JOURNAL OF ECONOMIC SURVEYS

doi: 10.1111/joes.12394

COMMON-ENEMY EFFECTS: MULTIDISCIPLINARY ANTECEDENTS AND ECONOMIC PERSPECTIVES

Kris De Jaegher*

Utrecht University School of Economics

Abstract. A disparate literature hypothesizes what can broadly be described as the common-enemy effect: the fact that the interaction with a common enemy (formed by Nature, an individual, or a group) increases cooperation. This review identifies the multidisciplinary antecedents of this effect, and then distinguishes between several strands of literature applying noncooperative game theory to account for it. A first strand argues that the threat posed by a common enemy makes each player's cooperative effort more critical. In a second strand a behavioral common-enemy effect caused by group interaction is studied experimentally. A third strand models the common-enemy effect as the formation of a coalition of players against another player in a contest. A fourth strand formalizes the principle that the 'enemy of my enemy is my friend', either in a model of social relations, interdependent altruistic preferences, or indirect reciprocity in repeated games. The connections between these strands of literature are investigated, and questions for future research are proposed.

Keywords. Balance theory; Coalition formation; Collective action; Common-enemy effects; Criticality; Group interaction

1. Introduction

It has been argued that one of the unique features of humans is their ability to cooperate (Rand and Nowak, 2013), and the question then arises what factors encourage cooperation (Chaudhuri, 2011). Several scientific disciplines have independently hypothesized what can be broadly referred to as the common-enemy effect, stating that the interaction with a common enemy (in the form of Nature, an individual, or a group) makes individuals more prone to cooperate; that this effect is also the subject of a folk theory is reflected in the old adage 'the enemy of my enemy is my friend'. The following examples, and the questions that they raise, show the importance of gaining insight into the common-enemy effect.

Example 1. The Intergovernmental Panel on Climate Change (2018) has argued that global warming of more than 1.5 °C compared to pre-industrial levels could create a tipping point, in the form of a major change in the global climate system. As global temperatures rise, each larger country's efforts to reduce greenhouse gas emissions may become critical in beating the common enemy of climate change. Does this mean that no single country will want to deviate from cutting emissions, as this may cause the tipping point to occur? Or does it on the contrary mean that no country will want to take the first step in cutting emissions, as this first step may not make any difference (for similar questions applied to a common-pool problem, see Vasi and Macy, 2003)?

*Corresponding author contact email: k.dejaegher@uu.nl. Tel. +31 30 2539964.

Journal of Economic Surveys (2021) Vol. 35, No. 1, pp. 3–33

Example 2. A manager of a firm needs her staff to cooperate on an innovative task. Instead of setting up a single team to work on this task, she sets up two separate teams to work independently on it. This is because she believes that within each team, the other team will be seen as a common enemy, fostering within-team cooperation and leading to a better outcome than with a single team (for examples along these lines, see e.g. Baer *et al*., 2010; Markussen *et al*., 2013). In order to induce such an effect, should the manager give a bonus to the team that comes up with the best innovative solution? Or does it suffice that teams are able to compare each other's performance (Böhm and Rockenbach, 2013)? Is there a risk of an unintended effect where teams sabotage each other's efforts (cf. Van Knippenberg, 2003)?

Example 3. Both in Brussels and Washington, lobbyists form ever changing ad hoc lobbying coalitions, depending on the precise issue at hand. This may lead to strange bedfellows, as witnessed by the lobbying coalition between EU environmental organizations and producers of electrical appliances, which successfully lobbied against legislation to make producers rather than member states responsible for achieving waste collection targets (Beyers and De Bruycker, 2018). In this concrete instance, environmental organizations and producers were able to set their differences aside, and pool their efforts by forming a coalition against a common enemy. In general, what characteristics of coalition members and of the common enemy lead to an outcome where a coalition is formed against a common enemy?

Example 4. Let Ann be an evangelical Christian in the United States, whose viewpoints on ethical matters such as abortion tend her towards the Republicans, but whose viewpoints on social welfare are somewhat closer to those of the Democrats. Ann's colleague Bill is a libertarian Republican, and is strongly opposed to big government, which leads to some disagreement between Ann and Bill on Obamacare. Enter Ann and Bill's new colleague Carl. Carl votes for the Democrats, but is a militant atheist (which causes Ann to strongly disagree with Carl on ethical matters) who calls himself a socialist (which causes Bill to strongly disagree with Carl on economic matters). Do these facts make Ann and Bill consider Carl as a common enemy, and does this moreover change Ann's viewpoints on Obamacare so that they become closer in line with those of Bill and/or change Bill's viewpoints on religion so that they become closer in line with those of Ann? More generally, do the type of local interactions described lead to political cleavages (Patrikios, 2008)?

These four examples are representative of four main strands of literature dealing with the common-enemy effect, which we will consecutively review (Sections 3 to 6).[1] Our focus is on literature rooted in noncooperative game theory, and on laboratory experiments that use noncooperative game theory as a benchmark to investigate behavior.[2] As the strands of literature reviewed have antecedents across several disciplines, we start by treating these antecedents in Section 2. Section 3 ('Criticality') treats a theoretical literature where the presence of a common enemy (Nature or an individual), or the intensity of the threat posed by this common enemy, makes each player's cooperative effort within a group more critical. Under specific circumstances, this leads to a higher probability of cooperation. Section 4 ('Group interaction') concerns an experimental literature where specifically the interaction with another group leads to more cooperation within a group. This literature presents a behavioral account of the common-enemy effect, as the experiments reviewed are designed such that groups can merely compare each other's performance, or obtain a group bonus scheme that does not change the nature of the game within a group, whatever the performance of the other group. Section 5 ('Coalition formation in contests') reviews literature within contest theory where players are involved in a multi-player contest for a resource. When the effort costs of a single player (= the common enemy) decrease, so that she can put more effort in the contest, the remaining players pool their efforts and compete as a group against the common enemy, dividing the benefits from the contest against the common enemy among them. Section 6 ('Balance') reviews three separate game-theoretic rationales for the prediction of so-called balance theory (see Section 2) that 'the enemy of my enemy is my friend'. A clique of players form friendly social relations, have positive social preferences towards each other, or assess each other's reputation positively, respectively because of their inimical relations with a common enemy (Section 6.1), because of the social preferences of the

common enemy towards them (Section 6.2), or because of the common enemy's past behavior towards them (Section 6.3).

By juxtaposing these four strands of literature, we will be able to answer the following questions in the Discussion (Section 7). What are the key differences between these four approaches to the common-enemy effect? Can these different strands of literature be seen as giving different rationales for the same hypothesized empirical effect? Or, do they instead operate at different levels of analysis, or in different contexts, such that the only aspect that they have in common is their reference to the same catch-all term or folk theory? To the extent that these approaches offer alternative rationales for the same type of hypothesized empirical effect, what different underlying factors are proposed to explain the effect? To the extent that the approaches operate at different levels of analysis, can these approaches still be usefully integrated in a unified framework, and if so, how could the common-enemy effects operating at different levels of analysis impact on each other? Alternatively, to the extent that such a unified framework is not useful, can the different approaches still learn from each other, by importing assumptions from one approach into the other?

2. Multidisciplinary Antecedents

Political science appears to have the earliest antecedents in hypothesizing a common-enemy effect, with Sallust proposing that the destruction of Carthage led to discord and eventually to civil war in Rome, whose citizens no longer had a common enemy to unite against (Evrigenis (2008) provides a history of this hypothesis in political thought). In the more contemporary scapegoat hypothesis, a government may actively seek an enemy to avoid discord among its citizens (Levy, 1989). Other applications in political science include the formation of military alliances by states against a common enemy (Crescenzi *et al*., 2012), and the backfiring effect, arguing that increased government repression makes dissidents more prone to unite (Muller and Opp, 1986; McLauchlin and Pearlman, 2012).

In biology, the common-enemy hypothesis dates back to at least Kropotkin (1902), who wanted to provide a counterweight to the idea that survival of the fittest can only lead to egoism, and who stressed examples of animal cooperation, in particular in harsh environments. If in a harsher environment an organism fails to cooperate within its group, it becomes to a larger extent the victim of its own defection, an effect that is referred to as the boomerang effect (Mesterton-Gibbons and Dugatkin, 1992).[3] For instance, among predators, cooperative hunting is hypothesized to evolve more frequently the larger the prey faced, as every predator's participation in the cooperative hunt then becomes pivotal for catching the prey (cf. Section 3 below). A separate literature in evolutionary biology related to the common-enemy effect is based on group selection arguments, where the presence of competing groups causes the evolution of so-called parochial altruism, with ingroup love and outgroup hate coevolving (Choi and Bowles, 2007; Bowles, 2009; Konrad and Morath, 2012; for an overview, see Rusch, 2014).

In sociology, a classical reference is Coser (1956), who building on the work of Simmel [1908] (1955) and Sumner (1906) notes that social conflict may have positive functions; in particular, the conflict-cohesion hypothesis states that the presence of a common enemy unites members in a group (Chapter V in Coser; for an early overview, see Stein (1976); Benard and Doan (2011) contains a more recent overview). As noted by Stein (1976), while literature citing Coser almost exclusively refers to the conflict-cohesion hypothesis, Coser also argued that in some circumstances (such as a lack of initial solidarity), conflict with an outgroup blocks ingroup cooperation, rather than promoting it.

A common research objective in social psychology is to identify the circumstances in which the combination of ingroup love and outgroup hate arises. Based on field experiments, realistic group conflict theory argues that the source of ingroup love and outgroup hate lies in competing group interests (Sherif, 1966; Jackson, 1993). While the reasons why groups compete are seen as objective and grounded in opposing interests, as pointed out by Bornstein and Ben-Yossef (1994), at the same time this theory does

not specify how conflicts of interest within the group are resolved. In social identity theory (Tajfel and Turner, 1986), a conflict of interest is not necessary to cause ingroup love and outgroup hate, and these may arise based even on trivial differences in characteristics. To stress this point, the so-called minimal group paradigm consists of a stream of psychological experiments where participants are assigned to groups based on trivial criteria (Hornsey, 2008). Participants are observed to attribute higher evaluative ratings, and to make higher other-other allocations (which are necessarily made without personal benefit), to ingroup than to outgroup members.

Finally, in cognitive psychology, balance theory argues that if two individuals both have negative attitudes towards the same object or person, then these two individuals experience cognitive dissonance if they also have negative attitudes towards each other (Heider, 1946, 1958; for an overview, see Hummon and Doreian, 2003). This imbalance is resolved and balance is locally achieved when these individuals become friends ('the enemy of my enemy is my friend'). By the same reasoning, the friend of your friend cannot be your enemy. Also, the friend of your enemy, as well as the enemy of your friend, cannot be your friend (e.g. when a befriended couple divorces, you would only stay friends with one of them (Antal *et al*., 2006)). As aptly summarized by Traag (2014, p. 129) '… friends should think alike about a third person, while enemies should disagree'. Cartwright and Harary (1956) formalized balance theory in graph-theoretic terms by means of signed networks, where a link between nodes may be positive or negative. The authors show that balance leads to polarization, in that any balanced graph consists of two cliques with negative links between the cliques (where by definition there are only positive links within a clique). By relaxing the conditions for balance and allowing for triads consisting of negative links, Davis (1967) obtains that balanced graphs can consist of multiple cliques. While balance theory has been applied beyond social relations between individuals to for example the balance of power in relations between countries (Healy and Stein, 1973), empirical support for this theory is mixed. For instance, using data on dynamic friendship networks, Doreian and Krackhardt (2001) find that specific balanced triad structures occur more frequently through time, but that this is also true for some unbalanced triad structures. Also, using post WWII data, Doreian and Mrvar (2015) find no support for a trend towards balance in the network representing relations between countries.

3. Criticality

In the models reviewed in this section, the starting point is that players in a preexisting group are in an asymmetric conflict with a common enemy. In a sequential game, the players first simultaneously decide whether or not to cooperate, and the common enemy next attacks the players. The common enemy is either Nature, or is a strategic player who is harmed by the total welfare achieved by the players. The common-enemy effect takes the form of players being more likely to coordinate on joint cooperation. In the attack-defense model in Section 3.1, the effect is caused by a larger number of attacks if the common enemy is Nature; in case the common enemy is strategic, the effect is caused by the common enemy being able to make more sophisticated attacks. The Supporting Information contains a review of literature that checks the robustness of this attack-defense model to alternative modeling assumptions. In the model of strategic network formation and disruption reviewed in Section 3.2, the players' cooperative efforts provide benefits beyond defense against the attacks of the common enemy, and the common-enemy effect is caused by the very presence of a common enemy. In both modeling variants, the mechanism underlying the common-enemy effect is that the common enemy makes each player's cooperative effort more critical. The criticality argument is reflected in the figurative use of the expressions 'closing the ranks', and 'circling the wagons'. As will be shown, such an account of the common-enemy effect is not trivial, as the effect of criticality on the probability of cooperation turns out to be ambiguous.

Table 1. Criticality in Terms of the Number of Random Attacks.

	Cooperate	Defect
Cooperate	$1-c, 1-c$	$0.5^A-c, 0.5^A$
Defect	$0.5^A, 0.5^A-c$	0, 0

3.1 *Attack-Defense Model*

The criticality argument can be formalized by the following simple game (De Jaegher and Hoyer, 2016a, 2016b). For concreteness, consider two neighboring countries that each have a nuclear power plant. An incident at the nuclear power plant of one country damages both countries to the same extent (so that safety from nuclear incidents is a public good). Incidents can be caused either by natural disasters (e.g. an earthquake), or by terrorist attacks. Each country can at costs c make its nuclear power plant immune to incidents; as one country's investment also benefits the other country, investing is tantamount to cooperating. The benefit to each country of being free from nuclear incidents in either country is normalized to 1; it follows that if both countries invest, each country obtains payoff $1-c$. The benefit to each country when there is an incident in one or both countries is 0. It is assumed that $0 < c < 1$, so that joint cooperation is efficient for the two countries. Denote by A the number of attacks faced by the countries (where an attack can refer both to a terrorist attack, or to an incident caused by Nature), with $A \geq 1$ (the analysis is not changed when a nonzero number of attacks only takes place probabilistically). Attacks take place through a process of random sampling with replacement of the two countries. For this reason, the probability that only one specific country's nuclear power plant gets attacked equals 0.5^A. When the common enemy is Nature, the cause of any common-enemy effect is simply a larger number of random attacks. When the common enemy is instead strategic, the cause of the common-enemy effect may be an increase in the sophistication of the strategic common enemy in being able to detect which country did not invest. Such an increase in sophistication has an analogous effect to an increase in the number of random attacks, since with a larger number of random attacks, any non-investing country is attacked with high probability (for a detailed account, see Section A.1 of the Supporting Information). We here limit ourselves to the countries' game against Nature, which is represented in Table 1.

The change in benefits to the individual country of cooperating rather than defecting when the other country cooperates, or added benefit of cooperating jointly, equals $1-0.5^A$. Similarly, the change in benefits to the individual country of cooperating rather than defecting when the other country defects, or added benefit of cooperating alone, equals 0.5^A. The nature of the game in Table 1 now depends on the relation between the cooperation costs c, and the two mentioned added benefits. These are represented in Figure 1 as a function of A. With one attack the two added benefits are equal. The added benefit of cooperating jointly increases in A, as the impact of unilaterally deviating from joint cooperation is more severe the higher the number of attacks (this fits the boomerang effect in evolutionary biology, see Section 2). At the same time, the added benefit of cooperating alone decreases as a function of A, reflecting the idea that when each country's investment becomes more critical, a single investing country also makes less of a difference.[4]

As indicated in Figure 1, when the cooperation costs exceed both added benefits, in the unique Nash equilibrium, neither country invests (Prisoner's Dilemma). When both benefits exceed the cooperation costs, the game has a unique Nash equilibrium where both countries invest (Harmony Game). Finally, when the cooperation costs exceed the added benefit of cooperating alone, but not the added benefit of cooperating jointly, the game has both a pure-strategy equilibrium where both countries invest, and one where neither country invests (Stag Hunt). When cooperation costs are now large ($c > 0.5$), an increase in A causes the game to turn from a Prisoner's Dilemma into a Stag Hunt. Thus, an increase in the number

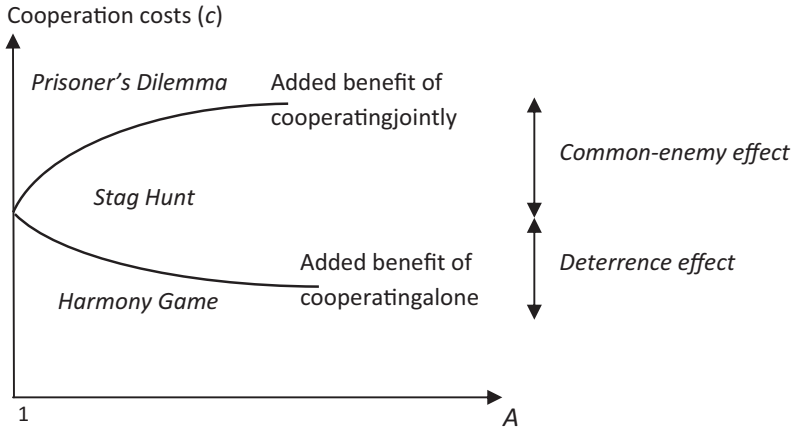

Figure 1. Added Benefit of Cooperating Jointly, and Added Benefit of Cooperation Alone for the Game in Table 1, as a Function of the Number of Attacks *A*

Note: For ease of representation, the added benefits are represented as continuous curves. Depending on the relation of these added benefits to the cooperation costs (c), the game is a Prisoner's Dilemma, a Stag Hunt, or a Harmony Game. The game change as *A* is increased leads to a common-enemy effect for large cooperation costs, and a deterrence effect for small cooperation costs.

of attacks makes it possible that both countries invest, whereas they did not before. Furthermore, when the game is a Stag Hunt and *A* is increased, the basin of attraction of the equilibrium where both countries invest becomes larger (De Jaegher and Hoyer, 2016a), so that a common-enemy effect is obtained.

Yet, if cooperation costs are small, this result is completely reversed: as indicated in Figure 1, in this case the game is a Harmony Game for small *A*, and becomes a Stag Hunt for large *A*. Thus, for large *A*, joint defection becomes possible as an equilibrium, whereas it was not before. An increased threat in the form of a larger number of attacks therefore has a deterrence effect on cooperation instead of a common-enemy effect (De Jaegher and Hoyer, 2016a, 2016b). Furthermore, for small cooperation costs, within the Stag Hunt, it can be checked that the basin of attraction of the joint defection equilibrium becomes larger as the threat level is increased. This shows that there is a flipside to the criticality argument: if a higher threat level makes each player's contribution more critical, it becomes more difficult for players to escape joint defection. The ambiguity of the effect of a larger threat is in line with Hirshleifer's (1983) point that harsh environments can both lead to heroism, or to social collapse, and does justice to the ambiguous effect of external conflict in the work of Coser (1956) (see Section 2).[5]

It should be noted that the model has wider applications than defense against attacks in a literal sense. For instance, in an alternative background story for the game in Table 1, the players are two hunters that surround a large prey by each taking up one of two sides of the prey. The prey repeatedly and randomly chooses one of the sides to try and escape. If the prey chooses a side that has been left unguarded by a hunter, the prey escapes and is lost to both hunters. As long as the prey chooses for its escape attempts a side that is guarded, the prey cannot escape. 'Attacks' are now interpreted as escape attempts, and 'defending' the public good is interpreted as contributing to the surrounding of the prey.

The Supporting Information reviews literature that checks the robustness of the results to modified modeling assumptions (e.g. multiple players, heterogeneity of the players, partial excludability of the public good), where as a rule both the common-enemy and deterrence effects continue to be predicted, with the incidence of these effects identified in more detail.

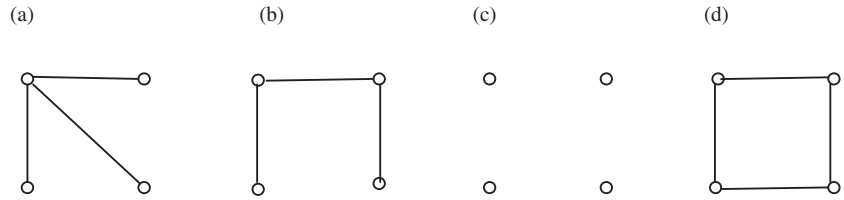

Figure 2. Undirected Networks: (a) Star, (b) Line, (c) Empty Network and (d) Circle. Networks (a), (b) and (d) Are Connected, Networks (a) and (b) Are Minimally Connected.

3.2 *Strategic Network Formation and Strategic Disruption*

In the model in Table 1, starting from an existing threat level, the countries may become better off if the threat level is increased, but they are still better off if they face no threat at all, as they do not need to invest in precautions then. As we now illustrate by means of a simple example, when a disruptor is added to a strategic network formation game, the very presence of this disruptor can make players better off. For instance, criminals would be able to form a cooperative network when they know that they face police forces that try to disrupt their network, but not if they do not face police forces.

We consider here a simplified version of the model of Hoyer and De Jaegher (2012), which is an extension of models of strategic network disruption and defense (Hoyer and De Jaegher, 2016; Dziubiński and Goyal, 2013), with the added feature that link formation is decentralized. Strategic network formation is modeled in the same way as in Jackson and Wolinsky's (1996) connections model. Let there be four players, who each possess one unit of information. Each pair of players is able to form a link with the purpose of sharing information. Each link comes at a total cost c; the cost of a link between two players i and j is equally shared, so that they each incur cost $c/2$ from the link. The set of all the links formed by the players constitutes an undirected network (examples are found in Figure 2, where nodes denote players, and lines denote links). If there is a path in a network between two players i and j, then player i benefits equally well from his own information as from the information of player j; the benefit of a player is simply 1 plus the number of players to which he has a path. Typical network structures that are considered are the following: in a connected network, there is a path between all players (Figures 2a, b and d), and in a minimally connected network, removal of any link means that the network is no longer connected (Figures 2a and 2b); the circle network is connected, but not minimally connected (Figure 2d); in the empty network no pair of players has links (Figure 2c). A network is pairwise stable (Jackson and Wolinsky, 1996) if no two players are better off by adding a link between them, and if no single player is better off when unilaterally deleting a link. The underlying reasoning is that any social link will only be formed if both players agree to form it.

We first look at the case without, and then at the case with a disruptor. For the case without a disruptor, any network that is connected but not minimally connected (e.g. Figure 2d) cannot be pairwise stable. This is because there is then always a player who becomes better off by removing a link, as he saves costs but is still connected to all other players. Further results depend on whether linking costs are small ($c < 2$) or large ($2 < c < 4$). With small linking costs, in any network where i and j are not connected (e.g. Figure 2c), both i and j gain one extra unit of information by forming a link, and incur a cost $c/2$ less than 1. It follows that only minimally connected networks are pairwise stable. With large linking costs, no network where a player i forms a link to a player j who herself does not have further links (a so-called end link) is pairwise stable, because player i only obtains benefit 1 from this link, but incurs cost $c/2$ larger than 1. Since at the same time, networks that are connected but not minimally connected cannot be pairwise stable either, it follows that only the empty network is pairwise stable. This is in spite

of the fact that with a minimally connected network, players' total welfare is larger than with the empty network.

A disruptor is now added in the following way. After the players have formed their network, the disruptor observes the network, and can disrupt a single link with the purpose of minimizing the sum of the informational benefits obtained by the players (the disruptor's ability to disrupt nodes rather than links leads to similar results). One can think of this as police forces preventing two criminals in a network of communicating with each other. Consider now the players forming a circle network. The disruptor cannot disconnect this network, meaning that the individual player obtains payoff $4 - c$. If a single player decides to remove a link, the disruptor can cut the remaining line in two parts, and the player obtains payoff $2 - c/2$. The circle is therefore pairwise stable as long as $c < 4$. At the same time, the empty network is always pairwise stable, because when a pair of players deviates from the empty network by forming a link, the disruptor automatically disrupts this link. It follows that for large linking costs, whereas only the empty network is pairwise stable without a disruptor, the circle network additionally becomes pairwise stable with a disruptor. The presence of a disruptor can thus make players better off and a common-enemy effect is obtained, because each link becomes critical. Contrary to what is the case in Section 3.1, the reason that each link becomes critical is not a change in the shape of the impact function (as the amount of information obtained continues to depend linearly on the number of players accessed), but lies in the structure of the network that players form in the presence of a disruptor, where removal of a single link means the disruptor can cut the network in two. For small linking costs, whereas only the minimally connected networks are pairwise stable without a disruptor, the empty network (on top of the circle network) also becomes pairwise stable with a disruptor. Here, the presence of a disruptor can make players worse off and a deterrence effect is obtained, because the disruptor makes it more difficult to escape joint defection, as any single link formed is automatically targeted. Section A.2 of the Supporting Information shows that the common-enemy effect in Hoyer and De Jaegher (2012) is not maintained in a model of one-sided link formation with strategic disruption (Hoyer and Haller, 2019), where such a model fits observation and imitation networks, rather than the social connections model treated here.

A weakness of this network model is the multiplicity of equilibrium networks, and the lack of an account of how players can reach non-empty networks (indeed, as an experiment by Mir Djawadi *et al*. (2019) shows, this problem arises even if a single participant designs the network). Hoyer and Rosenkranz (2018) contains a laboratory experiment based on the network disruption model of Hoyer and De Jaegher (2012) that addresses exactly this issue, in that it includes a dynamic network formation process (Kirchsteiger *et al*., 2016). In order to focus on the pure effect of the presence of disruption, the authors consider a game with a computerized strategic disruptor, and set up a control treatment without a disruptor where for each network formed, participants obtain with certainty the expected payoffs they would obtain in the presence of a disruptor. Two possible effects of the presence of a disruptor then remain, namely increased risk of forming a network, and an increase in the required farsightedness to reach a non-empty network. The authors hypothesize that from the perspective of farsightedness, there should be more frequent play of the circle network in the disruptor treatment, as participants are to a larger extent forced to think ahead. At the same time, they hypothesize that from the perspective of riskiness the empty network should be played more often in the disruptor treatment, as non-empty networks become riskier. The authors find that the empty network is played more often, and the circle network less often in the presence of a disruptor. While in general farsighted participants are found to reach the circle more often, the results suggest the disruptor has a larger effect on riskiness than on farsightedness.

4. Group Interaction

In the experimental literature reviewed in this section, the common enemy is necessarily a group. While the literature reviewed follows the methodology of experimental economics, the theoretical background

lies in social psychology, where a social group (or ingroup) exists by virtue of there being an outgroup (Dawes and Messick, 2000; Hogg, 2001; for an overview of intergroup interaction in social psychology, see Böhm *et al*. (2020)). Interaction with Nature, such as included in Section 3.1, is considered as insufficient to create a common-enemy effect: in the words of Bornstein (2003, p.130) 'Nature (…) never competes back'.

In the typical experiment reviewed in this section, within each group, players simultaneously set their cooperative efforts. In the control, they do so without interacting with another group, and in the treatment they interact with another group. The experiments are designed such that noncooperative game theory predicts that the investments in the public good in the treatment and in the control should not differ. This is because the focus is on identifying a behavioral common-enemy effect, in contrast with principal-agent literature that investigates how group competition can be designed to make it individually rational for members of a group to cooperate (e.g. Nalbantian and Schotter, 1997; Reuben and Tyran, 2010); as cooperation within groups then arises by design, this literature is outside of the scope of the literature reviewed here. Two lines of research can be identified, which differ according to how group interaction is implemented in the treatment, and which have their origins in two separate theories within social psychology, namely realistic group conflict theory and social identity theory (see Section 2). In Section 4.1, group interaction in the treatment takes the form of each group receiving a group bonus that depends linearly on the difference between the level of cooperation in the own group and the other group. In Section 4.2, group interaction in the treatment takes the form of groups merely being able to compare to each other.

4.1 *Group Competition*

In a research line initiated by Bornstein and Ben-Yossef (1994) (henceforth B&B in this subsection), the starting point is realistic group conflict theory, where there needs to be a genuine conflict between ingroup and outgroup for the presence of an outgroup to boost ingroup cooperation (see Bornstein (2003) for an overview). Yet, such an intergroup conflict may at the same time make it individually rational for group members to cooperate, in which case it becomes impossible to distinguish any behavioral common-enemy effect from a standard incentive effect.[6] Thus, the challenge is to construct an experimental design where there is a genuine intergroup conflict, and where at the same time the presence of an outgroup does not change individual incentives – a challenge which B&B meet with their Intergroup Prisoner's Dilemma. In each session of their experiment, $2n$ participants are randomly divided over two groups (the green group and the red group) of size n each. Each participant gets an endowment of e, and decides simultaneously with the other participants whether or not to invest this endowment (where investing is tantamount to cooperating). A bonus of size $2nF$ is divided over the two groups, depending on the difference between the number of cooperators in one group, and in the other; within each group, the bonus is always shared equally. If t more participants cooperate in group A than in group B, participants of group A each get a bonus of $\frac{(n+t)F}{n}$; if t fewer participants cooperate in group A than in group B, participants of group A each get a bonus of $\frac{(n-t)F}{n}$ (if the same number of participants cooperates in each group, each group gets bonus nF).

This linear difference-form bonus scheme means that all else equal, a participant who defects instead of cooperates always gains e and loses $\frac{F}{n}$, whatever the level of cooperation in the other group. Comparing the games played by members of group A for several fixed numbers of cooperators within group B, it becomes clear that all these games are strategically equivalent, in that the payoffs in one game differ from the payoffs in another game by a fixed amount (namely $\frac{F}{n}$). It is assumed that $e > \frac{F}{n}$, so that defecting is the only *individually rational* strategy and joint defection is the only Nash equilibrium. At the same time, a participant who defects instead of cooperates gains e, but makes his or her group lose F. It is assumed that $F > e$, so that for any given number of cooperators in the other group, the sum of group members'

Table 2. Variant of Bornstein and Ben-Yossef's (1994) Intergroup Prisoner's Dilemma. Prisoner's Dilemma Played by Members of Group A, as a Function of the Number of Cooperators m_B in Group B, Where Group B Is Symmetric.

$m_B = 2$			$m_B = 1$			$m_B = 0$		
	C	D		C	D		C	D
C	6, 6	3, 7	C	9, 9	6, 10	C	12, 12	9, 13
D	7, 3	4, 4	D	10, 6	7, 7	D	13, 9	10, 10

payoffs is maximized when all invest (*group rationality*). This tension between individual rationality and group rationality means that, for any given number of cooperators in the other group, group members are playing an n-player Prisoner's Dilemma. Finally, note that if the individual player invests e, this also means a decrease in payoffs of F for the members of the other group. It follows that from the perspective of the two groups jointly, total welfare is maximized when no participant invests (*collective rationality*). Thus, if the presence of another group makes group members focus on collective rationality, this should lead them to defect more often, ensuring that any behavioral common-enemy effect in the form of an apparent enhanced focus on group rationality, cannot be confounded with an enhanced focus on collective rationality.

In the set-up of B&B, it is the case that $n = 3$, $F = 9$, and $e = 5$. For ease of representation, in Table 2 we instead represent the example $n = 2$, $F = 6$, and $e = 4$, where as a function of the number m_B of cooperators in group B, the game facing the players in group A is represented. As Table 2 illustrates, the bonus scheme is indeed such that the three Prisoner's Dilemmas are strategically equivalent: moving from one game to the other, a participant is given 3 extra points whatever the initial ingroup strategy profile. From this perspective, game theory predicts that there should not be any difference between the entire game played in Table 2, and either of the Prisoner's Dilemmas in Table 2 played in isolation. B&B exploit this fact by using the separate Prisoner Dilemmas as controls in their experiment. To avoid that there are still effects of the absolute size of the payoffs, the authors use both the separate Prisoner's Dilemma with the lowest and the highest payoffs as controls (the authors do not find significant differences between these controls); moreover, to avoid that there is an effect from merely being divided in groups, the authors also divide the players into groups in the control treatments. Participants registered 10 decisions, knowing that one of these decisions would be implemented. The authors find that the fraction of cooperating players is almost twice as large in the group-competition treatment than in the controls, thus confirming the existence of a common-enemy effect. This effect is at the same time behavioral, and in line with realistic group conflict theory: while game theory predicts no difference between control and treatment, there still exists a genuine conflict between the two groups. In a follow-up study (Bornstein *et al*., 1996), the same experiment is performed over 40 rounds in a partner design, where participants receive only aggregate-level feedback (note that two participants per group would automatically mean that there is individual-level feedback); the effect is confirmed, but becomes smaller towards the last rounds. Rebers and Koopmans (2012) implement the design of B&B as a one-shot game, but add a stage where altruistic punishment is allowed, and find that such punishment is higher in the treatment than in the control.

Given the set-up of B&B, it is not clear whether their treatment effect of more investments is due to participants defending the interests of their ingroup, or attacking the interests of the outgroup. More recent experiments attempt to disentangle between such defending and attacking, first, by constructing a variant of B&B where participants decide whether to contribute to the ingroup benefit with or without hurting the outgroup benefit (where choosing to contribute with hurting of the outgroup is considered as attacking); second, by considering an asymmetric variant of B&B where one of the groups is subject to

the linear difference-form bonus scheme (= defending group) and the other is not (= attacking group); and third, by letting groups move sequentially, where in specific variants either the second-mover group is considered as retaliating in defense, or the first-mover group is considered as defending by preemptively attacking.

The first approach is found in Halevy *et al*. (2008), who find that participants almost exclusively prefer to invest in their own group without hurting the outgroup ('within-group investments'), rather than additionally hurting the outgroup ('between-group investments'). While this suggests that defending the ingroup is the predominant mechanism explaining increased cooperation, Weisel and Zultan (2016) point out that Halevy *et al*. (2008) also modify the experimental set-up of B&B by framing payoffs in terms of a negative externality imposed on the outgroup, rather than in terms of a comparison of investments of groups.[7] Indeed, Weisel and Zultan show that framing the exact same set up as B&B in this way leads to a significant large and *negative* effect on investments of the presence of an outgroup, suggesting that a similar framing effect is taking place in Halevy *et al*. (2008).[8] At the same time, Weisel and Böhm (2015) implement a modified version of the set-up of Halevy *et al*. (2008) where the choice is between contributing to a pool that benefits the ingroup, or to a pool that additionally benefits the outgroup, where the former can now be seen as attacking. These authors also find that most participants prefer to favor the other group (by choosing for a positive externality on the outgroup), but that at the same time a much larger fraction of participants than in Halevy *et al*. (2008) prefer to harm the outgroup (by not letting them benefit from the positive externality); yet contrary to other studies reviewed here, these authors consider natural groups, and not ones that are artificially created.

The second approach is found in Weisel and Zultan (2016), who set up a close variant of B&B, with the only difference that only one group (= defending group) is affected by the linear difference-form bonus scheme, whereas the other group (= attacking group) plays a standard Prisoner's Dilemma (so that investments within the attacking group impose a negative externality on the defending group). The authors find investments to be significantly larger in the defending group, and conclude that what encourages cooperation is the threat posed by an outgroup, rather than one's ability to harm an outgroup. A combination of the first and second approaches is found in Cacault *et al*. (2015), who have an attacking and defending group in the same way as Weisel and Zultan (2016), but additionally let the participants in the attacking group choose between investing in the ingroup without hurting the defending group, investing in the ingroup benefit and simultaneously hurting the defending group, or hurting the defending group without any benefits to the ingroup (where the latter two are considered asocial investments). Cacault *et al*. find that around a third of investments of groups in the attacking role are anti-social investments; an open-answer survey at the end of the experiment suggests that the motivation of more than half of these investments is to harm the other group. Also, compared to the standard Prisoner's Dilemma in the control, both having the option of anti-social investments, and facing the possibility of being victimized, increases investments. The authors attribute the absence of between-group investments in Halevy *et al*. (2008) to the symmetric design of this experiment, which they argue creates expectations of retaliation that are subdued in their asymmetric design.

The third approach is found in Böhm *et al*. (2016). Just as in Halevy *et al*. (2008), these authors let participants in all groups decide between a within-group and a between-group investment. However, using the set-up of Halevy *et al*. as a control, they let groups move sequentially, where the participants in the second-mover group are asked to plan their between-group investments as a function of the within- or between-group investment chosen by the first-mover group (strategy method); 'defense' here takes the form of retaliation against attacks (in the forms of between-group investments) by the first-mover group. While the authors do no find significantly different between-group investments in the first-mover group and the control that follows the set-up of Halevy *et al*., the strategy method reveals that the second-mover group is willing to make substantial between-group investments in retaliation to possible between-group investments of the first-mover group. In an additional treatment, rather than retaliating, the first-mover group can preempt the other group's between-group investments. This is done by between-group

investments that reduce the negative externality imposed by the second-mover group; 'defense' takes the form here of a preemptive attack. In the control to this treatment, the second-mover group is not allowed to make a between-group investment; in this way, any positive treatment effect on between-group investments reflects a desire to defend the own group's interests, rather than to attack the other group. A significant positive treatment effect is indeed found.

4.2 *Group Comparison*

Several related linear public-goods experiments observe that mere comparison of groups, without any bonuses attached to comparative group performance, suffices to boost intragroup cooperation. In Tan and Bolle (2007), in the control, three participants play a linear public goods game for 20 rounds, of which 10 rounds in a stranger design and 10 rounds in a partner design with counterbalancing, and where feedback is given at the end of each round about total ingroup investments. In the comparison treatment, participants are additionally informed on whether total investments in another group were higher or lower. In the bonus treatment, participants whose group has larger investments than another group receive a bonus. The authors find that in general, investments are higher in the partner design, and decrease over time. Moreover, investments are highest in the bonus treatment, followed by the comparison treatment and then the control. Böhm and Rockenbach (2013) come to the same conclusion for a partner design with 20 rounds, where in the treatment participants additionally receive information on the total investments in the other group, and where the authors vary group size and marginal per capita return.

Yet, purely the fact of being informed about a ranking in the treatment but not in the control could explain the observed effect. An experiment by Burton-Chellew and West (2012) deals with this problem. Four participants play a linear public goods game for 20 rounds in a stranger design, and again receive feedback on total investments within their group at the end of each round. In the within-group information treatment, participants are additionally informed about the ranking of investments within their group. In the among-group information treatment, participants are instead additionally informed about the ranking of total investments across four four-player groups, including their own. In this way, players get similar ranked information in both treatments. Participants play each treatment for 10 rounds with counterbalancing. The authors find that over all rounds, investments are significantly higher in the among-group information treatment, even though investments in the first rounds are not significantly different; investment levels decrease more slowly in the among-group information treatment.

These papers have links with experimental literature on discrimination between ingroups and outgroups; for a meta-analysis of economic laboratory experiments in this literature, see Lane (2016). Ingroups and outgroups are either artificially created in the laboratory, or are preexisting groups. The typical set-up in this literature differs from the papers reviewed above, and either looks at how a participant's allocations between two others (other-other allocations), or how participants' behavior in two-player games, depends on group membership. Yet, the results in this literature are not as clear-cut as in the public-good experiments reviewed above (e.g. Chen and Li (2009) and Hargreaves and Zizzo (2009) find that artificially-created groups suffice to induce discrimination, but Charness *et al.* (2007) and Zizzo (2011) do not). Lane (2016) reports that roughly a third of the studies included in his meta-analysis find significant discrimination. Furthermore, he finds that discrimination depends on context, in being much stronger in other-other allocations than in two-player games. Interestingly, he finds more discrimination between groups artificially created in the laboratory, than between groups based on nationality or ethnicity (though discrimination is strongest among groups based on social or geographical distinction). The author argues that this difference may be due to experimenter-demand effect in case of artificial groups (Zizzo, 2010) and/or to an unwillingness to be seen as discriminating in the case of ethnic or nationality groups.

We conclude from the group-interaction literature reviewed in this section that there is general agreement that group interaction framed as a conflict between groups leads to more cooperation within groups, even if the conflict is constructed in such a way that rationally it should not matter for behavior within groups. While a number of studies find the same effect when groups are merely able to compare to each other, the results of literature that directly tests for intergroup discrimination leads to much less clear-cut results. In Section B of the Supporting Information we additionally review experiments that attempt to identify what factors underlie the behavioral common-enemy effect observed in the context of the group-interaction experiments reviewed above. The proposed underlying factors are that group interaction changes participants' social preferences, changes their competitive preferences, or changes their perception of their decision problem. Specifically in the latter case, it is argued that participants perceive themselves to have a common fate (Messick and Brewer, 1983), perceive their efforts to be critical (De Cremer and Van Vugt, 1998), or perceive themselves as contributing to their individual reputation in line with the bounded generalized reciprocity model of Yamagishi *et al*. (1999). The conclusion coming forward here is that there is no agreement on what underlying factors explain the behavior observed in the experiments reviewed above.

5. Coalition Formation in Contests

In both the approaches in Section 3 and 4, groups are exogenously given, and the common-enemy effect takes the form of increased cooperation *within* a preexisting group. In the contest models reviewed in this section, by contrast, the common-enemy effect takes the form of a subset of players stopping to compete against each other and pooling their efforts to form a group, excluding a common enemy. Such games consist of two stages, where in the first stage players form groups in a process in line with noncooperative game theory, and where in the second stage they simultaneously set their efforts. As an initial intuition, we may think of players who are originally engaged in a multi-player contest over a prize. When the effort costs of one player (= the common enemy) are decreased so that she can invest more effort in the multi-player contest, the remaining players stop competing against each other, pool their efforts to compete against the common enemy, and thereby increase their chances of winning (where the players somehow divide the benefits of competing jointly against the common enemy). We will later correct this initial intuition.

We present as an example a simple three-player contest or 'truel', with the possibility of coalition formation. This example is not meant as a general model, but instead serves as a benchmark to analyze how modeling assumptions closer to the existing literature change the predictions obtained from this simple example. We consider three players 1, 2 and 3 competing for a prize of size 1 (for similar treatments of the truel, see Skaperdas (1998), Noh (2002) and Herbst *et al*. (2015)). At stage 1, players simultaneously announce with whom they want to form a coalition, where we restrict the strategy space such that only players 1 and 2 can announce coalitions with each other.

After coalitions have been formed, at stage 2, each player i simultaneously with the other two players decides on an effort $e_i \geq 0$. This effort is not productive but extractive, and increases the probability that i wins the prize (a so-called partial equilibrium model; Garfinkel and Skaperdas, 2007). The effort causes i to incur linear costs $c_i e_i$, where c_i is i's constant marginal cost (cf. Sánchez-Pagés, 2007). More in detail, we assume that 1 and 2 have identical marginal cost c ($c_1 = c_2 = c$), and that 3 possibly has different marginal cost c_3. For a player i who is not in a coalition, the probability of winning the contest equals $\frac{e_i}{e_i+e_j+e_k}$ (lottery contest success function; Tullock, 1980); this is also the expected benefit from the contest.[9] When they form a coalition, 1 and 2 produce their group effort according to a simple linear impact function, such that group effort equals $e_1 + e_2$. The expected benefit to 1 and 2 from the contest then equals $0.5\frac{e_1+e_2}{e_1+e_2+e_3}$. A coalition of 1 and 2 thus operates as a single player in a two-player contest against 3, where the coalition wins the prize with probability $\frac{e_1+e_2}{e_1+e_2+e_3}$; players within a coalition

divide the prize equally (egalitarian sharing rule; Nitzan, 1991). When 1 and 2 form a coalition, they are assumed to always play the symmetric Nash equilibrium at stage 2.

In a subgame perfect equilibrium, players form coalitions at stage 1, anticipating the Nash equilibrium that will be played at stage 2 as a function of their decisions at stage 1. As follows from Hart and Kurz (1983), to check whether coalitional structure [{1, 2}, {3}] is part of a subgame perfect equilibrium, it suffices to check that neither 1 nor 2 want to leave coalition {1, 2}; to check that coalitional structure [{1}, {2}, {3}] is part of a subgame perfect equilibrium, it suffices to check that player 1 and/or 2 are better off leaving coalition {1, 2}. We analyze this example by gradually increasing complexity, starting with exogenously given efforts (Section 5.1), then moving to the situation where only 3's effort is exogenous (Section 5.2), and finally to endogenous efforts for all (Section 5.3). As the results obtained deviate from our initial intuition, in Section 5.4, we treat literature with alternative modeling assumptions that come to the rescue of our initial intuition. In Section 5.5., we consider the few experiments related to this model.

5.1 *Exogenous Efforts*

In this case, efforts can be reinterpreted as endowments or abilities. It is easy to see that any coalition that contains a player with an effort strictly larger than those of the other coalition members cannot be part of an equilibrium coalitional structure, as this player is then always better off leaving the coalition. It follows that any coalition must consist of players with the same efforts, where the players are then just still willing not to leave the coalition (Skaperdas, 1998). [{1, 2}, {3}] is thus part of a weak subgame perfect equilibrium if 1 and 2 have equal efforts e, and this no matter whether they are stronger than ($e > e_3$), weaker than ($e < e_3$), or equally strong as 3 ($e = e_3$).

5.2 *Endogenous Efforts of Two Players, Exogenous Threat*

A useful starting point to analyze the case where the effort of 3 is exogenously given, is the case where all efforts are endogenous, as treated in Herbst *et al*. (2015). Following calculations by these authors, if all efforts are endogenous and the three players have identical marginal costs, then without a coalition all players set the same effort $\frac{2}{9c}$ and obtain expected payoff $\frac{1}{9}$. If 1 and 2 form a coalition, then in the symmetric Nash equilibrium at stage 2, they each set effort $\frac{1}{18c}$ and continue to obtain payoff $\frac{1}{9}$; 3 continues to set effort $\frac{2}{9c}$, but now obtains payoff $\frac{4}{9}$. It follows that [{1, 2}, {3}] is a weak subgame perfect equilibrium, as 1 and 2 are just still willing not to leave. The fact that there are only weak incentives to form a coalition in this case, is known as the 'paradox of alliance formation' (Konrad, 2009).

Using the results of Herbst *et al*. (2015), it follows that if the effort of player 3 is exogenously given at $e_3 = \frac{2}{9c}$, 1 and 2 are indifferent between forming a coalition or not forming one. It can be checked now that for $e_3 < \frac{2}{9c}$, [{1, 2}, {3}] is subgame perfect, and that for $\frac{2}{9c} < e_3 < \frac{1}{c}$, [{1}, {2}, {3}] is subgame perfect. In other words, contrary to the initial intuition at the start of this section, 1 and 2 form a coalition when facing a small threat (and when they are relatively strong), but not when facing a large threat (and when they are relatively weak).

To understand this result, note first that when 1 and 2 form a coalition, they reduce their efforts, both because they stop competing against each other and because of a tendency to free-ride within the group, as the individual player obtains only half of the prize won by the coalition. An intuition for why 1 and 2 may form a coalition in case of a small threat, is that in the absence of any threat, players who form a coalition continue to obtain the prize with probability $\frac{1}{2}$, but save costs by no longer wasting effort competing against each other. As long as the threat is small, these costs savings dominate, as forming a coalition only slightly reduces the expected benefit from the contest. For a large threat, forming a coalition and reducing efforts decreases the expected benefit from the contest too much compared to the cost savings, and a coalition is not formed.

5.3 *Endogenous Efforts, and Endogenous Threat*

We finally consider the case when all efforts are endogenous. The threat level posed by 3 is now reflected by how low her marginal costs c_3 are (as this means that she sets higher effort), where 3 poses a small threat when $c < c_3 < 2c$, and a large threat when $c_3 < c$. Applying the calculations of Herbst *et al.* (2015), one finds the following expected payoffs Π and efforts for the players (where the superscript A refers to coalitional structure [{1}, {2}, {3}], superscript B to coalitional structure [{1, 2}, {3}], and the subscript refers to the player): $\Pi_1^A = \Pi_2^A = (\frac{c_3}{c_3+2c})^2$ and $\Pi_3^A = (\frac{2c-c_3}{c_3+2c})^2$, with $e_1^A = e_2^A = \frac{2c_3}{(c_3+2c)^2}$ and $e_3^A = \frac{2(2c-c_3)}{(c_3+2c)^2}$; also, $\Pi_1^B = \Pi_2^B = \frac{c_3(c_3+c)}{2(c_3+2c)^2}$ and $\Pi_3^B = (\frac{2c}{c_3+2c})^2$, with $e_1^B = e_2^B = \frac{c_3}{2(c_3+2c)^2}$ and $e_3^B = \frac{2c}{(c_3+2c)^2}$.

Comparing the payoffs, it follows that 1 and 2 do *not* form a coalition when the threat is small/when they are relatively strong ($c < c_3 < 2c$, in which case $e_3^A < 2/(9c)$), and do form a coalition when the threat is large/when they are relatively weak ($c_3 < c$, in which case $e_3^A > 2/(9c)$). Endogenizing the threat therefore completely reverses the results of Section 5.2. The intuition for this reversal lies in the way in which 3 reacts to the efforts of 1 and 2. Independently of whether 1 and 2 form a coalition, as a function of a symmetric effort e of 1 and 2, the reaction curve of 3 takes the form $e_3 = \sqrt{\frac{2e}{c_3}} - 2e$, meaning that for small e, 3 sets more effort the higher the effort of 1 and 2, whereas for large e the opposite is the case (Bloch, 2012); the reason is that eventually, as e becomes large, matching 1 and 2's efforts becomes too expensive for 3.

Consider now a small threat of 3, with c_3 close to $2c$. Without a coalition, 1 and 2 now set e close to $1/(4c)$, and in response 3 sets e_3^A close to zero, so that we are in the decreasing part of 3's reaction curve. If 1 and 2 would form a coalition (which by the above means they would decrease their effort fourfold), they would therefore cause 3's effort to increase, which is why they prefer not to form a coalition. Consider next a large threat of 3, with c_3 close to 0. When forming a coalition, 1 and 2 now set effort close to 0, and in response 3 does the same, so that we are in the increasing part of 3's reaction curve. If 1 and 2 would *not* form a coalition (which means they would increase their effort fourfold), 1 and 2 would again cause 3's effort to increase, which is why they now do prefer to form a coalition.

Our initial intuition of two weak players forming a coalition against a strong player is therefore seemingly confirmed when all efforts are endogenous, yet the reason for this is somewhat unexpected. The weak players 1 and 2 form a coalition to de-escalate the conflict, thus avoiding that 3 sets a high effort: the fact that 1 and 2 form a coalition causes them to reduce their efforts, so that in response 3 reduces her effort as well. In fact, by forming a coalition, 1 and 2 decrease their chances of winning the prize, and increase the payoff of 3 (for a similar result of free-riding within a coalition benefiting all players, see Niou and Tan, 2005). But given that the costs of 1 and 2 are relatively large in this case, the fact that they save costs by forming a coalition, more than compensates. When instead we consider a strong 1 and 2 and a weak 3, 1 and 2 do not form a coalition in order to keep their efforts high, and prevent 3 from setting a high effort. It should be noted that when 1 and 2 are relatively strong and do not form a coalition, if the strength of 3 is now increased to such an extent that they do prefer to form a coalition, this will not make 1 and 2 better off.

5.4 *Rescuing the Initial Intuition*

We conclude from our analysis so far that in the truel, while two weak players may form a coalition against a strong player, this result is not in line with our initial intuition, as by forming a coalition the weak players decrease their chances of winning. Moreover, as shown in Section C of the Supporting Information, the result in Section 5.3 of two weak players forming a coalition against a strong player is not maintained when relaxing key simplifying assumptions by also allowing players to form a grand coalition, allowing for more than three players, or by considering other sharing rules within the coalition.

We now continue by reviewing several alternative assumptions that still come to the rescue of our initial intuition.

First, instead of a partial equilibrium, a general equilibrium may be considered where players' efforts are also productive (Noh, 2002). Instead of having a cost function, each player now has an endowment R_i, that can be allocated to a productive effort P_i (producing 'butter') or an extractive effort e_i (producing 'guns'), with $P_i + e_i = R_i$. The prize over which the contest takes place is equal to what the players jointly produce; in the simplest setting where production equals the sum of the players' productive efforts, the prize equals $R - \sum_{i=1}^{n} e_i$, where R is the sum of the endowments.[10] It follows that in an interior solution, heterogeneity in the endowments does not matter for the results (the 'paradox of power'; Hirshleifer, 1991). It is easily checked now that for $n = 3$, 1 and 2 are strictly better off forming a coalition, because this increases the size of the prize. Yet, once one allows 3 to also join coalitions, it becomes clear that for the same reason only the grand coalition is subgame perfect. Still, as modeled by Münster and Staal (2011), when there is also an internal contest within coalitions, using endowments for external competition may actually lead to more production, as fewer resources are then left for harmful within-coalition competition.

Second, coalitional effort could be aggregated according to a non-linear impact function. Following Sheremata (2011), consider the weakest-link impact function, where in order to have comparability with our initial case, we assume that 1 and 2 when forming a coalition produce group effort equal to $2 * \min(e_1, e_2)$. In this way, as long as 1 and 2 choose the same symmetric effort as before, group effort is exactly the same, and only the degree of complementarity between efforts is changed. In the case with symmetric costs, when 1 and 2 form a coalition, if among the continuum of Nash equilibria they coordinate on the Pareto-superior one, 1 and 2 each set effort $\frac{1}{8c}$ and each obtain payoff $\frac{1}{8}$; 3 sets effort $\frac{1}{4c}$, but now obtains payoff $\frac{1}{4}$. It follows that 1 and 2 are now strictly better off when forming a coalition – though they continue to decrease their expected benefit from the contest compared to the case where they do not form a coalition.

Third, forming a coalition could create synergies between the players' efforts. A tractable contest success function with this effect assumes that the probability of winning of a player who is not in a coalition is not changed, but that in the contest success function the efforts of two players who form a coalition are multiplied by a coefficient $\theta > 1$, so that their probability of winning becomes $\frac{\theta(e_1+e_2)}{\theta(e_1+e_2)+e_3}$ (Noh, 2002). For the case of symmetric costs, it is straightforward to calculate that in this case $e_1^B = e_2^B = \frac{\theta}{2c(\theta+2)^2}$, $e_3^B = \frac{2\theta}{c(\theta+2)^2}$, $\Pi_1^B = \Pi_2^B = \frac{\theta(\theta+1)}{2(\theta+2)^2}$, and $\Pi_3^B = (\frac{2}{\theta+2})^2$. It follows that 1 and 2 strictly prefer to form a coalition, and for a sufficiently large θ each increase their expected benefit from the contest. For θ just above 1, forming a coalition means that 1 and 2 increase their efforts, though for larger θ they reduce their efforts, as they then need very little effort to win the prize with high probability. Also, for sufficiently large θ, the fact that 1 and 2 form a coalition reduces 3's payoff.

While the effect on the results of combining several of these alternative assumptions deserves future attention, a preliminary conclusion is that our initial intuition at the start of this section is only confirmed, and the paradox of alliance formation is only solved, when forming a coalition leads to synergies. Yet, as argued by Bloch (2012), '[t]he paradox of alliance formation poses a challenge to economists, as one observes alliances forming even in the absence of synergies'.

5.5 *Experimental Evidence*

We end by looking at whether experimental literature on coalition formation can solve the paradox of alliance formation by means of behavioral explanations. Sheremata (2018) recently reviews the experimental literature on group contests, and reports as a key finding overprovision of effort compared to theoretical predictions. While underlying factors as proposed in the context of group competition could explain such overprovision (e.g. parochial altruism; see Supporting Information, Section B), as noted

by Sheremata, overprovision is also observed in two-player contests. Ahn *et al*. (2011) moreover show that overprovision takes place whether one faces an individual or a group. It is therefore not clear what part of overprovision is due to the fact of playing a contest as a group or as an individual, or is due to playing a contest against a group or an individual. Yet, in the experimental design of Abbink *et al*. (2010), within any group the prize won is a public good, in which case the theoretically predicted total effort is independent of group size. The authors find that groups of four players invest more in the contest than one player, and this independently of whether a single other player is faced, or another group. This suggests that there is a group-specific motive for overprovision, but that contrary to what one could conclude from the literature reviewed in Section 4, it does not matter for this motive whether one faces another group or an individual.

This raises the question: do players form coalitions, anticipating the high efforts that will take place within groups? As noted by Sheremata (2018), endogenous coalition formation in contests is treated in only a few experiments. Herbst *et al*. (2015) find that efforts within groups are higher when groups are formed endogenously, than when they are formed exogenously; this suggests that participants who are more prone to ingroup love self-select into coalitions. Deck *et al*. (2015) consider a sequential game where two players ('defenders') move first in setting efforts, which are observed by a third player who moves next ('attacker'). If the defenders do not form a coalition, then the attacker targets the defender with the lowest effort, and the attacker and this defender engage in a two-player contest over the defender's private good (there is therefore no contest between the defenders). If the defenders form a coalition, then their efforts are added up in a two-player contest between one random defender and the attacker. As shown by the authors, such defenders are always better off forming a coalition, because without a coalition they are involved in an all-pay auction that dissipates the profits from the contest; while forming a coalition reduces their probability of winning, this is more than compensated by cost savings. This theoretical prediction is confirmed experimentally, even though overprovision of effort is again observed.

6. Balance: The Enemy of My Enemy Is My Friend

All models in this section explain why as a rule, the enemy of one's enemy is one's friend. While these models are thus in line with the predictions of balance theory (see Section 2), contrary to what is the case in this theory, the explanation that the models provide for these predictions is not a behavioral one based on cognitive dissonance, but a rational one. In the three models covered, players have symmetric roles and move at the same stages, and the common enemy is formed endogenously. As far as one can talks of groups, these are cliques of players that all have friendly bilateral relations, whereas the bilateral relations with other players are inimical. The causes of the common-enemy effect lie in the social relations with the common enemy (Section 6.1), in the common enemy's social preferences (Section 6.2), or in the past behavior of the common enemy (Section 6.3).

In the strategic network formation model in Section 6.1, if Ann holds a negative bilateral relation with Carl (and engages in a bilateral contest with him), and if Bill has the same relation with Carl, then Ann and Bill necessarily hold a positive bilateral relation with each other (and in this way avoid a bilateral contest between each other). In the interdependent-utilities model in Section 6.2, if social preferences between three players Ann, Bill and Carl are such that spite between both Ann and Carl, and between Bill and Carl is large, then Ann and Bill behave altruistically towards each other. This altruistic behavior between Ann and Bill is not explicitly modeled, but is a logical consequence of the model. In Section 6.3, players repeatedly play a two-player game such as the Prisoner's Dilemma. While players are constantly re-matched in pairs, they observe the number of times other players cooperated in the past. Though one may in general expect that players who fail to cooperate build up a bad reputation, if Carl failed to

cooperate with Ann in the past, and if Bill now fails to cooperate with Carl, then this may benefit rather than harm Bill's reputation with Ann, and Ann may cooperate with Bill in the future.

6.1 *Strategic Network Formation with Positive and Negative Links*

Hiller's (2017) model of network formation differs from the graph-theoretic literature formalizing balance theory (see Section 2) by the fact that links are formed strategically. Each of n players simultaneously decide whether to form a positive or a negative *directed* link to each other player. When both directed links between players i and j are positive, they are friends (= positive *undirected* link between them) and do not engage in a bilateral contest with each other. When at least one of the two directed links between i and j is negative, they are enemies (= negative *undirected* link between them), and engage in a bilateral contest (a player who has x enemies will thus be involved in x bilateral contests). From such a contest between i and j, both players obtain a zero payoff when they have equal strengths; when i is stronger than j, i obtains a positive payoff and j a negative payoff, where the difference in these payoffs is larger the larger the difference in their strengths.

A player's strength equals her autonomous strength (normalized to 1) plus the sum of the autonomous strengths of her friends. Unless stated otherwise, we will focus on the case where autonomous strengths can differ only marginally among players. Contrary to what is the case in the model in Section 5, friends do not pool their strengths and compete together against the other players as a single unit, dividing the benefits. Instead, players always continue to engage in separate bilateral contests with their enemies, but do so benefiting from positive externalities (e.g. information sharing) obtained from their friends. Forming a positive link comes at a zero cost. Forming a negative link comes at a positive cost, and furthermore there is a cost of being someone's enemy (note that i incurs the latter cost without incurring the former cost, when i forms a positive directed link to j, but receives a negative directed link from j). A player's payoff therefore equals the proceeds from her bilateral contests with all her enemies, minus the costs of the negative links she maintains, and minus the costs of the number of enemies she has. Note that the added benefit of having a friend consists of the benefits from one's increased strength in one's bilateral contests with one's enemies, and of the benefit of avoiding a contest with the friend, if the friend is stronger (if the friend is weaker, the latter part becomes a loss in benefit). The cost savings of having a friend lie in avoiding the cost of being someone's enemy, and in avoiding the cost of forming a negative link. The equilibrium concept employed is the Nash equilibrium, as reflected by a Nash network describing for each pair of players their mutual directed links, with the characteristic that no player strictly prefers to change a directed link.

Several characteristics of Nash networks come forward. In any Nash network, two players i and j will never maintain bilateral negative directed links. If j forms a negative directed link to i, then i has nothing to gain from also forming a negative directed link to j: the link makes her incur costs, and j remains an enemy whether i's directed link is negative or positive. In a negative undirected link, the positive directed link will be formed by the weaker, and the negative directed link by the stronger player. In the example of three players, four qualitatively different undirected networks represented in Figure 3 remain possible, where a plus (minus) indicates players who are friends (enemies).

Networks in Figure 3 can be eliminated as Nash networks by the following principle. i does not want to maintain a negative directed link with a weakly stronger j, as this link is costly, and cannot increase i's benefits; given this fact, i also does not want to maintain a negative directed link to j if j is marginally weaker. It follows that in any Nash network, subsets of players who because of the network structure are about equally strong, are necessarily friends of each other. This principle immediately eliminates Figure 3(a): since no players are friends, they are necessarily about equally strong, and therefore the indicated negative undirected links cannot constitute an equilibrium. In Figure 3(b), the one pair of enemies have the same number of friends, and are therefore about equally strong, so that the negative

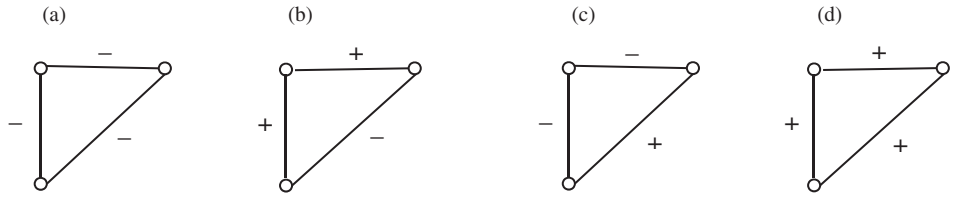

Figure 3. Four Cases of Possible Inimical (Indicated as –) and Friendly (Indicated as +) Relations between Three Players.

Note: In Hiller's (2017) account of balance theory as resulting from strategic network formation, (c) and (d) are the only Nash Networks. In the model of interdependent utilities, signed relations either represent how agents care about each other's social utilities in their structural-form utilities, or about each other's private utilities In their reduced-form utilities. In specific circumstances, only Figures 3(c) and 3(d) can represent reduced-form utilities.

undirected link between them can also not be part of an equilibrium. In Figure 3(c), however, in each of the two enemy pairs, there is a strong player (the one with a friend and an enemy), and a weak player (the one with two enemies). Specifically, Figure 3(c) is a Nash network if the top left player forms positive directed links to the other two players, and the other two players form negative links to the top left player; the top left player is then the weakest player and is exploited by the others. Finally, if as in Figure 3(d) all players are about equally strong because they all form positive directed links to each other, no player wants to change any relation into a negative one. Note that the two Nash networks (c) and (d) cannot be Pareto-ranked: the top-left player is better off in Figure 3(d), the other players in Figure 3(c). As shown by Hiller (2017), these features derived for three players above are generic for any number of players: in any Nash network, players are partitioned in cliques of friends, with enemy relations between the cliques. This is in line with balance theory, where the enemy of your enemy is your friend.

If in Figure 3(c) all three players have the same autonomous strength, then each player can be the common enemy of the rest. How is this changed if autonomous strengths are marginally different? Then in any Nash equilibrium fitting Figure 3(c), the common enemy is always the player with the smallest autonomous strength. In any situation where this is not the case, a player in the clique can switch her negative and positive links around, and at the same cost obtain a higher benefit in the contest with the player with the smallest autonomous strength. Starting from identical autonomous strengths, it is thus when the threat posed by one player in the form of her autonomous strength is *decreased*, that players cooperate against her. Yet, at the same time, when deviating from the case of autonomous strengths that can only differ marginally, as long as one player has an autonomous strength that is sufficiently large compared to the other players, this player forms a negative directed link to the remaining players, and becomes a common enemy to the remaining players.

6.2 *Interdependent Utilities*

We here show that the model of interdependent social preferences by Bergstrom (1999) can provide a rationale for balance theory, when it is extended to allow for spite (where players are rational in that they maximize following their social preferences). Consider an agent i who may either be altruistic or spiteful towards $(n-1)$ individual other agents. One can argue that if this agent is truly altruistic towards an agent j, then in a non-paternalistic manner she should care positively about the entire utility of agent j. In the same way, if agent i is truly spiteful towards j, her utility should negatively be affected by the

entire utility of j. Following the additively separable version of Bergstrom's model, the utility of i can now be written as

$$U_i = u_i + \sum_{j \neq i} \alpha_{ij} U_j \tag{1}$$

where U_i denotes agent i's entire utility, referred to as social utility, u_i denotes agent i's private utility, and α_{ij} is the weight that i puts on j's social utility (we do not follow Bourlès *et al*. (2017), where social utility also puts weight on the private utility of others). We assume that $|\alpha_{ij}| < 1$, so that agent i never puts more (negative or positive) weight on another agent's social utility, than on her own private utility. While (1) specifies how altruistic or spiteful agent i is to each other agent, it does not suffice to determine how agent i behaves in social contexts such as public-good production or charitable donation (Bourlès *et al*., 2017). For that purpose we need to represent agent i's utility in reduced form, as a function of all other $(n-1)$ agents' private utilities. Representing the system of n social utility functions in matrix form as $\boldsymbol{U} = \boldsymbol{u} + \boldsymbol{A}\boldsymbol{U}$, if $(\boldsymbol{I} - \boldsymbol{A})^{-1}$ exists, the system of reduced-form utility functions takes the form $\boldsymbol{U} = (\boldsymbol{I} - \boldsymbol{A})^{-1}\boldsymbol{u}$ (where $\boldsymbol{A}$ is the matrix of all weights that players put on each other's social utilities). We limit ourselves to studying the properties of the reduced-form utility functions as a function of the structural utility functions in (1), and do not apply the model to decisions in social contexts.

Consider as an example $n = 3$, and let agents 1 and 2 be identical agents, who equally care about each other and equally care about 3 ($\alpha_{12} = \alpha_{21} = \alpha, \alpha_{13} = \alpha_{23} = \beta$), and whom agent 3 equally cares about ($\alpha_{31} = \alpha_{32}$). Also, let agents 1 and 3, and 2 and 3 put the same weight on each other's social utilities ($\alpha_{13} = \alpha_{31} = \alpha_{23} = \alpha_{32} = \beta$). Then these agents' structural-form utility functions are presented as:

$$\begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} + \begin{bmatrix} 0 & \alpha & \beta \\ \alpha & 0 & \beta \\ \beta & \beta & 0 \end{bmatrix} \begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix} \tag{2}$$

Assuming that $1 - \alpha - 2\beta^2 > 0$, the agents' reduced-form utility functions can now be calculated as:

$$\begin{bmatrix} U_1 \\ U_2 \\ U_3 \end{bmatrix} = \frac{1}{(1+\alpha)\left(1-\alpha-2\beta^2\right)} \begin{bmatrix} 1-\beta^2 & \alpha+\beta^2 & (1+\alpha)\beta \\ \alpha+\beta^2 & 1-\beta^2 & (1+\alpha)\beta \\ (1+\alpha)\beta & (1+\alpha)\beta & 1-\alpha^2 \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix} \tag{3}$$

This reveals that when agents 1 and 2 care positively (respectively negatively) about agent 3's social utility in the structural-form utility function, or $\beta > 0$ (respectively $\beta < 0$), they also put positive (respectively, negative) weight on agent 3's private utility in the reduced-form utility function. Equally, when agent 3 cares positively (respectively negatively) about the social utilities of agents 1 and 2 in the structural-form utility function, or $\beta > 0$ (respectively $\beta < 0$), she also puts positive (respectively, negative) weight on the private utilities of agents 1 and 2 in the reduced-form utility function. Yet, agents 1 and 2 may care differently about each other's social utilities in the structural-form utility functions, than about their private utilities in the reduced-form utility functions, as we now go on to show.

Four cases can be distinguished, where Figure 3 can be used to represent these; the sign of a link is then interpreted as representing players who care positively or negatively about each other's social utilities in the structural-form utility functions (2). When $\alpha > 0$, $\beta > 0$ (Figure 3d), all agents care positively about each other's social utilities in the structural-form utility functions, and positively about each other's private utilities in the reduced-form utility functions. When $\alpha > 0$, $\beta < 0$ (Figure 3c, with agent 3 the top left agent), 1 and 2 care positively about each other's social utilities in (2), and about each other's private utilities in (3); also, both 1 and 3, and 2 and 3, care negatively about each other's social utilities in (2), and also about each other's private utilities in (3). Thus, in the first two cases, agents do not care differently about their private utilities in the reduced-form utility functions, and about their social utilities in the structural-form utility functions.

Yet, when $\alpha < 0$, $\beta > 0$ (Figure 3b with agent 3 the top left agent) and when $\beta^2 > -\alpha$, 1 and 2 care negatively about each other's social utilities in (2), but positively about each other's private utilities in (3) (where both 1 and 3, and 2 and 3, care positively about each other's social utilities in (2) and each other's private utilities in (3)). Intuitively, when the friendship between 1 and 3, and between 2 and 3 is sufficiently strong, while 1 and 2 may dislike each other, overall they still care positively about each other's private utilities, as these private utilities improve the plight of their mutual friend agent 3. In short, the friend of my friend is my friend; while the structural form for this case is represented by Figure 3(b), the reduced form is represented by Figure 3(d) (where the sign of the links then represents how players care about each other's private utilities). Yet, note that this reasoning only applies if the friendships between 1 and 3, and between 2 and 3, is sufficiently strong ($\beta^2 > -\alpha$); if this is not the case, in the reduced-form equation, the friend of my friend will still be my enemy.

Finally, when $\alpha < 0$, $\beta < 0$ (Figure 3a) and $\beta^2 > -\alpha$, 1 and 2 care negatively about each other's social utilities in (2), but positively about each other's private utilities in (3) (where both 1 and 3, and 2 and 3, care negatively about each other's social utilities in (2) and about each other's private utilities in (3)). This time, when the enmity between agent 1 and 3, and between agent 2 and 3 is sufficiently strong, while agents 1 and 2 may dislike each other, overall they still care positively about each other's private utilities, as these private utilities worsen the plight of their common enemy 3. In short, the enemy of my enemy is my friend; while Figure 3(a) represents the structural form, Figure 3(c) represents the reduced form. Again, this reasoning only applies if the spite between 1 and 3, and between 2, and 3, is sufficiently strong ($\beta^2 > -\alpha$); if this is not the case, in the reduced-form equations, the enemy of my enemy will still be my enemy. We conclude that a model of interdependent preferences provides a rationale for balance theory in specific instances. A general treatment beyond this specific example is due.

6.3 *Indirect reciprocity*

A basic rationale for cooperation is that the individual player builds up a reputation with the players with whom she interacts (direct reciprocity). Yet, in large populations, the probability of meeting the same player twice may be small, and therefore cooperation may only take place if one can also build up a reputation for having helped others, rather than only for having helped the person one is currently facing (indirect reciprocity). Nowak and Sigmund (1998) present simulations to show that cooperation is supported by the so-called image scoring rule where donors who donate to recipients increase their image score by one unit. Yet, following Sugden (1986), the reputation of a donor who fails to donate to a recipient who has not been donating in the past herself need not deteriorate, as this donor may be seen as punishing the recipient for his past behavior. The simulations of Leimar and Hammerstein (2001) suggest that the image scoring rule does not survive against a so-called standing strategy, where donors keep their reputation when they fail to donate to recipients who previously failed to donate, but gain in reputation when they donate to recipients who previously did donate.

Ohtsuki and Iwasa (2004) consider a binary reputation (good or bad), and within this framework, consider all possible reputation dynamics for players who, when matched, simultaneously decide whether or not to donate to each other. For each individual reputation dynamic, they derive the evolutionarily stable strategies (ESS), determining whether one should donate or not as a function of one's own reputation and that of the player to whom one is matched. In this manner, they obtain pairs consisting of a reputation dynamic and an ESS. They focus on the eight pairs that, across the parameter space, lead to the highest payoffs for the players (the 'leading eight'). The reputation dynamics in the leading eight have in common that one's reputation remains high/increases when cooperating with players with a good reputation, but decreases/remains low when defecting against them. Moreover, the leading eight have in common that players with a good reputation keep their reputation when not donating to players with a low reputation. Players can thus be seen as following reputation heuristics where friends (i.e. players

who just donated) of friends (i.e. players with a good reputation) are considered as friends (i.e. their reputation is increased or maintained), and enemies (i.e. players who just failed to donate) of friends (i.e. players with a good reputation) are seen as enemies (i.e. their reputation deteriorates). Finally, enemies (i.e. players who just failed to donate) of enemies (i.e. players with a bad reputation) are considered as friends (in the sense that their reputation does not deteriorate) specifically by players who have a good reputation themselves. As shown by the authors, those pairs among the leading eight where the latter reputation heuristic also applies to players with a bad reputation, achieve the highest payoff, but are also most vulnerable to errors.

In Gross and De Dreu (2019), players do not observe previous actions of others, but can exchange opinions and apply reputation heuristics in line with balance theory, based on their private experience and the private experience of the players from whom they obtain an opinion. The authors consider both players with 'friendship heuristics', who consider only the opinion of players who have been friendly to them ('the friend of my friend is my friend' and 'the enemy of my friend is my enemy'), and players with 'Heider heuristics', who additionally consider the opinion of players who have been unfriendly to them ('the friend of my enemy is my enemy' and 'the enemy of my enemy is my friend'). Their simulations show that the population cycles between these two types of players, and that the population polarizes into groups, with cooperation within groups and defection across groups – though groups constantly reconfigure.

Concluding this section, it is clear that rationalizations of balance theory can be provided, but that these are as different from the preference for belief consonance that originally underlay balance theory, as they are from each other.

7. Discussion

Having treated four approaches to the common-enemy effect separately, we are now ready to compare them. Table 3 summarizes the four approaches in terms of the cause of the common-enemy effect (i.e. what constitutes a common enemy), what moderates the effect (i.e. individual rationality, or a change in preferences or psychology); and what form the actual effect takes (an increase in cooperation in a specified form, or a change in one's social relations, social preferences, or assessment of others' reputations). As is clear from Table 3, the four approaches systematically differ both in what is considered as a common enemy, and in the form of the effect that this common enemy causes. Moreover, each approach has unique features that are not shared with any of the other approaches. The criticality approach is specific in allowing Nature to figure as a common enemy. When it does consider a strategic common enemy, the criticality approach focuses on an asymmetric conflict between first-mover defenders and a second-mover attacker. Cooperation specifically takes the form of coordination, and the approach allows for an opposite effect to the common-enemy effect, in the form of the deterrence effect. In the group-interaction approach, the common enemy takes the form of a group, rather than an individual, and the common-enemy effect considered is behavioral and rooted in laboratory experiments. The coalition formation approach is specific in endogenizing the process of group formation, rather than considering groups as given, as is the case in the criticality and group-interaction approaches. Finally, the balance-theory approach distinguishes itself by focusing on bilateral relations, in the form of social preferences or social relations between players, or in the form of players' assessment of each other's reputation. Considering a positive relation between two players as the equivalent of being in the same group (and a negative relationship as not being in the same group), this approach thus in principle allows for overlapping groups, though non-overlapping groups are still predicted to arise endogenously. A conclusion may therefore be that these four approaches have little more in common than their reference to the same catch-all term, or folk theory – justifying the plural use of common-enemy effects in the title of this review. Yet, we now consider several ways in which it is useful to consider these four approaches jointly.

Table 3. Comparison of the Four Models, According to Cause, Moderator, and Form of the Common-Enemy Effect.

Literature	Cause	Moderator of causal effect	Effect
Criticality	expected larger number of attacks by Nature, more sophisticated attacks by second-mover strategic attacker, or presence of a second-mover strategic network disruptor	individual rationality	higher probability of coordination on joint defense, or of coordination on the formation of a non-empty network (with possibility of lower probability of coordination: deterrence effect)
Group competition/group comparison	facing competing group (in a linear difference-form bonus scheme) or being able to compare to other group, rather than not facing any other group	change in preferences/change in psychology	more cooperation within each group (e.g. cooperation in within group Prisoner's Dilemma, or public-good investments)
Coalition formation in multi-player contests	lower effort costs of one contestant (= common enemy) in a multi-player contest	individual rationality	stopping to compete against each other, and pooling of efforts to compete against the common enemy, dividing the proceeds
Balance	negative social relations of one player with considered players, negative social preferences of one player towards considered players, or bad past behavior of one player towards considered players	individual rationality	forming of positive social relations between players, positive change in reduced-form social preferences towards each other, or positive assessment of each other's reputations

First, with some adaptations, the approaches of the criticality and group-interaction literatures can be integrated, and considered as providing alternative explanations for the same empirical effect. Suppose one would let participants play the defense-attack game (Section 3.1) in the laboratory, and would observe a positive impact on cooperation of the threat posed by a more sophisticated attacker. While this is in line with the theoretical predictions, one could not determine whether this is exclusively due to the predicted incentive effect where each player's contribution becomes more critical, or also to an unobserved behavioral effect. Following the same approach as in the group-interaction literature, to isolate a behavioral effect, the challenge is to construct experimental designs where the participants perceive a threat, even though their incentives are not changed. A recent paper in this direction is

Theelen and Böhm (forthcoming), whose experiment lets a second-mover attacker take away points from defenders in a way that is strategically neutral to the defenders. The authors still observe a positive effect of the presence of such a second-mover attacker on cooperation. Ideally, a related experimental design could serve as an intermediate treatment between a treatment without a threat, and one with a threat that changes players' incentives as in Section 3.1, thus allowing one to assess the relative importance of objective changes in the extent to which one's contribution is critical on the one hand, and perceived criticality on the other hand.

In the opposite direction, the group-interaction literature can be brought closer in line with the defense-attack game. Indeed, as shown in Section 4.1, several recent papers extend the experimental set-up of Bornstein and Ben-Yossef (1994) to asymmetric conflicts between groups, involving defense and attacks. An open question here is to what extent there is an effect of merely framing group interaction as the defense of a public good against attacks. Moreover, variants on the experimental design of Bornstein and Ben-Yossef could be constructed that allow one to establish whether there is a purely behavioral effect of facing a threat by an individual instead of Nature, and by a group instead of an individual (for an experimental design that achieves this purpose in the contest literature, see Abbink *et al*., 2010). Finally, the experimental design of Bornstein and Ben-Yossef could be extended to a coordination game as in the defense-attack model. Existing examples in this direction are experiments on intergroup Stag Hunt games (where joint cooperation continues to be Pareto-efficient within a group, but both joint cooperation and defection are Nash equilibria; Bornstein *et al*., 2002; Riechmann and Weimann, 2008); and on intergroup Snowdrift games (where it is both efficient, and an equilibrium, that a single one of the players in the group cooperates; Bornstein *et al*., 1997). While these experiments also find that the presence of an outgroup increases the level of coordination, contrary to what is the case in Bornstein and Ben-Yossef (1994), group interaction affects individual incentives, and the apparent behavioral common-enemy effect can be explained by a simple change in individual incentives as well (Jordan *et al*., 2017). The challenge remains here to design experiments that isolate a behavioral common-enemy effect.

Second, even if the approaches cannot be seen as giving different explanations for the same empirical effect, considering the four different approaches jointly is still useful as they can be seen as dealing with different stages of a single sequential game. In this view, the balance approach can be seen as dealing with the earliest stage of the sequential game, where players' social preferences or social relations are formed, or where the information obtained from previous play is assessed. The coalition-formation approach deals with the process of group formation. The group-interaction approach can be seen as investigating how incentive or behavioral effects affect individual investments in the public good of the group, once interacting groups have been formed. Finally, the criticality approach can be seen as dealing with the possibility that competing groups sabotage each other's public goods once these have been produced, which in turn prompts defensive cooperation to prevent such sabotage. Using team sports as a metaphor, competition between sports teams may be modeled as a contest between groups, where the pooled efforts of each team determines which team is more likely to win. Yet, this may only provide a complete model of team competition for specific sports such as team time trials in cycling, or tugs-of-war. In team field sports, competition additionally has asymmetric aspects, where one team tries to sabotage the efforts of the other team, with the other team defending against such sabotage (where sabotage may violate the rules of the game, but may be difficult to monitor; for empirical evidence on sabotage in sports, see Kempa and Rusch, 2019).

In first instance from a theoretical point of view, the question then arises how these different stages impact on each other, where by backward induction, players anticipate the outcome in the future stages. Players may thus form social relations anticipating how this will lead them to form groups. They may in turn form groups, anticipating how group formation will in turn alter behavior in the form of increased investments in the public good of the group (indeed, the coalition-formation approach in Section 5 already includes this type of analysis, as players form coalitions in contests, anticipating how this will affect their investments within the coalition). Finally, once groups have been formed, players may invest in

the public good of the group, anticipating how this will in turn impact on possible intergroup sabotage, and on defense against such sabotage. It should be noted that this avenue for future research is not straightforward, as the modeling assumptions of the different approaches are not typically compatible in their present form, and are not easily integrated in a single sequential game.

Third, apart from whether some approaches can be seen as different explanations of the same empirical effect, or whether the approaches can be integrated by representing them as different stages of a sequential game, jointly considering these four literatures is useful, as they can benefit from importing each other's assumptions or approaches, as we now illustrate. Consider the approach in the group-interaction literature of constructing a realistic conflict, which at the same time objectively speaking should not matter for the players' behavior. Such an experimental design does not serve so much to model any real-world situation, but to isolate behavioral effects that could otherwise never be observed. This approach can therefore be exported to investigate behavioral effects in the other approaches, where these effects may be behavioral translations of the rational effects described in the theories. After all, whereas adaptive evolution may provide rationales for common-enemy effects, in a contemporary environment such effects may appear as behavioral biases (e.g. Eaton *et al*., 2011).

Furthermore, consider the view of the common-enemy effect in the group-interaction literature where, for given efforts of the other group, the players of an individual group are better off the more they cooperate; yet, considered across both groups, players are better off if none of them cooperates. Such an argument may equally apply to the defense-attack model (Section 3.1), in a symmetrized version of this model where two competing groups both jointly produce defense against each other, and attacks. While defending may be welfare improving for the individual group given the attacking level produced by the other group, seen across all groups, players could be better off if no cooperation on attacking or defending would take place. Also, the defense-attack model could benefit from an extension to continuous efforts by the defenders and continuous attacking levels by the attacker, allowing for a more sophisticated analysis where the attacker spreads her attacking efforts over the defenders in any way that she sees fit. At the same time, the modeling assumptions of the attack-defense model make sense for modeling instances of coalition formation as well. The decision to preemptively form a defensive alliance (e.g. NATO) against a possible future enemy rather than to face this enemy alone, is a costly all-or-nothing decision that requires coordination; the players are then involved in an asymmetric conflict with the enemy, who is a second mover attacking after having observed whether or not an alliance has been formed (for an approach in this direction in the contest literature, see Deck *et al*., 2015). Finally, among the extensions of the attack-defense model considered (Section 3.1), heterogeneity is an important extension to consider in this and other approaches. The experiment of Theelen and Böhm (forthcoming) has a similar set-up as the attack-defense model, and includes a treatment where players are heterogeneously affected by the attacker; they find that in spite of this, the presence of the attacker continues to have a positive effect on cooperation. At the same time, in the group-interaction literature, Van Bunderen *et al*. (2018) experimentally induce egalitarian and non-egalitarian teams. The authors find that the lack of a common fate in non-egalitarian teams leads to a *reduction* in intragroup cooperation because of group comparison, rather than to an increase. The apparently conflicting conclusions from these two recent papers point to the importance of future research on the impact of heterogeneity on the common-enemy effect.

Notes

1. We do not claim to provide an overview of all game-theoretic accounts of the common-enemy effect. For instance, Kovenock and Roberson (2012) and Rietzke and Roberson (2013) study Colonel Blotto games where transfers may take place among players with a common enemy. Intuitively, a strong player may make a transfer to a weak player, forcing the common enemy to spend part of her resources in a contest with the weak player, which in turn creates a positive externality for the strong player.

2. Not treated in this review is literature that investigates whether war or natural disasters increase cooperation (Bauer *et al*., 2016; Cassar *et al*., 2017). While part of this literature is based on laboratory experiments, confounding factors other than any common-enemy effect may explain the results, which deserves separate attention. For this reason, when looking at empirical studies, this paper focuses on studies where the common enemy is induced in the laboratory.
3. In a variant of this argument, a harsher environment increases agents' uncertainty, and this drives the evolution of cooperation (Andras *et al*., 2007).
4. In terms of Rapoport (1967), when the added benefit of cooperating jointly does not exceed cooperation costs c, $0.5^A - 1 + c$ can be considered as the greed component of the social dilemma, namely the incentive to free-ride on joint cooperation. Similarly, when the added benefit of cooperating alone does not exceed c, $-0.5^A + c$ can be considered as the fear component of the social dilemma, namely the incentive to avoid being the victim of free-riding. In Rapoport's terms, the result of an increased number of attacks is ambiguous, because the greed component decreases, but the fear component increases.
5. De Waal (2006) coined the term 'veneer theory' for the theory that moral behaviour is not natural to humans, and stresses the long history of this theory in scientific thought. The logical consequence of this theory is that when for example a natural disaster makes society collapse, cooperation breaks down.
6. For instance, this is the case for Gunnthorsdottir and Rapoport (2006), who study an Intergroup Prisoner's Dilemma with a fixed bonus for the winning group instead of the linear difference-form bonus scheme in Bornstein and Ben-Yossef (1994).
7. E.g., the game in Table 2 can be reframed as a discrete-choice linear public-goods game with negative intergroup externalities as follows. Each participant of group A receives a fixed payoff of F, with $F = 6$, and an endowment of e, with $e = 4$, which he or she can either invest in the group account of group A, or keep for him- or herself. The sum of the investments in the group account of group A is multiplied by a coefficient F/e (the so-called marginal per capita return), and divided equally among the members of group A. This sum of the investments in the group account of group A, multiplied by the coefficient F/e, at the same time constitutes a negative externality to the members of group B, which is equally divided over them. Group B is completely symmetric. For an overview of framing effects in public-goods experiments, see Cartwright (2016).
8. As shown in Puurtinen and Mappes (2009), when group competition is added to such a reframed game by giving a bonus that is a function of the difference between the investment levels of the two groups, group competition again increases cooperation; yet, group competition affects individual incentives to cooperate as well in this case.
9. More precisely, this is the probability of winning when at least one of the efforts is positive; if all efforts of winning are zero, then the probability of winning equals ⅓. In the same way, when players 1 and 2 form a coalition but all efforts are zero, the probability that the coalition wins the contest equals $\frac{1}{2}$.
10. Alternatively, there is no joint production, but the individual player's efforts decrease the resources that other players win from him/her by winning the contest. In this case, it makes sense to distinguish between groups that are only defensive and try to defend their joint resources, or groups that are also offensive and additionally try to obtain the resources of other players (Niou and Tan, 2005). Just as it can pay off for groups to restrain their efforts in our case, it can pay off for groups not to be offensive.

Acknowledgements

I would like to thank Joyce Delnoij, Timo Hiller, Britta Hoyer, Florian Morath, two anonymous referees and the editor for helpful comments and feedback. Any remaining errors are my own.

References

Abbink, K., Brandts, J., Herrmann, B. and Orzen, H. (2010) Intergroup conflict and intra-group punishment in an experimental contest game. *American Economic Review* 100: 420–447.

Ahn, T.K., Isaac, R.M. and Salmon, T.C. (2011) Rent seeking in groups. *International Journal of Industrial Organization* 29: 116–125.

Andras, P., Lazarus, J. and Roberts, G. (2007) Environmental adversity and uncertainty favour cooperation. *BMC Evolutionary Biology* 7: 240.

Antal, T., Krapivsky, P.L. and Redner, S. (2006) Social balance on networks: the dynamics of friendship and enmity. *Physica D* 224: 130–136.

Baer, M., Leenders, R.T.A.J., Oldham, G.R. and Vadera, A.K. (2010) Win or lose the battle for creativity: the power and perils of intergroup competition. *Academy of Management Journal* 4: 827–845.

Bauer, M., Blattman, C., Chytilová, J., Henrich, J., Miguel, E. and Mitts, T. (2016) Can war foster cooperation? *Journal of Economic Perspectives* 30: 249–274.

Benard, S. and Doan, L. (2011) The conflict-cohesion hypothesis: past, present, and possible futures. *Advances in Group Processes* 28: 189–225.

Bergstrom, T. (1999) Systems of benevolent utility functions. *Journal of Public Economic Theory* 1: 71–100.

Beyers, J. and De Bruycker, I. (2018) Lobbying makes (strange) bedfellows: explaining the formation and composition of lobbying coalitions in EU legislative politics. *Political Studies* 66: 959–984.

Bloch, F. (2012) Endogenous formation of alliances in conflicts. In M.R. Garfinkel and S. Skaperdas (eds.), *Oxford Handbook of the Economics of Peace and Conflict*. New York: Oxford University Press.

Böhm, R. and Rockenbach, B. (2013) The inter-group comparison – intra-group cooperation hypothesis: comparisons between groups increase efficiency in public goods provision. *PLoS One* 8(2): e56152.

Böhm, R., Rusch, H. and Baron, J. (2020) The psychology of intergroup conflict: a review of theories and measures. *Journal of Economic Behavior and Organization* 178: 947–962.

Böhm, R., Rusch, H. and Gürerk, Ö. (2016) What makes people go to war? Defensive intentions motivate retaliatory and preemptive intergroup aggression. *Evolution and Human Behavior* 37: 29–34.

Bornstein, G. (2003) Intergroup conflict: individual, group, and collective interests. *Personality and Social Psychology Review* 7: 129–145.

Bornstein, G. and Ben-Yossef, M. (1994) Cooperation in intergroup and single-group social dilemmas. *Journal of Experimental Social Psychology* 30: 52–67.

Bornstein, G., Budescu, D. and Zamir, S. (1997) Cooperation in intergroup, N-person, and two-person games of chicken. *Journal of Conflict Resolution* 41: 384–406.

Bornstein, G., Gneezy, U. and Nagel, R. (2002) The effect of intergroup competition on group coordination: an experimental study. *Games and Economic Behavior* 41: 1–25.

Bornstein, G., Winter, E. and Goren, H. (1996) Experimental study of repeated team games. *European Journal of Political Economy* 12: 629–639.

Bowles, S. (2009) Did warfare among ancestral hunter-gatherers affect the evolution of human social behaviors? *Science* 324: 1293–1298.

Bourlès, R., Bramoullé, Y. and Perez-Richet, E. (2017) Altruism in networks. *Econometrica* 85: 675–689.

Burton-Chellew, M.N. and West, S.A. (2012) Pseudocompetition among groups increases cooperation in a public-goods game. *Animal Behaviour* 84: 947–952.

Cacault, M.P., Goette, L., Lalive, R. and Thoenig, M. (2015) Do we harm others even if we don't need to? *Frontiers in Psychology* 6: 729.

Cartwright, D. and Harary, F. (1956) Structural balance: a generalization of Heider's theory. *The Psychological Review* 63: 277–293.

Cartwright, E. (2016) A comment on framing effects in linear public good games. *Journal of the Economic Science Association* 2: 73–84.

Cassar, A., Healy, A. and Von Kessler, C. (2017) Trust, risk, and time preferences after a natural disaster: experimental evidence from Thailand. *World Development* 94: 90–105.

Charness, C., Rigotti, L. and Rustichini, A. (2007) Individual behavior and group membership. *American Economic Review* 97: 1340–1352.

Chaudhuri, A. (2011) Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics* 14: 47–83.

Chen, Y. and Li, S.X. (2009) Group identity and social preferences. *American Economic Review* 99: 431–457.
Choi, J.-K. and Bowles, S. (2007) The coevolution of parochial altruism and war. *Science* 318: 636–640.
Coser, L.A. (1956) *The Functions of Social Conflict*. Glencoe, IL: Free Press.
Crescenzi, M.J.C., Kathman, J.D., Kleinberg, K.B. and Wood, R.M. (2012) Reliability, reputation, and alliance formation. *International Studies Quarterly* 56: 259–274.
Davis, J.A. (1967) Clustering and structural balance in graphs. *Human Relations* 20: 181–187.
Dawes, R.M. and Messick, D.M. (2000) Social dilemmas. *International Journal of Psychology* 35: 111–116.
Deck, C., Foster, J. and Song, H. (2015) Defense against an opportunistic challenger: theory and experiments. *European Journal of Operational Research* 242: 501–513.
De Cremer, D. and Van Vugt, M. (1998) Collective identity and cooperation in public goods dilemma: a matter of trust or self-efficacy? *Current Research in Social Psychology* 3: 1–11.
De Jaegher, K. and Hoyer, B. (2016a) By-product mutualism and the ambiguous effects of harsher environments: a game-theoretic model. *Journal of Theoretical Biology* 393: 82–97.
De Jaegher, K. and Hoyer, B. (2016b) Collective action and the common-enemy effect. *Defence and Peace Economics* 27: 644–664.
De Waal, F. (2006) Morally evolved: primate social conflicts, human morality, and the rise and fall of "veneer theory. In Macedo, S. and Ober, J. (eds.), *Primates and Philosophers: How Morality Evolved*, pp. 1–58. Princeton, NJ: Princeton University Press.
Doreian, P. and Krackhardt, D. (2001) Pre-transitive balance mechanisms for signed networks. *Journal of Mathematical Sociology* 25: 43–67.
Doreian, P. and Mrvar, A. (2015) Structural balance and signed international relations. *Journal of Social Structure* 16: 2.
Dziubiński, M. and Goyal, S. (2013) Strategic network disruption and defense. *Journal of Public Economic Theory* 18: 802–830.
Eaton, B.C., Eswaran, M. and Oxoby, R.J. (2011) Us" and "Them": the origin of identity, and its economic implications. *Canadian Journal of Economics* 44: 719–748.
Evrigenis, I.D. (2008) *Fear of Enemies and Collective Action*. Cambridge: Cambridge University Presots.
Garfinkel, M. and Skaperdas, S. (2007) Economics of conflict: an overview. In Sandler, T. and Hartley, K. (eds), *Handbook of Defense Economics*, vol. II, pp. 649–709. Amsterdam: Elsevier Science.
Gross, J. and De Dreu, C.K.W. (2019) The rise and fall of cooperation through reputation and group polarization. *Nature Communications* 10: 776.
Gunnthorsdottir, A. and Rapoport, A. (2006) Embedding social dilemmas in intergroup competition reduces free-riding. *Organization Behavior and Human Decision Processes* 101: 184–199.
Halevy, N., Bornstein, G. and Sagiv, L. (2008) In-group love" and "out-group hate" as motives for individual participation in intergroup conflict – a new game paradigm. *Psychological Science* 19: 405–411.
Hargreaves, S.P. and Zizzo, D.J. (2009) The value of groups. *American Economic Review* 99: 295–323.
Hart, S. and Kurz, M. (1983) Endogenous formation of coalitions. *Econometrica* 51: 1047–1064.
Healy, B. and Stein, A. (1973) The balance of power in international history: theory and reality. *Journal of Conflict Resolution* 17: 33–61.
Heider, F. (1946) Attitudes and cognitive organization. *Journal of Psychology* 21: 107–112.
Heider, F. (1958) *The Psychology of Interpersonal Relations*. Wiley: New York.
Herbst, L., Konrad, K.A. and Morath, F. (2015) Endogenous group formation in experimental contests. *European Economic Review* 74: 163–189.
Hiller, T. (2017) Friends and enemies: a model of signed network formation. *Theoretical Economics* 12: 1057–1087.
Hirshleifer, J. (1983) From weakest-link to best-shot: the voluntary provision of public goods. *Public choice* 41: 371–386.
Hirshleifer, J. (1991) The paradox of power. *Economics and Politics* 3: 177–200.
Hogg, M.A. (2001) Social categorization, depersonalization and group behaviour. In Hogg, M.A. and Tindale, R.S. (eds.) *Blackwell Handbook of Social Psychology: Group Processes* (pp. 56–85). Oxford: Blackwell Publishers.
Hornsey, M.J. (2008) Social identity theory and self-categorization theory: a historical review. *Social and Personality Psychology Compass* 2: 204–222.

Hoyer, B. and De Jaegher, K. (2012) Network disruption and the common enemy effect. U.S.E. Discussion Paper Series No. 12-06, Utrecht University.
Hoyer, B. and De Jaegher, K. (2016) Strategic network disruption and defence. *Journal of Public Economic Theory* 18: 802–830.
Hoyer, B. and Haller, H. (2019) The common enemy effect under strategic network formation and disruption. *Journal of Economic Behavior and Organization* 162: 146–163.
Hoyer, B. and Rosenkranz, S. (2018) Determinants of equilibrium selection in network formation – an experiment. *Games* 9: 89.
Hummon, N.P. and Doreian, P. (2003) Some dynamics of social balance process: bringing Heider back into balance theory. *Social Networks* 25: 17–49.
Intergovernmental Panel on Climate Change (2018) *Global Warming of 1.5 °C above Pre-industrial Levels and Related Global Greenhouse Gas Emission Pathways, in the Context of Strengthening the Global Response to the Threat of Climate Change, Sustainable Development, and Efforts to Eradicate Poverty*. https://www.ipcc.ch/sr15/download/#full.
Jackson, J.W. (1993) Realistic group conflict theory: a review and evaluation of the theoretical and empirical literature. *Psychological Record* 43: 395–413.
Jackson, M.O. and Wolinsky, A. (1996) A strategic model of social and economic networks. *Journal of Economic Theory* 71: 44–74.
Jordan, M.R., Jordan, J.J. and Rand, D.G. (2017) No unique effect of intergroup competition on cooperation: non-competitive thresholds are as effective as competitions between groups for increasing human cooperative behavior. *Evolution and Human Behavior* 38: 102–108.
Kempa, K. and Rusch, H. (2019) Dissent, sabotage, and leader behaviour in contests: evidence from European football. *Managerial and Decision Economics* 40: 500–514.
Kirchsteiger, G., Mantovani, M., Mauleon, A. and Vannetelbosch, V. (2016) Limited farsightedness in network formation. *Journal of Economic Behavior and Organization* 128: 97–120.
Konrad, K.A. (2009) *Strategy and Dynamics in Contests*. Oxford, UK: Oxford University Press.
Konrad, K.A. and Morath, F. (2012) Evolutionarily stable in-group favoritism and out-group spite in intergroup conflict. *Journal of Theoretical Biology* 306: 61–67.
Kovenock, D. and Roberson, B. (2012) Coalitional Colonel Blotto games with application to the economics of alliances. *Journal of Public Economic Theory* 14: 653–676.
Kropotkin, P. (1902) *Mutual Aid*. London: Heinemann.
Lane, T. (2016) Discrimination in the laboratory: a meta-analysis of economics experiments. *European Economic Review* 90: 375–402.
Leimar, O. and Hammerstein, P. (2001) Evolution of cooperation through indirect reciprocity. *Proceedings of the Royal Society of London B* 268: 745–753.
Levy, J.S. (1989) The diversionary theory of war: a critique. In M. M. Midlarsky (ed.), *Handbook of War Studies*, pp. 259–288. London: Unwin Hyman, .
Markussen, T., Reuben, E. and Tyran, J.-R. (2013) Competition, cooperation and collective choice. *Economic Journal* 124: F163–F195.
McLauchlin, T. and Pearlman, W. (2012) Out-group conflict, in-group unity? Exploring the effect of repression on intramovement cooperation. *Journal of Conflict Resolution* 56: 41–66.
Messick, D.M. and Brewer, M.B. (1983) Solving social dilemmas. In Wheeler, L. and Shaver P.R. (eds.), *Review of Personality and Social Psychology*. pp. 11–44. Beverly Hills, CA: Sage.
Mesterton-Gibbons, M. and Dugatkin, L.A. (1992) Cooperation among unrelated individuals: evolutionary factors. *Quarterly Review of Biology* 67: 267–281.
M. Djawadi, B., E., A., Hoyer, B. and Recker, S. (2019) Network formation and disruption – an experiment: are equilibrium networks too complex? *Journal of Economic Behavior and Organization* 157: 708–734.
Muller, E.D. and Opp, K.D. (1986) Rational choice and rebellious collective action. *American Political Science Review* 80: 471–488.
Münster, J. and Staal, K. (2011) War with outsiders makes peace inside. *Conflict Management and Peace Science* 28: 91–110.
Nalbantian, H.R. and Schotter, A. (1997) Productivity under group incentives: an experimental study. *American Economic Review* 87: 314–341.

Niou, E.M.S. and Tan, G. (2005) External threat and collective action. *Economic Inquiry* 43: 519–530.
Nitzan, S. (1991) Collective rent dissipation. *Economic Journal* 101: 1522–1534.
Noh, S.J. (2002) Resource distribution and stable alliances with endogenous sharing rules. *European Journal of Political Economy* 18: 129–151.
Nowak, M.A. and Sigmund, K. (1998) Evolution of indirect reciprocity by image scoring. *Nature* 393: 573–577.
Ohtsuki, H. and Iwasa, Y. (2004) How should we define goodness? – reputation dynamics in indirect reciprocity. *Journal of Theoretical Biology* 231: 107–120.
Patrikios, S. (2008) American Republican religion? Disentangling the causal link between religion and politics in the US. *Political Behavior* 30: 367–389.
Puurtinen, M. and Mappes, T. (2009) Between-group competition and human cooperation. *Proceedings of the Royal Society B* 276: 355–360.
Rand, D.G. and Nowak, M.A. (2013) Human cooperation. *Trends in Cognitive Sciences* 17: 413–425.
Rapoport, A. (1967) A note on the 'index of cooperation' for prisoner's dilemma. *Journal of Conflict Resolution* 11: 101–103.
Rebers, S. and Koopmans, R. (2012) Altruistic punishment and between-group competition – evidence from n-person Prisoner's Dilemmas. *Human Nature* 23: 173–190.
Reuben, E. and Tyran, J.-R. (2010) Everyone is a winner: promoting cooperation through all-can-win intergroup competition. *European Journal of Political Economy* 26: 25–35.
Riechmann, T. and Weimann, J. (2008) Competition as a coordination device: experimental evidence for a minimum effort coordination game. *European Journal of Political Economy* 24: 437–454.
Rietzke, D. and Roberson, B. (2013) The robustness of "enemy-of-my-enemy-is-my-friend' alliances. *Social Choice and Welfare* 40: 937–956.
Rusch, H. (2014) The evolutionary interplay of intergroup conflict and altruism in humans: a review of parochial altruism theory and prospects for its extension. *Proceedings of the Royal Society B* 281: 20141539.
Sánchez-Pagés, S. (2007) Endogenous coalition formation in contests. *Review of Economic Design* 11: 139–163.
Sheremata, R.M. (2011) Perfect-substitutes, best-shot, and weakest-link contests between groups. *Korean Economic Review* 2: 5–32.
Sheremata, R.M. (2018) Behavior in group contests: a review of experimental research. *Journal of Economic Surveys* 32: 683–704.
Sherif, M. (1966) *In Common Predicament: Social Psychology of Intergroup Conflict and Cooperation*. Boston: Houghton Mifflin Company.
Simmel, G. [1908](1955) *Conflict*. Glencoe, IL: Free Press.
Skaperdas, S. (1998) On the formation of alliances in conflict and contests. *Public Choice* 96: 25–42.
Stein, A.A. (1976) Conflict and cohesion: a review of the literature. *Journal of Conflict Resolution* 20: 143–172.
Sugden, R. (1986) *The Economics of Rights, Co-operation and Welfare*. Oxford, UK: Blackwell.
Sumner, W.G. (1906) *Folkways: A Study of the Sociological Importance of Usages, Manners, Customs, Mores, and Morals*. Boston, MA: Ginn.
Tajfel, H. and Turner, J.C. (1986) The social identity theory of intergroup behaviour. In S. Worchel and W. Austin (eds.) *Psychology of Intergroup Relations* (pp. 7–24). Chicago: Nelson-Hall.
Tan, J.H.W. and Bolle, F. (2007) Team competition and the public goods game. *Economics Letters* 96: 133–139.
Theelen, M.M.P. and Böhm, R. (forthcoming) The conflict-cooperation effect persists under intragroup payoff asymmetry. *Group Processes and Intergroup Relations*
Traag, V. (2014) *Algorithms and Dynamical Models for Communities and Reputation in Social Networks*. Cham: Springer.
Tullock, G. (1980) Efficient rent-seeking. In R.D. Tollison and G. Tullock (eds.), *Toward a Theory of the Rent-Seeking Society*, pp. 97–112, College Station: Texas A&M University Press.
Van Bunderen, L., Greer, L.L. and Van Knippenberg, D. (2018) When interteam conflict spirals into intrateam power struggles: the pivotal role of team power structures. *Academy of Management Journal* 61: 1100–1130.

Van Knippenberg, D. (2003) Intergroup relations in organizations. In M.A. West, D. Tjosvold and K.G. Smith (eds.), *International Handbook of Organizational Teamwork and Cooperative Working* (pp. 381–399). Chichester, UK: Wiley.
Vasi, I.B and Macy, M. (2003) The mobilizer's dilemma: crisis, empowerment and collective action. *Social Forces* 81: 979–998.
Weisel, O. and Böhm, R. (2015) Ingroup love" and "outgroup hate" in intergroup conflict between natural groups. *Journal of Experimental Social Psychology* 60: 110–120.
Weisel, O. and Zultan, R. (2016) Social motives in intergroup conflict: group identity and perceived target of threat. *European Economic Review* 90: 122–133.
Yamagishi, T., Jin, N. and Kiyonari, T. (1999) Bounded generalized reciprocity: in-group boasting and in-group favoritism. *Advances in Group Processes* 16: 161–197.
Zizzo, D.J. (2010) Experimenter demand effects in economic experiments. *Experimental Economics* 13: 75–98.
Zizzo, D.J. (2011) You are not in my boat: common fate and discrimination against outgroup members. *International Review of Economics* 58: 91–103.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Table A1. Criticality in Terms of the Degree of Complementarity.