, collocation analysis helps in "limiting the interpretation to what is contextually appropriate or plausible" (Brown and Yule, 1982, p59).

In light of these limitations, various strategies have been employed to substantiate the co-occurrence frequency results. One commonly employed strategy to understand why certain words collocate more frequently than others involves the use of in-depth reading tools like Critical Discourse Analysis (CDA) (Potts, et al., 2015; Brezina, 2018; Brookes and Baker, 2022). In a different methodological trajectory, Salvini and Fabrikant (2016) have taken a unique approach by leveraging the content tags present on Wikipedia webpages; they extracted toponym co-occurrence and used these tags for classifying the results. Meijers and Peris (2019) have integrated machine learning techniques to initially categorize text topics and subsequently quantify the strength of relationships between these topics using toponym co-occurrence.

Despite these methodological advancements, the extraction of meaningful information from unstructured text remains a formidable challenge, underscoring the necessity for continued innovation in research methodologies especially for geographical analysis.

3. Research questions

The main research question driving this Ph.D. thesis is :

To what extent and how can collocation analysis be used to extract, categorize, analyze, and evaluate relations between cities?

This question aims to steer the two primary objectives of the Ph.D. research: methodologically, to explore the potential of this collocation analysis method in extracting relationships between cities; empirically, to leverage the results collected from collocation analysis to deepen the understanding of the role that inter-city relationships play in enhancing city performance and regional development.

To comprehensively address this main question, it is further unpacked into four interconnected sub-questions, each leading to an empirical case study that reveals the complexity of city networks. The four sub-questions start from the foundational step of toponym co-occurrence, namely 1) extracting relationships between cities from large textual datasets, 2) progressing to characterizing the multidimensional facets of city networks. This foundation paves the way for an empirical exploration of city network applications, specifically highlighting the significance of city relationships in 3) enhancing city performance and 4) formulating functional coherent megaregion plans.

Given the complexities of handling large databases, developing practical methods to extract collocation patterns from these vast data sources or obtaining ready-to-use datasets for social study domain researchers can be difficult. Therefore, my first sub-question is

1) *How can city relationships be effectively and practically extracted from large data for social domain researchers?*

Common approaches that use collocation patterns to proxy city relationships fall short in capturing the nuanced meanings of relationships between the cities. Therefore, my second sub-question is

2) *How can relations between cities as found through toponym co-occurrences be categorized?*

After collecting collocation-based city relationships, the study intends to address a longstanding empirical question about the importance of city network externalities, especially compared with agglomeration externalities. My third sub-question is

3) *What is the comparative importance of city relationships vis-à-vis agglomeration in influencing city performance?*

Given the important role that city relationships play in city performance, the advocacy for planning 'megaregions' by policymakers and urban planners, with the goal of leveraging network externalities, becomes a logical step. However, realizing this ambition demands that cities within megaregions be well-integrated across multiple dimensions, a concept referred to as 'functional coherence.' Consequently, to develop targeted

policy strategies for each megaregion, it is essential to first assess the varying degrees of functional coherence within these planned megaregions. Such an assessment sets the stage for the following sub-question:

4) *How can the functional coherence of proposed megaregions be scrutinized and categorized based on the strength of their intercity relationships?*

By tackling the four sub-questions, this thesis aims to respond to the main question by examining the capability and limitations of collocation analysis in the study of city networks and geography. From the capabilities aspect, it focuses on the practicality of using collocation analysis in identifying and categorizing city relationships. Regarding the limitations aspect, it intends to assess the depth and accuracy of the captured collocation-based city relationships.

4. Geographical focus

The empirical focus of the four research questions will be on China. The nation's unbalanced urbanization highlights the necessity of strengthening relationships between cities as a strategic means to foster balanced development (Phelps et al., 2020). Furthermore, the numerous large-scale regional and national development plans underscore the local and central government's commitment to enhancing inter-city relationships for coordinated regional planning (Fang and Yu, 2017; Harrison and Gu, 2019). This context presents a unique opportunity to examine the impact of city relationships on urban and regional development and hopefully, the analysis results can provide for future urban and regional planning in China.

The potential of collocation analysis to obtain large-scale data on city relationships is particularly significant, given the current scarcity of such city relational data in China. Despite much literature that explores the relationships between cities through diverse metrics—be it trade, commuting patterns, migration flows, tourism activities, goods transportation, or telecommunications—access to large-scale data remains limited. As a result, existing studies on China's multi-dimensional city networks often limit their scope either to a select number of top-tier cities at the national level (Lao et al., 2016) or to regional case studies, such as the Yangtze River Delta (Cao et al., 2021; Zhang et al., 2020). The

forthcoming results include new large-scale intercity relationship datasets, thereby hopefully the results will add value for future city network research.

One comparative advantage of using the Chinese language in our study lies in its avoidance of semantic ambiguity often encountered with Latin-based languages, as most Chinese characters are unique to the language itself. Additionally, the absence of plural or gender forms in Chinese words simplifies the process of text corpus searching, further enhancing methodological efficiency. In the end, the conclusion chapter will address the question of whether the collocation method, as applied within the context of this study, holds broader applicability across different languages and cultural contexts globally, examining its potential universal effectiveness in analyzing city networks.

5. Thesis outline

This Ph.D. thesis is structured around six interconnected chapters. The sequential workflow of these chapters is presented in Figure 1, providing a schematic overview that aids in understanding the logical progression and interconnectedness of the chapters.

Chapter 2 answers research question 1 by developing an efficient and user-accessible method to determine the co-occurrence frequency of cities. This work is founded on an empirical study extracting intercity relationships between 293 Chinese cities from a substantial 6.98 TB corpus amassed from Common Crawl, a web archive. This chapter has been published in Cybergeo: European Journal of Geography, where a ready-to-use dataset of these intercity relationships is offered.

Chapter 3 addresses research question 2, taking the co-occurrence methodology a step further. It employs a lexicon-based text mining method, facilitated by natural language processing, to categorize the toponym co-occurrences discerned in the previous chapter. This chapter aims to classify these frequencies into six distinct categories: industry, information technology (IT), finance, research, culture, and government. Each category displays different network patterns, and this multiplexity is mapped and analyzed. This chapter has been published in Regional Studies.

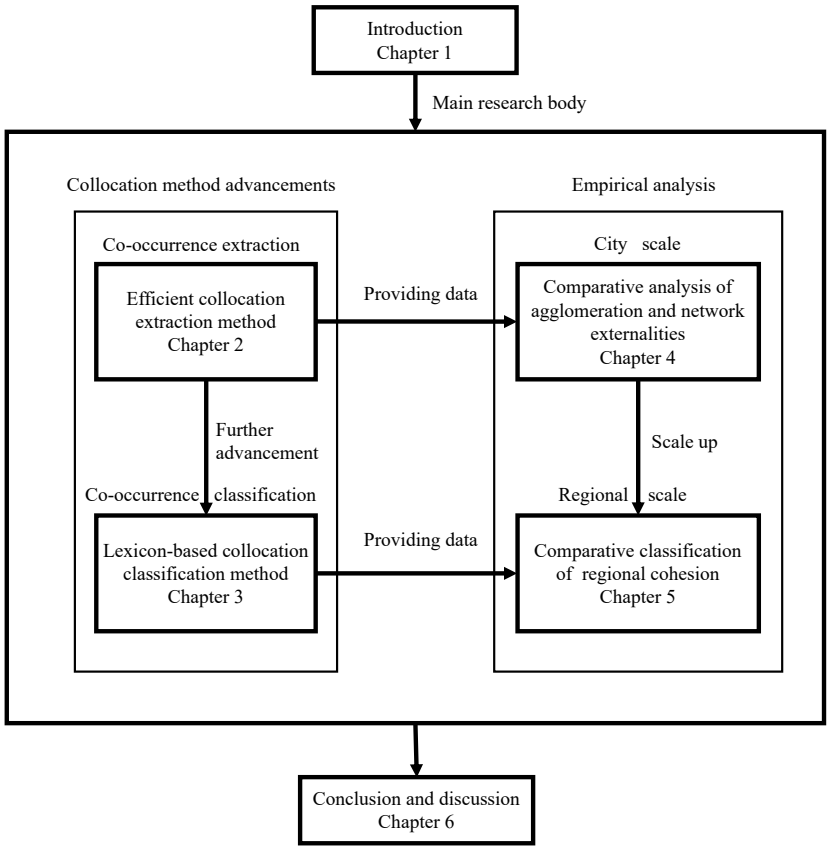
Chapter 4 responds to research question 3, initiating the computation and mapping of both absolute and relative network positions of each city

involved in the study. This step is followed by a comparative analysis of the implications city network externalities and agglomeration externalities hold for urban performance through a series of regression models. The empirical study employs the dataset derived in chapter 2. This chapter has been published in International Journal of Urban Sciences.

Chapter 5 answers research question 4 by proposing a systematic approach to assess the functional coherence of megaregions from three dimensions: inclusion, integration, and consistency. This examination is applied to fifteen government-delineated Chinese megaregions, each subjected to detailed evaluation and classification. This chapter has been published in Regional Studies.

Finally, Chapter 6 concludes by addressing the four subquestions and the main research question. Following this, it discusses the limitations of this study and proposes directions for future research.

Figure 1. Work flow of PhD thesis



References

Acs, Z. J., Braunerhjelm, P., Audretsch, D. B., & Carlsson, B. (2009). The knowledge spillover theory of entrepreneurship. Small business economics, 32, 15-30.

Alonso, W. (1973). Urban zero population growth. Daedalus, 191-206.

Balland, P. A., & Boschma, R. (2022). Do scientific capabilities in specific domains matter for technological diversification in European regions?. Research Policy, 51(10), 104594.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse & society, 19(3), 273-306.

Bathelt, H., & Storper, M. (2023). Related variety and regional development: A critique. Economic Geography, 1-30.

Bathelt, H., Malmberg, A., & Maskell, P. (2004). Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. Progress in human geography, 28(1), 31-56.

Batty, M. (2008). The size, scale, and shape of cities. science, 319(5864), 769-771.

Batty, M. (2013). The new science of cities. MIT press.

Boschma, R. (2005). Proximity and innovation: a critical assessment. Regional studies, 39(1), 61-74.

Boschma, R., & Iammarino, S. (2009). Related variety, trade linkages, and regional growth in Italy. Economic geography, 85(3), 289-311.

Boschma, R., Miguelez, E., Moreno, R., & Ocampo-Corrales, D. B. (2023). The role of relatedness and unrelatedness for the geography of technological breakthroughs in Europe. Economic Geography, 99(2), 117-139.

Brookes, G., & Baker, P. (2023). Cancer services patient experience in England: Quantitative and qualitative analyses of the National Cancer Patient Experience Survey. BMJ Supportive & Palliative Care, 13(e3), e1149–e1155.Burger, M. J., Meijers, E. J., & Van Oort, F. G. (2014). Multiple Perspectives on Functional Coherence: Heterogeneity and Multiplexity in the Randstad. Tijdschrift voor economische en sociale geografie, 105(4), 444-464.

Burger, M., & Meijers, E. (2012). Form follows function? Linking morphological and functional polycentricity. Urban studies, 49(5), 1127-1149.

Cao, H., Sankaranarayanan, J., Feng, J., Li, Y., & Samet, H. (2019). Understanding metropolitan crowd mobility via mobile cellular accessing data. ACM Transactions on Spatial Algorithms and Systems (TSAS), 5(2), 1-18.

Cao, J., Liu, X. C., Wang, Y., & Li, Q. (2013). Accessibility impacts of China's high-speed rail network. Journal of Transport Geography, 28, 12-21.

Cao, W., Feng, X., & Zhang, H. (2019). The structural and spatial properties of the high-speed railway network in China: A complex network perspective. Journal of Rail Transport Planning & Management, 9, 46-56.

Cao, Z., Derudder, B., & Peng, Z. (2019). Interaction between different forms of proximity in inter-organizational scientific collaboration: The case of medical sciences research network in the Yangtze River Delta region. Papers in Regional Science, 98(5), 1903-1924.

Capello, R. (2000). The city network paradigm: measuring urban network externalities. Urban Studies, 37(11), 1925-1945.

Capello, R., & Caragliu, A. (2018). Proximities and the intensity of scientific relations: synergies and nonlinearities. International Regional Science Review, 41(1), 7-44.

Cardillo, A., Zanin, M., Gómez-Gardenes, J., Romance, M., García del Amo, A. J., & Boccaletti, S. (2013). Modeling the multi-layer nature of the European Air Transport Network: Resilience and passengers re-scheduling under random failures. The European Physical Journal Special Topics, 215, 23-33.

Castells, M. (1996). The space of flows. The rise of the network society, 1, 376-482.

Derudder, B., & Witlox, F. (2005). An appraisal of the use of airline data in assessing the world city network: a research note on data. Urban Studies, 42(13), 2371-2388.

Derudder, Witlox, & Catalano. (2003). Hierarchical tendencies and regional patterns in the world city network: a global urban analysis of 234 cities. Regional Studies, 37(9), 875-886.

Ducruet, C., & Beauguitte, L. (2014). Spatial science and network science: review and outcomes of a complex relationship. Networks and Spatial Economics, 14(3-4), 297-316.

Evert, S (2008). Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), Corpus Linguistics. An International Handbook, article 58, pages 1212-1248. Mouton de Gruyter, Berlin.

Florida, R., Gulden, T., & Mellander, C. (2008). The rise of the mega-region. Cambridge journal of regions, economy and society, 1(3), 459-476.

Green, N. (2007). Functional polycentricity: A formal definition in terms of social network analysis. Urban studies, 44(11), 2077-2103.

Hadachi, A., Pourmoradnasseri, M., & Khoshkhah, K. (2020). Unveiling large-scale commuting patterns based on mobile phone cellular network data. Journal of Transport Geography, 89, 102871.

Hu, X., Wang, C., Wu, J., & Stanley, H. E. (2020). Understanding interurban networks from a multiplexity perspective. Cities, 99, 102625.

Huang, X., Zhao, Y., Ma, C., Yang, J., Ye, X., & Zhang, C. (2015). TrajGraph: A graph-based visual analytics approach to studying urban network centralities using taxi trajectory data. IEEE transactions on visualization and computer graphics, 22(1), 160-169.

Kim, K. (2019). Identifying the structure of cities by clustering using a new similarity measure based on smart card data. IEEE Transactions on Intelligent Transportation Systems, 21(5), 2002-2011.

Kloosterman, R. C., & Musterd, S. (2001). The polycentric urban region: towards a research agenda. Urban studies, 38(4), 623-633.

Li, M., Kwan, M. P., Wang, F., & Wang, J. (2018). Using points-of-interest data to estimate commuting patterns in central Shanghai, China. Journal of Transport Geography, 72, 201-210.

Li, Y., & Wu, F. (2012). The transformation of regional governance in China: The rescaling of statehood. Progress in Planning, 78(2), 55-99.

Liu, X., Derudder, B., & Wu, K. (2016). Measuring polycentric urban development in China: An intercity transportation network perspective. Regional Studies, 50(8), 1302-1315.

Liu, X., Gong, L., Gong, Y., & Liu, Y. (2015). Revealing travel patterns and city structure with taxi trip data. Journal of transport Geography, 43, 78-90.

Liu, Y., Wang, F., Kang, C., Gao, Y., & Lu, Y. (2014). Analyzing Relatedness by Toponym Co-O ccurrences on Web Pages. Transactions in GIS, 18(1), 89-107.

Marshall, A. (1920). Principles of economics, 8th ed. Palgrave Macmillan.

Meijers, E. J., & Burger, M. J. (2017). Stretching the concept of 'borrowed size'. Urban studies, 54(1), 269-291.

Meijers, E., & Peris, A. (2019). Using toponym co-occurrences to measure relationships between places: Review, application and evaluation. International Journal of Urban Sciences, 23(2), 246-268.

Meijers, E., Hoogerbrugge, M., & Cardoso, R. (2018). Beyond polycentricity: Does stronger integration between cities in polycentric urban regions improve performance?. Tijdschrift voor economische en sociale geografie, 109(1), 1-21.

Memon, I., Chen, L., Majid, A., Lv, M., Hussain, I., & Chen, G. (2015). Travel recommendation using geo-tagged photos in social media for tourist. Wireless Personal Communications, 80, 1347-1362.

Moyo, T., & Musakwa, W. (2019). Exploring the potential of crowd sourced data to map commuter points of interest: a case study of Johannesburg. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 42, 1587-1592.

Neffke, F., Henning, M., & Boschma, R. (2011). How do regions diversify over time? Industry relatedness and the development of new growth paths in regions. Economic geography, 87(3), 237-265.

Nooteboom, B., Van Haverbeke, W., Duysters, G., Gilsing, V., & Van den Oord, A. (2007). Optimal cognitive distance and absorptive capacity. Research policy, 36(7), 1016-1034.

Peris, A., Faber, W. J., Meijers, E., & Van Ham, M. (2020). One century of information diffusion in the Netherlands derived from a massive digital archive of historical newspapers: The DIGGER dataset. Cybergeo: European Journal of Geography.

Peris, A., Meijers, E., & van Ham, M. (2018). The evolution of the systems of cities literature since 1995: Schools of thought and their interaction. Networks and Spatial Economics, 18, 533-554.

Phelps, N. A., & Ozawa, T. (2003). Contrasts in agglomeration: proto-industrial, industrial and post-industrial forms compared. Progress in human geography, 27(5), 583-604.

Phelps, N. A., Fallon, R. J., & Williams, C. L. (2001). Small firms, borrowed size and the urban-rural shift. Regional studies, 35(7), 613-624.

Porter, C., Atkinson, P., & Gregory, I. N. (2018). Space and time in 100 million words: health and disease in a nineteenth-century newspaper. International Journal of Humanities and Arts Computing, 12(2), 196-216.

Potts, A., Bednarek, M., & Caple, H. (2015). How can computer-based methods help researchers to investigate news values in large datasets? A corpus linguistic study of the construction of newsworthiness in the reporting on Hurricane Katrina. Discourse & Communication, 9(2), 149-172.

Rutten, R. (2017). Beyond proximities: The socio-spatial dynamics of knowledge creation. Progress in Human Geography, 41(2), 159-177.

Salvini, M. M., & Fabrikant, S. I. (2016). Spatialization of user-generated content to uncover the multirelational world city network. Environment and Planning B: Planning and Design, 43(1), 228-248.

Song, Y., Long, Y., Wu, P., & Wang, X. (2018). Are all cities with similar urban form or not? Redefining cities with ubiquitous points of interest and evaluating them with indicators at city and block levels in China. International Journal of Geographical Information Science, 32(12), 2447-2476.

Taylor, P. J. (2004) World City Network: A Global Urban Analysis. Routledge.

Taylor, P. J. (2014). A Research Odyssey: from Interlocking Network Model to Extraordinary Cities. Tijdschrift voor economische en sociale geografie, 105(4), 387-397.

Taylor, P. J., & Derudder, B. (2022). Cities in Castells' Theorising of Social Space. Tijdschrift voor economische en sociale geografie, 113(3), 250-256.

Taylor, P. J., Ni, P., Derudder, B., Hoyler, M., Huang, J., & Witlox, F. (2012). Global urban analysis: A survey of cities in globalization. Routledge.

Wang, M. H., Schrock, S. D., Vander Broek, N., & Mulinazzi, T. (2013). Estimating dynamic origin-destination data and travel demand using cell phone network data. International Journal of Intelligent Transportation Systems Research, 11, 76-86.

Wang, Y., Zhang, W., Zhang, F., Yin, L., Zhang, J., Tian, C., & Jiang, W. (2020, July). Analysis of subway passenger flow based on smart card data. In 2020 6th International Conference on Big Data Computing and Communications (BIGCOM) (pp. 198-202). IEEE.

Werker, C., Korzinov, V., & Cunningham, S. (2019). Formation and output of collaborations: the role of proximity in German nanotechnology. Journal of Evolutionary Economics, 29, 697-719.

Wu, F. (2018). Planning centrality, market instruments: Governing Chinese urban transformation under state entrepreneurialism. Urban studies, 55(7), 1383-1399.

Wu, F., & Zhang, J. (2007). Planning the competitive city-region: The emergence of strategic development plan in China. Urban Affairs Review, 42(5), 714-740.

Wu, J., Feng, Z., Zhang, X., Xu, Y., & Peng, J. (2020). Delineating urban hinterland boundaries in the Pearl River Delta: An approach integrating toponym co-occurrence with field strength model. Cities, 96, 102457.

Yang, C., Xiao, M., Ding, X., Tian, W., Zhai, Y., Chen, J., Liu, L., & Ye, X. (2019). Exploring human mobility patterns using geo-tagged social media data at the group level. Journal of Spatial Science, 64(2), 221-238.

Chapter 2

Quantification of intercity relationships between 293 Chinese cities through toponym co-occurrence

This chapter is published as: Tongjing, W., Yin, Z., Bao, Z., & Meijers, E. (2024). Intercity relationships between 293 Chinese cities quantified based on toponym co-occurrence. Cybergeo: European Journal of Geography, No. 1057

Abstract: This paper presents relationships between 293 Chinese cities, derived using a toponym co-occurrence method. By employing this toponym co-occurrence analysis method, the strength of an intercity relationship is determined by the frequency at which both city names appear on the same webpage. The data was sourced from the Common Crawl web archive's 2019 April Corpus, which contains approximately 2.5 billion web pages. The primary aim of this dataset is to provide a fresh perspective on intercity relationships, thereby facilitating studies on city network analysis. The dataset not only encourages further research into comparing this innovative city relationship with other established networks but is also a showcase that presents a straightforward methodology that can be applied to other archives within Common Crawl. As such, it paves the way for longitudinal studies that probe the evolution of city networks.

Keywords: City networks, toponym co-occurrence, city relationship, geographical information retrieval

1. Introduction

Cities maintain varying degrees of relationships with a range of other cities (McCann and Acs, 2011; Pumain, 2021). These relationships shape the urban network system, offering opportunities for synergistic urban development (Meijers and Burger, 2017; Glaeser et al., 2016). The introduction of terms such as ‘regional externalities’ (Parr, 2002), ‘borrowed size’ (Meijers and Burger, 2017; Phelps et al., 2001), and ‘urban network externalities’ (Capello, 2000) underlines the external socioeconomic benefits that can be derived from these city relationships.

Previous research has acknowledged the role of intercity relationships in improving the urban system efficacy (Van den Berghe et al., 2022; Derudder et al., 2022) and in informing sound regional development policy (Meijers and Burger, 2010). However, many studies predominantly rely on mobility-related proxies, such as trains, MetroCards, and mobile signals to represent the strength between cities, or focus solely on large cities, due to constraints in data comprehension and accessibility. While there are some exceptions (e.g. Zhang et al., 2016; Pan et al., 2020), gathering data on diverse forms of city relationships, particularly on a large scale, remains a significant challenge.

Text mining methods offer a new feasible approach to address these limitations by extracting intercity relationships from the text. An effective and straightforward text mining method is collocation analysis (Mello, 2002), or so-called toponym co-occurrence in geography. This method suggests that the strength of the relationship between cities is associated with the co-appearance of city names in text. In other words, the more frequently two cities are mentioned together in the text, the more strongly they are related. This method has been employed at the scale of the USA, China, and Europe (Fize et al., 2016; Meijers and Peris, 2018; Liu et al., 2014), using a variety of text corpora, such as Wikipedia, newspapers, search engines and archives (Overell and Rüger, 2008; Zhong et al., 2017; Peris et al., 2018), which all show that the relationship strength mirrors reality to some extent.

Compared to machine learning-based text methods, which classify text with millions of features, this word collocation-based method focuses only on one textual feature: the co-appearance frequency of city names. This simplicity allows for straightforward interpretation of results and avoids reliability issues inherent in "black box" machine learning methods (De Graaf and van der Vossen, 2013). Additionally, collocation analysis allows for easy calibration of parameters, making it well-suited for unstructured text, such as webpages and social media texts.

While the Internet offers a vast, ever-growing source of unstructured text, an ideal source for collocation analysis, processing such a large text source can be difficult in practice. The immense size of digital text corpora, often in the Gigabyte (GB) or even Terabyte (TB) range, makes the use of conventional approaches challenging and necessitates mastery of cloud and distributed computing techniques.

The objective of this paper is twofold. First, it aims to develop an efficient and user-friendly method that is able to extract collocation analysis results from a large text corpus, thereby detailing the steps to be taken. Second, it aims to offer a ready-to-use dataset presenting how related cities are (broadly defined), thereby offering a valuable resource for researchers interested in urban systems.

The empirical focus will be on China, where access to large-scale data on city relationships remains limited. Therefore, our study aims to fill this gap

by providing a methodological approach and dataset that can serve as valuable tools in understanding the complex pattern of relationships between Chinese cities.

The remainder of this paper is organized as follows. Section 2 reviews the relevant literature on intercity relationships and text-mining methods, establishing the context for our study. Section 3 describes the methodology employed for corpus extraction and collocation analysis. Section 4 presents the results of our analysis, highlighting key patterns in intercity relationships among Chinese cities. Finally, Section 5 draws the conclusion of this paper.

2. Background

Collocation analysis was first proposed by Firth (1957), who used the word collocation pattern to identify common language applications. Initially, linguists applied this method focused on grammatical and lexical usage (Halliday, 1966; Firth, 1968; Greenbaum, 1970). Subsequent research extended the application of collocation analysis to include actual content relationships between collocated words, as later research discovered that keywords could also signify the "aboutness" of text (Scott, 1999), and the connection between words may arise from the content of the text itself (Tognini-Bonelli, 2001) or the author's selectional preference (McEnery, 2006).

Just considering the co-occurrence of words often faces criticism for its apparent simplicity, arguing that collocation may overlook the actual text content of each word co-occurence. While it is a valid point, Baker et al(2008) advocate an interpretative, rather than merely descriptive, approach to collocation analysis. This is because effective communication does not necessarily requires a context more complex than what is essential for interpretation (Brown and Yule, 1982; Tognini-Bonelli, 2001). As such, collocation analysis helps in "*limiting the interpretation to what is contextually appropriate or plausible*" (Brown and Yule, 1982, p59), and it does, in practice, enable the systematic analysis of large-scale texts (Baker, 2006). However, to better understand why some words collocate more frequently than others (Brezina, 2018), in-depth reading tools like Critical Discourse Analysis are often used to provide a deeper examination of each collocation occurrence (Baker and Vessey, 2018).

Geographers have also recognized the potential of collocation analysis early on. Tobler and Wineburg (1971) were pioneers in applying the collocation analysis to geography. They counted the co-occurrence frequency of placenames in the cuneiform tablets of the pre-Hittite Cappadocian towns and, inspired by what would become known as Tobler's law (Tobler, 1970, p236) --"everything is related to everything else, but close things are more related than distant things", they reconstructed the settlement system based on those co-occurrences.

As an empirical method, subsequent research has effectively applied collocation analysis in many different cases, revealing that place name co-occurrence can, to some degree, mirror the strength of connections between places (Devriendt et al., 2011; Meijers and Peris, 2019). This approach aligns with Kitchin's assertion about big data analysis that "*instead of solely testing a theory by analysing relevant data, big data analytics seeks to uncover insights born from the data itself*" (Kitchin, 2014, p2). Therefore, collocation analysis is viewed as a practical tool capable of quantifying the strength of relationships between cities as represented in the text.

Early applications of collocation analysis on geography focused on clarifying toponym ambiguities based on placename co-occurrence on Wikipedia (Overell and Rüger, 2008). More recent work has also tended to explore settlement systems, in particular how cities relate to each other through interactions between people, firms, and institutions through digital newspapers, search engines, Wikipedia, and web archives (Devriendt et al., 2011; Liu et al., 2014; Peris et al., 2020; Tongjing et al., 2023a). These studies generally indicate that toponym co-occurrence networks do reflect real-world spatial interactions to a certain extent. An interesting comparison of toponym co-occurrence data with transportation data was conducted by Lin et al. (2019) for cities in Guangdong, a province in China. They analyzed the toponym co-occurrence network generated by Sina, a Chinese news media outlet, and compared it with bus and rail service networks between these cities. Their results show a strong correlation between the overall connectivity of a city as represented by the toponym co-occurrence network and its connectivity as evidenced by the bus and rail service networks. Another interesting way of using toponym co-occurrence data is to incorporate the collocation results into existing models to increase their accuracy. For instance, Wu et al. (2019) incorporated toponym co-occurrence results to refine the field strength

model, thereby improving the identification of a metropolitan area's hinterland.

3. Data collection and processing

The primary dataset of this study was obtained from the Common Crawl, a web archive that has periodically crawled the Internet since 2008. Starting in 2017, the archive began crawling the Internet on a monthly basis, providing an up-to-date and longitudinal resource. As of 2019, more than 40 TB have been crawled monthly using the Web ARChive (WARC) format and stored on Amazon S3. We used the entire Common Crawl text archive from April 2019 for processing and conducting experiments (Common Crawl, 2022). This corpus contains approximately 6.98 TB of uncompressed text, comprising 2.5 billion web pages crawled between April 18th and 26th.

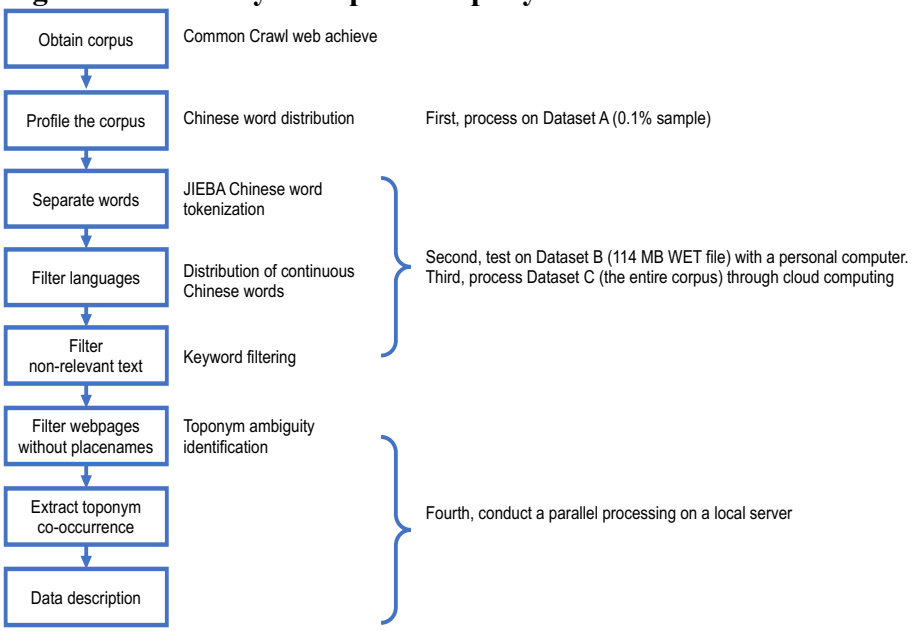
Processing the entire corpus can be time-consuming and costly. Therefore, we initially used a random 0.1% sample (Dataset A) of the corpus to examine the distribution of the share of Chinese tokens in each URL. Next, we downloaded a random subset of the corpus (44,018 URLs, 114 MB WET file, Dataset B) for manual inspection and local processing. Subsequently, we used the entire Common Crawl archive of April 2019 (Dataset C) for processing and conducting experiments using Amazon AWS cloud computing to obtain the filtered corpus. Lastly, we extracted the co-occurrence results from the filtered corpus from a local server. Our processing steps are summarized in Figure 1.

Clear Chinese word separation was required for further processing. Chinese words often consist of multiple Chinese characters, but a Chinese sentence is composed of consecutive characters without any clear separation. Accurate separation of Chinese characters requires Natural Language Processing techniques. Here we used a popular term frequency-inverse document frequency (TF-IDF) module named JIEBA for word separation (Sun, 2019). This module can separate Chinese sentences into tokens, which can be loosely considered as words. A token is a term in Natural Language Processing that represents a group of characters in the text as a useful semantic unit for processing.

Since our studies focused on Chinese cities, we opted to filter out non-Chinese webpages to reduce the total amount of data to process. However,

Chinese content is often mixed with English or other languages. We defined Chinese webpages as those in which Chinese tokens constitute a significant proportion.

Figure 1. Summary of step-wise toponym co-occurrence extraction



After using JIEBA to separate Chinese texts into tokens, we examined the distribution of Chinese tokens in the sample webpage data (Dataset A). As shown in Figure 2, we found two peaks. The left peak indicates a significant number of webpages with few Chinese tokens, while the right peak represents webpages with a majority of Chinese content. We included all webpages in the right peak, where the Chinese token ratio was 55% and higher. This approach considerably reduced the workload for subsequent processing while retaining the most relevant webpages.

Nonetheless, filtering the entire corpus based on each URL's Chinese token ratio still proved to be an insurmountable task. We decided to use an alternative approach to determine whether a webpage should be included or filtered out. We discovered that the Chinese token ratio is strongly related to the number of continuous Chinese tokens. Figure 3 displays the relationship between the Chinese token ratio with the longest Chinese tokens.

Figure 2. Chinese Token Ratio Distribution

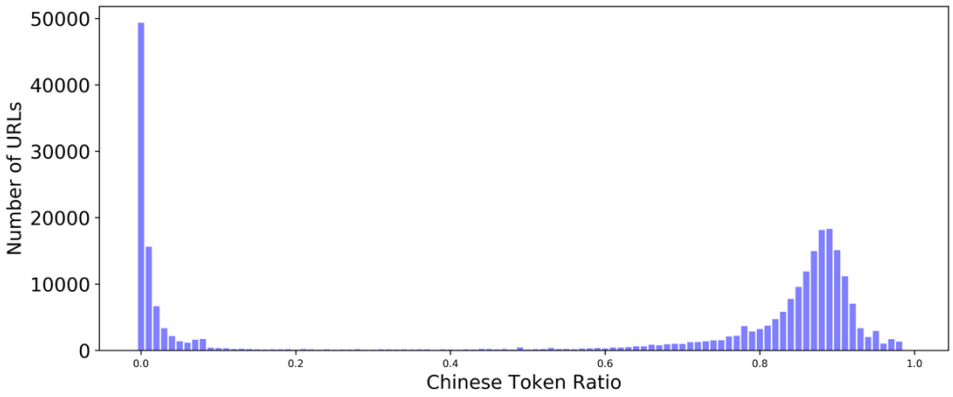
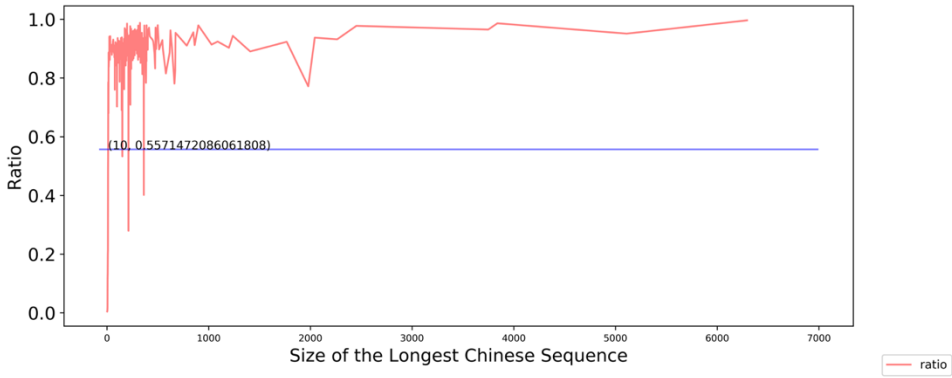


Figure 3. Relationship between the Chinese token ratio and the number of continuing Chinese tokens



In Figure 3, the red line illustrates the relationship between the Chinese token percentage on the webpage and the longest Chinese sequence. The blue line demonstrates that by setting the threshold at 10 continuous Chinese tokens, we can select almost the same number of URLs with Chinese token ratios higher than 55% as possible, while only a few URLs will have ratios lower than 55%. Choosing values lower than 10 would include more URLs with ratios lower than 55% (and fewer URLs with ratios above 55%). We further verified that the URLs filtered by 10 continuing Chinese tokens and 55% ratio were comparable and 96% of them being the same. This filtering process is much more efficient than calculating Chinese token ratios, as it does not require traversing all content of each URL, and results in very similar URLs being selected. We performed this filtering process on a random file, which is a 114 MB WET file (Dataset B). After filtering out non-Chinese languages, we obtained a

corpus of 8.3 Megabytes (MB) WET file, which is about 7.28% of the original file.

Upon examining the documents in the filtered Dataset B, we discovered that pornographic and gambling content made up a significant proportion of the total documents, biasing our data set and diminishing the importance of more relevant texts. As the objective of this paper is also to develop a relevant text corpus in an affordable way, striking a balance between accuracy and efficiency is necessary. To achieve high processing efficiency, we employed only simple processing methods, excluding machine learning-based approaches. Here, we used keyword filtering due to its high efficiency.

While keyword filtering can be subjective, it simplifies the censoring process and makes it easier to adapt to other circumstances. In general, it strikes a balance between accuracy and efficiency. To remove pornographic and gambling webpages, we used a dictionary with over 10,000 stars on GitHub that contains over 1000 sexual and 68 gambling words (GitHub, 2022). By using this keyword filtering dictionary, the 8.3 MB corpus was filtered down to a 3.8 MB size WET file.

To complete the entire task, we used a highly scalable Hadoop-based framework for processing the full Common Crawl corpus, utilizing a 1,080 CPU cluster on the Amazon Elastic Map/Reduce infrastructure. To lower costs, we employed AWS spot instances, which were only about 10% of the regular on-demand price. The price of spot instances varies depending on the AWS region and day of the week. During our study, we launched a cluster of 1 master and 29 spot instances of c4.8 xlarge nodes (36 CPUs each). The total operating time was about 58 minutes and the total cost was approximately $30. The resulting output corpus was about 202 GB.

As processing on a local server was more convenient than on a cloud server, we downloaded the whole filtered corpus (202 GB) from AWS for subsequent processing. Later processing was conducted on a local server.

The next step was to filter out webpages that did not contain placenames. Chinese placenames can be ambiguous. The co-occurrence of place names may be inaccurate due to toponym ambiguity issues, which means a place name may also refer to other entities. Wacholder et al. (1997) identified

multiple types of toponym ambiguities, but in Chinese, the most common toponym ambiguity falls into three categories:

- Structural: E.g., Beihai, is a city in Guangxi Province, but also means the "north of the sea".
- Semantic: E.g., Hezuo, is a city in Gansu Province, but also means "cooperation".
- Referent: E.g., Chaoyang, is a city in Liaoning Province, but is also a district name in Beijing.

To precisely disambiguate toponyms, it is essential to take the context into account, which, however, would require a lot of processing and is time-consuming. Due to the sheer size of the dataset, we decided to exclude these potentially ambiguous cities. Fortunately, most of the toponym ambiguity occurs at the low administration level (e.g. county, village, and district level). Conversely, cities at the prefecture administrative level or higher generally exhibit fewer risks of toponym ambiguity.

To reduce the risk of toponym ambiguity, we did a two-step process. Initially, we conducted a manual investigation to identify city names that were evidently ambiguous. City names such as "Baiyin" (translating to "silver"), "Dachang" ("big factory"), and "Dongfang" ("East") were systematically delisted to enhance the precision of the dataset.

Subsequently, a gravity model was used to conduct a comparative analysis between the absolute frequencies of co-occurrences and the predicted frequencies as calculated by the model. The gravity model can be formalized as follows:

$$I_{ij} = K\frac{M_i^{\beta_1} M_j^{\beta_2}}{D_{ij}^{\beta_3}}$$

where I_{ij} is the total relationship, K is the constant, M_i and M_j are the sizes of place i and j, respectively, D_{ij} is the physical distance between the two places, β_1 and β_2 reflects the ability of place i and j to attract flows, β_3 reflecting the rate of increase in the friction of distance.

We then fit the gravity model to the toponym co-occurrence results in order to obtain an estimated strength of intercity relationships. This established a benchmark against which the empirical co-occurrence data could be evaluated. Subsequent analyses focused on outliers—cases where the

observed relationships between cities are much higher than the model's predictions.

Such outliers were examined to verify whether their high estimation was attributable to naming ambiguities. For example, strong relationships were initially observed between "Beijing" and "Chaoyang," and "Shanghai" and "Siping"; however, upon further examination, it was found that "Chaoyang" and "Siping", two small cities, are also district names in Beijing and Shanghai, respectively, thereby clarifying the source of the apparent anomaly.

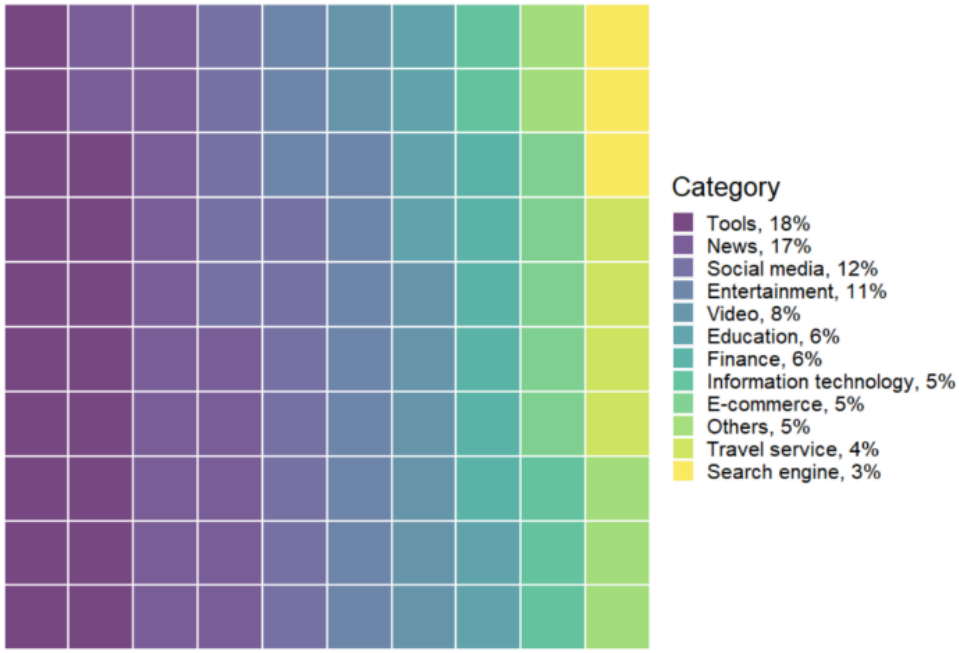
After checking the risk of toponym ambiguity, we selected 293 cities.

The webpages in this corpus that did not contain Chinese characters or the selected Chinese placenames were excluded. The size of the Chinese corpus was then further reduced from 202 GB to 139.5GB of text, which is about 69% of the filtered Chinses corpus. This confirms Hill's estimate that 70% of documents contain geographical information (Hill, 2009). The filtered Chinese corpus contains approximately 91 million pages from 1,067 top domains and 1,792,759 domain names. In total, the corpus contains around 110 billion tokens, as calculated by the JIEBA tokenizer.

On average, in the filtered corpus, the top 100 websites ranked by the Chinaz website have about 290 million tokens and 234,691 pages. The total number of tokens in the top 100 websites is about 26.4% of the filtered corpus. To gain a basic understanding of the content in this corpus, we classified the content distribution of these websites. This categorization was according to the primary content areas listed on the Chinaz website for each of these top 100 sites. The percentage distribution of Chinese website types is shown in Figure 4.

Lastly, we tallied the number of webpages where two placenames co-occur and saved the results as a toponym co-occurrence network dataset. The number of webpages where two placenames co-occur serves as a measure of the strength of the relationship between the respective cities.

Figure 4. The distribution of content on popular Chinese websites



4. Results

In this study, the relationship strength between cities was represented by the frequency of toponym co-occurrence. To better understand this data, we approached it from a network perspective, where each city is considered a node and the frequency serves as the link weight between two nodes. Table 1 provides an overview of the toponym co-occurrence network dataset.

Given the 293 cities in our analysis, there are 42,778 city pairs among which relationships are measured, and we found- toponym co-occurrences for all these pairs, which means it is a full graph.

Table 1 shows that the maximum frequency (7,391,725; Beijing - Shanghai) is more than thirty-two times larger than the third quartile (230,970), while the difference between the third (230,970) and the first quartile (175,844) is relatively small in comparison. The mean frequency (2,272,417) is the average co-occurrence frequency of the 42,778 relationships and is substantially larger than the median frequency (198,880). This substantial difference between the mean and median suggests the presence of a

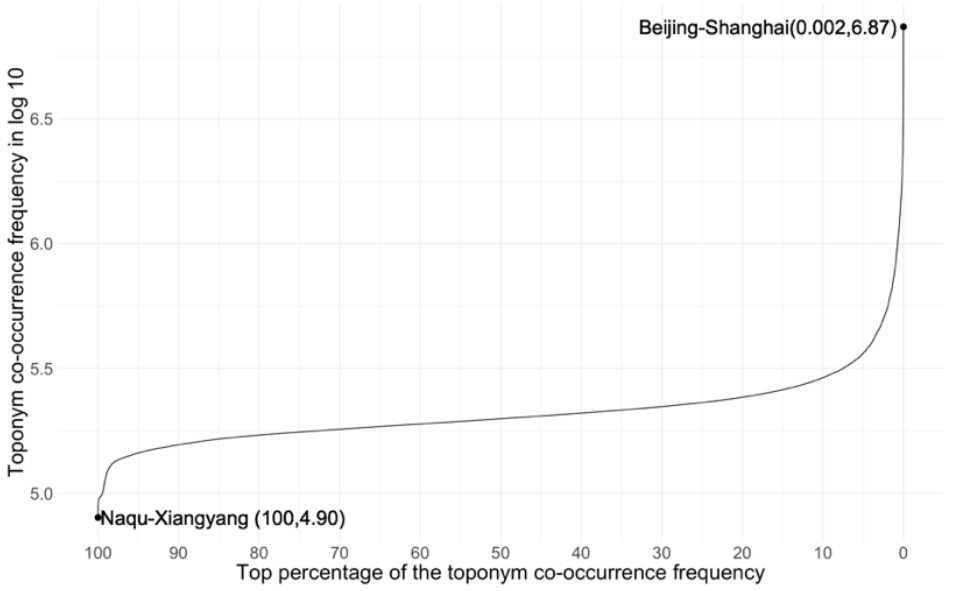
skewed distribution with extreme values. In other words, some city pairs are mentioned far more frequently than others.

Table 1. Statistics of the intercity relationships

Type of information	Value
Number of cities	293
Number of city pairs that co-occur	42,778
Mean frequency	2,272,417
Standard deviation	15,877
Minimum frequency	79,985
First quartile	175,844
Medium	198,880
Third quartile	230,970
Maximum frequency	7,391,725

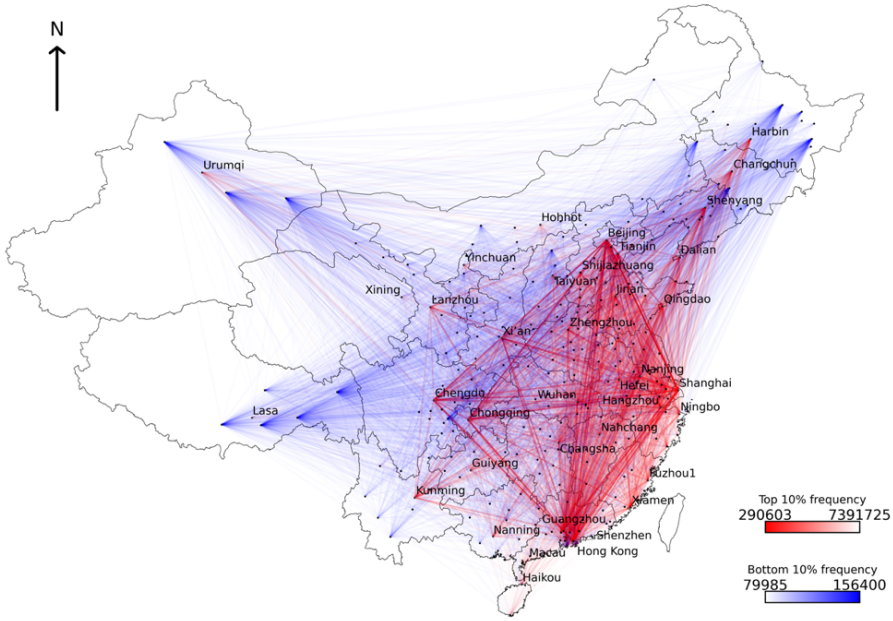
More detailed results can be observed in the distribution of the toponym co-occurrence frequency. Figure 5 presents the distribution of the frequency values (log 10) as the variation between toponym frequencies is quite significant—90% of the frequency values are between 100,000 and 300,000, while the top 10% extends from 300,000 to 7,391,725. Most of these high values represent relationships between cities with a high administrative level.

Figure 5. The distribution of the toponym co-occurrence frequency



Given that the majority of co-occurrence frequency is concentrated within a relatively narrow span, Figure 6 offers a more focused visualization by highlighting the top 10% and bottom 10% of results. This Figure shows that high frequency relationships are primarily situated in the central and southeastern regions of China. Conversely, the low frequency relationships are located in the northern and western areas.

Figure 6. The co-occurrence frequency map of China



To complement Figure 6, Table 2 lists the top 10 intercity relationships with the highest co-occurrence frequency. This highlights that the dominant relationships in China are between Beijing, Shanghai, Chongqing, Tianjin—the four direct-administered municipalities, which are at the highest administrative level, and Guangzhou and Shenzhen, which are among the most prosperous cities following the four direct-administered municipalities.

Table 3 lists the 10 intercity relationships with the lowest frequency, featuring more dispersed relationships of 12 cities compared to the 6 cities in Table 2. These cities include Naqu—a small city in Tibet—and Macau and Hong Kong, both special administrative regions in China, each appearing in more than one low-frequency relationship. Other cities in this

list are Xiangyang, Xining, Liaoyuan, Wuzhong, Qitaihe, Mayannur, Hegang, and Lincang, predominantly small cities located in northern China.

Table 2. Top 10 intercity relationships with the highest highest co-occurrence frequency

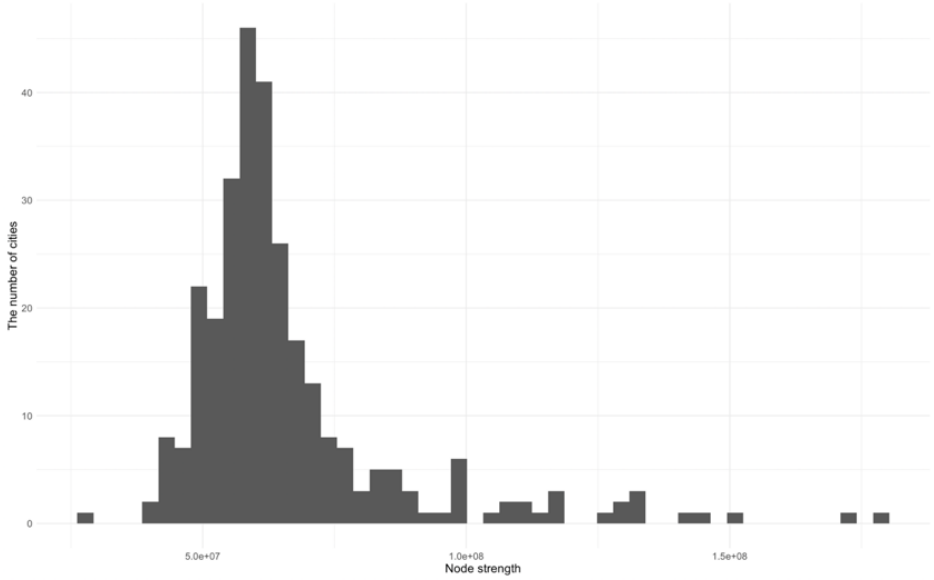
City1	City2	weight
Beijing	Shanghai	7,391,725
Beijing	Chongqing	4,617,074
Chongqing	Shanghai	4,355,687
Beijing	Tianjin	4,333,913
Beijing	Guangzhou	4,292,093
Beijing	Shenzhen	4,286,888
Shanghai	Shenzhen	4,251,592
Guangzhou	Shanghai	4,157,057
Shanghai	Tianjin	4,035,094
Guangzhou	Shenzhen	3,525,924

Table 3. 10 intercity relationships with the lowest frequency

City1	City2	weight
Naqu	Xiangyang	79,985
Macau	Naqu	81,756
Naqu	Xining	82,289
Liaoyuan	Macau	87,646
Naqu	Wuzhong	88,299
Macau	Qitaihe	88,606
Bayannur	Naqu	89,854
Hong Kong	Wuzhong	90,031
Hegang	Naqu	90,526
Hong Kong	Lincang	90,876

Comparable extreme values can also be observed when analyzing individual cities. We computed the total number of co-occurrence for each city, thus taking into account its relations with all other cities, a measure we refer to as node strength. This measure reflects the extent of each city's connectivity with other cities. Figure 7 presents the node strength distribution of the general toponym co-occurrence network. The findings suggest that most cities have a total co-occurrence frequency of approximately 60 million times, while only a few cities have been mentioned more than 150 million times.

Figure 7. Node strength distribution of the general intercity relationships



The top 10 cities with the highest overall co-occurrence frequency are presented in Table 4, where Beijing has its dominant presence in the network with the highest co-occurrence frequency of more than 160 million. This is followed by Shanghai, Shenzhen, Guangzhou, and Chongqing with frequencies over 100 million, highlighting their vital interconnectedness in this network. Further down the list are Chengdu, Tianjin, Hangzhou, Wuhan, and Nanjing, whose overall co-occurrence frequency is close to 100 million. The central role of these cities, as highlighted by the data, possibly reflects their economic, cultural, or political significance in the broader scenario under study. A more detailed analysis of how a city's position in the network associates with its performance can be found in the referenced paper (Tongjing et al., 2023b).

Table 4. The top 10 cities with the highest total co-occurrence frequency

Node	Total co-occurrence frequency
Beijing	160,781,860
Shanghai	146,879,869
Shenzhen	127,290,672
Guangzhou	112,619,991
Chongqing	107,572,587
Chengdu	99,832,245
Tianjin	98,143,611
Hangzhou	97,054,991
Wuhan	94,684,226
Nanjing	92,935,422

5. Conclusion

While it is widely recognized that intercity relationships play a significant role in the urban system, gathering such relational data remains challenging, and current approaches are often skewed toward accessibility or transportation. The collocation analysis method has potential to provide an alternative solution, as it facilitates the retrieval of a new type of large-scale relational data. Nonetheless, urban researchers using the collocation analysis method often encounter difficulties in processing terabyte-scale data, particularly when it necessitates the use of cloud computing. This study showcases an efficient and relatively straightforward approach to extracting intercity relationships from the vast amounts of unstructured text stored in the Common Crawl web archive.

From an empirical perspective, this study demonstrates how to extract pertinent information via cloud computing, whilst keeping costs to a minimum. The final operation was completed in less than an hour, costing under $100. By deploying a 1080 CPU cluster on the Amazon Elastic Map/Reduce infrastructure, a 202 GB single-language corpus was

successfully isolated from a 6.98 TB dataset via a cloud parallel computing method on AWS.

The toponym co-occurrence method presents several advantages, both practical and theoretical. On the practical front, this method allows for collecting large-scale city relational data that can be difficult to obtain from location data sources, such as mobile signals and metrocards in and out. This makes it a particularly useful method for researchers and policymakers interested in comprehensive, wide-ranging analyses. Theoretically, the method captures a new dimension of city relationships that is distinct from what is revealed through interaction (e.g. transportation) data. While transportation data primarily focuses on the tangible, material movements between cities—such as the flow of goods, services, or people—the toponym co-occurrence method sheds also light on the symbolic or representational associations between cities. These could include cultural, historical, or social linkages that are not necessarily tied to physical movement but are nonetheless significant in understanding the multifaceted relationships between cities. As such, 'relations' is a more broad concept than 'interactions'.

By extracting relationships from this large text corpus, we ensure a comprehensive and generalized overview of inter-city relationship patterns. However, we acknowledge that our dataset offers merely a temporal snapshot, subject to potential variations when extended to different time frames, languages, and specific genres of text. Also note that the Common Crawl archive is sample data – when taking two consecutive monthly data dumps, about 20% of the urls is similar (CommonCrawl, 2023). These potential variations should not be considered as limitations but opportunities. Such variations underscore the dynamic and evolutionary nature of inter-city relationships in text and pave the way for further investigations in this field.

While we are optimistic about the potential of the collocation analysis method in general, and more specifically the toponym co-occurrence method in revealing city networks, we acknowledge that the approach is still in its infancy. As such, by presenting our method and dataset, we aim to encourage further engagement on this topic. We have already implicitly suggested several useful future research avenues, including comparisons with interaction data, and an analysis through time. The latter is perfectly possible utilizing the same text source at multiple points in time. Very

promising in this regard is the increasing availability of digitalized historic text corpora. Additionally, we call for efforts to classify the types of relationships extracted through collocation analysis, as well as explorations of its wider applicability. Our approach also allows for comparative analyses across cultures or linguistic communities, to see how the perception of relations between cities differs. Finally, while our dataset covers the whole of China, it is obviously also possible to explore regional urban systems using a subset of the cities included.

Dataset description

This dataset includes two tables in CSV format. The first one presents toponym co-occurrence between 293 Chinese cities. It is structured into three columns, the first two columns list the Chinese city names in English, and the third column is the co-occurrence between each city pair. The co-occurrence relationship is non-directional. The second table provides the corresponding Chinese names for the cities listed in the first table. The co-occurrence was extracted from the Common Crawl web archive's 2019 April Corpus, which contains approximately 2.5 billion web pages.

Related Publication:

Tongjing, W., Meijers, E., Bao, Z., & Wang, H. (2023b). Intercity networks and urban performance: a geographical text mining approach. International Journal of Urban Sciences, 1-22.

Language:
English and Chinese

Time Period Covered:
2019 April

Kind of data:
Text tables

Data Source:
Common Crawl, https://commoncrawl.org/the-data/get-started/ (accessed 4 June 2023)

Geographical Coverage:

293 cities in Mainland China

Geographical Unit:
Cities at the prefecture administration level or above

Type of article:
Data paper

Repository location:
Cybergeo Dataverse warehouse
https://doi.org/10.7910/DVN/O0M59W

References

Antaki C., Billig M., Edwards D., Potter J., 2003, "Discourse Analysis Means Doing Analysis: A Critique of Six Analytic Shortcomings", Discourse Analysis Online, Vol.1, No.1. [Online] Available at:
https://repository.lboro.ac.uk/articles/journal_contribution/Discourse_analysis_means_doing_analysis_a_critique_of_six_analytic_shortcomings/9473747

Baker P., Gabrielatos C., Khosravinik M., Krzyżanowski M., McEnery T., Wodak R., 2008, "A Useful Methodological Synergy? Combining Critical Discourse Analysis and Corpus Linguistics to Examine Discourses of Refugees and Asylum Seekers in the UK Press", Discourse & Society, Vol.19, No.3, 273-306.

Baker P., Vessey R., 2018, "A Corpus-Driven Comparison of English and French Islamist Extremist Texts", International Journal of Corpus Linguistics, Vol.23, No.3, 255-278.

Baroni B., Bernardini S., 2004, "BootCaT: Bootstrapping Corpora and Terms from the Web", in Proc. 4th Int. Conf. on Language Resources and Evaluation, Lisbon.

Baroni M., Bernardini S., Ferraresi A., Zanchetta E., 2009, "The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora", Language Resources and Evaluation, Vol.43, 209-226.

Brezina V., 2018, "Statistical Choices in Corpus-Based Discourse Analysis", in Taylor C., Marchi A. (eds) Corpus Approaches to

Discourse: A Critical Review, London and New York: Routledge, 259-280.
Brookes G., Baker P., 2021, "Obesity in the News: Language and Representation in the Press", Cambridge University Press.
Campelo C. E. C., de Souza Baptista C., 2009, "A Model for Geographic Knowledge Extraction on Web Documents", Advances in Conceptual Modeling-Challenging Perspectives: ER 2009 Workshops CoMoL, ETheCoM, FP-UML, MOST-ONISW, QoIS, RIGiM, SeCoGIS, Gramado, Brazil, November 9-12, 2009. Proceedings 28, Springer Berlin Heidelberg, 317-326.
Common Crawl, 2023a, https://commoncrawl.org/the-data/get-started/ Accessed June 15, 2023.
Common Crawl, 2023b, https://commoncrawl.github.io/cc-crawl-statistics/plots/crawloverlap, Accessed November 7, 2023.
Derudder B., Feng X., Shen W., Shao R., Taylor P. J., 2022, "Connections between Asian and European World Cities: Measurement, Analysis, and Evaluation", Land, Vol.11, No.9, 1574.
Devriendt L., Boulton A., Brunn S., Derudder B., Witlox F., 2011, "Searching for Cyberspace: The Position of Major Cities in the Information Age", Journal of Urban Technology, Vol.18, No.1, 73-92.
Ediger D., Jiang K., Riedy J., Bader D. A., Corley C., Farber R., Reynolds W. N., September 2010, "Massive Social Network Analysis: Mining Twitter for Social Good", 2010 39th International Conference on Parallel Processing, IEEE, 583-593.
Egbert J., Wizner S., Keller D., Biber D., McEnery T., Baker P., 2021, "Identifying and Describing Functional Discourse Units in the BNC Spoken 2014", Text & Talk, Vol.41, Nos.5-6, 715-737.
Firth J., 1957, "A Synopsis of Linguistic Theory, 1930-1955", Studies in Linguistic Analysis, 10-32.
Firth J., 1968, "Selected Papers of J.R. Firth 1952–1959", (F. R. Palmer, Ed.), London: Longmans.
Fize J., Moncla L., Martins B., 2021, "Deep Learning for Toponym Resolution: Geocoding Based on Pairs of Toponyms", ISPRS International Journal of Geo-Information, Vol.10, No.12, 818
Glaeser E. L., Ponzetto G. A., Zou Y., 2016, "Urban Networks: Connecting Markets, People, and Ideas", Papers in Regional Science, Vol.95, No.1, 17-59.
Greenbaum S., 1970, "Verb-Intensifier Collocations in English", Janua Linguarum, Series Minor 86, The Hague: Mouton.

Ghani R., Jones R., Mladenić D., 2001, "Mining the Web to Create Minority Language Corpora", Proceedings of the International Conference on Information and Knowledge Management, 279-286.

GitHub repository, 2022, "Jieba", https://github.com/fxsjy/jieba Accessed Sept 14, 2022.

Halliday M. A., 1966, "Lexis as a Linguistic Level", In Memory of JR Firth, 148-162.

Hill L. L., 2009, "Georeferencing: The Geographic Associations of Information", MIT Press.

Karami A., Lundy M., Webb F., Dwivedi Y. K., 2020, "Twitter and Research: A Systematic Literature Review Through Text Mining", IEEE Access, Vol.8, 67698-67717.

Lewis S. C., Westlund O., 2015, "Big Data and Journalism: Epistemology, Expertise, Economics, and Ethics", Digital Journalism, Vol.3, No.3, 447-466.

Lin J., Wu Z., Li X., 2019, "Measuring Inter-City Connectivity in an Urban Agglomeration Based on Multi-Source Data", International Journal of Geographical Information Science, Vol.33, No.5, 1062-1081.

Liu Y., Wang F., Kang C., Gao Y., Lu Y., 2014, "Analyzing Relatedness by Toponym Co-Occurrences on Web Pages", Transactions in GIS, Vol.18, No.1, 89-107.

McCann P., Acs Z. J., 2011, "Globalization: Countries, Cities and Multinationals", Regional Studies, Vol.45, No.1, 17-32.

McEnery T., 2004, "Swearing in English: Bad Language, Purity and Power from 1586 to the Present", Vol.1, Routledge.

Meijers E. J., Burger M. J., 2010, "Spatial Structure and Productivity in US Metropolitan Areas", Environment and Planning A, Vol.42, No.6, 1383-1402.

Meijers E. J., Burger M. J., 2017, "Stretching the Concept of 'Borrowed Size'", Urban Studies, Vol.54, No.1, 269-291.

Meijers E., Peris A., 2019, "Using Toponym Co-Occurrences to Measure Relationships Between Places: Review, Application and Evaluation", International Journal of Urban Sciences, Vol.23, No.2, 246-268.

Mello R. A., 2002, "Collocation Analysis: A Method for Conceptualizing and Understanding Narrative Data", Qualitative Research, Vol.2, No.2, 231-243.

Minaee S., Kalchbrenner N., Cambria E., Nikzad N., Chenaghlu M., Gao J., 2021, "Deep Learning-Based Text Classification: A Comprehensive Review", ACM Computing Surveys (CSUR), Vol.54, No.3, 1-40.

Pan F., Bi W., Liu X., Sigler T., 2020, "Exploring Financial Centre Networks Through Inter-Urban Collaboration in High-End Financial Transactions in China", Regional Studies, Vol.54, No.2, 162-172.

Parr J. B., 2002, "Agglomeration Economies: Ambiguities and Confusions", Environment and Planning A, Vol.34, No.4, 717-731.

Peris A., Meijers E., van Ham M., 2018, "The Evolution of the Systems of Cities Literature Since 1995: Schools of Thought and Their Interaction", Networks and Spatial Economics, Vol.18, 533-554.

Peris A., Faber W. J., Meijers E., Van Ham M., January 2020, "One Century of Information Diffusion in the Netherlands Derived from a Massive Digital Archive of Historical Newspapers: The DIGGER Dataset", Cybergeo: European Journal of Geography.

Phelps N. A., Fallon R. J., Williams C. L., 2001, "Small Firms, Borrowed Size and the Urban-Rural Shift", Regional Studies, Vol.35, No.7, 613-624.

Pumain D., 2021, "From Networks of Cities to Systems of Cities", in Z. Neal, & C. Rozenblat (Eds.), Handbook of Cities and Networks, 16–40, Cheltenham: Edward Elgar Publishing.

Scott M., 1999, "WordSmith Tools Help Manual", Version 3.0, Oxford, UK: Mike Scott and Oxford University Press.

Song Y., Shi S., Li J., Zhang H., June 2018, "Directional Skip-Gram: Explicitly Distinguishing Left and Right Context for Word Embeddings", Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 175-180.

Sun J., 2012, "Jieba Chinese Word Segmentation Tool", https://github.com/fxsjy/jieba Accessed Nov 13, 2022.

Tobler W. R., 1970, "A Computer Movie Simulating Urban Growth in the Detroit Region", Economic Geography, Vol.46, Sup1, 234-240.

Tobler W., Wineburg S., 1971, "A Cappadocian Speculation", Nature, Vol.231, No.5297, 39-41.

Tongjing W., Meijers E., Wang H., 2023a, "The Multiplex Relations Between Cities: A Lexicon-Based Approach to Detect Urban Systems", Regional Studies, Vol.57, No.8, 1592-1604.

Tongjing W., Meijers E.J., Bao Z., Wang H., 2023b, "Intercity networks and urban performance: a geographical text mining approach", International Journal of Urban Sciences, 1-22.

Tognini-Bonelli E., 2001, "Corpus Linguistics at Work", Amsterdam: Benjamins.

Van den Berghe K., Peris A., Meijers E., Jacobs W., 2023, "Friends with

Benefits: The Emergence of the Amsterdam–Rotterdam–Antwerp (ARA) Polycentric Port Region", Territory, Politics, Governance, Vol.11, No.2, 301-320.

Wacholder N., Ravin Y., Choi M., March 1997, "Disambiguation of Proper Names in Text", Fifth Conference on Applied Natural Language Processing, 202-208.

Wu J., Feng Z., Zhang X., Xu Y., Peng J., 2020, "Delineating Urban Hinterland Boundaries in the Pearl River Delta: An Approach Integrating Toponym Co-Occurrence with Field Strength Model", Cities, No.96, 102457.

Zhang W., Derudder B., Wang J., Shen W., Witlox F., 2016, "Using Location-Based Social Media to Chart the Patterns of People Moving Between Cities: The Case of Weibo-Users in the Yangtze River Delta", Journal of Urban Technology, Vol.23, No.3, 91-11

Chapter 3

The multiplex relations between cities: A lexicon-based approach

This chapter is published as: Tongjing, W., Meijers, E., & Wang, H. (2023). The multiplex relations between cities: a lexicon-based approach to detect urban systems. Regional Studies, 57(8), 1592-1604

Abstract: Cities relate to other cities in many ways and much scholarly effort goes into uncovering those relationships. Building on the principle that strongly related cities will co-occur frequently in texts, we propose a novel method to classify those toponym co-occurrences using a lexicon-based text mining method. Millions of webpages are analyzed to retrieve how 293 Chinese cities are related in terms of six types: industry, information technology (IT), finance, research, culture, and government. Each class displays different network patterns, and this multiplexity is mapped and analyzed. Further refinement of this lexicon-based approach can revolutionize the study of inter-urban relationships.

Keywords: City networks, Urban systems, Multiplexity, Toponym co-occurrence, Text mining

1. Introduction

People and firms interact in many ways with each other, and by no means are such interactions confined to city boundaries (e.g. Capello, 2000; Derudder and Taylor, 2018). At an aggregate level, cities are embedded in fine-grained systems of relationships, making them interdependent to such an extent that what happens in one place has an impact on people and firms in other connected cities. Over the last decades, five different schools of thought have emerged that try to make sense of such networks between cities (Peris et al., 2018).

A continuous, common challenge in this literature is understanding how such city networks can actually be measured, to then apply network analysis. As Neal (2012, p3) observed, "whatever the relationships are and whatever they connect, networks have a specific and observable content that can be studied". It is generally acknowledged that such city networks are multiplex in that they are comprised of "multidirectional flows of not only economic but also social, cultural, and environmental activities" (Davoudi, 2008, p.51). Each type of flow can be recognized as a layer of the network, which coexists and interacts with other layers, thereby creating network multiplexity, and making the system of cities function in a complex synergetic way.

A proper specification and measurement of network multiplexity can provide a more comprehensive profile of the complexity of our urban domain and can act as a guide for urban policy and practice (Derudder and

Neal, 2018). While most studies focus on just a single layer of city networks, studies that do recognize multiplexity (e.g. Berroir et al., 2017; Burger et al., 2014a; Hu et al., 2020) show that patterns of specific types of networks are often markedly different from other types. Burger et al. (2014b) show that judging the strength of functional coherence in a region depends highly on the scope and type of functional links upon which such an analysis rests. In other words: our understanding of networks between cities, whether on a global or regional scale, is very dependent upon the lens through which we assess these networks.

One could argue that to address multiplexity in city networks, it is a simple matter of combining data sets on particular flows of people and goods. And no doubt this is well possible for some countries or regions, using for example, (micro-)data from travel surveys or nowadays point of interest (POI) data. But such datasets generally cover only large cities and do not facilitate precise cross-border comparisons due to different standards used, and since standards change regularly, the time period that can be covered is mostly limited. More importantly, these data only refer to tangible exchange or physical movements in space (e.g. commuting is commonly studied), leaving out non-physical flows, specifically, information flows.

Information flows refer to the exchange of knowledge, ideas and sentiments expressed by people living in different cities. Their shared information also represents an essential type of city relationship as it reflects how individuals think, communicate, behave, and respond to people living in other cities. While information flows can be hardly captured by typical physical movements, it is often obtainable from texts, such as books, narratives, and interviews.

Information flows can be extracted by content analysis. Very promising with regards to retrieving and exploring city networks is the summative content analysis method, particularly the collocation analysis (Mello, 2002). It involves summarizing the co-appearance of specific words or content in texts: the more words or phrases co-occur, the more they are related. A straightforward approach is to search for toponym co-occurrences, and that is based on the assumption that places or regions that are more strongly related will co-occur more often in a text corpus (Callon et al., 1983; Devriendt et al., 2011; Tobler and Wineburg, 1971). While some of these co-occurrences will capture actual interactions, other toponym co-occurrences should be considered partly as ‘symbolic’, as the

collocation of words represents sometimes abstract and non-tangible interactions, which are then hard to interpret (Watts, 2004). Admittedly, any individual co-appearance of city names may be trivial, but the aggregated pattern holds great value as their accumulation can capture, generalize, and highlight the strength of hidden relatedness between cities.

Burger et al. (2014b) stated that an important criterion to evaluate methods for identifying city networks is whether they are capable of dealing with multiplexity. So far, the toponym co-occurrence approach has mainly focused on establishing the strength of intercity relationships, but hardly on identifying different types of the relationships or their meaning (Meijers and Peris, 2019, Watts, 2004). The exact nature of the co-occurrence of place names remains unclear, therefore. Salvini and Fabrikant (2016) and Hu et al. (2017) are interesting exceptions, but can rely on classifications (shared links or tags) present in the underlying text corpus. However, being reliant on pre-existing classifications leaves little flexibility, while such classifications are often absent in unstructured texts. It seems fair to assume that this toponym co-occurrence method will only gain prominence if it is able to substantiate what relationships between cities are about. So far, the initial attempts by Meijers and Peris (2019) to classify place name co-occurrences on web pages by using a supervised machine learning algorithm trained to identify pre-defined categories were not yet very accurate. Such machine learning-based methods may take a long time to reach the human ability to understand texts (Baeza-Yates et al., 2019). This paper aims to fill this gap by exploring how toponym co-occurrences in unstructured texts can be classified, thereby providing new insights in the multiplexity of urban systems.

The objective of this paper is to develop and apply a novel approach to classifying relationships between cities that are derived from toponym co-occurrences in texts. The ambition of implementing such an approach is to give further meaning to the idea that city networks are multiplex, or in other words, that the pattern of relations between cities varies according to the type of relationship that is being considered. This lexicon-based approach substantiates the normal toponym co-occurrence method by adding a classification method of the otherwise relatively abstract relationships. In essence, our approach involves collocation analysis: we pair the toponym co-occurrence with words that capture particular types of relationships. To give a simple example: if two city names co-occur with words like ‘match’, ‘scored’ and ‘goal’, such a city relationship could be

labeled as ‘sports’. This method therefore combines the deductive and summative content analysis methods, as it uses a predefined collection of words and phrases to represent the topic of each text content (deductive content analysis), and counts the occurrence of these predefined words and phrases to identify content patterns (summative content analysis). As such, this lexicon-based method builds on the keyness concept (Scott, 1999), which aims to identify the aboutness of the text using a set of semantically related words. The underlying reasoning is that the content of a text can be interpreted by the use of words whose frequencies have a statistically significant difference from those same words that are used in another context (Culpeper, 2009).

Empirically, we will focus on China. We are curious whether the unbalanced development, strongly weighted toward the coastal regions is associated with unequal representation in particular relationships (Li and Wu, 2012; Zhang and Peck, 2016). Another important reason is that China does not yet have the geo-relational data on a national level as commonly used in the western world. While much literatures have explored city relationships using trade data to commuting, to migration flows, to leisure/tourism flows, to goods being transported, telecommunication, few data on how cities relate to each other are actually publicly accessible. Data availability makes previous studies on the multiplexity of China either focused on a limited number of top-level cities at the national level (Lao et al., 2016) or only on a more detailed regional level, such as the Yangtze River Delta (Cao et al., 2021; Zhang et al., 2020). A comparative advantage of using the Chinese language is that it avoids some of the semantic ambiguation that results from the use of Latin letters by different Indo-European languages, as Chinese letters are used in the Chinese language only. Second, Chinese words lack plural or gender forms, so it is practically easier for searching words in text corpora.

Next to filling this empirical gap of investigating city network multiplexity at the Chinese national scale, this paper aims to have broader theoretical and methodological relevance. Theoretically, we add to the emerging debate on the multiplexity of intercity networks. Methodologically, we further advance the toponym co-occurrence method by introducing a method to classify the types of relationships that are obtained.

The paper is structured as follows. First, we review Chinese city network multiplexity studies and applications of the toponym co-occurrence

method (section 2). Second, we present our method, detailing the steps taken in the research process (section 3). Third, we map and analyze the similarity between different network layers at the national level (section 4). Finally, we conclude with a discussion of the pros and cons of this lexicon-based method and reflect on how this method can be successfully implemented in future studies (section 5).

2. Literature review

2.1 Network multiplexity

Network multiplexity refers to the phenomenon that nodes (cities) are related to each other in many different ways, for instance in social, economic and cultural dimensions. Studies have evaluated city networks using multiple types of relationships at different scales. On the global scale, studies generally focus on air transport (Derudder and Witlox, 2005), export (Hidalgo and Hausmann, 2009), and GaWC interlocking firm models (Derudder et al., 2003; Taylor et al., 2012). At the national and regional scale, commuting transport flows, high-speed railway and scientific collaboration networks dominate (Burger et al., 2011; Cao et al., 2021; Liu et al., 2016). In studies at the city scale exploring the network between districts, POI location data are gaining popularity (Jiang et al., 2015), such as locations from smartphone applications (An et al., 2019) and telephone calls (Järv et al., 2012). Choices to use particular types of data are often driven by data availability.

Comparing empirical studies that focus on a single type of city relationships can shed light on the multiplexity of city networks. For instance, studies using the high-speed railroad network in China (Liu et al., 2020; Yang et al., 2019) show that the best connected cities are usually geographically close, e.g. Beijing-Tianjin (125km) and Guangzhou-Shenzhen (136km). However, Pan et al. (2020) who constructed a financial center network based on the interlocking network model show that Beijing-Shenzhen (2175km) and Beijing-Shanghai (1214 km) are the most strongly connected cities in China. It is exactly for these diverging outcomes that scholars have argued that studying only one type of relationship can hardly give a comprehensive profile of the relationships between cities, and this may ultimately have consequences for the effectiveness of policy (Burger et al., 2014a).

Another reason to consider the multiplexity of inter-urban relationships is because different factors explain the variation in network patterns. Zhang et al. (2020) examined the determinants of three types of city networks in the Yangtze River Delta using transportation infrastructure, business interactions, and mobility data. They found that different determinants exist in each network: the infrastructure network is significantly correlated with landform patterns, and the mobility network is more associated with population and distance, while GDP and administrative relationships are the main determinants of the business. In addition, studies also have shown that large cities are likely to be more dependent on international networks while smaller cities benefit more from regional networks (Meijers et al. 2016).

2.2 Content analysis methods

Extracting and analyzing information content is not new. The earliest content analysis can be traced back to the 18th century (Rosengren, 1981). Nowadays this type of method is a widely used analytic technique to investigate myriads of information generated in the digital era.

Classical content analysis methods involve assigning codes to words and phrases to capture the text topic, help summarize the results and interpret meaning from the text content. According to the way of coding, traditional content analysis can be roughly classified into three approaches: deductive, inductive, and summative (Hsieh and Shannon, 2005). A deductive approach applies a predetermined set of codes to identify the content of the text, but an inductive approach creates codes based on the text content itself, subject to the analyzer's understanding of the content. A summative approach involves statistical tools to count and compare certain keywords or content.

Previous research on toponym co-occurrence often uses the text whose content actually has been classified and labelled. This is more of an inductive approach, as the label is based on the generator understanding of the content. For instance, based on the similarity of user tag generated tags of the articles, Salvini and Fabrikant (2016) quantified the relationships between cities using the number of shared articles from Wikipedia articles and classified the city relationship. A similar classification approach was conducted in the news by Hu et al. (2017), who extracted the toponym co-occurrence from news articles of The Guardian, and then classified them

by combining semantic related news tags of the articles. Such method is useful when there exist tags identifying the topic, but such method is hard to conduct for unstructured text without tags.

A big advantage of the summative approach is its capability of analyzing the content of massive unstructured text by identifying the keyness of the text, a selection of semantically related words that appear significantly less often in other texts. This method assumes that the specific linguistic choices a producer of a text makes indicates what the producer think (Scott, 1999). For example, Schuckert et al. (2015) identify articles related to tourism online reviews through relevant keyword selection and Song et al. (2016) identify top topic public-private partnership projects based on keyword frequency. This method is especially suitable for massive unstructured internet texts. For instance, studies involving analyzing social media texts (Chew and Eysenbach, 2010) often use hashtags to select relevant messages and treat the number of hashtag occurrence as an indicator of the popularity of an event.

Similar analysis can also be applied to deduct word relationship by its co-occurrence (Gregory et al., 2015; Porter et al., 2015). Content in text does not consist of isolated words, but of coherent, interdependent utterances, forming a structured story (Baxter and Montgomery, 1996; Griffin, 1993). Thus, it is possible to infer relationship between words based on their tendency to co-occur in the same text. Frequency of co-occurrence is a reflection of the extent to which a bundle of words is stored and used as a prefabricated chunk (Biber et al., 1999). In other words, a bundle of words with higher frequency are more likely to be stored than those with lower frequency.

More specifically, this relationship analysis based on co-occurrence is called collocation analysis. It counts and interprets the co-appearance of two or more entity names as the degree of relevance between the entities. Three measures determine the meaning and the strength of the relationship between the entities (Baker et al., 2008), the frequency of the appearance of the entity names, the frequency of where the entities collocate, and the frequency of how they collocate. This is because the collocation of words provides not only ‘a semantic analysis of a word’ (Sinclair, 1991), but also risks conveying implicit messages (Hunston, 2002). Many studies have used this method to identify relationships, such as searching for co-appearance of two universities in academic searching engines (Fanelli and

Lariviere, 2015), searching co-appearance of city and disease-related keywords in the Registrar-General's Reports (Porter et al., 2015), and searching co-appearance of refugees and their concerns in British news articles (Baker et al., 2008). In conclusion, the lexicon-driven approach has been used extensively for content classification and relevance detection.

Turning to urban systems, counting the co-appearance of two city names has shown great potential to retrieve inter-urban relationships. Studies have found that the toponym co-occurrence frequency between two places mirrors their "relatedness" in the real world in some aspects (Vaughan and You, 2010). For example, Liu et al. (2014) used the toponym co-occurrence method to investigate the relationship between provinces with data collected from Baidu. They found proximate provinces have similar network patterns, and the network pattern classification of provinces generally agrees with the widely accepted economic zoning schema in China. Later on, Zhong et al. (2017) applied complex network methods on the toponym co-occurrence results to evaluate city network positions. They found that the frequency of toponym co-occurrence is only weakly correlated with the distance between the two cities, but strongly correlated with administrative hierarchical distance. Besides investigating the toponym co-occurrence relationship patterns, experiments that combine this relationship with traditional data have been explored as well.

3. Method

In this paper we apply content analysis methods, more specifically a lexicon-based classification approach, to classify relationships between cities that have been found using the toponym co-occurrence method. A big advantage of such an approach is that this classification method can be applied to any text corpus and be tailored to any specific demand. But the approach will always involves these four steps:

1) Text corpus selection;

2) Preparation of lexicons containing sets of words where each lexicon captures a particular type of relationship;

3) Extraction of the relative frequencies with which words from each type of relationship appear from the text in order to categorize each text included in the corpus. The relative frequency indicates the extent to which the content is related to that type. For example, if a text contains five words relating to finance, three words relating to politics, and two words relating

to culture, then the relative frequencies are considered 0.5 finance, 0.3 politics, and 0.2 culture;
4) Building the database on relationships between cities. For each pair of cities that co-occurs in a text, we sum these frequencies across all texts in which these two city names co-appear for each type of relationship, and assign these frequencies as the link weight between the city pair in that type.

Below, we detail steps 1 and 2. For steps 3 and 4, we will focus on the outcomes, and hence discuss these in section 4.

3.1 Step 1: Text corpus selection

An important question is which text corpus is being used, as this is likely to affect outcomes to a considerable extent. Recent applications of the toponym co-occurrence method have used Wikipedia (Neal, 2012; Salvini and Fabrikant, 2016), newspapers (Hu et al., 2017; Janc, 2015; Zhong et al., 2017), or websites found through search engines such as Google and Baidu to explore web content (Devriendt et al., 2008; Liu et al., 2014). However, as indicated by Meijers and Peris (2019), search engines or a single newspaper source all suffer from potential biases and are not necessarily sufficiently representative for obtaining a comprehensive view. Instead of focusing on a single text source, they preferred the Common Crawl web archive. Common Crawl is one of the largest publicly available Web Archives, a monthly updated snapshot of all websites, and providing terabytes of Internet data, including billions of raw web pages in more than a hundred languages. More than half of its web pages are from top-level domains.

Therefore, we used the Common Crawl web archive as our preliminary data source to avoid selection bias. We extracted a 139 Gigabyte text corpus containing 91 million Chinese web pages from the April 2019 Common Crawl database using a Hadoop-based frame on a 1080 CPU cluster from the Amazon Elastic Map/Reduce infrastructure. The detailed corpus extraction processing and corpus statistics are presented in a separate data paper (Tongjing, 2024), which also explains the choice for the 293 cities in China at prefecture administration or higher and includes a link to the (open) data we created. We consider all websites on which we find two or more of these 293 Chinese placenames.

3.2 Step 2: Types of relationships and their lexicons

Since our interest is in identifying inter-urban relationships related to economic activities, we select six types by building on terms from the China Standard Industrial Classification of Economic Activities (GB/T 4754—2017), a procedure that helps avoid the risk of subjective selection bias. This Standard classifies economic activities into three main sectors and each sector is divided into several subsectors. Each economic activity in the Standard is described in a single word or phrase.

Given our focus on cities, we selected the economic activities from the secondary and tertiary sectors, and omitted the primary sector (agriculture, forestry, and fishery), as they are largely non-urban activities. We selected all economic activities in the secondary sector as representative of 'industry' type relationships. To obtain greater detail and clearer classification results for the tertiary sector, we subdivided this sector into five subsectors of tertiary economic activities: information technology (IT); finance; research (scientific research and technical service); culture (publishing, sports, and entertainment); and government management. Some example words or phrases that are used for each subsector in the Standard are listed below to illustrate how our lexicon-based approach works and to provide an idea of what economic activities are included in each subsector:

IT: e.g. 'information transfer', 'telecom', 'Internet', 'software', 'information technology',
Finance: e.g. 'finance', 'monetary finance', 'capital market', 'insurance',
Research: 'scientific method, 'research, 'quality inspection', 'technological application',
Culture: 'sports', 'entertainment', 'news', 'media', 'publication', 'broadcast', 'movie',
Government: 'social security', 'social organization', 'social working', 'communist party'.

Admittedly, some of the words that are in the Standard may not be used commonly in China and economic activity can also be described often by more than the one word provided in the Standard, so we enlarged the keyword set by adding words that are semantically related to the keywords in the Standard, creating a more comprehensive collection. To find semantically related words of a given word, a natural language processing

(NLP) technique called word embedding is utilized. This technique measures the relatedness of words or phrases by mapping them into multidimensional vectors based on their relationships in the documents (Bengio et al., 2003). In this paper, we selected a word embedding corpus from Tencent AI Lab (Song et al., 2018), which is a 16.75 Gigabyte dataset that has measured the relationship of over 8 million Chinese words and phrases using a directional skip-gram model by covering a large amount of domain-specific words or slang terms from news, webpages, and novels, and by including phrases from Wikipedia and Baidu Baike (a Chinese version of Wikipedia). This model relies on word co-occurrence within the local context and word sequence. The semantic similarity between two words was calculated using the cosine similarity of their corresponding embedding vectors. To optimize the number of semantic words that relate to economic activities and reduce the probability of words appearing in more than one type, we selected the threshold of similarity at 0.5. Then we cleaned the resulting dataset, by making sure that keywords are associated with just one (in our view the most logical) type. Some descriptives on the number of words per type and the method of adding them to their lexicon is presented in Table 1. The Chinese terms we used are listed in appendix I (with translation).

Table 1. Number of words in the lexicons for types of relationships

Number of words	Industry	IT	Finance	Research	Culture	Government
From the Standard	81	10	5	7	11	12
After NLP expansion	629	71	54	56	119	111

As can be seen in Table 1, the NLP extension substantially increases the number of words for each category, but to different degrees. The size of the 'bag of words' for the finance sector increase by 10.8 times, for the industry sector it is 7.7 times. This does not make up for the considerable difference in sizes of the lexicons for the different categories. Nevertheless, this does not mean that it is more likely that we will find more toponym co-occurrences for the industry sector, as what matters is the frequency of appearance of words, which differs substantially. Using very specific, smaller text corpora will influence those frequencies more than a broad corpus like the WebArchive employed here. Nevertheless, some care should be taken when comparing, which is why we focus predominantly

on relative strengths, network patterns more generally and especially the spatial pattern of each type of network of relationships.

4. Result

4.1 General multiplexity between network layers

For each pair of cities, we considered all websites on which the names of these cities co-occur, and the text of these websites was used to determine whether it was about one of our six categories or not, and if so, to what degree. These relative frequencies for a particular type were then summed for each pair of cities, and this is the link weight of the city pair in that type. The resulting pattern of co-occurrences was captured as a network layer. To assess the multiplexity of the national city network, we calculate the correlation between any two layers. Specifically, we used Spearman's rank correlation coefficient to measure the strength and direction of the association between two layers (Table 2), rather than Pearson's linear correlation coefficient, which requires a normal distribution dataset.

Table 2. Spearman's rank correlation between different layers in weight

	Industry	**IT**	**Finance**	**Research**	**Culture**	**Government**
Industry	1.000					
IT	0.887	1.000				
Finance	0.757	0.754	1.000			
Research	0.893	0.878	0.753	1.000		
Culture	0.802	0.858	0.899	0.823	1.000	
Government	0.795	0.802	0.622	0.896	0.743	1.000

While correlation coefficients always leave some room for normative interpretation, we would consider those correlations to be not extremely high (>.9), and even rather low between several types of city relationships in some cases. This illustrates the necessity to consider the multi-layer city network instead of a single layer network. Each type of relationship usually shows a higher correlation with two or three other types and a lower correlation with the others. For example, 'industrial' relationships between cities are highly correlated with 'IT' and 'research' relationships, perhaps because technology is central to their functioning. The pattern of industry relationships is least correlated with relationships that are traditionally more associated with a service economy, e.g. 'finance', 'government' and 'culture'. The pattern of relationships for especially 'finance' and

'government' is particularly dissimilar, but the 'finance' relationships are also less similar to those in 'IT' and 'industry, suggesting that the finance network layer is on average most dissimilar to the other types, perhaps because the command and control centers of financial institutions are relatively concentrated in a handful of cities. Beyond finance, a low correlation is also found for 'culture' and 'government', which is perhaps surprising too as capital cities tend to have been 'consumer cities avant la lettre', endowed with a surplus of cultural amenities (Cardoso and Meijers, 2016). This analysis of relations involving all 42,778 city pairs studied here suggests already that different types of relations show quite different network patterns.

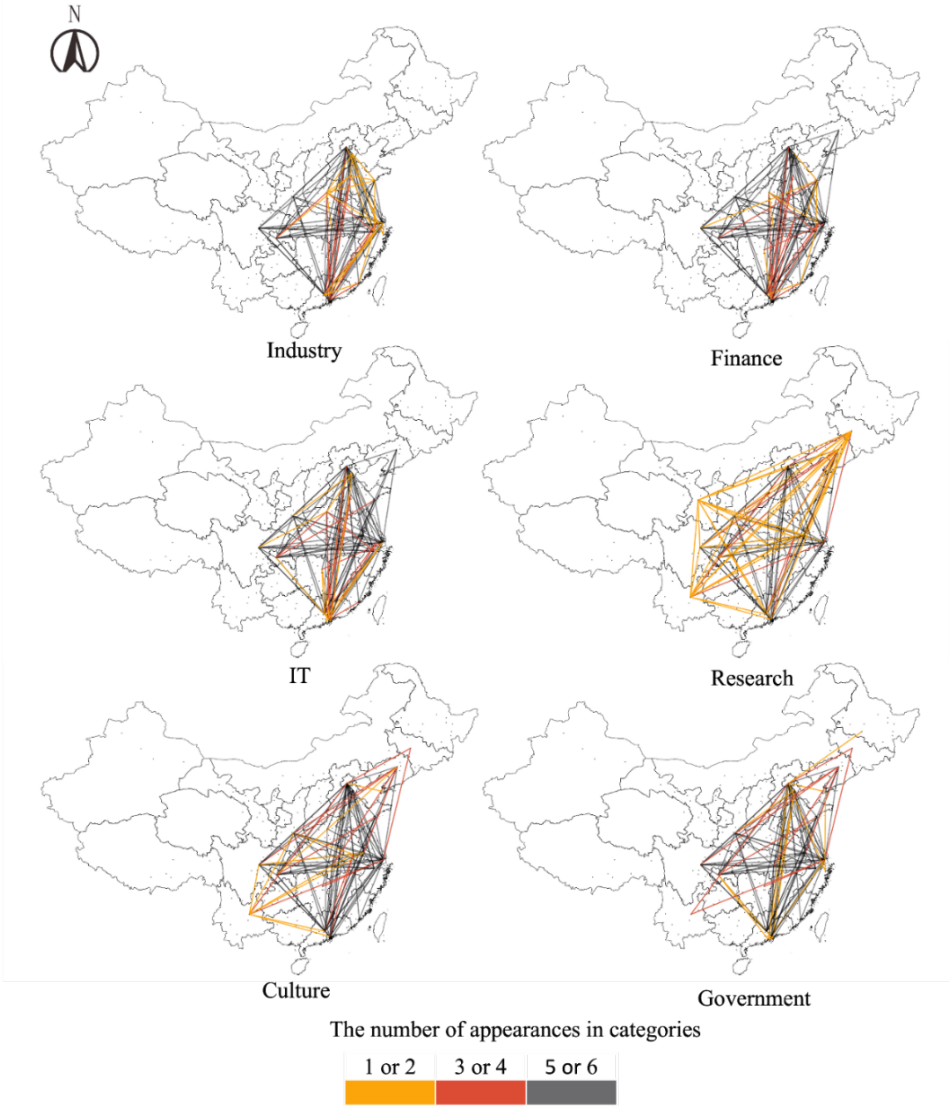
4.2 Multiplexity in the top relationships

To show more details of the multiplexity in the Chinese urban system, we narrowed it down to the top-100 relationships with the highest linkage weight in each layer, as mapping information on 42,778 pairs of cities would render these maps unreadable. The mapped results are presented in Figure 1, where the darker links re-occur more often in the top-100 of the six types of relationships. Orange links indicates that the pair of cities only occurs in the top-100 of just one or two types of relationship and a black line means that this city pair appears in the top-100 of all or five out of the six layers.

The main overlapping network pattern (black links) of the six types of relationships is between Yangtze river delta (YRD) in the east, Pearl River Delta (PRD) in the south, and the cities Beijing and Tianjin in the north, Chengdu, Chongqing in the west, and Wuhan in the middle. This aggregates to a diamond structure. This diamond structure is also observed in previous research, using top company headquarters-subsidiary relationships (Jiang et al., 2017) or airline transportation networks (Lao et al., 2016). The industry, IT, and finance layers are more concentrated in this overlapping area; and the research layer has more unique links than other types of relationships do, and is spatially more expanded, which includes the north-western city Lanzhou in Gansu province for instance. However, in general, cities in the western part are excluded, as they do not reach the threshold for inclusion probably due to their limited size and economic importance.

The absolute link weight also varies significantly within the top shared city relationships in the six categories. Table 3 shows the top-5 strongest relationships between cities in each category in absolute terms.

Figure 1. Networks between cities in different categories



As shown in Table 3, Beijing-Shanghai has the strongest relationship in all categories but the strength still differs considerably—their relationship in culture is roughly 13 times stronger than that in industry, despite the lexicon for industry being much larger.

Table 3. Top 5 strongest relationships between cities in each category

Industry	Value	IT	Value	Finance	Value	Research	Value	Culture	Value	Government	Value
Beijing - Shanghai	503644	Beijing -Shanghai	3604574	Beijing -Shanghai	1493076	Beijing -Shanghai	1437237	Beijing -Shanghai	6531389	Beijing -Shanghai	199418
Beijing -Chongqing	370011	Beijing -Tianjin	2393547	Beijing -Shenzhen	1137011	Beijing -Guangzhou	977433	Beijing -Chongqing	3894152	Beijing -Chongqing	145302
Beijing -Tianjin	360053	Beijing -Chongqing	2372491	Shanghai -Shenzhen	1100002	Shanghai -Guangzhou	933637	Beijing -Guangzhou	3798863	Shanghai -Chongqing	131115
Shanghai -Chongqing	351586	Shanghai -Tianjin	2362580	Beijing -Guangzhou	1036874	Beijing -Shenzhen	815386	Beijing -Shenzhen	3798179	Beijing -Tianjin	130195
Shanghai -Tianjin	331198	Shanghai -Chongqing	2357649	Shanghai -Guangzhou	990389	Beijing -Chongqing	808641	Shanghai -Shenzhen	3762768	Shanghai -Tianjin	116447

We also counted the number of cities in the top-100 relationships of each layer and calculated the network density, respectively. The network density is a measure of the number of relationships compared to the maximum possible number of relationships, which is defined as:

$$Density = \frac{2 \times m}{n(n-1)}$$

Where n is the number of cities connected, and m is the number of relationships.

Within the top-100 relationships, the patterns vary with the number of cities and network density. The result is shown in Table 4. The 'number of cities' presents how many different cities are included in the top-100 of relationships per network layer.

Table 4. The network density of the top 100 relationships

	Industry	IT	Finance	Research	Culture	Government
Number of cities	22	22	21	18	23	26
Density	0.43	0.43	0.47	0.65	0.40	0.31

The top-100 research and the government relationships are all between high-level administration cities, which generally are innovation and decision-making powerhouses (Ma, 2005; Li and Wu, 2012). However, they show opposite network patterns. The research relationships have the highest network density with the lowest number of cities included, but the government relationships are exactly the opposite, involving more cities, but with a lower network density. The rest fall in between.

4.3 Multiplexity in terms of population and distance effects

So far we have focused on the strongest relationships in absolute terms, but it is obviously not a surprising find to see that the largest cities top the rankings as this is not necessarily a reflection of strong network embeddedness, but may simply come forward from their size. To evaluate multiplexity in more relative terms, we use the gravity model to examine how the relationships relate with the population and distance. The gravity model in geography is a commonly-used method for estimating the pattern of intercity relationships by taking into account the population size and distance between cities, and this allows to compare the effects of population and distance on frequency of co-occurrences (Meijers and Peris, 2019).

The population data were gathered from the 2019 China provincial statistical yearbook. The results of the gravity model are shown in Table 5.

Table 5. Estimation results based on the gravity model[3]

	Industry	IT	Finance	Research	Culture	Government
	M1	**M2**	**M3**	**M4**	**M5**	**M6**
Intercept	9.180	9.472	10.849	6.219	11.252	4.938
	(0.040)***	(0.030)***	(0.034)***	(0.052)***	(0.032)***	(0.057)***
Pop.A (ln)	0.233	0.191	0.125	0.304	0.160	0.346
	(0.003)***	(0.002)***	(0.003)***	(0.003)***	(0.003)***	(0.004)***
Pop.B (ln)	0.216	0.163	0.118	0.260	0.147	0.268
	(0.003)***	(0.002)***	(0.003)***	(0.004)***	(0.003)***	(0.004)***
Distance (ln)	-0.055	-0.040	-0.032	-0.068	-0.041	-0.143
	(0.003)***	(0.002)***	(0.002)***	(0.004)***	(0.002)***	(0.004)***
Adjusted R^2	0.352	0.379	0.132	0.380	0.227	0.365
F-statistics	7745	8695	2177	8736	4198	8182
Root MSE	0.342	0.255	0.352	0.408	0.320	0.491
N	42778	42778	42778	42778	42778	42778

Standard error in parentheses. *P* values: 0 *** 0.001 ** 0.01 *

The results in Table 5 show that the network pattern of the six types of relationships are all explained by population sizes and distances between cities, but the overall level of the Adjusted R^2 is not that high for applications of the gravity model. Apparently, other factors beyond size and distance also play an important role in forging relationships. We can also see that there are quite considerable differences in the adjusted R^2 for

[3] For curiosity, we also ran the gravity model using GDP at city level rather than population sizes, which showed that levels of GDP explain relationship patterns slightly better than population size.

the different types of networks. Especially the financial and cultural relationships are much less explainable by population and distance than the others. Zooming in on individual factors, the coefficient of distance (ln) shows that the government relationships between cities decay much faster with distance than the other types of relationships, whereas the population factors are less important for finance relationships than they are for other relationships. Table 6 compares observed, absolute values with relative values as predicted by the gravity model, which is expressed in a percentage more or less than expected.

Table 6. Top 5 strongest relationships between cities in each category, accounting for size and distance

Industry	Value	IT	Value	Finance	Value	Research	Value	Culture	Value	Government	Value
Beijing - Shanghai	1704%	Beijing -Shanghai	1439%	Beijing -Shanghai	552%	Beijing -Shanghai	5862%	Beijing -Shanghai	1053%	Beijing -Shanghai	3362%
Beijing -Shenzhen	1230%	Beijing -Shenzhen	988%	Beijing -Shenzhen	462%	Beijing -Guangzhou	4782%	Beijing -Shenzhen	688%	Beijing -Macau	2937%
Beijing -Guangzhou	1211%	Beijing -Guangzhou	936%	Shanghai -Shenzhen	435%	Lanzhou -Kunming	4653%	Beijing -Guangzhou	677%	Hong Kong -Macau	2449%
Tianjin -Shanghai	1197%	Shanghai -Shenzhen	924%	Beijing -Guangzhou	415%	Lanzhou -Shenyang	4548%	Shanghai -Shenzhen	661%	Beijing -Hong Kong	2414%
Shanghai -Shenzhen	1192%	Beijing -Chongqing	892%	Shanghai -Guangzhou	385%	Lanzhou -Nanjing	4542%	Shanghai -Chongqing	614%	Beijing -Chongqing	2398%

As can be seen in Table 6, the gravity model underestimates the strength of the relationships between the main Chinese cities, and the strength of these relationships is much stronger than expected. Most of the top relationships in relative terms are the same as those in the top toponym co-occurrence relationships in absolute terms (see Table 3). However, it is interesting to note that some relatively small (by Chinese standards) and peripheral cities in the research category (Lanzhou, Kunming and Shenyang) are well embedded in networks. Newcomers in the government category include Macau and Hong Kong.

5. Conclusion

As a summative content analysis method, the toponym co-occurrence approach has great potential to reconstruct intercity network structures; however, the precise meaning of such relationships remains somewhat unclear, and one needs to embrace the basic underlying assumption that when two place names are frequently mentioned in one breath, this suggests that they are strongly related. Note that this is the standard assumption underlying all applications of text analysis. While the reliability of our approach primarily rests on following common and validated procedures in collocation analysis, the fact that the network

patterns for different types of relationships obtained corresponds to some extent with the gravity model is reassuring, as is our finding that general patterns found correspond with results of other studies.

Deriving relationships between cities from toponym co-occurrences in text corpora has huge potentials in that it can be applied in situations where traditional data is not even available. Moreover, it overcomes weaknesses of traditional data sources, for instance by providing a uniform and harmonized method to analyze cross-border relationships, or because it can be applied to places (or regions, or countries, or firms…) of any size, which sample data often cannot. This paper aimed to stimulate further discussion on the toponym co-occurrence method, and provides one of the first attempts to give more meaning to those co-occurrences by classifying them, and as a result capture and explore network multiplexity. Rather than suggesting machine learning approaches, we proposed a lexicon-based classification approach, which means that we considered texts in which two cities were both mentioned and tried to make sense of how they were related by examining the context in which they were mentioned. Using this lexicon-based approach we could zoom in on six broad categories of relationships. As broad or very specific lexicons can be developed and adapted to fit with every possible research problem, the main result of this paper is obviously showing that such an approach is possible indeed, which opens up a plethora of possible future research. Regarding the risk of toponym ambiguity, we have tried to use keywords that are less ambiguous, hence with only one meaning, or with just one very dominant meaning. If the chosen words have only one dominant meaning, the risk of content ambiguation will be low, as research (Guo et al., 2007; Lucas, 1999) shows that although the actual meaning of a word depends on the context, this context effect has only a limited effect if the word has a dominant meaning.

The principle of our method is based on the commonly-used collocation analysis in linguistic and data science analysis— in a large corpus the frequency of co-appearance of keywords can imply the relationship strength between these keywords, ignoring the actual text structure. This rule is not limited to a specific language as such method has been successfully applied in English (Gregory et al., 2015), Dutch (Meijers and Peris, 2019), French (Baker and Vessey, 2018) and Chinese (Liu et al., 2014) for instance. The methodological innovation of our paper is that we advanced the traditional collocation analysis by grouping it with a bag of semantically related keywords, and this tagging allows to classify a

relationship between a pair of cities. This basic principle can be applied in any language, and one can choose very specific keywords to detail relationships between places, regions or countries. Therefore, we believe this method has wide application potential.

Here, we were mainly interested in showing the multiplexity of interurban relationships in China, one of the many countries where national level data on relations between cities is largely missing. A diamond-shaped structure of relationships characterizes the urban system of China, and this structure is anchored on the Yangtze river delta (YRD) in the east, Pearl River Delta (PRD) in the south, the cities Beijing and Tianjin in the north, Chengdu, Chongqing in the west, and Wuhan in the center, which are the economic powerhouses. However, we also showed that network patterns of different types of relationships differ substantially from each other, and this calls for prudence in making bold claims on the structure of urban systems when these are based on the analysis of just one type of relationship. For instance, we found that city relationships in government decay much faster than other types which suggests that governmental collaborations are often restricted to nearby cities. Especially the patterns of research relationships differs, involving less cities, but creating stronger relationships. The government network pattern involves more cities, but these are more loosely related. Likewise, patterns in finance and culture follow the regularities of the gravity model much less than the other types. Yet, while this paper highlights the necessity of revealing the multiplexity of city networks, our results actually also show that each type of relationship is not completely independent from some other types of relations.

Toponym co-occurrences can revolutionize the study of city networks, certainly now that we can distinguish more specific types of relationships rather than just the overall strength of a relationship. While we used rather general categories here, future research needs to concentrate on the suitability of a lexicon-based approach using much more specific keywords to retrieve very specific types of relationships. As such, the lexicon-based approach provides more flexibility than supervised machine learning techniques. Research could also address an omission in our approach, namely not taking into account the text structure yet, such as the separation between the place where toponyms occur in a text; for instance, it probably makes a difference whether they are mentioned in the same sentence, or whether there are ten sentences in between. We also recommend using different types of text corpora, and pay more attention

to how texts develop. One drawback of our use of a Web Archive is that larger cities may be disproportionally more mentioned on web pages than smaller cities which could be an alternative explanation for our finding that the relative strength of relationships between the main Chinese cities is many times higher than we would expect based on gravity modelling. What the trend is in this regard is yet unknown, and it would be interesting to see time series of toponym co-occurrences for different types of relations. The quality of an analysis based on co-occurrences ultimately rests on the quality of a text corpus. We recommend using more specific or targeted text corpora now that we have tried broad Web Archives. The comparison should indicate whether results remain largely similar or not. Insofar these other measures are good proxies themselves, a certain correspondence of the results would further enhance credibility of the lexicon-based approach. Research could also concentrate on comparisons with network patterns obtained through more traditional methods, e.g., human migration data and capital flow data. Despite the challenges still ahead, we are convinced that the toponym co-occurrence method can spark a new perspective in studying city networks in the digital era.

ACKNOWLEDGEMENT
The authors would like to thank Zhao Yin and Ziyu Bao at Delft Technology of University for computer programming support. The paper also benefitted from feedback received at various international conferences, research seminars, as well as constructive feedback from external reviewers.

DISCLOSURE STATEMENT
No potential conflict of interest was reported by the authors.

FUNDING
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

An, D., Tong, X., Liu, K. and Chan, E. (2019) Understanding the impact of built environment on metro ridership using open source in Shanghai. Cities, 93: 177-187.

Baeza-Yates, R., Blanco, R. and Castellanos, M. (2019) Web text mining. In: Mitkov, R. (Ed.) The Oxford Handbook of Computational Linguistics, 2nd edition. Oxford: Oxford University Press.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyżanowski, M., McEnery, T., and Wodak, R. (2008) A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse and Society, 19(3): 273-306.

Baker, P., and Vessey, R. (2018). A corpus-driven comparison of English and French Islamist extremist texts. International Journal of Corpus Linguistics, 23(3), 255-278.

Baxter, L. and Montgomery, B. (1996) Relating: Dialogues and dialectics. Guilford Press.

Bengio, Y., Ducharme, R., Vincent, P. and Jauvin, C. (2003) A neural probabilistic language model. Journal of Machine Learning Research 3(Feb): 1137–1155. https://www.iro.umontreal.ca/~vincentp/Publications/lm_jmlr.pdf

Berroir, S., Cattan, N., Dobruszkes, F., Guérois, M., Paulus, F. and Vacchiani-Marcuzzo, C. (2017) Les Systèmes Urbains Français: une Approche Relationnelle. Cybergeo: European journal of geography, document 807.

Biber, D. and Finegan, E. (2008) Longman: grammar of spoken and written English. In Longman: grammar of spoken and written English. p.1204.

Burger, M., Meijers, E. and Van Oort, F. (2014a) Multiple perspectives on functional coherence: Heterogeneity and multiplexity in the Randstad. Tijdschrift voor Economische en Sociale Geografie 105(4): 444–464.

Burger, M., Van der Knaap, B. and Wall, R. (2014b) Polycentricity and the Multiplexity of Urban Networks. European Planning Studies 22(4): 816–840.

Burger, M., de Goei, B., Van der Laan, L. and Huisman, F. (2011) Heterogeneous development of metropolitan spatial structure: Evidence from commuting patterns in English and Welsh city-regions, 1981–2001. Cities, 28(2): 160-170.

Callon, M., Courtial, J., Turner, W. and Bauin, S. (1983) From translations to problematic networks: An introduction to co-word analysis. Social

Science Information, 22(2): 191-235.
Cao, Z., Peng, Z. and Derudder, B. (2021) Interurban scientific collaboration networks across Chinese city-regions. Environment and Planning A: Economy and Space, 53(1): 6-8.
Capello, R. (2000) The city network paradigm: Measuring urban network externalities. Urban Studies 37(11): 1925–1945.
Cardoso, R. and Meijers, E (2016) Contrasts between first-tier and second-tier cities in Europe: a functional perspective. European Planning Studies 24(5): 996–1015.
Chew, C. and Eysenbach, G. (2010). Pandemics in the age of Twitter: content analysis of Tweets during the 2009 H1N1 outbreak. PloS ONE, 5(11): e14118.
Culpeper, J. (2009) Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. International Journal of Corpus Linguistics, 14(1): 29-59.
Davoudi, S. (2008) Conceptions of the city-region: a critical review. Proceedings of the Institution of Civil Engineers-Urban Design and Planning 161(2): 51–60.
Derudder, B. and Neal, Z. (2018) Uncovering links between urban studies and network science. Networks and Spatial Economics 18(3): 441–446.
Derudder, B. and Taylor, P. J. (2018) Central flow theory: Comparative connectivities in the world-city network. Regional Studies, 52(8): 1029-1040.
Derudder, B. and Witlox, F. (2005) An appraisal of the use of airline data in assessing the world city network: a research note on data. Urban Studies, 42(13): 2371-2388.
Derudder, B., Taylor, P., Witlox, F. and Catalano, G. (2003) Hierarchical tendencies and regional patterns in the world city network: a global urban analysis of 234 cities. Regional Studies, 37(9), 875-886.
Devriendt, L., Derudder, B. and Witlox, F. (2008) Cyberplace and cyberspace: two approaches to analyzing digital intercity linkages. Journal of Urban Technology 15(2): 5–32.
Devriendt, L., Boulton, A., Brunn, S., Derudder, B., and Witlox, F. (2011). Searching for cyberspace: the position of major cities in the information age. Journal of Urban Technology, 18(1), 73-92.
Fanelli, D., Costas, R., and Larivière, V. (2015) Misconduct policies, academic culture and career stage, not gender or pressures to publish, affect scientific integrity. PloS ONE, 10(6): e0127556.
Gregory, I., Cooper, D., Hardie, A., and Rayson, P. (2015) Spatializing and

analyzing digital texts: Corpora, GIS and places. Spatial Narratives and Deep Maps, 150-78.
Griffin, L. (1993) Narrative, event-structure analysis, and causal interpretation in historical sociology. American Journal of Sociology, 98(5): 1094-1133.
Guo, J., Shu, H., and Li, P. (2007). Context effects in lexical ambiguity processing in Chinese: A meta-analysis. Journal of Cognitive Science, 8(1), 85-101.
Hidalgo, C. and Hausmann, R. (2009) The building blocks of economic complexity. Proceedings of the National Academy of Sciences, 106(26): 10570-10575.
Hsieh, H. and Shannon, S.(2005) Three approaches to qualitative content analysis. Qualitative Health Research, 15(9): 1277-1288.
Hu, X., Wang, C., Wu, J. and Stanley, H. (2020) Understanding interurban networks from a multiplexity perspective. Cities 99: 102625.
Hu, Y., Ye, X., and Shaw, S. (2017) Extracting and analyzing semantic relatedness between cities using news articles. International Journal of Geographical Information Science, 31(12): 2427-2451.
Hunston, S. (2002) Pattern grammar, language teaching, and linguistic variation. Using Corpora to Explore Linguistic Variation, 167-183.
Janc, K. (2015) Geography of hyperlinks—Spatial dimensions of local government websites. European Planning Studies, 23(5): 1019–1037.
Järv, O., Ahas, R., Saluveer, E., Derudder, B. and Witlox, F. (2012) Mobile phones in a traffic flow: a geographical perspective to evening rush hour traffic analysis using call detail records. PloS ONE, 7(11): e49171.
Jiang, X., Yang, Y., Wang, S., Wang, M. and Yang, Y. (2017) Spatial structure of Chinese intercity network based on the data of listed companies. City Planning Review 41(6): 18–26.
Jiang, S., Alves, A., Rodrigues, F., Ferreira, J. and Pereira, F. (2015) Mining point-of-interest data from social networks for urban land use classification and disaggregation. Computers, Environment and Urban Systems, 53: 36-46.
Lao, X., Zhang, X., Shen, T. and Skitmore, M. (2016) Comparing China's city transportation and economic networks. Cities, 53: 43-50.
Li, Y. and Wu, F. (2012) The transformation of regional governance in China: The rescaling of statehood. Progress in Planning, 78(2): 55-99.
Liu, S., Wan, Y. and Zhang, A. (2020) Does China’s high-speed rail development lead to regional disparities? A network perspective. Transportation Research Part A: Policy and Practice, 138:

299-321.
Liu, X., Derudder, B. and Wu, K. (2016) Measuring polycentric urban development in China: An intercity transportation network perspective. Regional Studies, 50(8): 1302-1315.
Liu, Y., Wang, F., Kang, C., Gao, Y. and Lu, Y. (2014) Analyzing Relatedness by Toponym Co-Occurrences on Web Pages. Transactions in GIS, 18(1): 89-107.
Lucas, M. (1999). Context effects in lexical access: A meta-analysis. Memory & Cognition, 27(3), 385-398.
Ma, L. (2005) Urban administrative restructuring, changing scale relations and local economic development in China. Political Geography, 24(4): 477-497.
Meijers, E., Burger, M. and Hoogerbrugge, M. (2016) Borrowing size in networks of cities: City size, network connectivity and metropolitan functions in Europe. Papers in Regional Science, 95(1): 181-198.
Meijers, E. and Peris, A. (2019 Using toponym co-occurrences to measure relationships between places: review, application and evaluation. International Journal of Urban Sciences, 23(2): 246-268.
Mello, R. (2002) Collocation analysis: A method for conceptualizing and understanding narrative data. Qualitative research, 2(2): 231-243.
Neal, Z. (2012) The connected city: How networks are shaping the modern metropolis. New York: Routledge.
Pan, F., Bi, W., Liu. X. and Sigler, T. (2020) Exploring financial centre networks through inter-urban collaboration in high-end financial transactions in China. Regional Studies 54(2): 162–172.
Peris, A., Meijers, E. and Van Ham, M. (2018) The evolution of the systems of cities literature since 1995: schools of thought and their interaction. Networks and Spatial Economics 18(3): 533–554.
Porter, C., Atkinson, P. and Gregory, I. (2015) Geographical Text Analysis: A new approach to understanding nineteenth-century mortality. Health & Place, 36, 25-34.
Rosengren, K. (1981) Advances in content analysis. Sage.
Salvini, M. and Fabrikant, S. (2016) Spatialization of user-generated content to uncover the multirelational world city network. Environment and Planning B: Planning and Design 43(1): 228–248.
Schuckert, M., Liu, X. and Law, R. (2015) Hospitality and tourism online reviews: Recent trends and future directions. Journal of Travel & Tourism Marketing, 32(5): 608-621.
Scott, M. (1999) WordSmith Tools Help Manual, Version 3.0.
Sinclair, J. and Sinclair, L. (1991) Corpus, concordance, collocation.

Oxford: Oxford University Press.
Song, Y., Shi, S., Li, J. and Zhang, H. (2018) Directional skip-gram: Explicitly distinguishing left and right context for word embeddings. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2: 175-180.
Song, J., Zhang, H. and Dong, W. (2016) A review of emerging trends in global PPP research: analysis and visualization. Scientometrics, 107(3): 1111-1147.
Taylor, P., Ni, P., Derudder, B., Hoyler, M., Huang, J. and Witlox, F. (2012) Global urban analysis: A survey of cities in globalization. Routledge.
Tobler, W. and Wineburg, S. (1971) A Cappadocian speculation. Nature 231(5297): 39–41.
Tongjing, W., Yin, Z., Bao, Z., & Meijers, E. (2024). Intercity relationships between 293 Chinese cities quantified based on toponym co-occurrence. Cybergeo: European Journal of Geography.
Vaughan, L. and You, J. (2010) Word co-occurrences on Webpages as a measure of the relatedness of organizations: A new Webometrics concept. Journal of Informetrics, 4(4), 483-491.
Watts, D. (2004) The “new” science of networks. Annual Review of Sociology, 30: 243-270.
Yang, H., Dijst, M., Witte, P., Van Ginkel, H., Wang, J. (2019) Comparing passenger flow and time schedule data to analyse High-Speed Railways and urban networks in China. Urban Studies 56(6): 1267–1287.
Zhang, J. and Peck, J. (2016) Variegated capitalism, Chinese style: Regional models, multi-scalar constructions. Regional Studies 50(1): 52–78.
Zhang, W., Derudder, B., Wang, J. and Witlox, F. (2020) An Analysis of the Determinants of the Multiplex Urban Networks in the Yangtze River Delta. Tijdschrift voor Economische en Sociale Geografie 111(2): 117–133.
Zhong, X., Liu, J., Gao, Y. and Wu, L. (2017) Analysis of co-occurrence toponyms in web pages based on complex networks. Physica A: Statistical Mechanics and its Applications, 466: 462-475.

Chapter 4

Intercity networks and urban performance: A geographical text mining approach

This chapter is published as: Tongjing, W., Meijers, E., Bao, Z., & Wang, H. (2024). Intercity networks and urban performance: a geographical text mining approach. International Journal of Urban Sciences, 28(2), 262-283.

Abstract: Our understanding of the relative significance between agglomeration and network externalities remains limited. This is largely due to limited data availability and we propose a general measure to proxy city network externalities based on toponym co-occurrences that indicate the relatedness between cities. This paper extracts intercity relationships based on the co-occurrence of Chinese place names on 2.5 billion webpages. We calculate and map absolute and relative network positions, which we use to explain urban labor productivity. We found that a stronger embeddedness in networks of cities is significantly and positively associated with urban productivity. Smaller cities benefit comparatively more from being well embedded in city networks, suggesting that these relations can compensate for a lack of agglomeration externalities. We also compare the importance for urban performance of city network externalities vis-à-vis agglomeration externalities. City network externalities turn out to be more important in explaining urban performance than agglomeration externalities. This calls for new theorizing on a relational approach to urban and regional development. Rather than stimulating further concentration of urbanization, our findings suggest that fostering relationships between cities is a viable alternative urban development strategy. We conclude with suggestions for a research agenda that delves deeper into city network externalities.

Keywords: city networks, urban system, labor productivity, China

1. Introduction

Agglomeration economies are considered key to urban growth, and much scholarly attention has consequently gone into studying this relationship. However, cities cannot be studied in isolation (Pumain, 2021). They are connected in varying degrees to a range of other cities through flows of labor, capital, and information (McCann and Acs, 2011). These various kinds of flows constitute the urban network system and open up possibilities for synergy in urban and economic development (Gordon and McCann, 2000; Johansson and Quigley, 2004; Parr, 2004). A growing awareness of cities' external economies is reflected in the introduction of new concepts such as 'regional externalities' (Parr, 2002), 'borrowed size' (Meijers and Burger, 2017; Phelps et al., 2001) and 'urban network externalities' (Capello, 2000). This means that studies of urban performance need to incorporate both agglomeration externalities and the externalities resulting from relationships of firms, households and

organizations with other cities (Burger and Meijers, 2016; Glaeser et al., 2016).

Nevertheless, although emphasis on the importance of city network externalities for city performance is not new (see for instance the urban systems literature of the 1960s and 1970s), and definitely on the rise, relatively little is still known about the comparative importance of city network externalities vis-à-vis agglomeration externalities. This is mostly due to challenges in quantifying city network externalities (Capello, 2000; Burger and Meijers, 2016; Van Meeteren et al., 2016). Yet, it is important to get an understanding of the relative importance of city network externalities and agglomeration externalities in improving urban performance, as benefiting from agglomeration externalities requires an entirely different urban and regional development strategy (focused on concentration) than enforcing network externalities (focused on establishing a myriad of connections and relationships). This paper therefore addresses an important planning debate on the future of urbanization. Should we focus on a more balanced development of national urban systems or rather put our eggs in one basket of a national champion city?

To answer such a fundamental question it is essential to solve the issue of how to measure city network externalities. This is easier said than done; the lack of city relational data was once considered to be the 'dirty little secret' (Short et al., 1996) of city network research and the measurement of city networks has been a continuous and common challenge ever since (Salvini and Fabrikant, 2016; Meijers and Peris, 2019). A typical challenge is that information on flows on higher geographical scales between cities is limited as traditional data collection is either focused on flows relevant for daily urban systems (e.g. commuting), or on inter-regional rather than inter-urban flows (e.g. trade, migration). Moreover, existing data often focuses on just one type of relation or flow (e.g. train timetables; participation in projects; intra-firm networks) which may be very useful for a particular analysis, but does not necessarily represent a comprehensive picture of city networks.

For these reasons, new methods of collecting intercity relational data are being explored. One promising method is geographic text analysis, which extracts information about places from text through keywords, structure, and content (Gregory et al., 2015; Porter et al., 2015). This type of methods

is of particular interest due to its accessibility and its ability of obtaining large-scale, intercity relational data. For instance, geo-tagged posts on social media (Fang et al., 2020; Hu et al., 2020) and photos (Yuan and Medel, 2016) have been extracted for analyzing intercity relationships. The potential of analysing text corpora for geographic purposes is very substantial, as estimates suggest that about 70% of our documents contain place references (Hill, 2009).

In geography, we often derive spatial relationships from co-occurrences. For instance, we search for co-occurrences of geo-located authors in scientific co-publication patterns (Dai et al., 2022; Ma & Xu, 2022) or patents (Petralia et al., 2016). At a higher, macro scale, we can also search directly for the co-occurrence of city names. This has the advantage of being more comprehensive than a micro-level approach, but the disadvantage that a specific co-occurrence of place names is harder to specify or categorise and does not necessarily imply a tangible relationship, nor an actual flow, as it can be more abstract and symbolic (Watts, 2004). Yet, as is the standard assumption underpinning every text mining analysis, co-occurrences indicate relatedness, and we assume that more than usual or expected co-occurrence of a pair of city names indicates that they are relatively more strongly related. We believe that, in order to proxy the broad concept of city network externalities, it makes sense to rely on such a more generic measure of relatedness, just as the broad concept of agglomeration externalities is typically proxied with generic proxies as ‘size’ or ‘density’.

The nature of the objective of this paper is both methodological and theoretical. Methodologically, we aim to identify intercity relationships and aggregate network patterns through applying the still rather novel toponym co-occurrence method. Theoretically, we aim to evaluate and compare the relative importance of city network externalities vis-à-vis agglomeration externalities on city performance, thereby using a Cobb Douglas production function. Empirically, we focus on China. This country’s unbalanced urbanization and heterogenous intercity relationships lead to a variety of city network positions (Phelps et al., 2020), which thus allows us to study their relevance for performance. The difficulty of finding city relational data in China is an additional reason why applying the toponym co-occurrence method to this country is of value. Another reason is that both the central and local governments have developed strategies to integrate cities within mega-regions through

reform and major infrastructure projects (Fang and Yu, 2017; Harrison and Gu, 2019), thereby prioritizing the development of networks over further concentration. This study will shed light on the desirability and feasibility of strategies aimed at enhancing the presence of city network externalities. As such, the relevance of this paper goes well beyond China.

This paper demonstrates the potential of the toponym co-occurrence method to identify intercity relationships and highlights the positive impact of a strong network position on a city's performance, especially for smaller cities. Network externalities are more crucial in explaining productivity levels than agglomeration externalities, suggesting that fostering relationships between cities can serve as a viable alternative urban development strategy to stimulating further agglomeration.

The paper is structured as follows. First, we provide a brief overview of how network externalities benefit city performance, and discuss the toponym co-occurrence method (section 2). Second, we present our experiment, detailing the steps taken in the process (section 3) to obtain the pattern of relationships between Chinese cities. Third, we present and map this toponym co-occurrence network of China and analyze how it compares to theoretical predictions of these patterns using the gravity model (section 4). Fourth, we analyze the impact of network externalities on city performance in comparison to agglomeration externalities (section 5). Finally, we discuss our findings and how they translate into a research agenda that is both of theoretical and methodological interest (section 6).

2. Literature review

2.1 Network externalities versus agglomeration externalities

The discussion of agglomeration economies often traces back to the economist Alfred Marshall (Marshall, 1920), who argued that agglomeration creates a more efficient labor market and reduces transportation costs and explained how input sharing, labor market pooling and knowledge spillovers were drivers of agglomeration. More recent studies also find other sources, including home market effects (Krugman, 1980, 1995), where concentration of demand encourages agglomeration, and consumption-related effects (Glaeser et al., 2001; Waldfogel, 2008).

While agglomeration benefits are still considered the main driving force for urban development, fear of agglomeration costs such as congestion, pollution, high housing prices and large-scale social unrest often deters politicians and urban planners from actively supporting high urban concentrations (Au and Henderson, 2006; Wei et al., 2015). An optimal balance between agglomeration benefits and costs is what urban policy-makers often strive for.

Although in many early empirical studies, agglomeration externalities are defined as being geographically constrained (Rosenthal and Strange, 2004), it it now now well known that agglomerations are not islands and most cities interact at least to a certain degree (McCann and Acs, 2011). Therefore, a possible strategy to avoid agglomeration costs but still enjoy agglomeration benefits is to strengthen the relations between cities. This idea was already at the heart of Howard's Garden City concept (1898) and conceptualized in Alonso's (1973) idea of 'borrowed size'. Recent studies further clarified this concept and find that borrowed size is enabled through the interactions in networks of cities (Capello and Camagni, 2000; Meijers et al., 2016; Meijers and Burger, 2017). The general idea is that the city interactions provide potentials for cities to exploit increasing returns through co-operative activities and complementary sectoral specializations. This effect is referred to as network externalities, to differentiate from agglomeration externalities.

Spatial proximity is certainly an important factor that encourages the interactions between cities. However, with the advance of communication technologies such as online meeting platforms, spatial forms of proximity can be complemented by other types of proximity (Capello, 2020; Johansson and Quigley, 2004), such as organizational, institutional, cognitive and social proximity (Boschma, 2005). Moreover, urban economies develop from manufacturing industries to services in which creativity and innovation play an important role (Florida, 2005; Glaeser, 2011), and for instance distant networks may have become an additional important source of performance next to local interactions (Bathelt et al., 2004), which may even make network externalities better at stimulating innovations (Basile et al., 2012; Galaso and Kovářík, 2021). For instance, research in the USA (Schilling and Phelps, 2007), Europe (Sebestyén and Varga, 2013) and China (Yao et al., 2020) and worldwide (Belderbos et al, 2022) all found that firms well embedded in interfirm collaboration networks have higher innovative output as the network facilitates the

circulation of knowledge across a larger pool of sources. Numerous scholars (Florida, 2008; Ross et al., 2016; Lang et al., 2020) have presented the notion of globally interconnected cities as the primary drivers of economic growth.

Camagni et al. (2015) found that second tier cities in Europe can overcome the lack of agglomeration through innovation and city networks. However, regarding cultural amenities, Burger et al. (2015) find that size of a city still matters most—larger cities actually profit more than smaller cities, and in fact, often cast an 'agglomeration shadow' over smaller cities. While agglomeration economies remain important, several studies (Camagni et al., 2016; Cicerone et al., 2020; Meijers et al., 2016; Huang et al., 2020) show that stronger relations and connectivity to other regions and (larger) cities fosters development, and makes agglomeration benefits spill over to nearby smaller places. Similar results were also found for Japan (Otsuka, 2020), USA (Chatman and Noland, 2014) and China.

In China, many studies have used the high-speed railway network as a proxy representing the strength of city relationships (Niu and Li, 2018; Jiao et al., 2020; Huang et al., 2020). Generally, these show that there is a positive effect on the economic growth of cities by improving cities' connectivity and accessibility in the high-speed railway network. Shi and Pain (2020) collected the passenger, freight, and intercity capital flows within the Mid-Yangtze River city region (MYR) in China. They found that a city's economic growth is significantly related with a city's internal capital stock, labor cost and technology advances, but is also significantly linked with a city's network position in the three types of flows. However, it must be noted that while there may be generative effect (a general positive effect of transport infrastructure improvement) for the economy as a whole, this may hide a distributive effect too in the sense that some cities profit, whereas others lose out due to improved accessibility, and in its wake, increased competition (Meijers et al., 2012). This would be the case particularly in non-targeted peripheral cities, as they are driven to specialize more in agriculture activities and lose industrial output (Faber, 2014; Baum-Snow et al., 2020). Here, we may draw an analogy with agglomeration benefits and costs, as there are also network benefits and costs.

However, evaluating the relationship between agglomeration and network externalities remains challenging for several reasons: firstly,

agglomeration and network externalities, conceptually, are fuzzy concepts (Van Meeteren et al., 2016), making it difficult to disintangling due to their interconnected nature. Secondly, there are a various types of agglomeration and network externalities (Burger et al., 2014; Gross and Ouyang, 2021; Tongjing et al., 2022). Drucker (2012) posits that the rate at which agglomeration externalities diminish with distance depends on the industry, type of agglomeration, and type of externality examined. Thirdly, the significance of agglomeration (size) and network connectivity varies with metropolitan functions (Meijers et al., 2016; Phelps, 2021). From certain perspectives, network connectivity is more critical than size in determining the existence of specific metropolitan functions in cities, implying that networks could effectively substitute for size.

2.2 Geographic text analysis: toponym co-occurrences

Documents often contain geolocated-information such as place names, addresses, postal codes, etc. Enabled by the increasing digitalization of texts, geographical information retrieval and text analysis allows to summarize and analyze geographical information in texts (Gregory et al., 2015). By searching words that frequently co-occur with a place name, it is possible to identify certain characteristics of a place. For instance, one can combine place names and disease-related keywords to map mortality patterns (Porter et al., 2015). The toponym co-occurrence method searches for the co-occurrence of two or more place names. Two place names are considered as 'co-occurring' if both are mentioned in a predefined textual context. When cities are often mentioned together, it is assumed they are strongly related. The roots of the method in geography go back to the famous geographer Waldo Tobler who reconstructed the urban system of 119 pre-Hittite towns in Capadoccia 4,000 years ago based on toponym co-occurrences on cuneiform tablets (Tobler and Wineburg, 1971), which inspired his first law of geography "Everything is related to everything else, but near things are more related than distant things".

Toponym co-occurrences have been explored in various types of corpora, such as Wikipedia (Overell and Rüger, 2008; Devriendt et al., 2011; Salvini et al., 2016), newspapers (Liu et al., 2014), and web archives (Meijers and Peris, 2019). The frequency of the toponym co-occurrence is found to be positively correlated with the strength of these relationships (Ballatore et al., 2014; Liu et al., 2014), however, most of the applications are limited in toponym disambiguation (Overell and Rüger, 2008). Patterns found

correspond with known patterns of interaction, and a certain overlap with predictions based on gravity modelling adds to its plausibility (Meijers and Peris, 2019).

Regarding the application of toponym co-occurrence in China, some pioneering studies have shown the potential this method has for analyzing intercity relationships. Liu et al. (2014) applied this method to a search engine to investigate the relationship between geographical entities with data collected from Baidu, a Chinese Internet search engine. They found this method can be used to find similarities between neighboring provinces and to study the spatial organization of China. Zhong et al. (2017) further developed this method by applying complex network theory to evaluate the topological structures of the toponym occurrence network which was extracted daily from a newspaper over the course of a year. They found that the network showed strong cluster characteristics, and the frequency of toponym co-occurrence was negatively correlated with the administrative hierarchy, but less so with geographic distance. Guo et al. (2022) calculated a city's total appearance with other Chinese cities on the search engine Baidu, and examined the factors that can contribute to the frequency of this appearance, finding that factors such as GDP, administration level, tourism and the number of enterprises all significantly increase a city's appearance on the Internet.

3. Method

3.1 Data

The accuracy of the toponym co-occurrence method depends on the choice of the text corpus to which it is applied. This choice depends on the purpose of the analysis, as very specific text corpora can be used. Here, we want to obtain a very general picture of the Chinese urban system, which is why we chose to focus on all Chinese webpages as our text corpus. While it is possible to search for toponym occurrences using a search engine like google, results returned are vulnerable to bias (Meijers and Peris, 2019), and like Meijers and Peris we prefer using corpora from the CommonCrawl Archive of webpages as our text corpus. We used the entire April 2019 database for processing and conducting experiments. The original database we extracted contains about 6.98 TB of uncompressed text containing 2.5 billion web pages crawled between 18 and 26 April 2019. We selected all pages using at least 10 Chinese characters. The

filtered corpus contains approximately 110 billion Chinese words on 91 million pages from 1,067 different domains. Over 91% of the tokens are from websites registered under the four top-level domains (TLD): .com (62.23%), .cn (14.80%), .net (7.86%), and .org (2.68%). The four TLDs make up about 87.57% of pages.

3.2 Corpora preprocessing

To identify Chinese place names, accurate separation of Chinese words is a prerequisite. This is because Chinese words are often composed of more than one Chinese character, and a Chinese sentence is formed by consecutive Chinese characters, without any clear separation. Here we used a popular Chinese word-segmentation module named JIEBA for accurate word separation. The principle of this word-separation module is based on frequency-inverse document frequency (TF-IDF) method, which assigns an importance to words according to their frequency in a given corpora relative to the frequency of these words across the complete document set. The detailed processing is explained in Tongjing et al, 2024.

3.3 Gravity model benchmark

We selected the 293 cities in China at prefecture administration or higher. We subsequently retrieved the number of webpages where each pair of cities co-occur, studying 42,778 pairs in total (which means it is a fully-connected network). As could be expected, the frequency of toponym co-occurrences is highly correlated with population size of the places involved. For our later analysis it is essential to disentangle city network and agglomeration effects, and measure the relative rather than absolute strength of relationships a city has. We decided to follow Meijers and Peris' approach, thus comparing absolute frequencies of co-occurrences with expected frequencies as predicted by the gravity model.

In the simplest form of the gravity model, the interaction of place i and place j is proportional to the product of place i and j, and inversely proportional to the distance between the two places:

$$I_{ij} = K\frac{M_i^{\beta_1}M_j^{\beta_2}}{D_{ij}^{\beta_3}}$$

where I_{ij} is the total relationship, K is the constant, M_i and M_j are the population sizes of place i and j, respectively, D_{ij} is the physical distance

between the two places, β_1 and β_2 reflects the ability of place i and j to attract flows, β_3 reflecting the rate of increase in the friction of distance.

We first fit the gravity model with the toponym co-occurrence results to estimate the strength of intercity relationship by controlling the size and distance. And then the estimated relationship is set as benchmark to determine whether the toponym co-occurrence between two cities is stronger or weaker than the gravity model estimation. The link weight of the intercity relationship is defined as the ratio of toponym co-occurrence to gravity model predicted results:

$$w_{ij} = \frac{T_{ij}}{I_{ij}}$$

where w_{ij} is the relative strength of intercity relationship between city i and city j (which we refer to as 'link weight'), T_{ij} is the actual toponym co-occurrence of city i and city j, and I_{ij} is the estimated co-occurrence of city i and city j in the gravity model. After transformation to percentages, a value higher than 100% means the relationship is stronger than expected given size and distance between the cities. Obviously, a value less than 100% indicates that the relationship is weaker than expected.

To measure a city's network position, we use a network topological attribute, namely average relatedness. Average relatedness (AR) is a concept of average node strength borrowed from network theory (Newman, 2010). It is the sum of link weight associated with city i divided by the number of links city i connects with, indicating the average relationship level of a city:

$$AR_i = \frac{\sum_{j \in N(i)} w_{ij}}{N - 1}$$

where $\mathcal{N}(i)$ is the set of the link weights of city i, N is the number of links city i connects with in the network, and w_{ij} is the relative strength of intercity relationship between city i and city j, as defined in 3.3. A value higher than 1 indicates that on average, the city is relatively more related to other cities than expected according to the gravity model.

3.4 City performance assessment

Next, we proceed to models investigating if there is a significant association between certain kinds of city network positions and city productivity when controlling for factors such as land (L), human capital (H), and investment (K). A city's productivity is measured as the city's product (Q) divided by the city's population (P). The Cobb-Douglas production function is a widely used standard approach for estimating city performance, as noted in previous studies (Melo et al., 2009; Meijers and Burger, 2010; Bird et al., 2020). Based on the linear Cobb-Douglas production function, the city productivity (Q/P) is specified as:

$$In\left(\frac{Q}{P}\right) = \theta_0 + \kappa In\left(\frac{K}{P}\right) + \theta In\left(\frac{H}{P}\right) + vIn\left(\frac{L}{P}\right) + hIn(IS) + \alpha In(\mathrm{AR}) + \beta In(P) + \mathrm{r}$$

In this function, average relatedness (AR) is the network externality variable, population size (P) captures agglomeration externalities, and investment per capita $\left(\frac{K}{P}\right)$, human capital per capita $\left(\frac{H}{P}\right)$, land per capita $\left(\frac{L}{P}\right)$ and industrial structure (*IS*) are control variables, while 'r' represents the remaining differences in total factor productivity. We added the industrial structure variable, measured by the ratio of secondary-tertiary industry (Barro, 1996), to account for the impact of industrial composition on productivity. Their measurement is discussed in the following subsection.

3.5 Data and variables

To estimate our productivity model, data was gathered from the 2019 China Statistical Yearbook for Cities, provincial statistical yearbooks, Hong Kong 2019 Annual Digest of Statistics, and Macau 2019 Yearbook of Statistics, which all provided city statistics for 2018.

The dependent variable nominal output (Q) was measured as the 2018 GDP in real Chinese yuan (CNY) of each individual prefecture-level city and its administered counties. The GDP of Hong Kong and Macau were converted into Chinese yuan based on the 2018 yearly average exchange rate.

As is common (Melo et al., 2009), we measure the presence of agglomeration externalities with a variable capturing city size (P), which is measured as the size of the total year-end permanent residents in 2018. There are four control variables in our model. This includes human capital per capita $\left(\frac{H}{P}\right)$, which is often based on the average years of schooling (Benhabib and Spiegel, 1994), but such data is generally not available in China. Another important factor is that China's big cities' populations have a large proportion of migrants. For instance, of Shanghai's 24 million residents, 9 million (or over 39%) are long-term migrants. Thus, tabulating the number of years in school for local residents only provides a fraction of the actual education level in these big cities. Instead, a positive relationship between patents and human capital has been identified and extensively researched (Li and Jiang, 2016; Li and Phelps, 2019). Here we use the number of patent authorizations as a proxy for human capital (H) from the statistical yearbooks. A second control variable is data on the investment per capita $\left(\frac{K}{P}\right)$, which is based on the government expenditure to reflect the internal intervention of local governments (Li et al., 2019). The third control variable is land per capita $\left(\frac{L}{P}\right)$, which is captured through the size of the built-up area of a city. The fourth control variable is the industrial structure, which is measured by the ratio of secondary-tertiary industry (Barro, 1996). Descriptive statistics are provided in Table 1 and correlations in Figure 1.

Table 1. Descriptive statistics

All cities

Variable	Obs	Mean	S.D	Min	Max
GDP/population (ln)	293	10.88	0.59	8.99	13.32
Investment/population (ln)	293	8.30	0.85	4.65	11.87
Human capital/population (ln)	293	1.79	1.28	-3.56	4.68
Land/population (ln)	293	-10.43	0.71	-15.15	-8.31
Industrial structure (ln)	293	0.11	0.45	-1.01	3.13
Agglomeration: City size (ln)	293	5.84	0.75	3.13	8.04
Network position: Average relatedness (ln)	293	0.03	0.17	-0.36	0.65

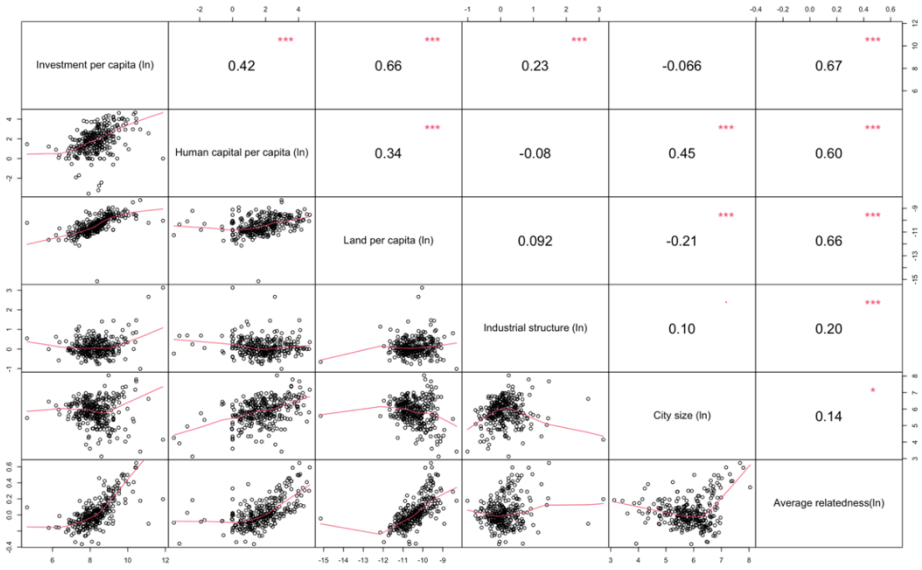
Cities in three major megaregions

Variable	Obs	Mean	S.D	Min	Max
GDP/population (ln)	51	11.39	0.61	10.32	13.32
Investment/population (ln)	51	8.92	0.94	7.33	11.87
Human capital/population (ln)	51	3.12	1.09	0.00	4.68
Land/population (ln)	51	-10.24	0.54	-11.41	-9.08
Industrial structure (ln)	51	0.24	0.65	-0.44	3.13
Agglomeration: City size (ln)	51	6.29	0.70	4.15	7.79
Network position: Average relatedness (ln)	51	0.15	0.20	-0.11	0.65

Cities outside megaregions

Variable	Obs	Mean	S.D	Min	Max
GDP/population (ln)	242	10.77	0.53	8.99	12.59
Investment/population (ln)	242	8.17	0.77	4.65	10.67
Human capital/population (ln)	242	1.51	1.14	-3.56	3.95
Land/population (ln)	242	-10.47	0.73	-15.15	-8.31
Industrial structure (ln)	242	0.08	0.40	-1.01	1.45
Agglomeration: City size (ln)	242	5.75	0.73	3.13	8.04
Network position: Average relatedness (ln)	242	0.00	0.15	-0.36	0.50

Figure 1. Correlation matrix.



Significance codes: 0 '***' 0.001 '**' 0.01 '*'

4. The Chinese urban system

4.1 Absolute patterns of toponym co-occurrences

We explored how often pairs of Chinese city names are mentioned 'in one breath' on websites, the resulting network is presented in Figure 2. For readability, we only present the top20% most frequently occurring pairs of cities.

Figure 2 shows that, in absolute terms, most of the relationships concentrate in the southeastern side of China where the Chinese population also concentrates, and the well-known "Hu Huanyong Line" divide is visible (Chen et al., 2016). Our results reveal that the strongest relationships exist between five city clusters (thus, also clusters of relationships) that are highlighted in yellow. These relationships together form a diamond shape where the outline is formed by five city clusters: Yangtze River Delta (Shanghai, Hangzhou, and Nanjing) in the east; Pearl River Delta (Guangzhou, Shenzhen, and Hong Kong) in the south; Chengyu Region (Chongqing and Chengdu) in the west; the Beijing-Tianjin-Hubei Region (Beijing and Tianjin) in the north; and the Middle-Yangzte River Region (Wuhan) in the center. It is noteworthy that all five city clusters are classified as top-tier city clusters in the latest China fourteen-five-year plan, which emphasizes the importance of developing interconnected and interdependent city clusters to promote sustainable economic growth and regional development. Importantly, this diamond pattern is also found in the gravity economic index map plotted by Lao et al. (Lao et al., 2016), in the social connection map plotted by Feng et al. (2012) using Sina microblog data, a social medium or in the patterns between city pairs using Baidu search engine (Guo et al., 2022).

We also summed up the toponym co-occurrences a city has with the other cities to represent a city's "attraction" . As shown in Figure 3, there is a positive correlation between city "attraction" and its GDP rank. This verifies the previous study activities (Guo et al., 2022) that cities with more affluence or higher populations appear more frequently in news reports, social media, and business, and apparently are also mentioned more often in the same breadth with other cities. However, we also observed that cities with high administration levels tend to have higher "attraction" than normal prefectural cities with similar GDP rankings. Notably, Hong Kong and Macau, two highly prosperous cities, had relatively lower aggregated

toponym co-occurrence compared to cities with similar GDP rankings. One possible explanation is that both cities are Special Administrative Regions, characterized by independent legal and custom systems, resulting in fewer connections with mainland China. Additionally Figure 3 also shows that cities with higher populations usually have more relations. The top ten cities in terms of relationship strength are listed in Table 2. As can be seen, the network position of those cities is not very dissimilar.

Figure 2. The visualization of top 20% toponym co-occurrence in China

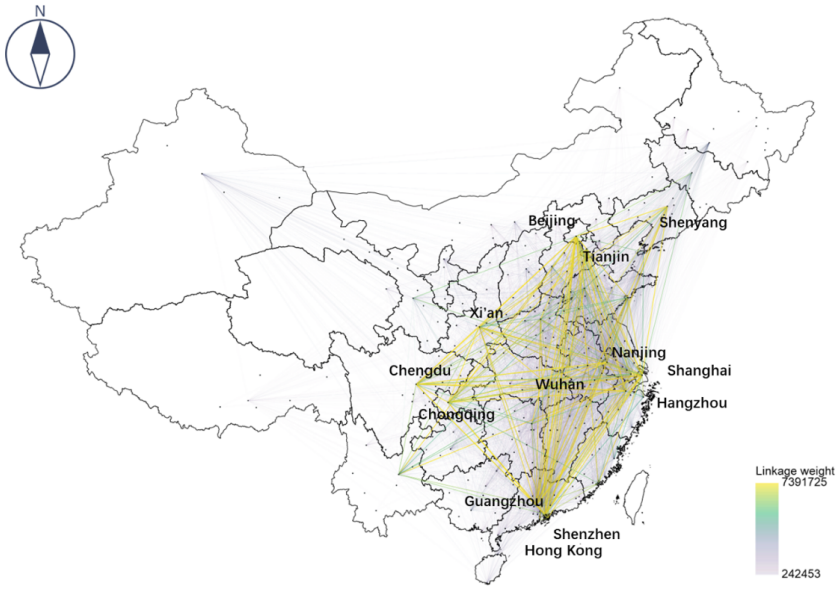


Figure 3. Cities' aggregated toponym co-occurrence versus GDP rank and population size

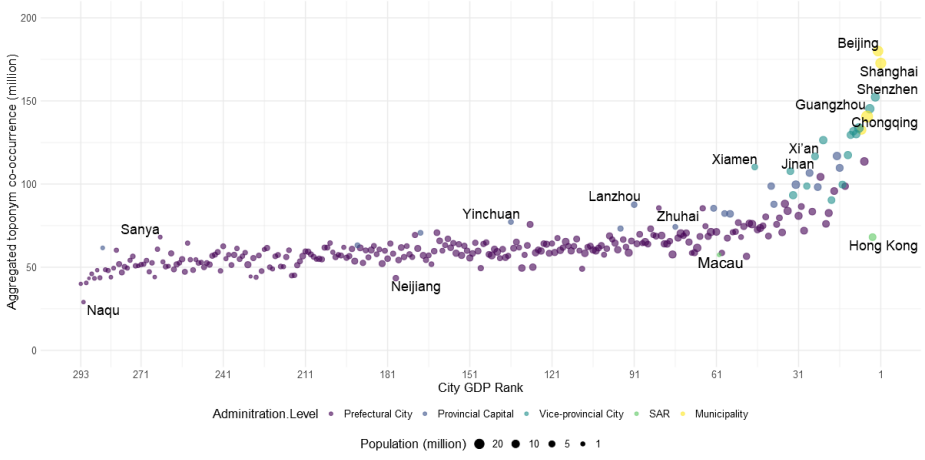


Table 2. Top 10 China aggregated toponym co-occurrence

City	City aggregated toponym co-occurrence
Beijing	180,015,350
Shanghai	172,740,260
Shenzhen	152,324,365
Guangzhou	145,335,760
Chongqing	140,614,295
Chengdu	133,734,425
Tianjin	132,567,892
Hangzhou	131,810,662
Wuhan	130,094,692
Nanjing	129,555,664

4.2 Relative pattern of toponym co-occurrences

In 4.1 we found that the population size appears to have a significant impact on the absolute strength of intercity relationship. Building upon this finding, our more comprehensive analysis in Table 3 indicates that the gravity model reasonably captures the absolute pattern of toponym co-occurrences.

Table 3. Gravity model result.

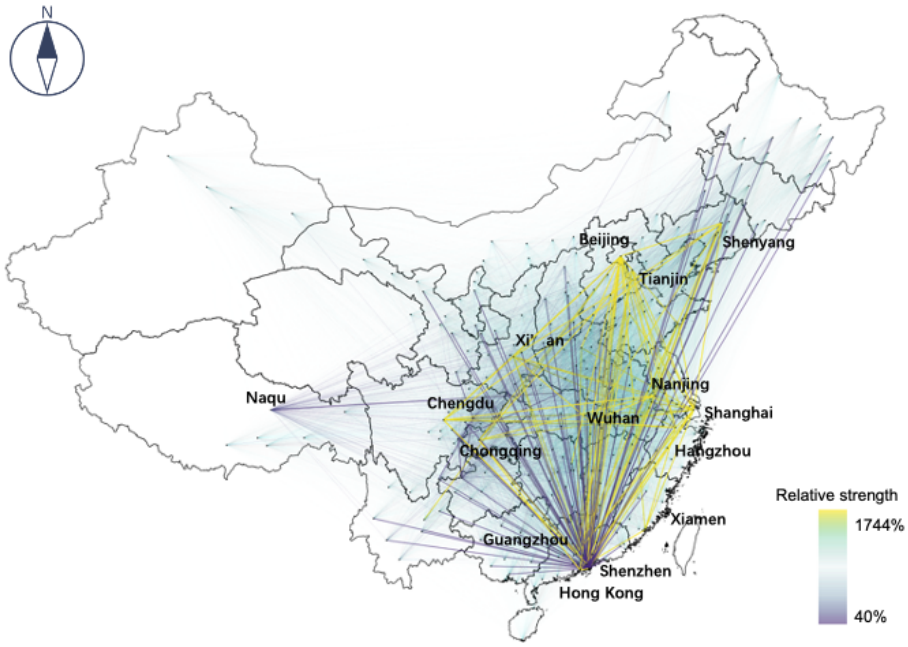
Gravity Model	
Pop. i (ln)	0.192 (0.002) ***
Pop. j (ln)	0.180 (0.002) ***
D_{ij} (ln)	-0.046 (0.002) ***
Number of observations	42,774
Adjusted R^2	0.397
P value	<2.2e-16
F-statistics	9370
Root MSE	0.257

Significance codes: 0 '***' 0.001 '**' 0.01 '*'. Standard error in parentheses.

The adjusted R^2 indicates that about 40% of the variation of the toponym co-occurrence frequency can be explained by the two cities' populations and the distance. For comparison, Meijers and Peris (2019) found that the gravity model can explain 56% of the variation in toponym co-occurrences of places with populations over 10,000 in the Netherlands.

The visualization of the relative strength of relationships is shown in Figure 4. For each pair of cities, we calculated the ratio between the observed actual toponym co-occurrence value and the expected value (based on the gravity model). For interpretation: 100% means that both are similar, 50% means that only half of the expected co-occurrences were found. A similar diamond pattern between the top linkages can still be identified. It means the relationships between those top administration cities cannot be fully explained just by population or distance.

Figure 4. Relative strength of relationships between Chinese cities.

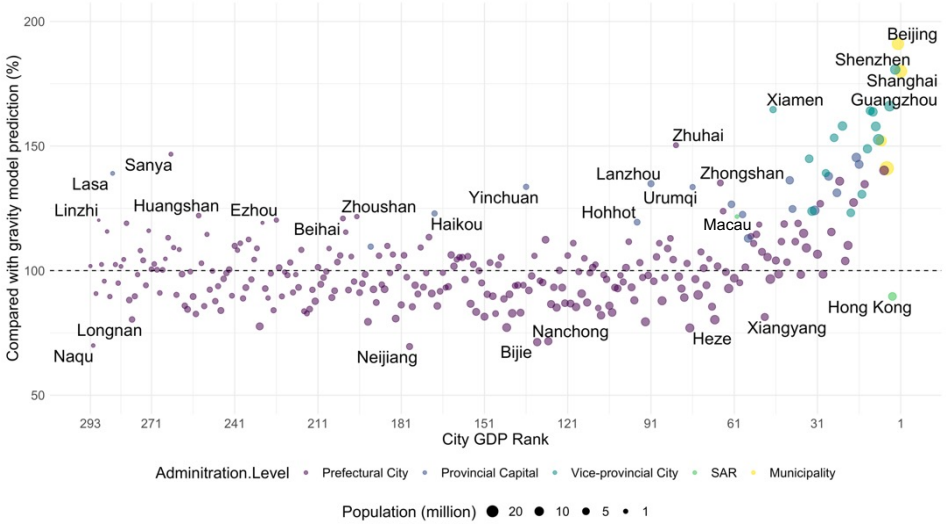


Before comparing city network externalities with agglomeration externalities, we first visualized the relationship between a city's GDP rank, population and average relatedness in Figure 5. Recall that a city's average relatedness (AR) is defined at the city (not city pair) level and concerns the ratio of observed toponym co-occurrences to gravity model predicted values. For interpretation: 100% means that both are similar, 50% means that only half of the co-occurrences expected was found. As presented in Figure 5, a positive correlation between network position and GDP rank can be observed. However, it also shows that cities with high administration levels generally have higher network strength than normal prefecture cities with similar GDP rankings.

It is interesting to compare Shenzhen, Hong Kong and Macau in this Figure. Shenzhen and Hong Kong are both top tier cities in China, however, their relationships with other cities vary significantly. Some of the largest negative residuals are linked with Hong Kong, which indicates a relatively weak position of Hong Kong in the Chinese urban system. In sharp contrast, most linkages with Shenzhen are positive. It is also interesting to compare between Hong Kong and Macau, both special administrative regions.

This shows that Hong Kong, a special administration region in China with its own judicial power and customs, is not yet fully integrated in the Chinese urban system: the strength of relationships is clearly moderated by border effects (Capello et al., 2018; Sohn et al., 2022). However, when compring Hong Kong and Macau, both are Special Administrative Regions, Macau seems to be much more integrated than Hong Kong, probably due to the size.

Figure 5. Observed toponym co-occurrences compared with gravity model prediction



4.3 City networks vis-à-vis agglomeration

In this section we turn to the discussion whether the position in networks between cities matters for the performance of cities, and assess the importance of such network externalities in comparison to agglomeration externalities. In our models 1-2-3-4-5 (see Table 4), we include all cities and all variables, but also focus on just our variables of interest capturing

agglomeration and network externalities, including the interaction effect variable between agglomeration and network externalities (because of some collinearity between our network variable and controls - recall Figure 1). Next to the overall picture for all cities, and as a robustness check, we explore whether effects differ for subregions, thereby running regressions for cities in the three megaregions (Yangtze River Delta, Pearl River Delta and Jingjinji) and for all other cities in the remainder of China in models 6-9.

In general, as shown in Model 1-4, average relatedness is positively associated with city performance. The positive sign indicates that being well related to other cities enhances productivity. This effect is repeatedly found in nearly all other model specifications which suggests that it is a rather robust finding.

Nevertheless, when comparing Model 1-2-3-4-5, the network variable shows significant association in all models, while the agglomeration variable is only significant when it is entered just by itself and can only explain negligible variance of productivity when compared with network variable results. Surprisingly, though, this effect runs counter to what the literature on agglomeration externalities predicts, as the agglomeration effect is commonly considered key for productivity (Camagni et al., 2016; Cicerone et al., 2020). A possible explanation is that agglomeration costs often outweigh agglomeration benefits in the Chinese cities, which is actually a key reason for the Chinese government to pursue a polycentric development strategy in combination with fostering relations between cities. Rather than agglomeration externalities driving productivity, it seems that network externalities are a much more relevant productivity enhancing mechanism.

Regarding the interaction effect between agglomeration (city size) and network externalities, Model 5 indicates a negative and significant relationship. When city size increases, the importance of relations for performance decreases. Vice versa, when population size decreases, the importance of relatedness for performance becomes stronger. It confirms previous studies that the benefits of network externalities diminish in large cities (Li et al., 2019; Meijers and Burger, 2010). At the same time, relatively small cities that have stronger than expected relations with other cities have higher productivity — they probably can use their networks to compensate for a (comparative) lack of agglomeration externalities. These

results suggest that relations with other cities allow smaller and medium-sized cities to 'borrow size' through those relations.

Table 4. Results of regression models on metropolitan productivity

Model	Model 1 Network	Model 2 Agglomeration	Model 3 Network & Agglomeration	Model 4 All included	Model 5 Interaction
Intercept	10.82 *** (0.03)	10.02 * (0.39)	10.38 *** (0.32)	8.08 *** (0.91)	8.40 *** (0.98)
Investment – per capita (ln)				0.26 *** (0.07)	0.27 *** (0.07)
Human capital – per capita (ln)				0.17 *** (0.03)	0.16 *** (0.03)
Land– per capita (ln)				0.06 (0.05)	0.06 (0.05)
Industrial structure (ln)				-0.08 (0.12)	-0.08 (0.11)
Population size (ln)		0.15 * (0.06)	0.07 (0.05)	0.04 (0.04)	0.08 * (0.04)
Average relatedness (ln)	2.29 *** (0.14)		2.24 *** (0.15)	0.52 * (0.28)	2.86 * (1.13)
Population size (ln)* Average relatedness (ln)					-0.36 * (0.18)
Adjusted R^2	0.45	0.03	0.46	0.65	0.65
F-statistic	241.30	10.52	124.6	89.54	91.43
Root MSE	0.44	0.58	0.43	0.35	0.34
VIF Mean			1.02	2.14	
N	293	293	293	293	293
Robust standard error in parentheses, Significant codes: 0 '***' 0.001 '**' 0.01 '*'					

As a robustness check, we split the sample to compare cities in megaregions (Table 5, models 6-7) and cities that are not located in megaregions (Models 8-9).

The results show that in models 6-7, including just cities in the three megaregions, the average relatedness has no significant association with the city performance, but for cities in the remainder of China it has a positive effect (models 8-9). A possible explanation is that the average relatedness represents an overall relatedness at the national level. For most Chinese cities a greater nation-wide relatedness brings higher performance. However, cities in megaregions seem less dependent on their national network position. There are several large cities in each megaregion, such as Shanghai, Nanjing, Hangzhou in Yangtze River Delta, Beijing and Tianjin in Jingjinji, which may function as gateway to other cities at the national scale. Probably, for the smaller cities in megaregions, the relation with their larger neighbour is primarily of importance, allowing them to

borrow their agglomeration benefits, whereas relations with cities beyond their megaregion are only secondary. Another contrast between cities in megaregions and those outside the three main megaregions concerns the effect of population size (or, agglomeration). It tends to be not significant for cities in megaregions, but population size is positively associated with productivity for cities outside those megaregions. One possible explanation is that in more highly related megaregions (recall Table 1, with a higher average relatedness for cities in megaregions), the complex interplay between agglomeration and networks produces the 'borrowed size'-effects and 'agglomeration shadows' discussed in section 2, which disrupt the more traditional patterns that we still see for cities outside of megaregions.

Table 5. Results of regression models by regions

Model	Model 6	Model 7	Model 8	Model 9
	Cities in Megaregions – all included	Cities in Megaregions – interaction	Cities in remainder China – all included	Cities in remainder China – interaction
Intercept	-0.08 (0.07)	-0.12 (0.08)	-0.05 (0.07)	-0.06 (0.07)
Investment – per capita (ln)	0.64 ** (0.18)	0.61 *** (0.16)	0.23 *** (0.07)	0.23 *** (0.07)
Human capital – per capita (ln)	0.41 *** (0.10)	0.40 *** (0.09)	0.27 *** (0.07)	0.26 *** (0.07)
Land– per capita (ln)	-0.01 (0.08)	-0.00 (0.11)	0.24 ** (0.09)	0.23 ** (0.08)
Population size (ln)	-0.12 (0.13)	-0.05 (0.13)	0.13 * (0.05)	0.15 ** (0.06)
Industrial structure (ln)	0.28 * (0.16)	0.30 * (0.13)	-0.23 *** (0.04)	-0.23 *** (0.04)
Average relatedness (ln)	-0.10 (0.12)	0.39 (0.66)	0.23 *** (0.06)	0.53 * (0.25)
Population size (ln)* Average relatedness (ln)		-0.53 (0.70)		-0.29 (0.25)
Adjusted R^2	0.78	0.81	0.64	0.64
F-statistic	30.88	26.30	69.54	59.94
Root MSE	0.09	0.09	0.09	0.09
VIF Mean	3.37		1.97	
N	51	51	242	242
Robust standard error in parentheses, Significance codes: 0 '***' 0.001 '**' 0.01 '*'				

5. Conclusion

Around the world, there is growing interest in development strategies aimed at integrating cities more strongly. The underlying assumption is

that a city's performance is not only determined by agglomeration externalities but also, and perhaps increasingly by its network position (Sassen, 2007; Huang et al., 2020). While many studies have used rather specific data to capture city network externalities, we proposed and used a generic measure of city network embeddedness derived from toponym co-occurrences in an enormous text corpus of all websites in Chinese. Previous studies have shown that the toponym co-occurrence method is able to (re)construct networks of relationships between cities, and given the increasing availability of text corpora, be it digitalized historical archives or archives of contemporary resources like Web Archives, such methods are very promising. Using this method, and computational social science, we extracted intercity relationships at an unprecedented large scale, detailing over 42,000 intercity relationships in China. Although this method is used in China as we studied the Chinese city relationship, it is also suitable to study other countries by searching their local languages.

The toponym co-occurrence method is presented in this paper, and its results mapped and analyzed. Gravity modelling is used to gain an understanding of the relative network position of cities in China. This all served the bigger ambition of this paper, namely to explore the importance of network externalities in comparison to agglomeration externalities. In line with Huang et al. (2020), a strong network position was found to be positively associated with a city's performance, emphasizing the existence of network externalities. This importance was greater for the relatively smaller Chinese cities among the almost 300 largest cities studied. In comparison to agglomeration externalities, network externalities were much more important in explaining productivity levels.

This finding begs for more theorizing as the urban and regional development literature is dominated by urban triumph narratives in which agglomeration benefits are considered the key driver of growth. Quite in contrast, we find that a relational perspective on urban growth and decline is a more promising avenue, and this calls for a better understanding of urban network externalities. What we have seen in this paper is that larger cities profit less from their network position, whereas well-performing smaller cities seem to compensate a lack of size with a good network position. Of course, in this debate we need to discuss whether our findings only hold for China, or have wider relevance, but findings for Europe do not seem to be dissimilar (Meijers and Burger, 2016). An essential theme for further research is how agglomeration externalities and network

externalities relate to each other, and perhaps even may be dependent on each other. Disentangling a potential recursive relationship should be a key concern for future studies.

This also has important implications for policy making. Strategies to make Chinese cities more competitive and productive should not be foremost oriented at a further concentration of people and firms in space. Instead, our results suggest that in particular smaller- and medium-sized cities can better gain competitiveness from being strongly related to other cities in the country. Those that are better embedded in all kinds of functional, political, cultural, economic, academic and social networks do perform better. Policies could target the institutions and infrastructures that allow for such networks to develop. There is no a priori reason to assume that such a general policy strategy should be different in other countries. Further research is needed to specify which types of relationships are most important for positive city network externalities (see Schweitzer, et al., 2009).

Methodologically speaking, the toponym co-occurrence method is still in its infancy in geography and needs to be more widely tested and applied as a way to measure interurban relatedness. Conceptually, it would be good to know which share of co-occurrences generally involves some physical transfer of for instance people or goods, and which share is more symbolic. Particular attention needs to be devoted to how relationships between cities that co-occur frequently in texts should be interpreted. A logical next step is to classify such relationships. Also the application of the method to different text corpora is recommended. A comprehensive database, such as Common Crawl, may indicate a general city-network pattern but detailed analysis requires an understanding of the actual meaning of the relationships. Perhaps a more targeted data source, such as newspapers, social media, or Wikipedia, may provide a clearer city network. Alternatively, the use of machine-learning techniques to classify linkages, as suggested by Meijers and Peris (2019), deserves attention. In terms of modelling, while we included the main factors in urban productivity, future studies could expand the range of factors included to see whether our results still hold, thereby including evolutionary, behavioural and institutional factors. Also the endogeneity issue deserves more attention. A complex interplay between agglomeration and network externalities on the one hand, and labor productivity on the other exists. While we theoretically argue that being well positioned in networks allows to profit (‘borrow size’)

from production factors located in other cities with which one city is well-connected, it is not unthinkable that highly productive cities develop more relations, because of their productivity.

While the research agenda still ahead call for some caution in drawing final conclusions, it is nevertheless obvious that city network externalities deserve much more attention, both in theory and in empirical research. Methodologically, the text mining of digital text corpora, still a nascent research approach in economic geography, seems to hold much promise and needs to be further developed potentially enabling a better understanding of urban and regional development.

References

Alonso, W. (1973). Urban zero population growth. Daedalus, 102: 191–206.

Au, C-C., & Henderson, J.V. (2006). How migration restrictions limit agglomeration and productivity in China. Journal of Development Economics, 80: 350–388.

Ballatore, A., Bertolotto, M., & Wilson DC (2014). An evaluative baseline for geo-semantic relatedness and similarity. GeoInformatica, 18: 747–767.

Basile, R., Capello, R., & Caragliu, A. (2012). Technological interdependence and regional growth in Europe: Proximity and synergy in knowledge spillovers. Papers in Regional Science, 91(4), 697-722.

Bathelt, H., Malmberg, A., & Maskell, P. (2004). Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation. Progress in Human Geography, 28: 31–56.

Barro, R.J. (1996). Determinants of economic growth: A cross-country empirical study. MIT Press Books.

Baum-Snow, N., Henderson, J.V., Turner, M.A., Zhang, Q., & Brandt, L. (2020). Does investment in national highways help or hurt hinterland city growth? Journal of Urban Economics, 115, 103124.

Belderbos, R., Benoit, F., & Derudder, B. (2022). World City Innovation and Service Networks and Economic Growth. Papers in Regional Science. DOI: 10.1111/pirs.12687.

Benhabib, J., & Spiegel, M.M. (1994). The role of human capital in economic development evidence from aggregate cross-country data. Journal of Monetary Economics, 34: 143–173.

Bird, J., Lebrand, M., & Venables, A. J. (2020). The belt and road initiative: Reshaping economic geography in Central Asia?. Journal of Development Economics, 144, 102441.

Boschma, R.A. (2005). Proximity and innovation: A critical assessment. Regional Studies, 39: 61–74.

Burger, M., & Meijers, E. (2016). Agglomerations and the rise of urban network externalities. Papers in Regional Science, 95: 5–15.

Burger, M., Meijers, E., Hoogerbrugge, M., & Masip Tresserra, J. (2015). Borrowed size, agglomeration shadows and cultural amenities in North-West Europe. European Planning Studies, 23: 1090–1109.

Burger, M. J., Meijers, E. J., & Van Oort, F. G. (2014). Multiple Perspectives on Functional Coherence: Heterogeneity and Multiplexity in the Randstad. Tijdschrift voor economische en sociale geografie, 105(4), 444-464.

Camagni, R., Capello, R., & Caragliu, A. (2015). The rise of second-rank cities: what role for agglomeration economies? European Planning Studies, 23: 1069–1089.

Camagni, R., Capello, R., & Caragliu, A. (2016). Static vs. dynamic agglomeration economies. Spatial context and structural evolution behind urban growth. Papers in Regional Science, 95: 133–158.

Capello, R. (2000). The city network paradigm: Measuring urban network externalities. Urban Studies, 37: 1925–1945.

Capello, R. (2020). Proximity and regional competitiveness. Scienze Regionali, 19(3), 373-394.

Capello, R., & Camagni, R. (2000). Beyond optimal city size: an evaluation of alternative urban growth patterns. Urban studies, 37: 1479–1496.

Capello, R., Caragliu, A., & Fratesi, U. (2018). Breaking down the border: Physical, institutional and cultural obstacles. Economic Geography, 94(5), 485-513.

Chatman, D., & Noland, R. (2014). Transit Service , Physical Agglomeration and Productivity in US Metropolitan Areas. Urban Studies, 51: 917–937.

Chen, M., Gong, Y., Li, Y., Lu, D., & Zhang, H. (2016). Population distribution and urbanization on both sides of the Hu Huanyong Line: Answering the Premier's question. Journal of Geographical Sciences, 26(11), 1593–1610.

Cicerone, G., McCann, P., & Venhorst, V. (2020). Promoting regional growth and innovation: relatedness, revealed comparative advantage and the product space. Journal of Economic Geography, 20: 293–316.

Dai, L., Derudder, B., Cao, Z. & Ji, Y. (2022). Examining the evolving structures of intercity knowledge networks: the case of scientific collaboration in China, International Journal of Urban Sciences, DOI: 10.1080/12265934.2022.2042365

Devriendt, L., Derudder, B., & Witlox, F. (2008). Cyberplace and cyberspace: two approaches to analyzing digital intercity linkages. Journal of Urban Technology, 15(2), 5-32.

Faber, B. (2014). Trade integration, market size, and industrialization: evidence from China's National Trunk Highway System. Review of Economic Studies, 81(3), 1046-1070.

Fang, C., & Yu, D. (2017). Urban agglomeration: An evolving concept of an emerging phenomenon. Landscape and Urban Planning, 162: 126–136.

Fang, C., Yu, X., Zhang, X., Fang, J., & Liu, H. (2020). Big data analysis on the spatial networks of urban agglomeration. Cities, 102, 102735.

Feng, Z., Bo, W., & Yingxue, C. (2012). China's city network characteristics based on social network space: An empirical analysis of sina micro-blog. Acta Geographica Sinica, 67: 1031-1043.

Florida, R. (2005). Cities and the Creative Class. New York: Routledge.

Galaso, P., & Kovářík, J. (2021). Collaboration networks, geography and innovation: Local and national embeddedness. Papers in Regional Science, 100(2), 349-377.

Glaeser, E. (2011). Triumph of the City. The Penguin Press.

Glaeser, E., Kolko, J., & Saiz, A. (2001). Consumer city. Journal of Economic Geography, 1: 27–50.

Glaeser, E., Ponzetto, G., & Zou, Y. (2016). Urban networks: Connecting markets, people, and ideas. Papers in Regional Science, 95(1), 17-59.

Gross, J., & Ouyang, Y. (2021). Types of urbanization and economic growth. International Journal of Urban Sciences, 25(1), 71-85.

Gordon, I., & McCann, P. (2000). Industrial clusters: complexes, agglomeration and/or social networks? Urban studies, 37: 513–532.

Gregory, I., Cooper, D., Hardie, A., & Rayson, P. (2015). Spatializing and analyzing digital texts: Corpora, GIS and places. Spatial narratives and deep maps, 150-78.

Guo, H., Zhang, W., Du, H., Kang, C., & Liu, Y. (2022). Understanding China’s urban system evolution from web search index data. EPJ data science, 11(1), 20.

Harrison, J., & Gu, H. (2019). Planning megaregional futures: spatial imaginaries and megaregion formation in China. Regional Studies, 55: 77-89.

Hill, L. (2009). Georeferencing: The Geographic Associations of Information. Cambridge, MA: MIT Press.

Howard, E. (1898). Garden Cities of To-morrow. London: Swan Sonnenschein and Co. Ltd.

Hu, X., Wang, C., Wu, J., & Stanley, H. E. (2020). Understanding interurban networks from a multiplexity perspective. Cities, 99, 102625.

Huang, Y., Hong, T., & Ma, T. (2020). Urban network externalities, agglomeration economies and urban economic growth. Cities, 107, 102882.

Jiao, J., Wang, J., Zhang, F., Jin, F., & Liu, W. (2020). Roles of accessibility, connectivity and spatial interdependence in realizing the economic impact of high-speed rail: Evidence from china. Transport Policy, 91, 1-15.

Johansson, B., & Quigley, J. (2004). Agglomeration and networks in spatial economics. Papers in Regional Science, 83:165–176.

Krugman, P. (1980). Scale economies, product differentiation, and the pattern of trade. The American Economic Review, 70: 950-959.

Krugman, P. (1995). Increasing returns, imperfect competition and the positive theory of international trade. Handbook of International Economics, 3: 1243-1277.

Lang, R. E., Lim, J., & Danielsen, K. A. (2020). The origin, evolution, and application of the megapolitan area concept. International Journal of Urban Sciences, 24(1), 1-12.

Lao, X., Zhang, X., Shen, T., & Skitmore, M. (2016). Comparing China's city transportation and economic networks. Cities, 53: 43-50.

Li, J., & Jiang, Y. (2016). Calculation and Empirical Analysis on the Contributions of R&D Spending and Patents to China's Economic Growth. Theoretical Economics Letters, 6, 1256.

Li, W., Sun, B., & Zhang, T. (2019). Spatial structure and labour productivity: Evidence from prefectures in China. Urban Studies, 56: 1516-1532.

Li, Y., Phelps, N. (2019). Megalopolitan glocalization: the evolving relational economic geography of intercity knowledge linkages within and beyond China's Yangtze River Delta region, 2004-2014. Urban Geography, 40: 1310-1334.

Liu, Y., Wang, F., Kang, C., Gao, Y., & Lu, Y. (2014). Analyzing Relatedness by Toponym Co-Occurrences on Web Pages. Transactions in GIS, 18: 89-107.

Ma, H., & Xu, X. (2022). The effects of proximities on the evolving structure of intercity innovation networks in the Guangdong–Hong Kong–Macao Greater Bay Area: comparison between scientific and technology knowledge, International Journal of Urban Sciences, DOI: 10.1080/12265934.2022.2085154

Marshall, A. (1920). Principles of Economics. 8th Edition, Macmillan, London.

McCann, P., & Acs, Z. (2011). Globalization: countries, cities and multinationals. Regional Studies, 45: 17-32.

Meijers, E., & Peris, A. (2019). Using toponym co-occurrences to measure relationships between places: review, application and evaluation. International Journal of Urban Sciences, 23: 246-268.

Meijers, E., & Burger, M. (2010). Spatial structure and productivity in US metropolitan areas. Environment and Planning A, 42: 1383-1402.

Meijers, E., & Burger, M. (2017). Stretching the concept of 'borrowed size'. Urban Studies, 54: 269-291.

Meijers, E., Burger, M., & Hoogerbrugge, M. (2016). Borrowing size in networks of cities: City size, network connectivity and metropolitan functions in Europe. Papers in Regional Science, 95: 181-198.

Meijers, E., Hoekstra, J., Leijten, M., Louw, E., & Spaans, M. (2012). Connecting the periphery: Distributive effects of new infrastructure. Journal of Transport Geography, 22, 187-198.

Melo, P., Graham, D., Noland, R. (2009). A meta-analysis of estimates of urban agglomeration economies. Regional Science and Urban Economics, 39: 332-342.

Newman, M. (2010). Networks: An Introduction. Oxford University Press.

Niu, F., & Li, J. (2018). Visualizing the intercity railway network in Mainland China. Environment and Planning A: Economy and Space, 50(5), 945-947.

Otsuka, A., (2020). Inter-regional networks and productive efficiency in Japan. Papers in Regional Science, 99: 115–133.

Overell, S., & Rüger, S., (2008). Using co-occurrence models for placename disambiguation. International Journal of Geographical Information Science, 22, 265–287.

Parr, J. (2004). The polycentric urban region: a closer inspection. Regional studies, 38: 231-240.

Parr, J. (2002). Agglomeration economies: ambiguities and confusions. Environment and Planning A, 34: 717-731.

Petralia, S., Balland, P-A., & Rigby, D. (2016). Unveiling the geography of historical patents in the united states from 1836 to 1975. Scientific Data, 3.

Phelps, N. (2021). City systems research: From morphology to relationality and positionality. International Journal of Urban Sciences, 25(4), 480-500.

Phelps, N., Fallon, R., & Williams C. (2001). Small firms, borrowed size and the urban-rural shift. Regional Studies, 35: 613–624.

Phelps, N., Miao J., & Zhang, X. (2020). Polycentric urbanization as enclave urbanization: a research agenda with illustrations from the Yangtze River Delta Region (YRDR), China. Territory, Politics, Governance: 1–20. DOI: 10.1080/21622671.2020.1851750.

Porter, C., Atkinson, P., & Gregory, I., (2015). Geographical Text Analysis: A new approach to understanding nineteenth-century mortality. Health & Place 36: 25–34.

Pumain, D. (2021). From networks of cities to systems of cities. In: Neal, Z. & Rozenblat, C. (Eds.) Handbook of Cities and Networks. Edward Elgar Publishing, pp. 16-40.

Rosenthal, S. S., & Strange, W. C. (2004). Evidence on the nature and sources of agglomeration economies. In Handbook of regional and urban economics (Vol. 4, pp. 2119-2171). Elsevier.

Ross, C., Woo, M., & Wang, F. (2016). Megaregions and regional sustainability. International Journal of Urban Sciences, 20(3), 299-317.

Salvini, M., & Fabrikant, S. (2016). Spatialization of user-generated content to uncover the multirelational world city network. Environment and Planning B: Planning and Design, 43(1), 228-248.

Sassen, S. (2007) Megaregions: Benefits beyond sharing trains and parking lots. The economic geography of megaregions. Policy Research Institute for the Region Princeton, NJ: 59–83.

Schilling, M., & Phelps, C. (2007). Interfirm collaboration networks: The impact of large-scale network structure on firm innovation. Management science, 53(7), 1113-1126.

Schweitzer, F., Fagiolo, G., Sornette, D., Vega-Redondo, F., Vespignani, A., & White, D. (2009). Economic networks: The new challenges. Science, 325(5939), 422-425.

Sebestyén, T., & Varga, A. (2013). Research productivity and the quality of interregional knowledge networks. The Annals of Regional Science, 51(1), 155-189.

Shi, S., & Pain, K. (2020). Investigating China's Mid-Yangtze River economic growth region using a spatial network growth model. Urban Studies, 57:2973-2993.

Short, J., Kim, Y., Kuus, M., & Wells, H. (1996). The dirty little secret of world cities research: data problems in comparative analysis. International Journal of Urban and Regional Research, 20: 697-717.

Sohn, C., Licheron, J., & Meijers, E. (2022). Border cities: Out of the shadow. Papers in Regional Science, 101(2), 417-438.

Tobler, W., & Wineburg, S. (1971). A Cappadocian speculation. Nature, 231: 39–41.

Van Meeteren, M., Neal, Z., & Derudder, B. (2016). Disentangling agglomeration and network externalities: A conceptual typology. Papers in Regional Science, 95(1), 61-80.

Waldfogel, J. (2008). The median voter and the median consumer: Local private goods and population composition. Journal of urban Economics, 63: 567–582.

Watts, D. (2004). The "new" science of networks. Annual Review of Sociology, 30, 243-270.

Wei, Y., Huang, C., Lam, P. T., & Yuan, Z. (2015). Sustainable urban development: A review on urban carrying capacity assessment. Habitat International, 46: 64-71.

Yao, L., Li, J., & Li, J. (2020). Urban innovation and intercity patent collaboration: A network analysis of China's national innovation system. Technological Forecasting and Social Change, 160, 120185.

Yuan, Y., & Medel, M. (2016). Characterizing international travel behavior from geotagged photos: A case study of flickr. PloS one, 11: e0154885.

Zhong, X., Liu, J., Gao, Y., & Wu, L. (2017). Analysis of co-occurrence toponyms in web pages based on complex networks. Physica A: Statistical Mechanics and its Applications, 466: 462-475.

Chapter 5

Imagined, emerging and real "Chinese Dragons": Functional coherence of Chinese megaregions

This chapter is published as: Tongjing, W., & Meijers, E. (2024). Imagined, emerging and real 'Chinese dragons': analysing the functional coherence of Chinese megaregions. Regional Studies, 1-14.

Abstract: Overcoming fragmentation by strengthening the functional coherence of megaregions is deemed essential to reaping the benefits associated with megaregions. However, few studies have examined whether planned megaregions are functionally coherent. We assess the functional coherence of megaregions from three dimensions: inclusion, connectivity and consistency. Fifteen government-defined Chinese megaregions are evaluated and qualified. For nine of these cases, our analysis reveals discrepancies with the government's defined development level. Our approach allows to identify which dimension of megaregional functional coherence needs targeted policy attention to further develop megaregions and can be implemented to explore other megaregions across the world.

Keywords: megaregions, urban system, city networks, spatial structure, toponym co-occurrence, text mining, China

1. Introduction

While individual cities are still important drivers of economic growth, a belief in regionalized urban-economic units being the future motors of the global economy has become deeply engrained in political and planning discourses over the last decades (Gottmann, 1957; Florida et al., 2008; Harrison and Hoyler, 2015). As highlighted by Harrison and Hoyler (2014, p2), "there can be little doubt as to the importance currently being attached to the megaregion concept by its many advocates". Across the globe, governments have been advocating the development of such megaregions. Well-known examples of megaregions include the USA's Bos-Wash Corridor, Europe's Dutch Randstad, and China's Pearl River Delta, but megaregions are also envisioned in South Asia, Middle East, and Latin America, such as Delhi-Lahore, Cairo-Tel Aviv, and around Sao Paolo.

Despite many examples, precise definitions of the megaregion vary, as do the concepts capturing such megaregions. Regarding the latter, next to the simple 'megaregion', one can find similar concepts like 'Megacity-regions' (e.g. Sorensen, 2020), 'Megalopolis' (Gottman, 1957), 'City Clusters' (Yao et al., 1992), 'the Endless City' (Burdett and Sudijc, 2007) and varieties employing the adjective 'polycentric', such as the 'Polycentric Metropolis' (Hall and Pain, 2006). While these concepts' definitions slightly vary, certain keywords are common, such as "polycentricity, interconnectivity and other tangible and intangible forms of cohesion" (Lang and Knox,

2009, p119)'. These indicate that cities are connected by "environmental systems and topography, infrastructure systems, economic linkages, settlement patterns, and land use, and shared culture and history" (Regional Planning Association, 2006, p8). It is almost impossible to draw meaningful boundaries around them, as these are a matter of degree more than a dichotomy of in/out. As Contant and de Nie (2009, p15) have noted, "the boundaries of megaregions are not fixed but are as dynamic as the functional relationships that create the megaregion".

It is believed that those selections of close-by cities and metropolitan areas together form interrelated larger regional entities. As Sassen (2007) suggests, the advantages of a 'single economic space' imply a variety of complementary agglomeration economies and geographic settings. Others have claimed that megaregional benefits include economic synergies for regional development, inclusive growth, social and spatial cohesion, and collaborative environmental protections (Meijers et al, 2018), arguably making megaregions globally competitive. Those benefits, however, only occur when such planned megaregions have become truly integrated through a variety of consistently strong functional relationships (Harrison et al, 2023; Meijers et al., 2016). In other words, megaregions exist by the grace of their internal functional coherence. Beyond achieving the associated megaregional benefits, functional coherence is also essential for megaregions to effectively deal with many socio-economic challenges on the megaregion scale, as challenges related to infrastructure accessibility, economic development, social welfare, and government operation are interdependent and need to be addressed holistically (Cardoso and Meijers, 2021). For instance, effectively developing advanced transportation networks, establishing megaregional organizing capacity, fostering cooperation in R&D, and collaborating with respect to environmental protection, all require what Cardoso and Meijers (2021) call 'metropolitanisation' – an intertwined process of political-institutional, cultural-symbolic and functional integration in which the latter plays a key role. Therefore, achieving functional coherence not only brings megaregional benefits but also helps megaregions tackle these challenges.

However, in practice, most megaregion planning still follows what could be called a 'cartographic approach', in which a group of nearby large cities is delineated on a map and the cohesion between them is simply assumed (Harrison and Hoyler, 2015). A critical issue with this approach is that just two cities being close does not necessarily mean that they function as a

cohesive unit (Burger et al., 2014a). For instance, Manchester and Liverpool may be close, but plenty of evidence shows that they are rather disconnected (Harrison and Hoyler, 2014). In particular, also China has so zealously embraced this megaregion concept, that for many local governments, being included in the political definition of a megaregion is considered a key achievement, thereby neglecting the actual purpose of megaregional planning (Fang and Yu, 2016) aimed at obtaining the associated benefits through integration. And if there is attention for strengthening the relations between the cities making up a megaregion, most planning strategies are oriented toward one specific dimension of integration and do not address other dimensions (Fu and Zhang, 2020). It is still uncommon to find comprehensive planning that encompasses multiple aspects of socioeconomic development.

Despite its importance, relatively few research has evaluated the functional coherence of megaregions. This is mainly due to a lack of accessible information on intercity flows at this specific megaregional scale, and in fact the absence of appropriate indicators or proxies that can reveal city relationships (Meijers and Peris, 2019). Also, the evaluation of functional coherence requires collecting data on multiple types of intercity relationships, since the pattern and strength of the variety of intercity relationships are not necessarily identical: "a region may appear well-integrated on the basis of one type of relationship but loosely connected on the basis of another type" (Burger et al., 2014b, p818). In other words, it is hard to capture the 'multiplexity' of city networks (Tongjing et al., 2022), but any evaluation of functional coherence in megaregion planning is likely to be biased if the consistency in network patterns is not taken into account.

The objective of this paper is to propose and apply an evaluation framework that examines the functional coherence of megaregions from three dimensions: inclusion, connectivity, and consistency. The three dimensions respond to three questions, namely 1) whether the cities are actually included within a delineated megaregion, 2) how strong the intercity relationships are within a delineated megaregion, and 3) how consistent the patterns of intercity relationships of different types are. We apply this framework to fifteen Chinese megaregions using a dataset on relationships between cities that was obtained through the still rather novel toponym co-occurrence method. This dataset was extracted from webpages by using a lexicon-based text mining method which retrieves

and classifies intercity relationships based on collocation analysis (Mello, 2002) —the general assumption being that the more often words or phrases co-locate in texts, the more they are related. This method uses the co-occurrences of placenames of Chinese cities that constitute megaregions to unveil relationships. The co-occurrence of city names will subsequently be categorized using keywords that indicate whether the relationships between two cities can be categorized as 'industry', 'finance', 'culture', 'IT', 'research', or 'government'.

Hence, this paper aims to be novel in terms of methods by means of an alternative way of measuring relationships between cities and conceptually by proposing a multi-dimensional, standardized classification scheme for levels of functional coherence of megaregions that allows for easy comparison between megaregions. Empirically, we will use our method and multidimensional approach to compare the functional coherence of fifteen Chinese megaregions, thereby examining whether the delineated megaregions fit the alleged development level as defined in the latest fourteenth Five-Year central plan, which must be seen as the foundation for concrete megaregion planning policies in China.

The structure of the paper is as follows. First, we review the literature about megaregion development and Chinese megaregion studies (section 2). Second, we present our method and propose our three-dimensional evaluation framework to assess functional coherence (section 3). Third, we compare the functional coherence of fifteen Chinese megaregions based on these dimensions (section 4). We conclude with a discussion of the results and their implications for the planning of Chinese megaregions, and more generally, megaregions worldwide (section 5).

2. Megaregion theory

2.1 The goal of planning megaregions

While there is no single definition or shared standard for determining whether a large-scale sprawling urbanized landscape is a megaregion (Dewar and Epstein, 2007; Fang and Yu, 2017), the aim of employing the megaregion concept is generally quite clear, namely to foster the competitiveness, cohesiveness, and sustainability of a regional territory (CEC, 2011). The assumption here is that a large proportion of global

economic growth is and will continue to be concentrated in such megaregions.

The early approaches to defining megaregions often departed from a morphological perspective (Harrison and Hoyler, 2015), marking out space by observing the physical landscape, selecting a small number of proximate large cities, and then including the surrounding cities (Fang, 2015; Florida et al., 2008). This early morphological perspective focused on stock indicators such as the number of cities included, the population of cities within the delineated area, the whole built-up area, and the total economic size, etc. (Fang and Yu, 2017). However, later research shows that just proximity is not enough to guarantee a strong metropolitan economy (Ahrend et al., 2017).

'Mega' generally refers to the size of a region and implicitly suggests that having a substantial critical mass translates into the presence of many advantages related to size – better known as agglomeration economies. But summing small cities does not necessarily make a successful megaregion, as fragmentation looms: "institutional and spatial fragmentation, functional redundancies, uncoordinated transport planning, disconnected housing and labor markets, imbalanced distribution of investment, unwillingness to cooperate by local authorities in the absence of a metropolitan government, and a lack of common historical, cultural or political references able to shape a joint identity and shared strategic priorities" characterize many megaregions (Cardoso and Meijers, 2020, p362). Such fragmentation, except for some physical barriers, is not visible on the map, and renders the early cartographic/morphology-based approach to megaregions a rather limited one.

A more functional approach is needed to complement morphological approaches. It is nowadays widely acknowledged that to achieve the agglomeration economies associated with a large megaregional critical mass, the cities constituting megaregions should be integrated well, and along multiple dimensions (Ahrend et al., 2017; Meijers et al., 2018). Such integration processes take a long time, and are characterized by periods of rapidly increasing integration when the interests of economic and governmental actors as well as the public at large align, but can also be characterized by throw-backs if this is not the case (Cardoso and Meijers, 2021).

Governmental actors play an important role in this integration process, as their actions can either speed up or slow down such integration. But a megaregion is composed of a wide variety of places whose governments may have different interests that do not necessarily align. The interest of smaller cities can, for instance, be to obtain better access to urban functions in large cities, and/or to acquire a higher political status for negotiating with higher-level government by being part of a larger consortium (Meijers and Burger, 2017; Cardoso and Meijers, 2017; Derudder et al, 2022). Large cities in megaregions may want to extend the support base for their metropolitan functions, or use their relations with other cities to mitigate negative returns (Camagni et al, 2017). Beyond the local government level, national governments see the development of megaregions as a way for certain regions to obtain a favorable investment and business environment and consequently global competitiveness (Harrison and Hoyler, 2014; Wu, 2020a). In addition, and beyond economic imperatives, a series of urban challenges also calls for recognizing the megaregion as a coherent unit and scale for better collaboration and overcoming transterritorial issues, such as economic inequalities (Farole et al., 2011), competition between adjacent cities (Pan et al., 2017), inter-regional infrastructure services (Wang et al., 2021), and environmental hazards and pollution (Chen et al., 2023).

Megaregion plans comprise more than what metropolitan plans at the smaller scale normally focus on in functional terms (e.g. strengthening accessibility for daily commuting), as the megaregion plans are often envisioned as ‘associational repertoire’ (Cooke and Morgan, 1994), a framework that can facilitate effective interactions. As indicated by Cardoso and Meijers (2020), urban planning should contribute to a process of metropolisation to make ‘institutionally, functionally, and spatially fragmented urbanized regions become integrated along various dimensions and emerge as connected systems at a higher spatial scale’. In other words, actively developing a coherent region is what most megaregion plans have to strive for to reap the benefits associated with megaregions (Meijers et al., 2018).

2.2 Megaregion planning and analysis in China

In contrast with Western countries, megaregion development in China is predominantly led by the government (Wu, 2016). While the growth of individual large cities remained under strict central control all the time

until the tenth Five-Year Plan (2001-2005), the importance of collaboration between cities was recognized rather early by the central government, as it believed that such collaboration could forster the development of small and medium-sized cities (Zhang, 2019). In the 1980s, more than 100 economic zones were established under the central government (Xu, 2008). A notable example was the Shanghai Economic Region which included ten cities and five provinces in the Yangtze River Delta (YRD) and had its own regional development framework, despite being under the central government's control (Li and Wu, 2012).

The hierarchical central control on city growth weakened in the 1990s, and since then the development control of large cities was relaxed. During the time of the tenth Five-Year Plan (2001-05), megaregion planning started taking off. The fundamental urban development principle switched towards prioritizing regionally coordinated planning—forming horizontal connections rather than the previous top-down approach (Chung, 2015). The national government let a group of coastal regions pilot this development approach to advance the nation's global competitiveness. The resulting collaboration-focused planning approach achieved huge success and became the prototype of megaregion planning in China. However, prioritizing development in the coastal regions also widened the development gap between coastal and inland regions. Wu and Zhang (2022) state that the Chinese megaregion development demonstrates the key role of the central government in rescaling rather than a simple decentralizing process, therefore integration is not an outcome of 'market' forces as in the West (Harrison and Hoyler, 2015; Wu, 2018). While the integration of cities became a new strategy for economic growth development, Wu (2018) warned that the deep involvement of the central government can also be an obstacle to further integration.

To stimulate higher economic growth, local governments were given more freedom to encourage their 'entrepreneurial spirit' (Chien and Gordon, 2008). However, it resulted in fierce city competition and a series of redundant infrastructure development and environmental degradation (Wu, 2018). To mitigate these issues, in the next Five-Year Plan (2006-2010), the central government went one step further, considering megaregions as 'the main entity for driving urbanization' and emphasizing the importance of city collaboration rather than competition (Li and Jonas, 2019). Wu (2016, p1146) also emphasizes that "the only practical way to set up regional collaboration seems to be through city-based linkages". These

city-based linkages can be a series of collaborations in air pollution control (Chen et al., 2023), as well as issues related to the overall spatial distribution of economic growth, such as prioritizing balanced development and reducing the adverse economic competition between local governments (Li and Wu, 2012).

A paradigmatic shift is signaled in the 2014–20 National Plan on New Urbanization where planning is no longer dominated by the traditional planning scale of provinces. Planning on the scale of megaregions surfaced as a central policy concept, used by the central government to reassert and enhance its governance capacity in social-economic terms (Wu, 2016).

In the thirteenth Five-Year Plan (2016–20), the megaregion planning concept remains dominant in the national spatial plan. Three megaregions, the Yangtze River Delta (YRD), Pearl River Delta (PRD), and Jing Jin Ji (JJJ) were aimed to be well-integrated globally competitive megaregions in the central plan. During this period, the idea of megaregion planning became zealous. The inclusion in a megaregion plan is used as a marketing tool for some local governments to capture investment and showcase their strategic importance (Wu and Zhang, 2007). And for the public, inclusion in a megaregion is seen as a precondition for the city's prosperity, whereas exclusion would lead to stagnation (Fang, 2015).

While planning megaregions is often initiated by the central state in a top-down nature, local governments and business communities sometimes also take the lead in bottom-up ways (Li and Wu, 2012). The bottom-up mechanism is motivated by competition with other regions. For instance, to promote local business connections within the region, city governments in the Yangtze River Delta initiated a communication platform called *The Joint Conference of Directors of Coordination Offices of the Yangtze River Delta Region*. However, as indicated by Yeh and Xu (2010), such a cooperation mechanism is largely imaginary and enterprise-oriented only, and few general regional development plans can be achieved. Many local governors in China create their own megaregion plans and hope to get approvals from the central government, in which case the megaregion plan becomes a bargaining tool to attract investment (Harrison and Gu, 2021). The top-down and bottom-up approaches are not necessarily identical and mismatches between the delineation of administrative and economic regions were also possible (Wu, 2016b).

The latest fourteenth Five-Year plan (2021-2025) approved and classified national development plans for nineteen megaregions, focusing on the improvement of regional integration and coordinated development. The plan separates the nineteen megaregions into three levels: optimizing (优化提升), expanding (发展壮大), and nurturing and developing (培育发展). These English translations may actually sound rather similar, but the subtle differences actually indicate three distinct development stages. ‘Optimize’ refers to well-developed, ‘real’ megaregions. For instance, the Pearl River Delta is at this level. ‘Expanding’ refers to emerging megaregions, which are at a medium development level, and ‘ nurturing and developing’ refers to imagined megaregions, which are still at the incubating level.

Note that the central plan does not provide a reasoning behind this classification. Nonetheless, functional coherence should be a fundamental consideration across all stages of megaregion development. As highlighted in Chapter 28 of the Fourteenth Five-Year Plan, the overarching efforts include “improve the mechanisms for ensuring integrated and coordinated development and for sharing costs and benefits in city clusters based on coordinated infrastructure networks, industrial collaboration, shared access to public services, and joint contributions to ecological conservation and environmental governance (National Development and Reform Commission, 2023, p3).

In conclusion, Chinese megaregion planning is an evolving political, economic and spatial process moving from a centralized and hierarchic operation to more horizontal regional cooperation (2001-2005), from economy—focused to holistic planning (2006-2010), and from prioritizing coastal regions to considering megaregion planning as relevant for the entire national urban planning framework (2016-now). It is “a new imaginary” to reach beyond economic-oriented development ambitions (Wu, 2020b).

3. Method

In this chapter, we describe our approach to measuring the level of functional coherence of megaregions. First, we discern three dimensions of functional coherence that are relevant for megaregions, and second, we explain how we measure these dimensions, thereby building on a still rather novel approach to quantify relationships between cities.

3.1 Three dimensions of functional coherence

There is no widely acknowledged approach of how to measure the functional coherence of a regional network system, but similar to many studies using network theory to study city relationships, we propose that the megaregion network system must be viewed as a multilayer network where each layer is a type of intercity relationship, and functional coherence analysis can be conducted departing from three dimensions: inclusion, connectivity, and consistency.

The first dimension is 'inclusion', which addresses the question of whether cities included in the planned megaregion are actually related to the other cities. As the identification of megaregions in China was a normative government-led rather than analytical process, also involving lobbies by cities, it may be that cities are included in megaregions even though they are hardly functionally linked to the other cities. Including non-related cities often results from too ambitious megaregion plans. For instance, Liu et al., (2016) analyzed 22 Chinese urban region plans acknowledged by the National Development and Reform Commission (NDRC) and found that just four regions have relatively strong transportation connections within their region, while the rest are more or less arbitrary groupings. Obviously, the answer to the question of inclusion is not a simple yes or no, but requires to consider the extent of inclusion—specifically, the proportion of cities within a defined megaregion that are actually related. We assume that to qualify as included, a city must, at a minimum, demonstrate one functional relationship with at least one other city within the megaregion. Consequently, to quantify the extent of inclusion, our first dimension is to determine the number of cities that are actually included in the network.

Clearly, a city can be related to more than just one city and generally, the greater the number of relationships a city maintains, the more it stands to gain in terms of development (Bathelt and Glückler, 2017; Meijers et al., 2018; Phelps, 2021; Tongjing et al., 2023b). Building on this foundation, we assume that in a fully developed megaregion, all cities should be connected to each other. In other words, it is better to have as many as possible relationships within the megaregion. Consequently, our analysis progresses to a second dimension 'connectivity'—the proportion of relationships present within a megaregion.

The third dimension, 'consistency', addresses the potential disparity in integration levels of a megaregion when viewed through different types of relationships (Burger et al., 2014b; Fang and Yu, 2017). This dimension evaluates the extent to which intercity relationships maintain a uniform character across all relationship categories. Many megaregion plans for instance only focus on advancing infrastructure development to strengthen accessibility and connectivity, while neglecting the important role played by economic and political relationships in strengthening megaregion integration (Smart, 2018). This dimension also aligns with the objectives of megaregion planning outlined in the Fourteenth Five-Year Plan, which emphasizes that megaregions should be "a well-ordered layout with urban areas providing a full range of functions and collaborating with each other" (National Development and Reform Commissions, 2023, p1).

Given that all cities share some level of relationship with each other, we measure the three dimensions at various thresholds to give a detailed examination of a megaregion's functional coherence. This approach is also systematic and multi-scalar in that it progressively assesses megaregion functional coherence from basic city-level connections to network structures, and ultimately to the overall consistency of these connections. This approach is instrumental in identifying and addressing potential shortcomings in current megaregional planning strategies.

3.2 Measuring dimensions of functional coherence

Since cities are always connected to some extent (Green, 2007), it is essential to consider the strength of these connections when examining the three proposed dimensions of functional coherence. However, our dataset presents a challenge with the presence of a few disproportionately strong or weak relationships. If included without modification, these outliers could significantly distort the results, leading to potential biases. To mitigate this issue, we have transformed the weighted network into a non-weighted network. After all, no matter how strong one relationship is, it is just one among many within a megaregion.

Determining an appropriate threshold certainly poses its challenges, as setting a single threshold could introduce subjectivity. To address this, we measure the three dimensions at varying cut-off levels. By only retaining those relationships that surpass a specific strength value and treating them as equivalent in a non-weighted network, we effectively reduce the

influence of extreme outliers. This approach ensures that our analysis captures both the quantity and the quality of intercity relationships within megaregions.

For the inclusion dimension, we define the degree of inclusion as the ratio of cities that are included in the megaregion relative to the total number of cities constituting that megaregion, which can be formulated as :

$$Inclusion_m^t = \frac{City_m^t}{City_{m,total}} \times 100\%$$

$Inclusion_m^t$ is the inclusion degree of megaregion *m* at threshold *t*, $City_m^t$ is the number of cities in megaregion *m* that have at least one relationship at threshold *t*, $City_{m,total}$ is the total number of cities in megaregion *m*. The closer the value is to 100%, the higher the proportion of cities included, leaving few cities excluded.

To measure the connectivity dimension, we use the concept of network density from network science to indicate the connectivity level of a megaregion. This is measured as the ratio of the number of intercity relationships within the megaregion network relative to the maximum possible number of intercity relationships a megaregion could have. The latter is obtained when every city is connected to all other cities. This dimension can be formulated as:

$$Connectivity_m^t = \frac{Relationship_m^t}{Relationship_{m,total}} \times 100\%$$

$Connectivity_m^t$ is the connectivity degree of megaregion *m* at threshold *t*, $relationship_m^t$ is the number of relationships in megaregion *m* that are above threshold *t*, $Relationship_{m,total}$ is the total number of relationships megaregion *m* can have. The closer the value is to 100%, the denser the network is and the more cities are connected to each other in the megaregion.

The third dimension, consistency, involves measuring the extent to which all categories exhibit strong connectivity levels. This is done by counting the number of categories where the level of connectivity is high and then dividing that by the total number of categories. It is formulated as:

$$Consistency_m^t = \frac{Categories_m^t}{Categories_{total}}$$

$Consistency_m^t$ is the consistency degree of megaregion m at threshold t, $Categories_m^t$ is at threshold t the number of categories in megaregion m categories where the level of connectivity is high, $Categories_{total}$ is the total number of categories this analysis included. A higher value indicates a greater level of consistency in terms of how well a megaregion is interconnected.

3.3 Intercity relationship retrieval

In this paper we used a classified intercity relationship dataset created by Tongjing et al. (2023a) that was made open access. It presents six categories of intercity relationships between 293 Chinese cities using a lexicon-based toponym co-occurrence method. The data source of this paper is the 2019 April Corpus of the Common Crawl web archive, initially containing multiple languages. Through a preprocessing step, non-Chinese and irrelevant web pages (e.g. gambling, porn) were then excluded, resulting in a refined corpus, which contains approximately 91 million pages from 1,067 top domains and 1,792,759 domain names. The details of how data was extracted can be found in Tongjing et al. (2024).

This method searches for the co-occurrence of two toponyms (cities in this case) on a webpage, and uses topic keywords in the surrounding text to make sense of what this co-occurrence is about, which allows to classify the relationship that is found between cities. The underlying assumption of this method, as in every text mining exercise, is that the more frequent entities in different texts co-occur, the more related they are. The number of co-occurrences is used to indicate the relationship strength between a pair of cities and the topic keywords are used to categorize the relationship. The 293 cities included in the dataset are all Chinese cities at prefecture level or above. The topic keywords used to classify each toponym co-occurrence found were selected from the China Standard Industrial Classification of Economic Activities (GB/T 4754—2017) and enlarged by identifying semantic-related words through a natural language processing method. The six categories of intercity relationships include 'industry', 'finance', 'culture', 'IT', 'research', and 'government'. For further details on the classification process, please see Tongjing et al. (2023a).

Fifteen out of the nineteen megaregions approved and classified by the State Council of China are studied in this paper. As this dataset only includes cities at the prefecture level or higher, four of the nineteen megaregions had to be excluded as they mostly contain county-level cities. The fifteen megaregions are mapped in Figure 1, with the colors indicating the different levels of development according to the Chinese government, and some key statistics on their composition are presented in Table 1. Note that we prefer alternative names for the three development stages to present a more clear differentiation in stage. So, development level 1 (optimizing) is the most developed, which we term 'real' megaregions; development level 2 (expanding) is the intermediate stage and is referred to as 'emerging' megaregions; development level 3 (nuturing and developing) is the least developed stage, which we refer to as 'imagined' megaregions.

Figure 1. Chinese megaregions.

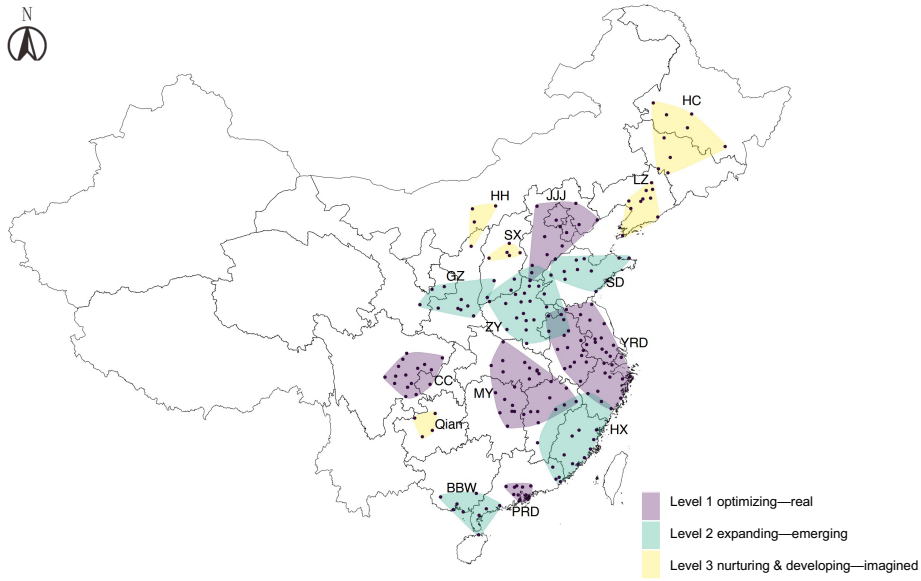


Table 1. Megaregions included

Region	YRD	PRD	JJJ	CC	MY	SD	HX	ZY	GZ	BBW	HC	LZ	SX	QIAN	HH
Planned level	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
Number of Cities included	41	11	13	16	28	12	20	30	11	10	10	10	5	4	4
Number of high administrative cities	5	4	3	2	3	2	2	1	1	2	2	2	1	1	1
Number of intercity relationships	820	55	78	120	378	66	190	435	55	45	45	45	10	6	6

Note: Yangtze River Delta (YRD); Pearl River Delta (PRD); Beijing-Tianjin-Hebei Regin (JJJ); Chengdu-Chongqing Region (CC); Middle Yangtze Region (MY); Shangdong Penisula (SD); Yue-Min-Zhe coastal region (HX); Zhongyuan Region (ZY); Guanzhong Region (GZ); Beibuwan Region (BBW); Ha’erbin-Changchun Region (HC); Liaoning Central and South Region (LZ); Shanxi Taiyuan Central Region (SX); Guizhou Central Region (Qian); Hohhot-Baotou-Ordos-Yulin Region (HH)

3.4 Absolute versus relative strength of an intercity relation

Of course, the largest cities have the strongest relationships in absolute terms with nearby large cities, but this is to be expected given their size and short distance. To reveal the relative intensity of the network strength, we used the gravity model to examine how strong the different relationships are after controlling for population sizes of, and distance between cities.

This gravity model analysis is conducted using all 293 cities in the dataset in order to create a national benchmark. We first fit the gravity model to estimate the strength of intercity relationship by controlling for size and distance.

Then, the estimated relationship is set as benchmark to determine whether the toponym co-occurrence between two cities is stronger or weaker than the gravity model estimation. The link strength of the intercity relationship is defined as the ratio of absolute number of toponym co-occurrences for a given pair of cities to the gravity model predicted number, multiplied by 100, which can be formulated as:

$$w_{ij,C} = \frac{T_{ij,C}}{I_{ij,C}} \times 100$$

where $w_{ij,C}$ is the relative strength of the intercity relationship between city i and city j in categoory C, $T_{ij,C}$ is the absolute relationship strength between city i and city j in category C, and $I_{ij,C}$ is the estimated co-relationship strength of city i and city j based on in the gravity model in category C. A $w_{ij,C}$ value above 100 means that a relation is stronger than expected, while lower values suggest that the relationship between two cities is weaker than expected.

4. Results

Here we will first evaluate and categorize the three dimensions of functional coherence for the fifteen Chinese megaregions. Eventually, in line with the central Five-Year-Plan, we summarize this information by categorizing each megaregion into one of three levels of coherence (low, medium, high) for each of the three dimensions. Then we will compare our classification results to those of the central Five-Year-Plan classification.

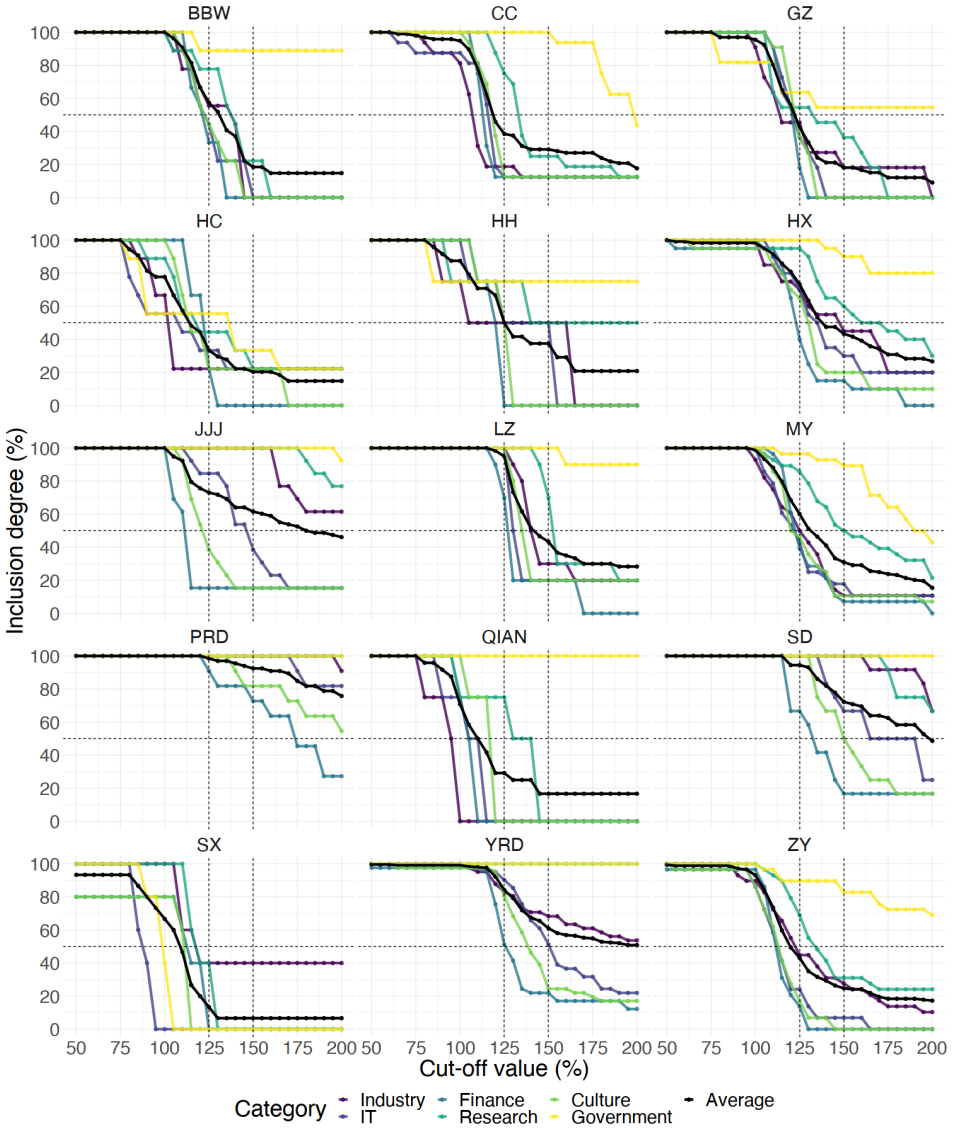
4.1 Analysis of inclusion

Figure 2 shows the degree of inclusion for various cut-off values. The x-axis presents the cut-off value and the y-axis is the degree of inclusion. This is the ratio of cities that are still included in the network given the threshold applied. The start value of the x-axis is set at 50%, which means the relationship strength is only half as strong as one would expect given the population sizes of cities and the distance between them. At this level, cities in most megaregions are connected. However, we obviously expect cities in megaregions to be more strongly connected than expected, otherwise, it makes little sense to define it as a coherent region. However, with an increase of the cut-off value to 125% (so, assuming that cities in the megaregion are 25% more strongly connected than expected), cities in many megaregions become disconnected. When more than 50% of the cities are disconnected at the 125% threshold, a megaregion has a 'low'-level of inclusion. This group consists of Chengdu-Chongqing (CC), Guanzhong (GZ), Harbin-Changchun (HC), Guizhou Central (Qian), Shanxi Taiyuan (SX), and Zhongyuan (ZY). Megaregions defined as having a 'medium'-level of inclusion have more than 50% of cities getting disconnected when we raise the cut-off value with another 25%, namely from 125% to 150%. This group includes Beibuwan (BBW), Hohhot-Baotou-Ordos-Yulin Region (HH), Yue-Min-Zhe (HX), Liaoning (LZ) and Middle Yangtze River (MY).

The remaining four megaregions remain well connected at this 150% threshold and include Jing-Jin-Ji (JJJ), Yangtze River Delta (YRD), Pearl River Delta (PRD), and, perhaps surprisingly, Shandong Peninsula (SD). SD is interesting, as the other three are generally considered as the main, and most well-developed megaregions in China, but SD is only categorized as a level 2 (emerging) megaregion in the latest central plan. Of interest to

note are also the different patterns for the six types of relationships in the dataset that we use. The strength of intercity relationships varies across the six categories analyzed in our study, with government relationships generally being the strongest. The inclusion level drops fastest when considering cultural and finance relationships. The former confirms Smart’s claim (2018) that Chinese urban development often lacks cultural interactions.

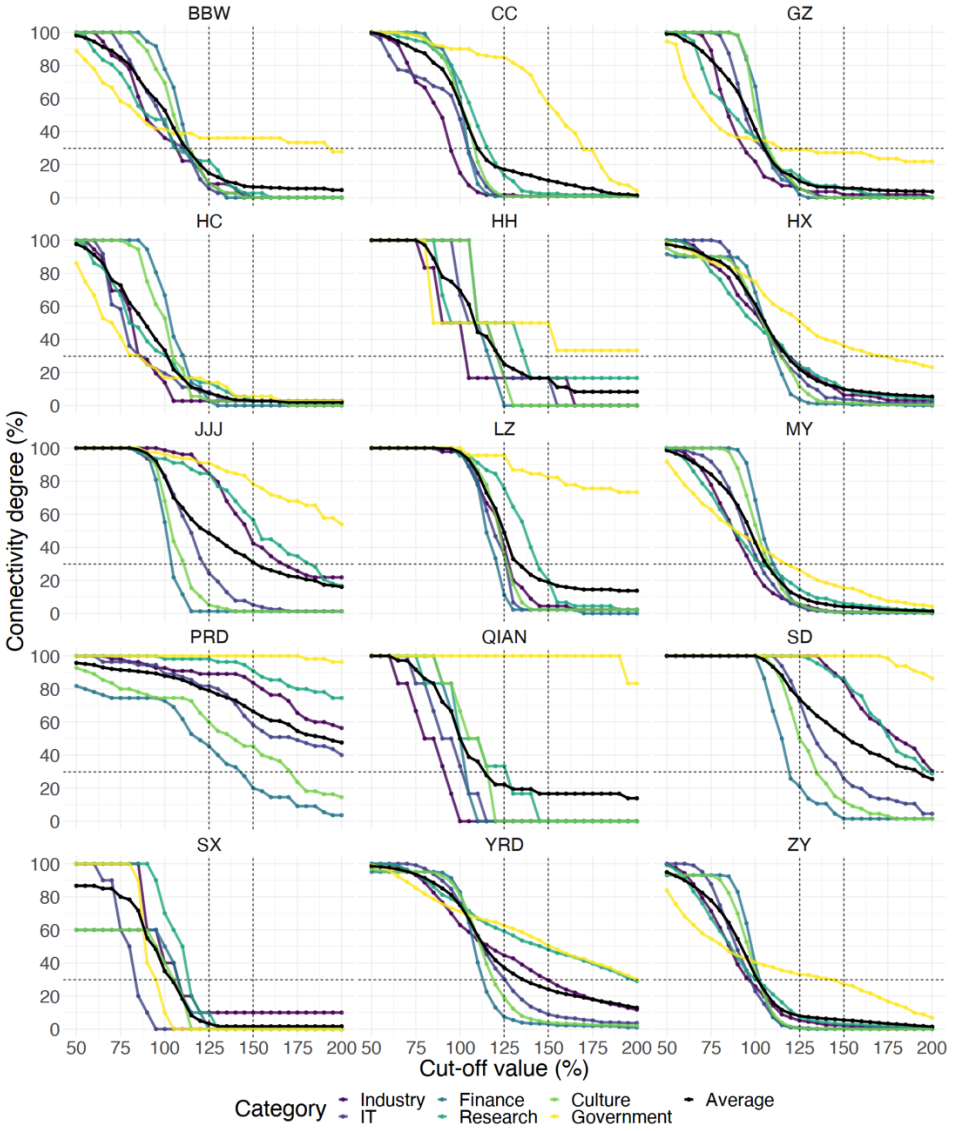
Figure 2 Functional coherence of Chinese megaregions: Inclusion



4.2 Analysis of connectivity

Similar to the analysis of inclusion, we also examine the connectivity level of a megaregion on a sliding scale. Figure 3 shows the change in the connectivity level with the increase of the cut-off value. Again, the x-axis presents the cut-off value, and the y-axis the connectivity level.

Figure 3 The functional coherence of Chinese megaregions: Connectivity



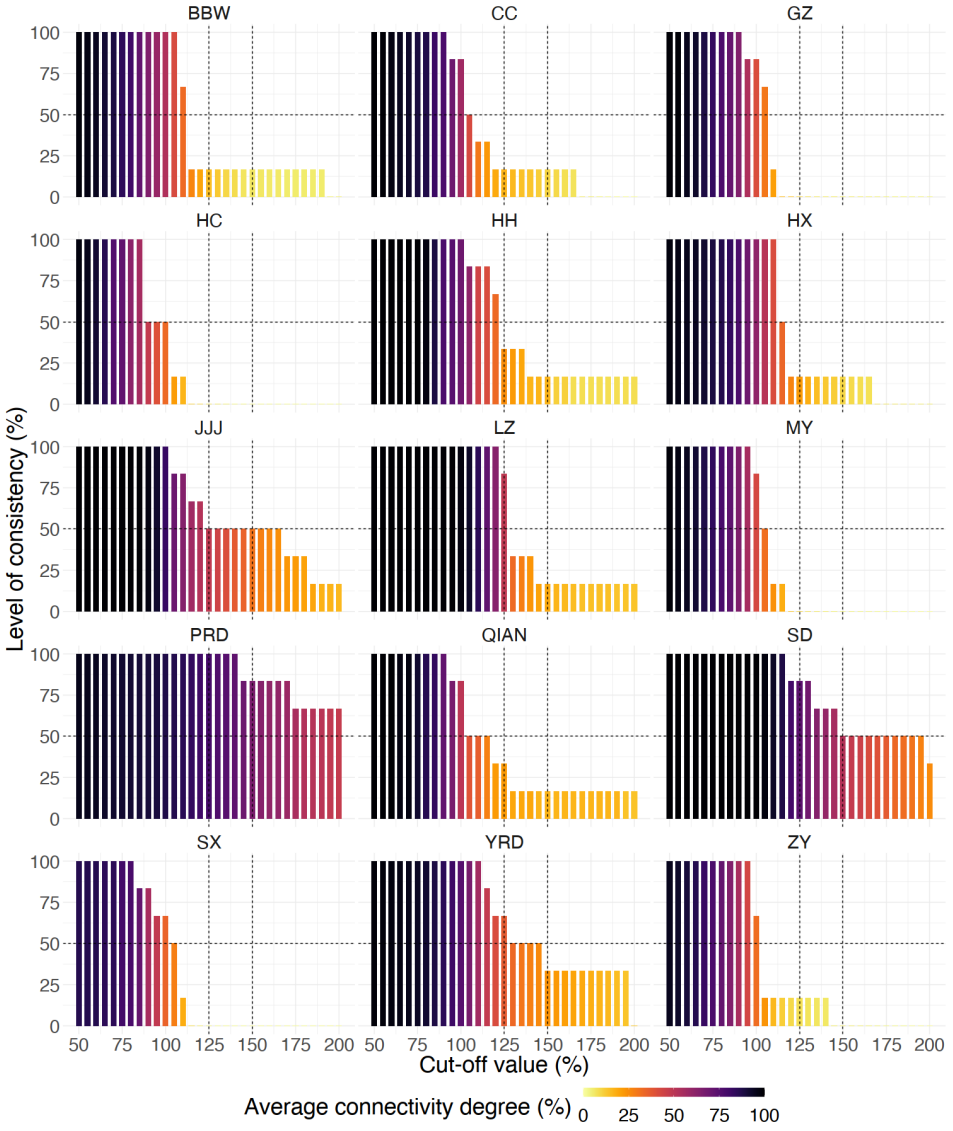
Also a megaregion’s connectivity can be roughly classified into three levels, based on Figure 3. The megaregions classified as having a ‘low’ level of connectivity comprise those that have less than 30% of city pairs reaching the 125% cut-off point. Eight megaregions are at this level, including Beibuwai (BBW), Chongdu-Chongqing (CC), Guanzhong (GZ), Harbin-Changchun (HC), Middle-Yangtze River Delta (MY), Guizhou central (Qian), Shanxi Taiyuan (SX) and Zhongyuan (ZY). For the medium level, we raise the cut-off point to 150% and again consider how many city pairs reach that threshold in each megaregion. This share is above 30% in Jing-Jin-Ji (JJJ), Pearl River Delta (PRD) and Shandong Peninsular (SD); which therefore score ‘high’ on the connectivity dimension. For the following megaregions, the share drops below 30% when replacing the 125% threshold with the 150% threshold. These score ‘medium’ on the connectivity dimension and include Hohhot-Baotou-Ordos-Yulin (HH), Yue-Min-Zhe coastal region (HX), Liaoning Central and South Region (LZ), and surprisingly, Yangtze River Delta (YRD) are at this level. .

Figure 3 also shows that the government connectivity is generally higher than the connectivity of other types. This seems to corroborate previous findings on the role of government in China’s urbanization (Wu, 2016; Harrison and Gu, 2021), perhaps in contrast to Western countries, where megaregions often lack governmental collaboration (Glass, 2015).

4.3 Analysis of consistency

Figure 3 provides an initial impression of consistency, as it reveals varying levels of connectivity across the six different categories. In this section, we further examine the consistency of patterns in Figure 4. As before, the x-axis represents the cut-off value, while the y-axis shows the consistency level. This is represented by the percentage of categories achieving a high level of connectivity. Like before, we regard a connectivity level of 30% (that is, 30% of the city pairs are on or above the threshold value displayed on the x-axis in their category). The graph is color-coded to indicate the average connectivity of the six categories at this value, with darker colors signifying higher average connectivity.

Figure 4 The functional coherence of Chinese megaregions: consistency



Similar to our previous analysis, we concentrate on cut-off values at 125% and 150%. Figure 4 reveals that most megaregions, including Beibuwai (BBW), Chengdu-Chongqing (CC), Guanzhong (GZ), Harbin-Changchun (HC), Hohhot-Baotou-Ordos-Yulin (HH), Yue-Min-Zhe coastal region (HX), Middle-Yangtze River Delta (MY), Guizhou central (Qian), Shanxi Taiyuan (SX), and Zhongyuan (ZY), do not exhibit strong connectivity in any category at the 125% cut-off. These megaregions are characterized as

having low consistency. Megaregions that display at least 50% consistency at the 125% cut-off value but less than 50% at a 150% cut-off are classified as medium level. Only the Yangtze River Delta (YRD) falls into this category. The remaining megaregions, where at least half of the types of relationships have high connectivity at the 150% cut-off values, are considered to have high consistency. These include Jing-Jin-Ji (JJJ), Pearl River Delta (PRD), and the Shandong Peninsula (SD).

4.4 Classification results

In the previous section, we rated the functional coherence of fifteen megaregions from the three dimensions and results are summarized in Table 2. The megaregions are ranked from most-developed to least developed, and alphabetically when they score similarly. To obtain a megaregion's overall functional coherence level, we use the most common classification level in the three dimensions.

Table 2 Megaregion classification result

Region	Government-proposed development level	Inclusion	Connectivity	Consistency	Research-based development level
JJJ	High	High	High	High	Real
PRD	High	High	High	High	Real
SD	Medium	High	High	High	Real
YRD	High	High	Medium	Medium	Developing
HH	Low	Medium	Medium	Low	Developing
HX	Medium	Medium	Medium	Low	Developing
LZ	Low	Medium	Medium	Low	Developing
BBW	Medium	Medium	Low	Low	Imagined
MY	High	Medium	Low	Low	Imagined
CC	High	Low	Low	Low	Imagined
GZ	Medium	Low	Low	Low	Imagined
HC	Low	Low	Low	Low	Imagined
QIAN	Low	Low	Low	Low	Imagined
SX	Low	Low	Low	Low	Imagined
ZY	Medium	Low	Low	Low	Imagined

Our evaluation learns that there is a considerable discrepancy between the government-defined level of functional coherence and the results of our analysis of coherence. Nine out of 15 megaregions were classified differently. Of course, one can discuss about the exact thresholds we have been using in our analysis, but our impression is that we set the bar certainly not too high, this bar being 25% and respectively 50% higher than

the gravity model estimation, given that we may expect the coherence between cities in the same megaregion to be higher than between cities not part of a megaregion. The functional coherence of six megaregions is overestimated: Yangtze River Delta (YRD); Zhongyuan Region (ZY); Guanzhong Region (GZ); Beibuwan Region (BBW); Chengdu-Chongqing Region (CC) and Middle Yangtze Region (MY). The Chengdu-Chongqing Region (CC) and Middle Yangtze Region (MY) were even considered to be most highly integrated by the government, while our analysis places them actually in the lowest category of functional coherence, namely 'imagined'. Apparently, having megacities (Chengdu and Chongqing in CC, Wuhan and Changsha in MY) does not mean you are a megaregion too. The relatedness of secondary cities in those regions needs attention. On the other hand, three megaregions are considerably more integrated than estimated by the government. This applies definitely to Shandong Peninsula (SD), which scores highest on all dimensions of functional coherence, and also holds for Liaoning Central and South Region (LZ) and Hohhot-Baotou-Ordos-Yulin Region (HH). Note that in contrast to Middle Yangtze Region (MY) and Chengdu-Chongqing Region (CC), Shandong Peninsula lacks top national cities, but its cities are overall well-connected. What helps in this case is that they are all located within the same province. As noted by Wu (2018), the coordinated development of megaregions that exceed their provincial boundaries and coordination remains difficult and problematic.

5. Conclusion

The development of megaregions has been associated with enhanced global economic competitiveness and improved quality of life. In many parts of the world, governments actively foster their development, and we argued that these attempts should primarily be focused on strengthening the functional coherence of those regions, as it allows to reap benefits of agglomeration at unprecedented spatial scales (although the term 'agglomeration externalities' can better be replaced with 'city network externalities' in this case), while it is also necessary to overcome the negative effects of (governmental, spatial, social, economic, cultural) fragmentation that by definition characterizes many megaregions.

Using novel data to determine the relatedness of cities, and a new framework to assess functional coherence of megaregions based on the dimensions of inclusion, connectivity and consistency, we explored the

functional coherence of 15 Chinese megaregions. We showed that there is actually a substantial difference between normative assessments of coherence by the Chinese government and the actual functional coherence of those megaregions; particularly also because the presence of a megacity cannot be equated with being part of a coherent megaregion. This is problematic as it may direct planning interventions and spatial investments in the wrong direction. Unlike many other countries, the Chinese government actually has a strategy for the development of megaregions and it certainly is not our aim to disqualify it. Rather, we aim to add nuance to the policy debate that we think is necessary for more targeted development strategies.

The charm of our three-dimensional framework and our operationalization, if we may say so ourselves, lies in the policy guidance it may provide to further develop megaregions, whatever development stage they are in. The three dimensions of inclusion-connectivity-consistency are not unrelated; there actually is a certain built-up in the sense that megaregions that score low on inclusion, will also score low on connectivity, while those that score low on connectivity are also likely to score low on consistency. Vice versa, those that score high on consistency also score high on the other dimensions, and those that score higher on connectivity do score well on inclusion too. For each megaregion it became clear in what dimension of cohesion it needs to develop. For instance, the Yangtze River Delta (YRD) is an exceptionally ambitious megaregion plan that includes a significantly higher number of cities compared to most other delineated megaregions. Although well-developed in terms of inclusion, it appears that connectivity levels need to be improved. Secondary cities seem primarily connected to major cities, but not to each other. The latter is much more the case in those other large megaregions, the Jing-Jin-Ji (JJJ) and Pearl River Delta (PRD) megaregions. For others, inclusion will be a prime concern, as without a reasonably high score on this dimension, it is hard to call this rather arbitrary collection of cities a megaregion at all. Fostering inclusion may mean that in this initial stage, a core city in the megaregion will get priority because of its potential bridging function in the region, and as such, the ability to make sure that also smaller cities in that megaregion become part of the megaregional network through this core city.

It is not the case that the three functionally highly coherent megaregions we identified can 'sit back and relax'. They are highly coherent in comparative terms, but could perhaps be much more coherent. The

question is whether they are also coherent in political-institutional and cultural-symbolic dimensions. These dimensions complement the functional one and are understood to be intertwined and part of a process of 'metropolization' (Cardoso and Meijers, 2020) that is always in progress but never finished. Future implementation of our analysis framework would profit from taking those political-institutional and cultural-symbolic dimensions more into account. The lexicon-based toponym co-occurrence method offers plenty of opportunities here, and the challenge will be to detail lexicons that capture these dimensions to a greater extent than we have done here."

This is also one of the main advantages of applying the toponym co-occurrence method in combination with a lexicon-based approach to classify the relationships found. A significant advantage of this approach is also its applicability in scenarios where traditional data sources are unavailable, as there will always be relevant text corpora to which it can be applied (Tongjing et al., 2023; Meijers and Peris, 2019). Their availability, and through digitalization of historic archives also the increasing availability of text corpora for the past, hence allowing for longitudinal analysis, is another important advantage.

Critiques of the toponym co-occurrence method often focus on its lack of definitive interpretations of tangible interactions between cities (Watts, 2004). Indeed, this method, like other text mining-based approaches for quantifying relationships, does not directly measure physical exchanges or concrete interactions. However, its value lies more in also providing insights into the symbolic or representational aspects of city relationships. These aspects, though less tangible, are crucial for a comprehensive understanding of the complexities inherent in urban relationships. However, as an emerging methodology, it certainly invites further exploration to realize its full potential. This exploration could involve comparative analysis with other datasets, such as railway networks or capital flow networks (Pan et al., 2020; Zhang et al., 2020), or integration into existing models to assess and enhance accuracy (Overell and Rüger, 2008; Wu et al., 2019).

Beyond a necessary and quite straightforward extension of our work such as using other data revealing functional relationships, we can imagine four main logical extensions of our work. First, we need to delve into those political-institutional and cultural-symbolic dimensions just mentioned,

and perhaps accommodate these in an extended framework moving beyond functional coherence to analyze coherence more generally (Du et al., 2024). Second, while our framework was designed to be applicable to megaregions all across the world, and the data on intercity relations we used based on toponym co-occurrences can in principle be obtained in a uniform and harmonized way for all megaregions across the world (taking language bias into account), this needs to be empirically tested. Third, a potential extension of our three-dimensional framework inclusion-connectivity-consistency involves taking into account the division of labor between the cities constituting a megaregion. A question then is whether increasing integration will lead to increased specialization in megaregions, or vice versa, whether it is more a matter of complementarities fueling integration. Whether these processes can be adequately captured through a lexicon-based approach applied to toponym co-occurrences remains to be seen. Fourth, long-term analysis of the evolution of the coherence of megaregions, and the mechanisms driving the formation of relations (Li and Phelps, 2019), and how this evolution has impacted their performance across a wide range of indicators is essential to substantiate the many claims made for the megaregion.

References

Ahrend, R., Farchy, E., Kaplanis, I., & Lembcke, A. C. (2017). What makes cities more productive? Evidence from five OECD countries on the role of urban governance. Journal of Regional Science, 57(3), 385-410.

Bathelt, H., & Glückler, J. (2017). Toward a relational economic geography. In Economy (pp. 73-100). Routledge.

Burdett, R., & Sudjic, D. (Eds.). (2007) The endless city: The urban age project by the London School of Economics and Deutsche Bank’s Alfred Herrhausen Society. Phaidon Press.

Burger, M., Meijers, E., & Van Oort, F. G. (2014a). Multiple perspectives on functional coherence: Heterogeneity and multiplexity in the Randstad. Tijdschrift voor Economische en Sociale Geografie, 105(4), 444-464.

Burger, M. J., Van Der Knaap, B., & Wall, R. S. (2014b). Polycentricity and the multiplexity of urban networks. European Planning Studies, 22(4), 816-840.

Camagni, R., Capello, R., & Caragliu, A. (2017). Static vs. Dynamic Agglomeration Economies: Spatial Context and Structural Evolution

Behind Urban Growth. In Capello, R. (Ed) Seminal Studies in Regional and Urban Economics: Contributions from an Impressive Mind (pp. 227-259).

Cardoso, R. V., & Meijers, E. J. (2017). Secondary yet metropolitan? The challenges of metropolitan integration for second-tier cities. Planning Theory & Practice, 18(4), 616-635.

Cardoso, R. V., & Meijers, E. (2021). Metropolisation: The winding road toward the citification of the region. Urban Geography, 42(1), 1-20.

Cardoso, R. V., & Meijers, E. (2020). The process of metropolization in megacity-regions. In D. Labbé & A. Sorensen (Eds.), Handbook of Megacities and Megacity-Regions (pp. 360-375). Edward Elgar Publishing.

Commission of the European Communities [CEC] (2010). Territorial Agenda of the European Union 2020: Towards an Inclusive, Smart and Sustainable Europe of Diverse Regions, Retrieved from https://vb.nweurope.eu/media/1216/territorial_agenda_2020.pdf (accessed 4 June 2023).

Chen, S., Zhao, X., & Zhou, L. (2023). Which works better? Comparing the environmental outcomes of different forms of intergovernmental collaboration in China's air pollution control. Journal of Environmental Policy & Planning, 25(1), 16-28.

Chien, S. S., & Gordon, I. (2008). Territorial competition in China and the West. Regional Studies, 42(1), 31-49.

Chung, H. (2015). Unequal regionalism: Regional planning in China and England. Planning Practice & Research, 30(5), 570-586.

Contant, C. K., & de Nie, K. L. (2009). Scale matters: Rethinking planning approaches across jurisdictional and sectoral boundaries. Megaregions: Planning for global competitiveness, 11-17.

Cooke, P., & Morgan, K. (1994). Growth regions under duress: Renewal strategies in Baden-Württemberg and Emilia-Romagna. Globalization, institutions, and regional development in Europe, 91-117.

Derudder, B., Meijers, E., Harrison, J., Hoyler, M., & Liu, X. (2022). Polycentric urban regions: conceptualization, identification and implications. Regional Studies, 56(1), 1-6.

Dewar, M., & Epstein, D. (2007). Planning for “megaregions” in the United States. Journal of Planning Literature, 22(2), 108-124.

Du, Y., Cardoso, R. V., & Rocco, R. (2024). The challenges of high-quality development in Chinese secondary cities: a typological exploration. Sustainable Cities and Society, 105266.

Fang, C. (2015). Important progress and future direction of studies on China’s urban agglomerations. Journal of Geographical Sciences, 25, 1003-1024.

Fang, C., & Yu, D. (2016). Spatial pattern of China’s new urbanization. In Springer Geography (pp. 179-232). (Springer Geography). Springer.

Fang, C., & Yu, D. (2017). Urban agglomeration: An evolving concept of an emerging phenomenon. Landscape and urban planning, 162, 126-136.

Farole, T., Rodríguez-Pose, A., & Storper, M. (2011). Human geography and the institutions that underlie economic growth. Progress in Human Geography, 35(1), 58-80.

Florida, R., Gulden, T., & Mellander, C. (2008). The rise of the mega-region. Cambridge Journal of Regions, Economy and Society, 1(3), 459-476.

Fu, Y., & Zhang, X. (2020). Mega urban agglomeration in the transformation era: Evolving theories, research typologies and governance. Cities, 105, 102813.

Glass, M. R. (2015). Conflicting spaces of governance in the imagined Great Lakes megaregion. In J. Harrison & M. Hoyler (Eds.), Megaregions: Globalization’s New Urban Form? (pp. 119-145). Edward Elgar.

Gottmann, J. (1957). Megalopolis or the Urbanization of the Northeastern Seaboard. Economic Geography, 33(3), 189-200.

Green, N. (2007). Functional polycentricity: A formal definition in terms of social network analysis. Urban studies, 44(11), 2077-2103.

Hall, P. G., & Pain, K. (Eds.). (2006). The polycentric metropolis: Learning from mega-city regions in Europe. Routledge.

Harrison, J., & Hoyler, M. (2014). Governing the new metropolis. Urban Studies, 51(11), 2249-2266.

Harrison, J., & Hoyler, M. (2015). Megaregions: foundations, frailties, futures. In Megaregions (pp. 1-28). Edward Elgar Publishing.

Harrison, J., & Gu, H. (2021). Planning megaregional futures: Spatial imaginaries and megaregion formation in China. Regional Studies, 55(1), 77-89.

Harrison, J., Hoyler, M., Derudder, B., Liu, X., & Meijers, E. (2023). Governing polycentric urban regions. Territory, Politics, Governance, 11(2), 213-221.

Sorensen, A. (2020). Urbanization and developmental pathways: Critical junctures of urban transition. In D. Labbé & A. Sorensen (Eds.), Handbook of Megacities and Megacity-Regions (pp. 47-64). Edward Elgar Publishing.

Lang, R., & Knox, P. K. (2009). The new metropolis: Rethinking megalopolis. Regional studies, 43(6), 789-802.

Li, Y., & Jonas, A. E. (2019). City-regionalism as countervailing geopolitical processes: The evolution and dynamics of Yangtze River Delta region, China. Political Geography, 73, 70-81.

Li, Y., & Phelps, N. A. (2019). Megalopolitan glocalization: the evolving relational economic geography of intercity knowledge linkages within and beyond China's Yangtze River Delta region, 2004-2014. Urban Geography, 40(9), 1310-1334.

Li, Y., & Wu, F. (2012). The transformation of regional governance in China: The rescaling of statehood. Progress in Planning, 78(2), 55-99.

Liu, X., Derudder, B., & Wu, K. (2016). Measuring polycentric urban development in China: An intercity transportation network perspective. Regional Studies, 50(8), 1302-1315.

Meijers, E. J., Burger, M. J., & Hoogerbrugge, M. M. (2016). Borrowing size in networks of cities: City size, network connectivity and metropolitan functions in Europe. Papers in Regional Science, 95(1), 181-198.

Meijers, E. J, & Peris, A. (2019). Using toponym co-occurrences to measure relationships between places: Review, application and evaluation. International Journal of Urban Sciences, 23(2), 246-268.

Meijers, E. J., & Burger, M. J. (2017). Stretching the concept of 'borrowed size'. Urban studies, 54(1), 269-291.

Meijers, E. J, Hoogerbrugge, M., & Cardoso, R. (2018). Beyond polycentricity: Does stronger integration between cities in polycentric urban regions improve performance?. Tijdschrift voor economische en sociale geografie, 109(1), 1-21.

Mello, R. A. (2002). Collocation analysis: A method for conceptualizing and understanding narrative data. Qualitative research, 2(2), 231-243.

National Development and Reform Commission. (2023). The 14th five-year plan-chapter 28-Improving urban spatial distribution. Accessed January 6, 2024, from https://en.ndrc.gov.cn/policies/202209/P020220923494700193893.pdf

Overell, S., & Rüger, S., (2008). Using co-occurrence models for placename disambiguation. International Journal of Geographical Information Science, 22, 265–287.

Pan, F., Bi, W., Liu, X., & Sigler, T. (2020). Exploring financial centre networks through inter-urban collaboration in high-end financial transactions in China. Regional Studies, 54(2), 162– 172.

Pan, F., Zhang, F., Zhu, S., & Wójcik, D. (2017). Developing by borrowing? Inter-jurisdictional competition, land finance and local debt accumulation in China. Urban Studies, 54(4), 897-916.

Regional Plan Association. (2006). America 2050: A prospectus. Retrieved from https://s3.us-east-1.amazonaws.com/rpa-org/pdfs/2050-Prospectus.pdf (accessed 4 June 2023).

Sassen, S. (2007). A sociology of globalization. Análisis político, 20(61), 3-27.

Smart, A. (2018). Ethnographic perspectives on the mediation of informality between people and plans in urbanising China. Urban Studies, 55(7), 1477-1483.

Tongjing, W., Meijers, E., & Wang, H. (2022). The multiplex relations between cities: a lexicon-based approach to detect urban systems. Regional Studies, 1-13.

Tongjing, W., Meijers, E., Bao, Z., & Wang, H. (2023). Intercity networks and urban performance: a geographical text mining approach. International Journal of Urban Sciences, 1-22.

Wang, L., Xue, X., Zhou, X., Wang, Z., & Liu, R. (2021). Analyzing the topology characteristic and effectiveness of the China city network. Environment and Planning B: Urban Analytics and City Science, 48(9), 2554-2573.

Wang, Y., Sun, B., & Zhang, T. (2022). Do polycentric urban regions promote functional spillovers and economic performance? Evidence from China. Regional Studies, 56(1), 63-74.

Watts, D. (2004). The “new” science of networks. Annual Review of Sociology, 30, 243-270.

Wu, F. (2016). China's emergent city-region governance: a new form of state spatial selectivity through state-orchestrated rescaling. International Journal of Urban and Regional Research, 40(6), 1134-1151.

Wu, F. (2018). Planning centrality, market instruments: Governing Chinese urban transformation under state entrepreneurialism. Urban studies, 55(7), 1383-1399.

Wu, F. (2020a). Adding new narratives to the urban imagination: An introduction to ‘New directions of urban studies in China’. Urban Studies, 57(3), 459-472.

Wu, F. (2020b). The state acts through the market:‘State entrepreneurialism’beyond varieties of urban entrepreneurialism. Dialogues in Human Geography, 10(3), 326-329.

Wu, F., & Zhang, J. (2007). Planning the competitive city-region: The emergence of strategic development plan in China. Urban Affairs Review, 42(5), 714-740.

Wu, J., Feng, Z., Zhang, X., Xu, Y., & Peng, J. (2020). Delineating urban hinterland boundaries in the Pearl River Delta: An approach integrating toponym co-occurrence with field strength model. Cities, 96, 102457.

Xu, J. (2008). Governing city-regions in China: Theoretical issues and perspectives for regional strategic planning. The Town Planning Review, 157-185.

Yao, S., Xu, X. Q., Wu, C. C., Huang, G. Y., Wang, L. P., Sun, J. S., ... Hou, X. H. (1992). Zhongguo de Chengshi Qun (The Urban Agglomerations of China). The Science and Technology University of China Press, Hefei, China.

Yeh, A. G., & Xu, J. (Eds.). (2010). China's Pan-Pearl River Delta: regional cooperation and development (Vol. 1). Hong Kong University Press.

Zhang, W., Derudder, B., Wang, J., & Witlox, F. (2020). An analysis of the determinants of the multiplex urban networks in the Yangtze River Delta. Tijdschrift Voor Economische en Sociale Geografie, 111(2), 117–133.

Zhang, X. (2019). Transformation of Chinese cities and city-regions in the era of globalization. In R. Yep, J. Wang & T. Johnson (Eds.), Handbook on Urban Development in China (pp. 137-154). Edward Elgar.

Chapter 6

Conclusion and discussion

1. Conclusions

This thesis explored the potential of collocation analysis in the field of city network and geography, focusing on the main research question:

To what extent and how can collocation analysis be utilized to extract, categorize, analyze, and evaluate relationships between cities?

To provide comprehensive and detailed answers, the main research question was subdivided into four sub-questions. Each sub-question addressed a specific yet connected aspect of collocation analysis. This section will start by answering each sub-question, which is then followed by the answer to the main question.

1.1 Answers to each sub-question

1) *How can city relationships be effectively and practically extracted from a large database for social domain researchers?*

This sub-question stemmed from the difficulty of handling large textual datasets, which pose several challenges for domain researchers, particularly in the social studies field.

The first challenge is about data processing management. Typically, computers temporarily store the entire dataset in memory during processing. However, this approach can lead to system overload when handling large datasets, due to their significant memory requirements.

The second challenge involves developing suitable processing algorithms for large-scale data, which often demands substantial computing power. This becomes particularly critical when machine learning methods are often data-intensive, as those methods require extensive computational resources that exceed the availability and capacity in many social studies.

The third challenge is that processing large datasets often exceeds the computational power and memory capabilities of standard computers. Thereby, it calls for the use of cloud computing platforms, like Amazon Cloud or Google Cloud. Transitioning to these cloud-based solutions introduces a set of complex challenges, including ensuring compatibility between pre-existing data processing frameworks and the cloud

infrastructure, managing the associated costs, and addressing the technical complexities inherent in cloud migration.

The three challenges are more from conceptual perspectives rather than technical, underscoring the need for researchers to recognize that, although traditional data processing methods may appear intuitive, they actually necessitate substantial computing resources. This requirement renders them impractical for analyzing large datasets.

Regarding the first challenge, Chapter 2 introduced an efficient and easy-to-follow method. This method first divided the entire dataset into smaller, manageable segments. These segments were then processed in parallel, reducing the strain on memory resources. Subsequently, the processed data from these segments were cohesively reassembled to form the complete dataset. As traditional programming languages like Python or R are not ideally suited for these types of memory-intensive, parallel processes, Scala, a programming language specifically designed for large-scale data processing, was used. To effectively use Scala, the distributed computing environment Hadoop was then developed.

To address the second challenge, this study used keyword-based filtering, as it is a far more efficient method than machine learning-based methods, requiring much less computing power. The initial step in this approach involved significantly narrowing down the dataset by filtering out irrelevant content, including pages in languages other than Chinese, as well as pages predominantly containing gambling or sexual content. Specifically, this study started by examining a sample representing 0.1% of the total dataset. This examination revealed that a majority of the webpages in Chinese contain over 90% Chinese characters. However, verifying each page for this specific percentage was computationally demanding. This study then identified a consistent feature among these web pages - the presence of at least 10 continuous Chinese characters. Consequently, the methodology adopted this feature as a criterion to determine whether a webpage is predominantly in Chinese. Besides, the process involved filtering out web pages with gambling or explicit content. This was achieved through keyword filtering, utilizing a widely-referenced censor keyword dictionary sourced from GitHub.

To mitigate the third challenge, the study initially processed small samples on a local computer to validate each step and then transitioned these small

samples to the cloud. Only after everything worked well, was the entire dataset processed. The first two steps were critical to ensure that the tasks can be successfully executed before they are scaled up to a larger operation on a cloud service. This way provides a safeguard against potential processing failures when handling the full dataset on a cloud platform.

This method was demonstrated through a case study focused on extracting Chinese toponym co-occurrence from a substantial 6.98 TB dataset obtained from the Common Crawl web archive. The streamlined parallel processing approach was implemented using Scala within a Hadoop environment, specifically on the Amazon Elastic Map/Reduce infrastructure. This operation was conducted using a robust 1080 CPU cluster. Remarkably, the efficiency of this method is underscored by its cost-effectiveness, with a total cost of less than $100. This economical approach led to the filtering out of approximately 93% of the initial dataset, effectively condensing it to a more manageable size of 202 GB. The ability to conduct such extensive data processing at a minimal cost highlights the feasibility of applying advanced computational platforms but simple programming techniques in the field of geography.

This research greatly benefited from the collaboration with two computer science engineering students, given my initial lack of familiarity with the practical challenges associated with processing large datasets and coding up as of 2019. However, as noted earlier, effectively handling large datasets hinges more on a conceptual understanding that traditional data processing methods fall short in big data contexts, rather than on the ability to produce complex code.

In terms of actual coding, it was challenging at the time of practice. However, the emergence of Large Language Models (LLMs), such as ChatGPT, has provided a valuable resource for automatically generating high-quality code that meets specific requirements, which can ease this challenge for social study researchers significantly.

Another potential difficulty was to operate the dataset processing methods in the AWS cloud environment. Initially, several efforts failed, and no solution was found. Subsequently, we sought assistance from AWS customer service, which proved invaluable in overcoming these hurdles. The customer serivce implemented our methods and found the error and corredted our coding. This experience shows that, with the right support,

the execution of such technologically demanding tasks is attainable for social science researchers, even those with a limited background in computer science.

2) *How can relations between cities as found through toponym co-occurrences be categorized?*

While Chapter 2 demonstrated the feasibility of extracting city relationships from a large database, the relationship patterns this method extracted were broadly defined and did not take context into account. Recognizing that intercity relationships can be diverse and not uniformly identical, the second sub-question delved deeper into identifying a method that can classify the relationships by considering the context where the place name keywords co-appear.

Classifying city relationships extracted from texts essentially involves classifying the text itself. However, text classification, especially when dealing with unstructured text, presents significant challenges, even to computer science researchers. Typically, text classification employs machine learning-based methods, which categorize text by extracting various textual features and comparing 'similarities' among them through supervised or unsupervised methods. Regardless of whether supervised or unsupervised learning methods are used, there is always a risk of relying on irrelevant textual features and misclassifying them. This issue is particularly crucial when analyzing intercity relationships through unstructured text, where clarity and relevance in classification are key to obtaining meaningful insights.

After recognizing this difficulty, Chapter 3 decided to choose textual features that are clear and meaningful, instead of trying to capture as many as possible textual features. Therefore, it introduced a lexicon-based classification approach, grounded in a rationale similar to that of collocation analysis. This lexicon-based approach posits that the co-occurrence of two city names alongside a specific keyword implies a relationship that is not only interconnected between the cities themselves but is also characterized by the context provided by the accompanying keyword. This method simplifies the classification process while aiming to retain the clarity of the relationships being extracted.

The lexicon dictionary for this method was derived from the China Standard Industrial Classification of Economic Activities (GB/T 4754-2017). Keywords from this standard served as the initial seed words. These were then augmented using a natural language processing model to identify and include semantically related terms. The classification process began by classifying each webpage based on the proportion of keywords from the dictionary present. Subsequently, city name appearances in each classified text were counted and aggregated. For instance, if a text contained 10 keywords from the dictionary—five pertaining to finance, three to government, and two to research—it was classified as 50% finance-related, 30% government-related, and 20% research-related. If two city names appeared in this text, their relationship was categorized accordingly. These results were then aggregated to represent the strength of the relationship between a pair of cities for each category.

Then, this lexicon-based approach was applied to the intercity relationships dataset extracted, as described in Chapter 2, and successfully showed substantial variations in the network patterns of different relationship types. Such findings call for caution in making assertions about the structure of urban systems based solely on a single type of relationship. For example, governmental relationships decayed more rapidly than other types, implying that such collaborations are typically more localized. The government network pattern encompassed more cities but with looser connections. Additionally, financial and cultural intercity relationships varied more significantly from the expected patterns of the gravity model compared to other relationship types.

3) *What is the comparative importance of city relationships vis-à-vis agglomeration in influencing city performance?*

Following the extraction of various types of city relationships in Chapter 2 and 3, the third sub-question was developed to apply the new knowledge on relationship patterns through exploring the impact of these relationships on city performance. This sub-question was designed to assess whether cities should prioritize expanding in size in order to enhance agglomeration benefits or focus on strengthening intercity relationships to leverage network externalities. While much theoretical research has made significant contributions to this topic, empirical evidence is still limited. This is partly due to the historical scarcity of large-scale data on intercity relationships. Now, with the application of collocation analysis, a

substantial dataset has been created, providing the necessary empirical foundation to undertake this evaluation.

Apart from the challenge of data availability, there were three major challenges in addressing this sub-question. The first challenge lies in evaluating whether the city relationships, as indicated by collocation patterns, align with common sense. Since these patterns might represent a novel type of city relationship, directly comparing them with other types of city relationships can be challenging.

The second challenge was to assess the impact that population size has on city relationships. Larger cities are mentioned more often in text, which could potentially undermine the statistical significance of population size in comparing a city's network position with its agglomeration benefits, a phenomenon known as the collinearity issue.

The third challenge was the identification of appropriate metrics that can accurately represent a city's position within a network. While network science provides a wide range of metrics to describe a city's position in a network from different perspectives, their applicability in the context of city networks, especially regarding social and economic implications, can be restricted. This challenge was particularly acute in the context of fully connected weighted networks, which often pose 'wicked' problems in network science. In these scenarios, it is critical to develop metrics that are capable of providing clear and direct insights into a city's network position.

In addressing the first challenge, this study opted for a qualitative assessment to validate whether the novel patterns identified are logically consistent with our existing understanding of city relationships. To this end, the dataset from Chapter 2 was visualized as a city network. In this visualization, it was observed that relationships with the highest co-occurrences are primarily among major cities, forming a diamond-shaped pattern. This pattern mirrors the backbone structure of national networks as identified in other intercity relationship data, such as railway and flight routes. Notably, the majority of strong relationships were concentrated in the southeast part of China, dividing the country into two parts. This pattern aligns with the fact that China's major economic activities are also predominantly concentrated in this region. Further analysis, including the calculation and ranking of each city's connections with others, showed that the top-tier cities identified correspond with findings from other studies.

Therefore, while direct verification of the validity of collocation-based patterns with other types of city relationships might be challenging, the fact that the observed general pattern aligns with known realities lends the credibility and applicability of this method.

In response to the second challenge, and indirectly also the first challenge, this study utilized the gravity model as a benchmark to validate whether the extracted city relationship patterns also follow the gravity model estimation that distance and populations are significantly important factors for determining city relationship strength.

The findings revealed that while population indeed plays a significant role in determining the strength of collocation-based relationships between cities, its impact varies notably across different city relationships—Cities of comparable sizes can exhibit vastly different relationship strengths with others, suggesting the involvement of additional influential factors. Although this research did not identify what these other factors are, it highlights the existence of additional elements on city relationships by contrasting actual collocation patterns with predictions derived from the gravity model: the ratios that exceed 1 indicated that the strength of the relationships surpasses what can be estimated by just geographical proximity or population size, suggesting the existence of additional factors in strengthening intercity connections. Therefore, the study adopted this ratio as a proxy to represent the relative strength of intercity relationships.

Concerning the third challenge, the analysis began by evaluating the correlation among potential network metrics. This examination revealed that most of these metrics are, in fact, highly correlated. Consequently, to minimize the risk of collinearity, the study decided to use a single metric: the average relationship strength that a city maintains with other cities.

After addressing the identified challenges, regression models were conducted to evaluate the importance of city relationships on a city's performance. The analysis results showed a strong correlation between a city's strong network position and its overall performance, indicated by GDP per capita, underscoring the critical role of network externalities. This correlation was particularly notable among smaller Chinese cities. Intriguingly, the analysis also suggested that, for the largest 50 Chinese cities, the costs associated with agglomeration tend to outweigh the benefits. In contrast, network externalities emerged as a more significant

factor in explaining productivity level variations across cities than agglomeration.

4) *How can the functional coherence of proposed megaregions be scrutinized and categorized based on the strength of their intercity relationships?*

Having explored the importance of city relationship data in analyzing network externalities at the city level, this sub-question was designed to showcase the importance of having multiple types of city relationships in developing planning strategies at the megaregional scale. The motivation for this question arose from the growing global interest in developing megaregion plans, particularly in China. However, methods to assess the development stage of functional coherence were often missing. This chapter posited that for megaregion plans to achieve their claimed benefits, cities within these regions should have strong relationships across various socioeconomic dimensions, a concept termed "functional coherence". Accordingly, planning strategies for developing megaregions should aim to enhance different dimensions of this functional coherence. A critical first step in this direction was the assessment of the actual level of functional coherence within the planned megaregions in China.

In addressing this subquestion, three main challenges were identified. The first challenge was to find suitable metrics that could proxy a megaregion's functional coherence. The concept of functional coherence in megaregions is abstract, making it challenging to determine metrics that can accurately and comprehensively reflect the extent of functional coherence.

The second challenge concerned the computational intricacies involved in analyzing weighted, fully connected networks. This task was more complex than the one outlined in the third sub-question, as it involved dealing with multiple types of intercity relationships, as opposed to just one. A particular concern was the presence of relationship strengths that were significantly higher than the average, which posed a risk of introducing biases into the analysis.

The third challenge was the categorization of the development levels of each megaregion. The difficulty lay in the absence of a clear standard or benchmark for defining what constitutes functional coherence. This

ambiguity presented difficulties in uniformly assessing and comparing the integration and effectiveness of different megaregions.

To address the first challenge, Chapter 5 proposed a multi-dimensional framework for assessing the functional coherence of planned megaregions. This framework includes on three key dimensions: inclusion (the extent to which cities in a planned megaregion are interconnected), connectivity (the degree of connection among these cities), and consistency (the uniformity of connections across different categories of relationships). The analysis adopted a progressive approach, starting from the city scale, advancing to each relationship layer, and concluding with an evaluation of overall consistency. This evaluation identified specific areas on which each planned megaregion should focus to enhance its functional cohesion.

To tackle the second challenge, the study transformed the original weighted network into a series of non-weighted networks based on various cut-off thresholds. In each of these non-weighted networks, the functional coherence metrics were then recalculated. This approach reduces potential biases that could arise from a few disproportionately high relationship strengths or the limitations of setting a single cut-off point, but meanwhile, it could still provide a complete profile of a region's functional coherence.

To overcome the third challenge related to the subjectivity of setting absolute standards, the study adopted an empirical comparative approach. Instead of imposing a fixed standard to categorize development levels, this approach involved a comparative assessment of differences across fifteen Chinese megaregions. By evaluating these regions relative to each other, the analysis avoided the pitfalls of setting arbitrary benchmarks.

1.2 Answer to the main research question

After the discussion of the individual sub-questions, this chapter now turns to the main research question:

To what extent and how can collocation analysis be utilized to extract, categorize, analyze, and evaluate relationships between cities?

The main question will be addressed from two perspectives: capabilities and limitations. The capabilities aspect examines the practicality of using collocation analysis in identifying and categorizing city relationships. The

limitations aspect assesses the depth and accuracy of the captured collocation-based city relationships.

1.2.1 Capabilities of using collocation-based patterns in analyzing city relationships

Extraction: Collocation analysis was demonstrated to be a feasible method for extracting intercity relationships, with the results mirroring city relationships to a certain extent: most of the strong relationships concentrate in the southeastern side of China where the Chinese population also concentrates, and the well-known "Hu Huanyong Line" divide is visible and the strongest relationships together form a diamond shape, which is also found in many other studies by other city relationship data, where the outline is formed by five city clusters: Yangtze River Delta (Shanghai, Hangzhou, and Nanjing) in the east; Pearl River Delta (Guangzhou, Shenzhen, and Hong Kong) in the south; Chengyu Region (Chongqing and Chengdu) in the west; the Beijing-Tianjin-Hubei Region (Beijing and Tianjin) in the north; and the Middle-Yangzte River Region (Wuhan) in the center. These collocation-based patterns also align with the gravity model, affirming that population size and distance are significant factors influencing the strength of intercity relationships, as they do in other forms of intercity relationships.

Regarding actual practice, for processing small text samples, standard computing methods utilizing Python and R prove adequate. However, extracting relationships from larger text samples, specifically those in gigabyte or terabyte sizes, can be challenging due to the memory limitations of standard computers. To address this, Chapter 2 developed a keyword-based parallel cloud computing method that can efficiently extract collocation patterns from a large text corpus.

Categorization: collocation analysis also proved its potential to classify the relationships by capturing certain categorical keywords with city name co-occurrences. To address this challenge, Chapter 3 proposed a lexicon-based classification method. The principle of this method is to use textual features that are straightforward and easily comprehensible, thereby delivering clear and interpretable results. This is in contrast with machine learning-based methods, which are capable of analyzing a vast array of textual features, but their effectiveness in yielding socially relevant and

widely applicable classification results is not guaranteed, nor transparent and results hard to replicate and compare.

Evaluation: Network science has been applied to analyze collocation-based city relationships and two primary challenges have been identified. The first challenge, discussed in Chapter 4, arises from the observation that collocation patterns are significantly influenced by the population of cities and the distance between them. To discern the significance of intercity relationships beyond these basic factors, Chapter 4 introduced a ratio that compares actual collocation patterns with gravity model estimations, to represent the relative relationship strength between two cities. Using this metric, this chapter proceeded to evaluate the relative importance of a city's network position vis-à-vis its population size.

The second challenge is that collocation-based relationships are weighted and tend to be fully connected, but only a few relationships are notably strong, typically among large cities. Therefore, a direct application of network metrics in such cases would disproportionately highlight these dominant relationships while diminishing the still relatively crucial relationships between smaller cities. To counteract this, Chapter 5 proposed a method to transform weighted networks into a series of non-weighted ones. This transformation ensured that the analytical results were clear, comprehensive, and objective while mitigating the risk of distortion caused by overemphasizing a few significant ones.

1.2.2 Limitations of using collocation-based patterns in analyzing city relationships

Extraction: The collocation method has faced criticism for its assumption that the frequency of word co-occurrences directly reflects the importance or significance of the relationship between those words, which overlooks the nuances and context surrounding each instance of co-occurrence. Consequently, this can lead to a lack of accuracy and depth in the analysis, particularly when it comes to unraveling the complexities of relationships.

However, it's also important to acknowledge the reality that contemporary information consumption habits tend towards superficiality, with many individuals skimming newspaper headlines and watching video clips without delving into deeper analysis. In this context, while co-mentions

may not fully unravel the complexities of word relationships, frequent mentions can indeed serve as indicators of public interest or attention.

As a method capable of quantifying city relationships from textual data on a large scale, it is actually already better than many other types of text analysis methods in terms of clarity and objectivity. Close-reading methods, while providing detailed analyses, struggle with processing text on a large scale and the objective quantification of city relationships. On the opposite side, machine learning methods can capture a broader range of information about city relationships and quantify them in objective ways. However, the process of compressing various textual features into a singular numerical representation often results in a loss of essential social relevance. Therefore, the collocation method is a balanced alternative between close-reading and machine learning-based methods for quantifying the strength of text-based city relationships.

The challenge of quantifying relationships from textual data is essentially about defining the nature of these relationships as they manifest themselves in text. The task differs from traditional datasets that capture city relationships through tangible metrics such as the flow of goods, people, or the volume of collaborative scientific publications. When relationships between cities are expressed in textual forms, they often acquire an subjective dimension, demanding personal interpretation that is beyond the scope of straightforward numerical quantification.

For example, consider the headline from the Washington Street Journal on March 14, 2024, “U.S and China Extend Landmark Bilateral Deal, Very Quietly.” This headline refers to the U.S.-China Science and Technology Agreement, an umbrella agreement for cooperation in research that was first signed in 1979 but expired on August 27, 2023. While the agreement itself promises no specific funding or concrete outcomes, it serves as a longstanding symbol of cooperation between the U.S. and China. The decision to extend the agreement for only six months, rather than the usual five years, and the low-key announcement, reveal the complexity of ongoing geopolitical competition between the U.S. and China. The interpretation of this headline can vary widely depending on the perspective of the reader. Some may view it as a strong positive sign of continued cooperation in research and politics between U.S. and China, while others may perceive it as less significant.

One of the key issues practically affecting the precision of extracting collocation-based relationships is the toponym ambiguity problem. The precision can be significantly influenced by whether the place names counted actually correspond to the intended geographical entities. Compared to close-reading and machine learning-based methods, the collocation method generally delivers less accurate results in this respect, as it takes into account less contextual information when identifying place names. This thesis does not thoroughly address this issue; it simplifies the approach by eliminating cities with a high risk of toponym ambiguity instead of verifying each instance to determine if a city name in the text truly corresponds to the specific entity.

To better address this issue within the collocation method, an effective strategy is to incorporate more contextual information into the analysis process. For example, instead of merely searching for two place names, the analysis could also consider additional keywords in the surrounding text. By searching for combinations such as “London”, “Toronto”, and “Canada” within the same sentence, the method can more accurately determine that London” refers to the city in Canada rather than in the UK.

A subjective element in collocation analysis is to decide under what circumstances two placenames are considered to co-occur. It could be when they appear within the same textual unit—like a sentence, paragraph, or article—or within a specified word window (e.g., three, five, or more words). Another subjectivity issue involved is the frequency of these co-occurrences within a single textual unit, for instance, whether three occurrences within one paragraph should be given more weight than if they appear only once. These issues were not explored within the scope of this thesis but could be a direction for future research, highlighting areas where further refinement and definition could enhance the precision and applicability of collocation analysis in studying city relationships.

Categorization: Categorization in text analysis fundamentally involves extracting insights from texts and classifying them based on common characteristics. Collocation methods, while efficient, typically capture less contextual information than close reading and machine-learning techniques. In this sense, classification results of collocation methods tend to be more superficial than other methods. It often highlights the relative importance of a topic in comparison to a general context.

However, to be practical, it is also difficult for humans to gain insights from complex documents such as government policy papers, legal documents, and corporate reports let alone webpages. While machine learning methods, particularly Large Language Models (LLMs) such as ChatGPT, which incorporates millions of textual features and even historical background information, seemingly provide more objective answers, human input remains essential. These LLMs still rely on human input, with thousands of engineers in Kenya labeling texts for ChatGPT, providing guidance on classifying and identifying important text features. Even if such models can classify and interpret such complex texts, the inherently opaque nature of these methods' analysis processes raises concerns about their ability to maintain social relevance. This opacity brings into question the extent of credibility that we can lend to their classification results.

Classification fundamentally is about interpreting the text, which requires human judgment in the end. Therefore, the unique advantage of the lexicon-based approach proposed in Chapter 3, which focuses on counting keyword appearances, is straightforward for human understanding.

Nevertheless, there is potential for further development by integrating collocation methods with close-reading techniques. This strategy would begin by collecting collocation patterns from extensive documents to establish a general understanding. Attention would then be directed toward identifying the patterns that are significantly higher or lower than expected. These outliers would undergo close-reading analysis to extract deeper insights. In this case, the collocation method serves as a navigational tool to highlight specific contexts warranting detailed examination. This methodology underscores the indispensable role of human interpretation, ensuring that the nuanced comprehension of documents is not lost.

Evaluation: this thesis presented two case studies to demonstrate the utility of collocation-based results—one focusing on the influence of city relationships on urban development at a city scale, and another examining the functional coherence of a delineated megaregion plan at a regional scale. While these two studies underscore the importance of having and utilizing data on relations between cities, an evaluation of the quality of the collocation-based relational data has not been explicitly addressed within these case studies. This is because evaluating collocation-based data is challenging, as collocation-based relationships are a new type of city

relationship analysis. Thereby, a mere similarity of this data to other city relational data does not necessarily affirm its quality.

Nonetheless, two approaches can be considered for future data validation: backward (retrospective) validation and forward (prospective) validation. A practical strategy for retrospective validation is to assess the collocation patterns with qualitative analysis methods, such as Critical Discourse Analysis (CDA). This method can be employed to examine whether the identified frequent co-occurrences and the results of lexicon-based classification are logical and meaningful, thereby ensuring the validity of the data extraction and categorization processes. Additionally, another approach for evaluating classification results is to apply this lexicon-based method to texts previously categorized by machine learning-based methods. In this context, the results derived from machine learning methodologies can act as a benchmark against the outcomes obtained through the lexicon-based method. While identical outcomes may not be achieved due to methodological differences, if the lexicon-based method is reliable, there should not be a significant disparity in the findings obtained through these distinct approaches.

Forward validation, on the other hand, assesses the practical utility of the method. This approach validates a dataset by its capability to contribute meaningful improvements to existing models. Several studies have pursued this direction, showcasing the potential benefits of incorporating collocation-derived data. However, it's critical to approach these enhancements with caution. An improvement in a model's accuracy, while beneficial, does not inherently validate the quality of the data used. Without a thorough understanding of the data's specific attributes and context, there's a risk of its inappropriate or theoretically unsound application within existing models. Such issues could potentially undermine the long-term reliability and validity of the improved models. Consequently, while such integrations might yield initial success, there is a risk that the results may not be reliable or accurate in the long term.

In summary, collocation analysis, as a keyword-based statistical method, stands out among text mining techniques through its capability to process large volumes of documents while maintaining interpretability and objectivity. The results tend to be more qualitatively robust and socially relevant than other text mining methods. However, broadening the applicability of this method necessitates improvements in several aspects:

developing a clearer definition of relationships as determined through collocation, validating the accuracy of the extraction process, and evaluating the relevance and precision of the identified patterns. Addressing these aspects is crucial for advancing collocation analysis as a more effective and reliable tool in the interpretation and quantification of text-based relationships.

1.2 policy implications

The policy implications derived from Chapters 4 and 5 underscore the significance of city relationships on urban performance at both city and regional levels, which highlights the necessity for more in-depth and ongoing study into the intercity relationships of urban systems.

Chapter 4 calls for a critical analysis of the traditional urban and regional development literature, where agglomeration benefits are considered the key driver of growth. The results of this chapter suggest that a city's relationship with others might be more important than the traditional thesis believes, in particular, smaller- and medium-sized cities can better gain competitiveness from being strongly related to other cities in the country, whereas larger cities profit less from their network position. These results imply to make Chinese cities more competitive and productive policy strategies should not be foremost oriented at a further concentration of people and firms in space. Policies could target the institutions and infrastructures that allow for such networks to develop.

Chapter 5 reveals a significant discrepancy between the normative classification of megaregions by the Chinese government and the actual functional coherence of those megaregions. The analysis of fifteen megaregions shows that in nine instances, the functional coherence does not align with the development levels defined by the central government. This divergence primarily stems from the analysis's focus on the overall connections within the planned megaregions, whereas the government's classification often emphasizes areas dominated by one or two megacities, which could potentially direct planning interventions and spatial investments in the wrong way. Instead, it is the overall intercity relationships within the planned region that are crucial for achieving functional coherence. This insight urges a reevaluation of how megaregions are classified and developed, advocating for a more holistic approach that considers the broader network of city relationships.

2. Future work direction

This thesis made an initial exploration into the contributions of collocation analysis in understanding intercity relationships, positioning it as a crucial component of urban and regional planning discourses. While progress has been made in this thesis, such as showcasing the extraction of collocation-based city relationships, providing an open-access dataset, proposing a lexicon-based classification method, and analyzing collocation-based relationships within the context of city and regional strategies, several critical questions remain open for further exploration. Accordingly, this section is structured into three subsections, each dedicated to outlining prospective research avenues: potential methodological advances; practical applications in revealing text elements; and implications for gaining policy insights.

2.1 Potential methodological advances

The method used in this thesis primarily focuses on identifying patterns and associations among a predetermined group of cities, which is referred to as the collocation-based approach. Distinct from this, there is another method called the collocation-driven approach. It initiates with a single keyword and subsequently proceeds to examine all possible terms that have strong associations with it. This type of method is referred to as the collocation-driven approach. For instance, it can start with the keyword “Amsterdam”, and then find all possible collocated words with “Amsterdam”, offering insights into whether the associations are novel, unexpected, or conform to established stereotypes, thereby illuminating how cities are conceptualized and discussed in the public sphere.

Within the collocation-driven framework, the scope of collocated terms is not confined to keywords alone but extends to encompass various types of elements. For instance, the elements can be sentiments identified by natural language processing, emojis in SMS communications, or tonal variations in audio files. By integrating these diverse elements into the collocation analysis, researchers can uncover richer, more detailed narratives and patterns that contribute to a broader understanding of the semantic landscapes within which cities and other subjects are embedded.

The traditional collocation method focuses on the mutual dependency of the two elements, characterizing their association as non-directional.

However, in reality, most city relationships are not symmetric. A case in point is the relationship between Amsterdam and Schiphol. The mention of Schiphol significantly increases the likelihood of Amsterdam being referenced, a pattern that might not be as pronounced in the opposite direction. This imbalance arises because Amsterdam maintains a wide range of strong relationships with other cities, in contrast to Schiphol, which is primarily acknowledged for its role as Amsterdam's airport. Therefore, a more precise assessment can be achieved by examining the likelihood of the occurrence of keyword B given the presence of keyword A. perspective on conditional probability provides a more accurate understanding of the directional nature of relationships between keywords, highlighting the asymmetric nature of their association, where one entity (Schiphol) is more likely to be mentioned in the context of another (Amsterdam) than vice versa.

It is also possible to expand the scope of collocation analysis to include triadic associations and beyond. This extension highlights the distinct nature of a triadic relationship (A-B-C) compared to a series of separate dyadic relationships (A-B, B-C, C-A), introducing the concept of higher-order networks within network science. This perspective is important in the context of city network analysis, where collaborations frequently extend beyond simple bilateral partnerships to encompass multi-party efforts. Such is often the case with scientific research projects collaborated by multiple institutions or metropolitan development plans engaging several cities. Therefore, this advanced method can capture a more accurate representation of city relationships, overcoming the limitations posed by traditional analyses focused solely on dyadic relationships.

The collocation method can also provide promising proxies for other forms of relationships, especially in contexts where direct measurement presents challenges. For instance, surveys are often conducted to understand public travel preferences between cities and perceptions about traveling patterns, but such surveys are costly and time-consuming. An alternative approach is to capture the city associations among the cities and examine the context for positive or negative sentiments in media. Media outlets often feature comment sections below their articles, allowing readers to provide feedback. This practice is prevalent not only in user-generated content platforms such as Twitter and Reddit, but also in traditional news websites like The New York Times. By comparing the commonly associated words and sentiments in the comments with those of the original article,

researchers can gain insights not only into the associated words in the content but also how readers respond to it. Consequently, researchers can gain preliminary insights into the association of these cities and the media and public's perception of their relationships.

More importantly, by comparing these results with survey data, researchers can evaluate the reliability of using media and online opinions to estimate reality obtained through surveys. This comparative analysis has the potential to provide a cost-effective and efficient means to measure public sentiment and city associations without the need for extensive surveys.

2.2 Analyzing text through comparing text elements

Given the reality that not every thought or action is captured in writing, word selection choice can reflect the writer's preferences. This is true even in today's digital age, where social media bots, capable of automatically generating and responding to messages, still operate with a degree of selectivity in their content generation. Therefore, collocation methods, by highlighting the frequency of specific word pairings, reveal preferences in word choice, thus providing insights into the subtle biases that shape text production and reception. For instance, it can be applied to indicate linguistic differences (English vs. French), genre (political news vs. leisure news), political ideologies (CNN vs. Fox News), the creators of content (influencers with different follower demographics), audience backgrounds, and distribution channels (national vs. local newspapers).

By taking into account text generation time, the collocation method enables tracking of how associated characteristics with a certain word change over time. This method can be employed to investigate the association between cities and specific events, examining the duration of such associations. For example, it enables the tracing of the connection between the COVID-19 pandemic and Wuhan, where the disease was first identified. Initially, the association between Wuhan and COVID-19 could be very strong. However, over time, this association is expected to weaken. The collocation method offers a quantitative approach to assess the lasting impact of positive and adverse events on locations, thereby deepening our insight into the evolution of public perception over time.

Moreover, by comparing associated words from different text sources, this method can also be used to measure the influence of text creators on the

degree of association. For example, researchers can gather tweets from both local government accounts and key opinion leaders focusing on local tourist attractions or events. Then by collecting metrics such as the number of followers of these accounts, and the number of comments, likes, and retweets each tweet receives, researchers can compare the effectiveness of local government marketing to attract tourists versus the impact of key opinion leaders promoting visits for sightseeing.

Integrating temporal factors into collocation-based patterns also allows for causal inference of specific factors on stimulating associations between cities. For instance, by examining the changes in the strength of associations between cities before and after the introduction of policies aimed at fostering metropolitan integration, researchers can assess the efficacy of such initiatives. Through this analysis, collocation methods can provide insights into their long-term effectiveness of policy interventions in shaping metropolitan integration strategies.

2.3 Potentials for further policy research

Words contain enormous knowledge worth investigating. This thesis showcased the potential to extract and classify collocation-based relationships from one corpus dataset sourced from an open-access web archive, Common Crawl. Given the fact that Common Crawl extracts such corpus on a monthly basis, it is then possible to create a time series of multiple types of city relationship data, which allows to characterize the evolution of intercity relationships. By discerning which type of intercity relationships emerges first and how it forsters the emergence of other relationship types, policymakers and urban planners can better understand whether a diversified approach, promoting multiple relationship types concurrently, would be more beneficial, or to prioritize the development of specific relationships according to a city's development stage.

Further analysis can extend beyond examining collocation-based relationships to include comparisons with other types of city relational data, such as transportation and migration flows. Such comparative studies can illustrate the unique characteristics of each type of collocation-based relationship, providing a more comprehensive understanding of the multiplexity of city relationships. Moreover, exploring collocation-based relationships can provide additional insights into the dynamics of urban

systems. By incorporating these insights, the accuracy of existing migration and city growth predictive models may be enhanced.

In Chapter 4, this thesis conducted an initial comparison of the relative impacts of agglomeration and network externalities on city performance. Building on this foundation, further analysis can continue studying the dynamic interplay between these factors over time, aiming to understand not only how agglomeration and network externalities collectively contribute to city performance but also how enhanced city performance, in turn, bolsters a city's capacity to attract people and strengthen its relationships with other cities.

The gravity model estimate results from Chapter 4 show that relationships between mainland Chinese cities and the Special Administrative Regions (SARs) of Hong Kong and Macau are significantly weaker compared to intra-mainland city relationships. This suggests the significant impact that national borders can have on city relationships. Future research can aim to identify and quantify factors that affect the strength of cross-border city relationships, which is crucial for developing strategies for promoting international-scale regional integration.

While the investigations in this thesis concentrate on China—marked by its distinctive political, economic, and social landscape—there remains a significant opportunity for future research to ascertain if the observed patterns and analyzed results hold in other diverse contexts, such as in the developed regions of Europe and the United States or the evolving developing regions in Southeast Asia, the Middle East, and Africa.

3. Closing

In closing, this thesis highlighted the potential of collocation patterns in city network analysis. The insights from two research papers underscore collocation patterns are useful and adaptable for revealing the multiplexity of relationships between cities. Additionally, two research papers illustrate the value of data-driven insights obtained from collocation patterns in informing policy decision-making at city and regional levels.

Summary

Cities form a variety of relationships that facilitate cooperative development. While it is well-recognized that intercity relationships play critical roles in boosting city performance and shaping regional policy, conducting large-scale empirical analyses of these intercity relationships presents substantial challenges. A primary obstacle is the scarcity of city relational data.

Collocation analysis, also known as toponym co-occurrence in geography, presents a novel approach to this challenge. This method quantifies the strength of the relationship between cities based on how frequently their names appear together in texts. While preliminary investigations have shown collocation patterns are effective in capturing city relationships, there remain several unanswered questions. This thesis explored the potential of collocation analysis to extract and interpret city relations from a large textual database, guided by the central question:

To what extent and how can collocation analysis be used to extract, categorize, analyze, and evaluate relations between cities?

To address the central question, this thesis is organized around four sub-questions. The first two sub-questions concentrated on methodological innovations and were addressed by proposing methods that demonstrated the utility and practicality of collocation analysis in capturing and categorizing city relationships. The latter two sub-questions pertained to the empirical applications of collocation patterns. The findings from the two empirical studies highlighted its potential to offer insights into strategic city and regional development.

The main contents of each chapter can be summarized as follows:

Chapter 1 is the introduction, presenting the motivation, theoretical background, research questions, and outline of the thesis.

Chapter 2 introduced an efficient and straightforward method to extract collocation patterns from a large text database, and illustrated through a case study on extracting collocation patterns of 293 Chinese cities from a

Common Crawl web archive. The collocation patterns were observed and analyzed. The results showed that the high-frequency collocation patterns were predominantly between cities with high administrative levels and concentrated in the central and southeastern regions of China, while low-frequency relationships were more prevalent in the northwestern areas. This proposed method is adaptable to other Common Crawl archives, and the generated dataset offers a rich resource for subsequent research in city network analysis.

Advancing from the groundwork laid in Chapter 2, Chapter 3 delves into the multiplexity of city relationships, acknowledging that the general collocation patterns identified earlier provide a broad overview. Practical policy implications require more specific information about city relationships, as there are multiple types of intercity relationships, which are not necessarily identical. To address this, Chapter 3 introduced a lexicon-based method to refine the collocation-based city relationships captured in Chapter 2, classifying them into six categories: industry, information technology (IT), finance, research, culture, and government. The results are mapped and analyzed, which showed that each category displayed different network patterns.

Chapter 4 highlighted the value of collocation-based relationships in addressing empirical challenges at the city scale. It highlighted that due to the lack of city relational data, there were few empirical studies on the comparative advantages of a city's network position versus its agglomeration benefits. Using collocation patterns derived from Chapter 2, an intercity network was constructed and analyzed, revealing that the retrieved collocation patterns reflected reality to some extent, especially mirroring the backbone structure of national networks identified through other intercity relationship data. Then gravity modeling was applied to the collocation patterns to identify the relative positions of these cities within the network. Subsequently, regression models assessed the impact of network externalities against agglomeration benefits on urban productivity. The results found that stronger embeddedness in networks of cities was significantly and positively associated with a city's productivity. The analysis also showed that city network externalities were more important in explaining urban performance than agglomeration externalities, with smaller cities benefiting more from a strong network position. These results suggested that strengthening city relationships could potentially offset the disadvantages of limited agglomeration externalities.

Chapter 5 highlighted the potential advantages of using collocation patterns in confronting planning on regional and national scales, with a focus on the development of large-scale regional plans, or "megaregions." It pointed out that there was an empirical gap in creating effective policies for megaregion planning, attributed to a lack of understanding of the functional coherence within these vast planned areas. To address this gap, Chapter 5 developed a three-dimensional evaluation framework to assess the functional coherence of planned megaregions: inclusion, integration, and consistency. Using the collocation patterns classified in Chapter 4, this framework was applied to fifteen government-defined Chinese megaregions. For nine of these cases, the results revealed discrepancies with the government's defined development level. Given its comparative analysis nature, this approach could be easily implemented to evaluate the functional coherence of other megaregions across the world.

Chapter 6 summarizes the main conclusions and discusses policy implications. It also discusses the limitations of this dissertation and suggests directions for future research.

Nederlandse Samenvatting

Steden zijn ingebed in een grote verscheidenheid aan relaties met andere steden die hun ontwikkeling bevorderen. Alhoewel vrij algemeen erkend wordt dat interstedelijke relaties een cruciale rol spelen bij het stimuleren van de welvaart van steden en richting geven aan regionaal ontwikkelingsbeleid, is empirisch inzicht in deze rollen beperkt doordat het moeilijk is deze interstedelijke relaties te meten, waardoor er weinig nuttige data voor dit soort onderzoek beschikbaar is.

Collocatieanalyse, en meer specifiek het samen voorkomen van toponiemen, biedt een nieuwe aanpak voor deze uitdagingen. Deze methode kwantificeert de sterkte van de relatie tussen steden op basis van hoe vaak hun namen samen in teksten voorkomen. Hoewel eerdere onderzoeken hebben aangetoond dat collocatiepatronen effectief zijn in het beschrijven van relaties tussen steden, zijn er nog veel onbeantwoorde vragen. Deze thesis onderzocht het potentieel van collocatieanalyse om relaties tussen steden uit een grote tekstuele database te extraheren en te interpreteren. De hoofdvraag van deze dissertatie was dan ook:

In hoeverre en hoe kan collocatieanalyse worden gebruikt om relaties tussen steden te extraheren, te categoriseren, te analyseren en te evalueren?

Om deze overkoepelende vraag te beantwoorden, is deze thesis georganiseerd rond vier deelvragen. De eerste twee deelvragen concentreerden zich op methodologische innovaties en werden aangepakt door methoden voor te stellen die het nut en de praktische toepasbaarheid van collocatieanalyse in het vastleggen en categoriseren van relaties tussen steden aantoonden. De laatste twee deelvragen hadden betrekking op de empirische toepassingen van de op basis van collocatiepatronen verworven inzichten in relaties tussen steden. Zodoende geven deze laatste twee deelvragen inzicht in praktische toepassingen van deze methode voor strategische stads- en regionale ontwikkeling.

De belangrijkste inhoud van elk hoofdstuk kan als volgt worden samengevat:

Hoofdstuk 1 is de inleiding, waarin de motivatie, theoretische achtergrond, onderzoeksvragen en opzet van de thesis worden voorgesteld.

Hoofdstuk 2 introduceerde een efficiënte en betrekkelijk eenvoudige methode om collocatiepatronen uit een grote tekstuele database te extraheren, en illustreerde dit door middel van een casestudy over het extraheren van collocatiepatronen van 293 Chinese steden uit een Common Crawl-webarchief. De collocatiepatronen werden waargenomen en geanalyseerd. De resultaten toonden aan dat de collocatiepatronen met hoge frequentie voornamelijk tussen steden met hoge bestuursniveaus voorkwamen en geconcentreerd waren in de centrale en zuidoostelijke regio's van China, terwijl relaties met lage frequentie vaker voorkwamen in de noordwestelijke gebieden. Deze voorgestelde methode is aanpasbaar aan andere Common Crawl-archieven, en de gegenereerde dataset biedt een rijke bron voor vervolgonderzoek in de analyse van stedelijke netwerken.

Voortbouwend op de basis gelegd in hoofdstuk 2, gaat hoofdstuk 3 dieper in op de multiplexiteit van relaties tussen steden. Bouwend op de bredere collocatiepatronen die daar werden geindentificeerd is de opgave in dit hoofdstuk om deze relaties tussen steden te categoriseren. Immers, praktische beleidsimplicaties vereisen vaak specifiekere informatie over relaties tussen steden, aangezien er meerdere soorten interstedelijke relaties bestaan, die niet noodzakelijkerwijs identiek zijn. Om dit aan te pakken, introduceerde Hoofdstuk 3 een op lexicons gebaseerde methode om de resultaten van hoofdstuk 2 verder te verfijnen door relaties tussen steden te classificeren in zes categorieën: industrie, informatietechnologie (IT), financiën, onderzoek, cultuur en overheid. De resultaten zijn in kaart gebracht en geanalyseerd, wat aantoonde dat elke categorie verschillende netwerkpatronen vertoonde.

Hoofdstuk 4 benadrukte de waarde van collocatiegebaseerde relaties bij het aanpakken van empirische uitdagingen op stadsniveau. Het hoofdstuk constateerde dat door het gebrek aan gegevens over relaties tussen steden er weinig empirische studies waren over hoe de positie van een stad in de netwerken met andere steden bepalend is voor het presteren ervan, en hoe zich dit verhoudt tot de makkelijker meetbare en daardoor veel onderzochte agglomeratievoordelen. Uit een analyse van de data over relaties tussen steden uit hoofdstuk 2 bleek dat de patronen in het netwerk tamelijk overeenkwamen met bestaande kennis over het Chinese netwerk

van steden, wat een zekere validatie van de methode inhoudt. Vooral de zogenaamde 'ruggengraatstructuur' van nationale netwerken die geïdentificeerd werden op basis van andere gegevens over interstedelijke relaties kwam terug. Vervolgens werd zwaartekrachtmodellering toegepast op de collocatiepatronen om de relatieve posities van deze steden binnen het netwerk te identificeren. Daarna beoordeelden regressiemodellen de impact van netwerkexternaliteiten op stedelijke productiviteit, waarbij ook de vergelijking met n agglomeratie-externaliteiten werd gemaakt. Het bleek dat een sterkere inbedding in netwerken van steden significant en positief geassocieerd was met de productiviteit van een stad. De analyse toonde verder aan dat netwerkexternaliteiten belangrijker waren bij het verklaren van stedelijke prestaties dan agglomeratieexternaliteiten, waarbij vooral kleinere steden meer profiteerden van een sterke netwerkpositie. Deze resultaten suggereren dat het versterken van relaties tussen steden potentieel de nadelen van beperkte agglomeratie-externaliteiten kan compenseren.

Hoofdstuk 5 belichtte de potentiële voordelen van het gebruik van collocatiepatronen voor planningopgaven op regionale en nationale schaal. De focus lag hierbij op de strategische ontwikkeling van "megaregio's" waarvoor momenteel veel aandacht is in China. De planning en ontwikkeling van dit soort megaregio's wordt gekarakteriseerd door een gebrekkige kennis van de functionele samenhang van de steden die verondersteld worden gezamenlijk een megaregio te vormen. In die zin is sprake van een empirische kennislacune die het creëren van effectief beleid voor de ontwikkeling van megaregio's belemmert. Om deze lacune aan te pakken, ontwikkelde hoofdstuk 5 een driedimensionaal evaluatiekader om de functionele coherentie van geplande megaregio's te beoordelen gebaseerd op inclusie, integratie en consistentie. Met gebruik van de in hoofdstuk 3 geclassificeerde collocatiepatronen, werd dit kader toegepast op vijftien door de overheid gedefinieerde Chinese megaregio's. Voor negen van deze regio's onthulden de resultaten verschillen met het door de overheid gedefinieerde ontwikkelingsniveau, en de drie onderzochte dimensies geven concrete handvaten om hier beleidsmatig op in te spelen. De methode is toepasbaar op andere megaregio's in de wereld.

Hoofdstuk 6 vat de belangrijkste conclusies samen en bespreekt beleidsimplicaties. Het bespreekt ook de beperkingen van dit proefschrift en suggereert richtingen voor toekomstig onderzoek.

Curriculum Vitae

Wang Tongjing was born on July 7th, 1992, in Hangzhou, China. He began his college education in 2011 at Chongqing Jiaotong University, majoring in civil engineering. After obtaining his bachelor's degree in 2015, he pursued graduate study at Tongji University, majoring in communication and transportation engineering. After completing his master's degree in 2018, he then moved to the Netherlands to study at the Amsterdam Institute for Advanced Metropolitan Solutions, undertaking a joint master's program in Metropolitan Analysis, Design, and Engineering with Delft University of Technology and Wageningen University and Research. In 2019, Tongjing started his PhD at the Department of Urbanism, Architecture School at Delft University of Technology, supervised by Evert Meijers and co-supervised by Huijuan Wang from the Department of Intelligent Systems. In February 2021, he continued his PhD studies at Utrecht University, following his supervisor Evert Meijers. During his PhD program, Tongjing has presented his work at various international conferences in Europe, the USA, and China. He also participated in an exchange program at Northeastern University for four months, sponsored by the MSCA 2020 Trend project.