Games and Economic Behavior 146 (2024) 255–276

Contents lists available at ScienceDirect

Games and Economic Behavior

journal homepage: www.elsevier.com/locate/geb

How alliances form and conflict ensues

Lu Dong [a], Lingbo Huang [b,*], Jaimie W. Lien [b], Jie Zheng [b]

[a] SUSTech Business School, Southern University of Science and Technology, Shenzhen 518055, China
[b] Center for Economic Research, Shandong University, Jinan 250100, China

ARTICLE INFO

JEL codes:
C72
C92
D74
D85
F51

Keywords:
Network formation
Conflict
Alliance
Bully
Peace

ABSTRACT

In a social network in which friendly and rival bilateral links can be formed, how do alliances between decision-makers form, and what determines whether a conflict will arise? We study a network formation game between ex-ante symmetric players in the laboratory to examine the dynamics of alliance formation and conflict evolution. A peaceful equilibrium yields the greatest social welfare, while a successful bullying attack transfers the victimized player's resources evenly to the attackers at a cost. In within-subject and between-subject laboratory experiments, we find that the relative frequency of peaceful and bullying outcomes increases in the cost of attack. We further examine the dynamics leading to the final network and find that groups tend to coordinate quickly on a first target for attack, while the first attacker entails a non-negligible risk of successful counter-attack. These findings provide insights for understanding social dynamics in group coordination.

1. Introduction

Resource-seeking and factional dynamics have driven much of the bloodshed and transference of resources observed throughout human history. Since the end of World War II, the total number of state-based conflicts (predominantly small-scale) occurring in any given year has generally increased, and more than doubled on average, in recent years.[1] When explaining conflicts, factors such as power struggles, initial resource distribution, tribal psychology of "us" versus "them," and individual leaders, among other path dependent factors, are nearly always invoked as the primary reasons for conflict. However, all of these reasons, which are to varying degrees driven by prior environmental, social conditions, or idiosyncratic characteristics of leaders, lead naturally to the following question: Will conflict still spontaneously arise in a highly neutral social environment with ex-ante homogenous individuals who have no prior rivalry or social interaction?

Interdisciplinary scientific evidence suggests that such spontaneous conflicts are not uncommon in a range of social contexts. Bullying among teenagers provides an analogous scenario among individuals of seemingly similar social status (see non-experimental studies, Salmivalli et al. 1997, O'Connell et al. 1999, Huitsing et al. 2012). Siding with the majority in a social clique could be a strategy for safety in numbers and higher social status, while the consequences for left-out or bullied individuals can be very

* Corresponding author.
E-mail addresses: donglu@sustech.edu.cn (L. Dong), lingbo.huang@outlook.com (L. Huang), jaimie.academic@gmail.com (J.W. Lien), jie.academic@gmail.com (J. Zheng).

[1] Included in the state-based conflicts are the following categories: Colonial or imperial conflicts, Conflicts between states, Civil conflicts, and Civil conflicts with foreign state intervention. Civil conflicts account for the greatest number of conflicts in the latest year available (2016), while Civil conflicts with foreign state intervention is the second most frequent category. See Roser (2016); retrieved from: https://ourworldindata.org/war-and-peace; "State-based conflicts since 1946, 1946 to 2016" data collected by Uppsala Conflict Data Program, last time accessed at June 23, 2021.

https://doi.org/10.1016/j.geb.2024.05.009
Received 28 July 2022;
Available online 8 June 2024

unpleasant. Intragroup several-against-one lethal attacks in the form of coordinated attacks by members of one alliance towards a targeted victim have even been observed in the wild among chimpanzees (e.g., Pruetz et al. 2017), as well as in some tribal societies (e.g., Macfarlan et al. 2014). Such observations also raise a possibility that bullying based on alliances could be an evolved behavior derived from broader survival strategies in the interaction between rival groups. Finally, a relevant historical example is that while the formation of military alliances in 19th century Europe was in flux for a long period historically, it eventually stabilized and led up to the First World War which still shapes the international landscape to the present day (e.g., Antal et al. 2006).

In this study, we focus on the interaction of decision-makers in a group, aiming to study the origins of conflicts in a network context. By implementing a laboratory experiment in which no participant is ex-ante different from the others, while players are allowed to freely form friend and enemy links in real-time, our study is able to reveal the extent to which groups converge towards conflict versus peaceful states, and furthermore, the path that groups take to arrive there.[2] What are the underlying processes through which alliances grow, and what determines whether conflict ensues? When a universal alliance can guarantee peace, equality and the highest social welfare, is group conflict still inevitable? Our simple, neutral and symmetric setting serves as a natural first step to answer these questions, which would be otherwise hard to address in a field setting.

We explore the dynamics of alliance formation and conflict in two laboratory experiments (a within-subject and a between-subject version for different costs incurred by attacking) on a signed network game in which players can either befriend or fight against others, obtaining our theoretical predictions by extending Hiller (2017) to allow for neutral links between players. Since link formation is critical to our experiment design, models such as hawk-dove, war of attrition or standard contest frameworks, lack key features in explaining how humans fight and make peace, because they do not consider the detailed process of alliance formation. A few studies that address endogenous alliance formation often do so in a structured manner involving sequential steps, each of which allows only a subset of strategies.[3]

Compared to prior studies, the network game we implement here provides a rich and flexible context to explore the dynamics of alliance formation. A player's instantaneous behavioral strategy is to send either a friendly link or a rival link to another player, make no change to the link status, or to remove an existing link of either type. In our setup, an alliance requires mutual consent while rivalry only requires unilateral aggression. Forming an alliance of more than two players requires pairwise mutual consent of each pair in the alliance. Players are free to initiate or break an alliance, and to extend or retract a rivalry.

Previous studies suggest that individuals often find it difficult to coordinate in network games due to the relatively large set of possible strategies to implement. To facilitate coordination, we implement a continuous-time setup in which participants can freely make links to each other while the network structure and hypothetical momentary payoffs are updated in real-time. We allow participants randomly assigned into groups of four, to freely form friend and enemy links with other participants in the group, with each round lasting between 75 and 105 s. Only the final network configuration determines the experimental participants' actual payments. This design provides ample time for participants to coordinate their decisions. It also allows participants to learn about the payoff consequences of their link choices before they are finalized.

Our theoretical analysis of the network formation game generates predictions about the relative likelihoods of final network structures based on equilibrium concepts. Two networks stand out as being most robust to a series of increasingly stringent equilibrium refinement criteria. One equilibrium network structure, which we call Peace, is the situation in which every pair of players in the network are mutually connected by friendly links. The other network, which we call Bully, is the situation that three members form an alliance (via mutual friendly links) and all attack (by each sending rival links to) the fourth member. While the Peace equilibrium provides equal payoffs among all the players and obtains the highest possible welfare outcome, the Bully equilibrium favors the attackers at the expense of the victim, but at a total welfare loss due to the cost of attacking. The theory predicts that there exists a threshold cost of attacking, beyond which Peace becomes more robust to equilibrium selection criteria, and therefore, we expect to observe more Peace final networks beyond that cost threshold. The predictions generated by the theoretical analysis of the finalized network game are strongly supported in the experimental data.

We then explore the details of the network formation, which reveals rich dynamics in the alliance and conflict incidence. Although from a theoretical standpoint, analyzing the dynamics of the network formation game is highly complicated, and a fully dynamic model is beyond the scope of our work, our key question of interest is why some groups converge to a bullying network while others obtain a peaceful configuration in the end. Group-level analysis shows that bullying networks not only form a three-member alliance quickly but also coordinate on a common rival early. Furthermore, individual-level data reveal that a player who receives the first attack from any other player in the group is the most likely (75.4% and 76.9% of the time in the within-subject and between-subject experiments, respectively) to become the final victim in Bully scenarios. Analysis also shows that coordinating upon which player to attack is highly path dependent, in that players who have become salient in the link formation process via a negative link directed towards them are the most vulnerable. Thus, the initiation of a first attack serves as a costly coordination device on behalf of the other two players.

Notably, we find that aside from the very first attacked player being most likely to be the final bullied victim, the initial *attacker* also bears a substantial risk of ultimately being bullied. First attacks pose a puzzle to some extent, because our experimental data imply that

[2] Experimental studies can be especially helpful in uncovering the origins of such social dynamics by isolating the targeted factors for study, as well as eliminating potential confounding factors such as inter-personal histories, which can be prevalent in field studies. Thus, while both types of studies are crucial for our deeper understanding of conflicts, laboratory experiments can contribute distinctively in identifying the origins of social conflicts.

[3] See for example, Bloch (2012) and Konrad (2014).

being the first attacker does not pay off empirically. Further analysis shows that the choice of being first attackers is also highly path dependent on whether the same individual was the initiator or the first victim in the previous round. Thus, the initiation of a first attack is likely to be a reactive choice to avoid being bullied again and may not be driven by considerations about underlying costs and benefits of initiating an attack.

Our experiments, conducted in a controlled laboratory environment, investigate alliance and conflict formation in a four-player network game featuring homogeneous players, no discernible leaders, and equal initial resource allocations. These experiments demonstrate that participants frequently and efficiently coordinate group attacks on a single, arbitrary player to capture and distribute their resources. In real-life situations, individuals are often heterogeneous. For example, in the case of school bullying, bullies and victims usually form distinct peer groups with shared roles and behaviors strongly associated with their peers (Salmivalli et al., 1997). As such, observing (and potentially recurring) bullying in these heterogeneous settings becomes even more plausible. Our theoretical and empirical findings suggest that a higher cost of attacks leads to a greater likelihood of peace, presenting a possible intervention against bullying. However, our results also warn that peace is more likely to deteriorate into conflict than to be reestablished once conflict arises.

Further insights emerge when analyzing the dynamic results, with the understanding that our specific observations may be influenced by the chosen laboratory setting. We find that alliance formation generally precedes the coordination on a common rival, and the process of targeting victims is highly path-dependent. These observations align with empirical observations in school bullying. In particular, studies indicate that bullies are often supported by peers who either actively participate or passively reinforce the bullying behavior through observation, while peers are less inclined to defend victims (O'Connell et al., 1999). We also find that former victims tend to become aggressors, which aligns with the notion that bullied victims may seek retribution by bullying others in return (Wolke and Lereya, 2015). These insights enhance our understanding of aggressive behaviors and inform potential strategies for mitigating real-world conflicts such as bullying.

2. Related literature

Our study primarily contributes to the experimental literature about network formation games. To the best of our knowledge, our study is the first to test the predictions of a signed network formation game in the laboratory, and is thus also the first to examine how the decision dynamics of positive and negative links in the network lead to equilibrium outcomes.[4] By contrast, most previous studies focused on settings in which players can essentially link with each other as "friends" but not as "rivals" in non-signed networks (Kosfeld, 2004; Callander and Plott, 2005; Berninghaus et al., 2006; Berninghaus et al., 2007; Burger and Buskens, 2009; Goeree et al., 2009; Falk and Kosfeld, 2012; Rong and Houser, 2015; Goyal et al., 2017; van Leeuwen et al., 2020). The underlying incentives to link typically resemble those in (local) public goods, following the theoretical work by Bala and Goyal (2000) and Galeotti and Goyal (2010).

Other previous studies are more closely connected to ours in the sense that they also test equilibrium selection concepts. For example, Kirchsteiger et al. (2016) test myopic and farsighted notions of pairwise stability in a network formation game in the laboratory. They find that the selection among multiple pairwise stable networks crucially depends on the level of farsightedness. Teteryatnikova and Tremewan (2020) also examine the role of farsightedness in a non-signed network formation game. An interesting feature of their game is that a stable network outcome is either a complete network which leads to an egalitarian outcome, or a one-link network in which the payoff distribution resembles that of the bullying outcome in our setting. However, the similarity in the payoff distribution is superficial. While the payoff distributions in their game are solely motivated to allow for a better separation of different concepts of stability, in our game the payoff distributions arise naturally from a signed network formation model which is explicitly about alliance formation and conflict. These interpretations about dynamics are also more natural under our setting which explicitly models different types of links (friendly versus rival) beyond their eventual consequences for the payoff distribution.[5]

In terms of the substantive research theme, our study is related to both the theoretical and experimental literature on alliance formation and conflict.[6] Our work is built directly on Hiller (2017) which applies a novel network approach to endogenous alliance

[4] In contrast to our study, in which network structure rises endogenously, other strands of literature study conflict within a network structure of *exogenous* nature. Specifically, one line of studies examines conflict on exogenous networks (Franke and Öztürk, 2015; König et al., 2017; Cortes-Corrales and Gorny, 2018; Xu et al., 2022). For example, Franke and Öztürk (2015) consider several classes of networks in which rivals invest effort to attack their neighbors, and study equilibrium properties for each class. The second line of research focuses on optimal network design when faced with an external threat (see Goyal et al. (2016) for a review). In a typical setting, a defense network designer chooses a network and an allocation of defensive resources, and an adversary then allocates offensive resources on nodes with the goal of minimizing the value of the network.

[5] Purely in terms of the payoff distribution, the peaceful and bullying outcomes in our game also resemble those in three-person bargaining games with a majority rule, where the stable outcome is either equal-split or a minimum winning coalition (ganging up on an excluded member). Tremewan and Vanberg (2016) implement a continuous-time bargaining game in the laboratory, and examine the transitional dynamics between different stable outcomes, which are generally consistent with myopic payoff maximization.

[6] More broadly, alliance formation is a type of endogenous group formation which has also been studied in settings such as team production or public goods provision (see Guido et al. (2019) for a review). Our study is also related to a social phenomenon known as the "common enemy effect," in which group members cooperate more with the presence of a common enemy (see De Jaegher (2021) for a review). Haller and Hoyer (2019) model a network formation game which predicts how linkage costs affect the impact of an external threat on group cooperation. While this strand of literature typically assumes that a common enemy is externally imposed, our setting can be viewed as a case in which a common enemy is endogenously determined.

formation, establishes and characterizes equilibrium results.[7] In contrast to our network approach, most of the previous literature has investigated alliance formation using non-network games. A class of theoretical work studies the stability and structure of coalition formation in contests (Ray and Vohra, 1999; Garfinkel, 2004; Bloch et al., 2006; Sánchez-Pagés, 2007; Ray, 2007; Acemoglu et al., 2008). Another strand of literature models the dynamics of alliance formation with the possibility of intergroup conflict and intra-alliance conflict (see Bloch (2012) and Konrad (2014) for reviews). Several papers have also put related models into experimental tests (Ke et al., 2013, 2015; Herbst et al., 2015).

A network study addressing fundamentally similar motivating questions of interest as ours is Jackson and Nei (2015), which models and empirically tests predictions on stable networks using historical field data on international trade and wars. However, by using historical field data, their study differs from ours in its objectives and methodology. While applying network models to field data provides valuable insights with regard to the real world, identifying the sources of different network structures is challenging due to data limitations and historical path dependency. While not a replacement for estimation using field data, an experimental approach can more easily pinpoint particular sources of conflict and peace.

Two non-network experimental studies on alliance formation and conflict are most closely related to the research questions proposed in our work. The first is Smith et al. (2012). In their experiment, players can form or break up alliances; they can decide how to use their endowment for production, defense and offence; furthermore, alliance members can determine jointly whether to pool offensive capacities, choose independently whom to attack, make transfers of endowment, and chat with each other. The design of rich interactions in their study is a deliberate effort to explore whether and how players cooperate and resolve conflict in an anarchic situation. Our study shares a similar spirit to their work in that we allow endogenous alliance formation: players are free to decide when and whom to interact with. However, by limiting the strategy to forming relationships and thus keeping the decision domain relatively simple, we remove other possible factors from consideration in explaining conflicts, such as trade and diplomacy. Our setup and findings share with their study in the insight that group formation can lead to or worsen conflicts.

The second closely related experimental study on alliance formation is Abbink and Doğan (2019). In their experiment, players can freely and simultaneously nominate one of the other three players as a victim. If they successfully coordinate on a common victim, they mob the victim's payoff. The game is repeated for 20 periods with the same players, which allows them to study dynamics in coordination over periods. Their study is mainly concerned about how different features such as payoff asymmetry, color differences and immunity to mobbing affect focality and coordination success. Unlike in their design, we allow players to form alliances and engage in conflict simultaneously in a continuous-time decision environment. Allowing these two processes to interact during the game offers us further insights into how alliances and conflict arise, rather than focusing only on how conflict (mobbing) occurs. Despite clear differences in our experimental design compared to theirs, some of the equilibrium behavior in our network game is similar to those found in their study: In the Bully equilibrium, three players coordinate on a common rival and grab the rival's payoff. Our study can arguably be viewed as providing a network-based foundation of victim selection in more subtle situations when direct nomination is not conventional or feasible. For example, one of the patterns we find is that alliances generally precede bullying, which is not a feature detectable in their design.

Finally, we recognize the extensive literature on signed networks and structural balance in fields such as sociology, international relations, and applied physics (Heider, 1958; Harary et al., 1965; Snyder, 1997; Easley and Kleinberg, 2010). Researchers have developed various dynamic models, though not micro-founded, to study how networks evolve towards structural balance. One common dynamic (Antal et al., 2006) involves: (1) select a random triad; no action is taken if it is balanced; (2) in an imbalanced triad with one rival link, the rival link may become friendly with probability p, and the friendly link may turn rival with probability $1 - p$; and (3) if an imbalanced triad has three rival links, alter one rival link to a friendly link to restore balance. Researchers then conduct simulations, referred to as "experiments," to examine the conditions and rate at which a balanced state is achieved. In our context, groups can attain either the Bully or Peace equilibrium, both of which represent structurally balanced states. This suggests that forces present in non-game-theoretical dynamics may also be applicable here. However, our continuous-time experimental setup is not conducive for direct comparisons with theoretical dynamics due to the numerous interaction possibilities among human participants. Instead, discrete-time experimental approaches may be more appropriate for such comparisons.

3. Theoretical framework

3.1. Model setup

We consider a network game with four ex-ante homogeneous agents, based on the setup of Hiller (2017). The key difference between our model and Hiller (2017) is that we allow for the possibility of neutral relationship between any two agents, whereas in Hiller (2017), each pair of agents must have either a positive or a negative relationship. As a practical matter in our experiment, two players with a neutral link between them is a possibility given the action space in our design, which calls for a modification of the original setup.

An agent's strategy is defined as a row vector $\boldsymbol{g}_i = \left(g_{i,1}, g_{i,2}, g_{i,3}, g_{i,4}\right)$, of relationships with each agent in the group, where $g_{i,j} \in \{1, 0, -1\}$ for each $j \in N/i$, and $g_{i,i} = 0$. Agent i extends a positive (friendly) link to j if $g_{i,j} = 1$, a negative (rival) link if $g_{i,j} = -1$, and no

[7] He finds that every Nash equilibrium obeys the property of structure balance for n-player and for a general class of payoff functions. Furthermore, strong Nash equilibrium selects the equilibrium in which a single player is in a rival relationship with everyone else.

link if $g_{i,j} = 0$. The resulting network of relationships is denoted by $\boldsymbol{g} = (\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3, \boldsymbol{g}_4)$. We define the undirected network $\bar{\boldsymbol{g}}$ in the following way: $\bar{g}_{i,j} = 1$ if $g_{i,j} = g_{j,i} = 1$; $\bar{g}_{i,j} = -1$ if $\min\{g_{i,j}, g_{j,i}\} = -1$; $\bar{g}_{i,j} = 0$ otherwise.[8] Thus, a pairwise friendship or alliance is successfully formed only if both agents agree, while a rivalry is formed if at least one agent picks a fight.

Define the following set, $N_i^+(\boldsymbol{g}) = \left\{ j \in N \;\middle|\; \bar{g}_{i,j} = 1 \right\}$, as the set of agents that agent i establishes a friendship with by reciprocating a positive link. The number of friends agent i has can then be denoted by $n_i(\boldsymbol{g}) = \left|N_i^+(\boldsymbol{g})\right|$. Similarly, we define $N_i^-(\boldsymbol{g}) = \left\{ j \in N \;\middle|\; \bar{g}_{i,j} = -1 \right\}$ as the set of agents with whom agent i forms rival relationships. A subset of $N_i^-(\boldsymbol{g})$, defined as $N_i^{e-}(\boldsymbol{g}) = \left\{ j \in N \;\middle|\; g_{i,j} = -1 \right\}$, is the set of agents to whom agent i *extends* a negative link, and we denote by $e_i(\boldsymbol{g}) = \left|N_i^{e-}(\boldsymbol{g})\right|$ the number of negative links agent i extends.

Extending a positive link is assumed to be costless. Extending a negative link, i.e., attacking, however, is assumed to incur a cost of $c > 0$. An agent draws strength from his friends whenever a conflict arises. The potential fighting strength of agent i is thus given by the number of effective friendships that agent i has, n_i. The payoff of agent i from extending an attack to agent j are determined by the payoff function $h_i(n_i, n_j)$, and the payoff of agent i from receiving an attack from agent j is determined by the payoff function $h_i(n_j, n_i)$. We assume what the winning side receives is exactly equal to what the other side loses, except that the attacking agent must incur a cost of c. Therefore, this means that $h(n_i, n_j) = -h(n_j, n_i)$. For simplicity and consistency with our experimental implementation, we additionally assume that the payoff function is linear in the difference between the two sides' strengths derived from their friendships, i.e., $h_i(n_i, n_j) = k(n_i - n_j),\ k > 0$. k thus represents the gross payoff benefit to the attacker from having a marginal friendship advantage compared to the attack target.

Agents do not receive any direct payoffs from a friendship. Thus, the only directly payoff-relevant purpose of a friendship or alliance is to increase agents' strengths in a fight, which in turn increases the payoffs from a rival relationship.

An agent's utility in network $\boldsymbol{g}$ is defined as the total payoff from being involved in negative relationships minus the total costs of initiating negative relationships, given by

$$u_i(\boldsymbol{g}) = \sum_{j \in N_i^-(\boldsymbol{g})} k(n_i - n_j) - e_i(\boldsymbol{g})c.$$

3.2. Equilibrium analysis

A Nash equilibrium of the four-agent network formation game satisfies the following definition:

Definition (Equilibrium): A strategy profile $\boldsymbol{g}^* = (\boldsymbol{g}_1^*, \boldsymbol{g}_2^*, \boldsymbol{g}_3^*, \boldsymbol{g}_4^*)$ constitutes a Nash equilibrium if for any agent i, for any $\boldsymbol{g}_i' \neq \boldsymbol{g}_i^*$, $u_i(\boldsymbol{g}_i', \boldsymbol{g}_{-i}^*) \leq u_i(\boldsymbol{g}^*)$, where $\boldsymbol{g}_{-i}^*$ represents the equilibrium strategy profile of all agents other than i.

As mentioned earlier, in our study of alliance and conflict dynamics, it is natural to allow for the possibility that agents have no particular relationship in our experiment – in other words, agents need not necessarily be either friends or enemies. A theoretical consequence of this more flexible setup, however, is that the set of Nash equilibria is larger compared to that in Hiller (2017). Therefore, in our analysis we also consider three equilibrium refinement criteria, which also serve as our evaluations of predicted equilibrium robustness: pairwise stability, no pairwise profitable deviation condition, and no 3-person profitable deviation condition, which are formally defined as follows:

Condition 1 (Pairwise Stability): $\bar{\boldsymbol{g}}^*$ is pairwise stable if (i) for any link $ij \in \bar{\boldsymbol{g}}^*$ such that $\bar{g}_{i,j} = 1, u_i(\boldsymbol{g}^*) \geq u_i(\boldsymbol{g}^* - \boldsymbol{ij})$ and $u_j(\boldsymbol{g}^*) \geq u_j(\boldsymbol{g}^* - \boldsymbol{ij})$; (ii) for any link $ij \notin \bar{\boldsymbol{g}}^*$ such that $\bar{g}_{i,j} = 0$, if $u_i(\boldsymbol{g}^*) < u_i(\boldsymbol{g}^* + \boldsymbol{ij})$ then $u_j(\boldsymbol{g}^*) > u_j(\boldsymbol{g}^* + \boldsymbol{ij})$.[9]

Condition 2 (No Pairwise Profitable Deviation): Equilibrium $\boldsymbol{g}^*$ is robust to pairwise profitable deviation if for any agent pair (i, j), for any $\boldsymbol{g}_i' \neq \boldsymbol{g}_i^*, \boldsymbol{g}_j' \neq \boldsymbol{g}_j^*$, if $u_i(\boldsymbol{g}^*) < u_i\left(\boldsymbol{g}_i', \boldsymbol{g}_j', \boldsymbol{g}_{-(i,j)}^*\right)$ then $u_j(\boldsymbol{g}^*) > u_j\left(\boldsymbol{g}_i', \boldsymbol{g}_j', \boldsymbol{g}_{-(i,j)}^*\right)$, where $\boldsymbol{g}_{-(i,j)}^*$ represents all agents other than i and j 's equilibrium strategy profile..

Condition 3 (No 3-Person Profitable Deviation): Equilibrium $\boldsymbol{g}^*$ is robust to 3-person profitable deviation if for any three agents (i, j, k), for any $\boldsymbol{g}_i' \neq \boldsymbol{g}_i^*, \boldsymbol{g}_j' \neq \boldsymbol{g}_j^*, \boldsymbol{g}_k' \neq \boldsymbol{g}_k^*$, if there exists an agent $s \in \{i, j, k\}$ such that $u_s(\boldsymbol{g}^*) < u_s\left(\boldsymbol{g}_i', \boldsymbol{g}_j', \boldsymbol{g}_k', \boldsymbol{g}_{-(i,j,k)}^*\right)$ then there must exist another agent $t \in \{i, j, k\} (t \neq s)$ such that $u_t(\boldsymbol{g}^*) > u_t\left(\boldsymbol{g}_i', \boldsymbol{g}_j', \boldsymbol{g}_k', \boldsymbol{g}_{-(i,j,k)}^*\right)$, where $\boldsymbol{g}_{-(i,j,k)}^*$ represents agent other than i, j and k 's equilibrium strategy profile.

[8] $\min\{g_{i,j}, g_{j,i}\} = -1$ means either $g_{i,j} = -1$ or $g_{j,i} = -1$, or $g_{i,j} = g_{j,i} = -1$.

[9] The network formation literature considers various concepts about pairwise stability, including myopic and farsighted versions. Since we mainly focus on equilibrium refinement rather than stability refinement, we only adopt the most standard concept about pairwise stability (Jackson and Wolinsky, 1996).

Note that pairwise stability has been commonly used in the network literature as a condition for undirected network structures, by considering the stability of the network with respect to bilateral links, rather than entire strategies. In addition, since we consider a continuous-time network formation process, it is also possible that a group of agents coordinates to form coalitions through their dynamic interactions. Thus, we also consider two stronger conditions on the robustness of an equilibrium with regard to subsets of individual players' entire network strategies: no pairwise profitable deviation, and no 3-person profitable deviation. Note that an equilibrium that is robust to 3-person profitable deviation is also by definition robust to pairwise (2-person) profitable deviation, and an equilibrium that is robust to pairwise profitable deviation is also pairwise stable.[10] The conditions are thus conceptually nested, which facilitates comparison. An equilibrium $\boldsymbol{g}^*$ is *more robust* than another equilibrium $\boldsymbol{g}^{*'}$ if the set of conditions that $\boldsymbol{g}^*$ is robust to is a subset of conditions that $\boldsymbol{g}^*$ is robust to.

Recalling that k represents the marginal benefit of attacking with one unit of additional strength (friendship) advantage over the target, and c represents the marginal cost of attacking, $c = k$ is a natural threshold at which the set of equilibrium outcomes changes. Furthermore, in our 4-person setting, the most that any player can gain from attack arises from a situation in which they have a 2-unit (friendship) strength advantage over the victim, hence for costs beyond $2k$, attacking will no longer be worthwhile for any player, so it suffices to examine the cost ranges $c < k$ and $k < c < 2k$. Fig. 1 provides a detailed depiction of all Nash equilibria, and Table 1 summarizes equilibria that survive the equilibrium refinements for attack cost levels $c < k$ and $k < c < 2k$, respectively.[11]

In general, attacking another player may be profitable as long as one has successfully formed an alliance with at least one other player. However, a player must balance the potential benefits of attacking another player against the benefit of instead forming an alliance with them, which can then serve as a mutual aid device in attacking another player. Among the set of equilibria, two emerge as the most robust and intuitive. In the *Peace* equilibrium, all possible links in the network are positive. In the *Bully* equilibrium, three agents reciprocate friendly links with each other and each of the three agents also extends a rival link to the fourth agent. These two equilibria are illustrated in Fig. 2.[12]

In the current setup, the Peace and Bully equilibria coexist. For $c < k$, the Peace equilibrium is pairwise stable, neither robust to pairwise profitable deviation, nor to 3-person profitable deviation; the Bully equilibrium is robust to all refinements. For $k < c < 2k$, the Peace equilibrium is pairwise stable, robust to pairwise profitable deviation, but not robust to 3-person profitable deviation. The Bully equilibrium is again robust to all refinements. This means that as the cost increases from below k to above k, the Peace equilibrium becomes more robust in the sense that more robustness conditions are satisfied. In contrast, the Bully equilibrium satisfies all three robustness conditions for both low-cost levels ($c < k$) and high-cost level ($k < c < 2k$).

These theoretical results have two empirical implications. First, since the Peace equilibrium is more robust for $k < c < 2k$ than for $c < k$, we expect the Peace network to result more frequently for $k < c < 2k$. Second, since the Bully equilibrium is more robust than the Peace equilibrium, for both $c < k$ and $k < c < 2k$, we expect the Bully network to result more frequently than the Peace network for both cost ranges. Furthermore, the Bully network, once instantaneously formed, is expected to be more stable and thus less likely to disintegrate into other network formations than the Peace network. We will formally state our experimental hypotheses resulting from these implications in the next section.

Finally, we briefly mention other equilibria here. There exist "quasi-peace" equilibria in which no negative link exists in the network while some bilateral pairs have no particular relationship (equilibria 3–8 in Fig. 1). However, none of the quasi-peace equilibria are robust to 3-person profitable deviation. Furthermore, for $c < k$, there exists a "quasi-bully" equilibrium (equilibrium 9) and two other equilibria (equilibria 11 and 12) that are neither classifiable as quasi-peace nor quasi-bully. These three equilibria are relatively non-robust, not even being robust to pairwise stability. Finally, for $k < c < 2k$, there also exists a "quasi-bully" equilibrium (equilibrium 10). While this equilibrium is robust to pairwise stability, it is not robust to pairwise profitable deviation. In summary, all these "other" equilibria are less robust than the Peace and Bully equilibria and therefore less likely to be observed as outcomes in the experiment.

4. Experimental design and implementation

4.1. Basic setup

We implement a continuous-time experimental design in which participants can freely make links to each other, and both the network structure and momentary payoff implications are updated in real-time (Goyal et al., 2017; Rezaei et al., 2024).[13] Within each round, participants can freely adjust their linking decisions for a period lasting between 75 and 105 s and ending at an unknown

[10] Our robustness criteria share some differences and similarities with the concept of Strong Nash equilibrium in the literature, in the sense that Strong Nash equilibrium is a much stronger criterion that requires not only Conditions 1, 2 and 3 to hold, but also a no 4-person profitable deviation condition, in our setup. We adopt Conditions 1, 2, and 3 to successively measure different degrees of robustness for equilibrium refinement, as described in the subsequent Propositions. Furthermore, the No 3-Person Profitable Deviation condition succeeds in refining the set of equilibria down to a single equilibrium, therefore further refinement criteria are unnecessary.

[11] In Online Appendix A, we provide the complete equilibrium characterization. Also note that the equilibrium network type (11) is not structurally balanced. This is different from the equilibrium characterization in Hiller (2017) who proves that all equilibria must be balanced.

[12] Both Peace and Bully equilibria satisfy the property of structural balance (Cartwright and Harary, 1956).

[13] Previous experimental studies on networks show that individuals often find it difficult to coordinate in network games due to the complex interaction (Rosenkranz and Weitzel, 2012; Falk and Kosfeld, 2012). Our continuous-time design can help to facilitate coordination.

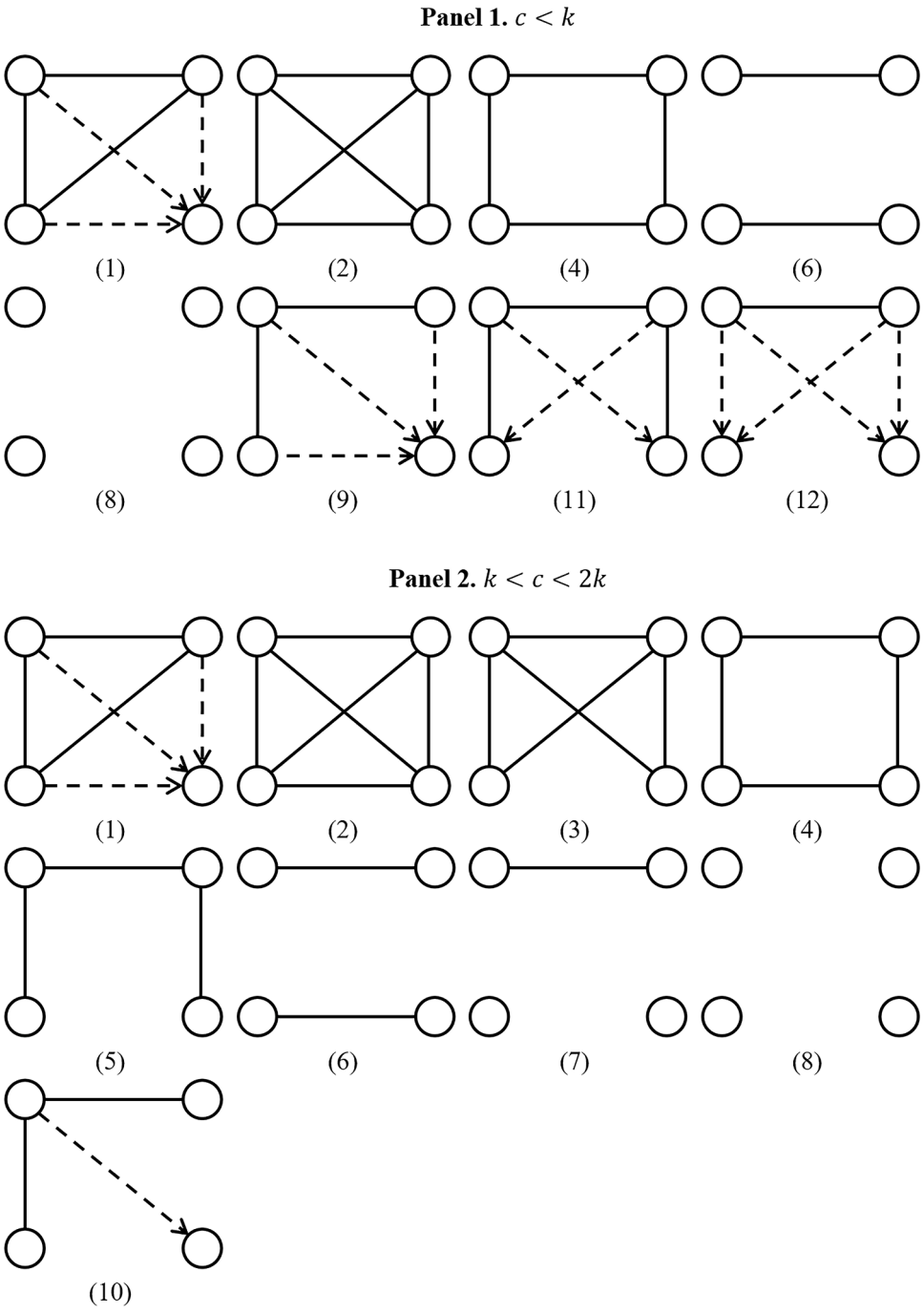

Fig. 1. All Nash equilibria by cost ranges.

Table 1
Equilibrium refinements.

Equilibrium refinement	Outcome ($c < k$)	Outcome ($k < c < 2k$)
Nash (single profitable deviation)	1,2,4,6,8,9,11,12	1,2,3,4,5,6,7,8,10
Pairwise stability	1,2,4,6,8	1,2,3,4,5,6,7,8,10
No Pairwise profitable deviation	1	1,2,6,7,8
No 3-person profitable deviation	1	1

Fig. 2. Peace and Bully equilibria.
Notes: Mutual friendly links are represented by solid lines and rival links are dashed lines with arrows indicating the direction of attack.

moment, and this random termination feature is known by all participants. The random termination design is used to minimize potential end-game effects so that participants are less likely to change decisions at the last few seconds.

Each subject's decisions are continually updated on screens of all other group members. Full information about momentary hypothetical payoffs of all group members based on the current set of links is also continually updated on the screen, also indicated by the size of the circle representing each player. The continuous-time decision environment is designed to facilitate learning and to observe how networks converge to their final states without income effect considerations. Thus, the payoff consequences of players' decisions depend only on the network structure of the very *end* of a round, a feature which is made clear to participants.[14]

Fig. 3 depicts a screenshot of the decision screen. The green circle represents a participant's own position while the black circles represent the three other participants in their group. Each participant can extend a friendly or rival link to another participant using the computer mouse. One extends a friendly link by left-clicking on one of the black circles. A blue link with an arrow pointing to that participant will then appear. Left-clicking again on that participant, the blue link will be removed. Alternatively, one may extend a rival link by right-clicking on a black circle. A red link with an arrow pointing to that participant will appear. Right-clicking again on that participant, the red link will be removed. In all of our treatments, extending blue links is free and each red link costs some points (while a retracted red link costs nothing).

As in our model, forming a pairwise alliance (named "partnership" in the instructions) requires mutual consent. Therefore, only when both sides extend a friendly link to each other does their alliance become *effective*. An effective friendship is depicted as a thickened blue link with double-headed arrows on the decision screen. On the other hand, establishing a rival relationship (referred to as a "competitive relationship" in the instructions) only requires a unilateral decision. We say that a rival relationship is effective as long as at least one side initiates a red link. When both sides extend a red link to each other, a thickened red link with double-headed arrows will appear. Note that one cannot initiate both blue and red links to the same other participant at the same time.

To facilitate coordination and to assist with the calculation of payoffs, we also add other helpful information to participants' screens. The number on top of each circle indicates that participant's current points. A larger circle indicates that participant has more points. The bottom number in each circle indicates the number of effective pairwise alliances that a participant currently has.[15]

Each subject's hypothetical momentary payoff is determined by

$$\pi_i = 70 + 10 \sum_{j \in N_i^-(\boldsymbol{g})} (n_i - n_j) - e_i(\boldsymbol{g})^* c$$

such that each player's initial endowment is 70 and the parameter k from the model is 10. The cost parameter c varies by treatment, and is deducted for each rival link extended by player i. Note that their payoff will only be materialized at the end of the round.

4.2. Treatments

Our experiment includes both within-subject and between-subject versions. We begin with a within-subject design comprising five treatments, which correspond to the varying costs of forming negative links, as outlined in our theoretical hypotheses. In the experiment, the payoff function parameter k is set at 10, while the parameter c assumes one of the following five values: $\{3, 5, 7, 9, 11\}$. Consequently, in four of the five cost values, $c < k = 10$, whereas in one case, $c = 11 > k = 10$. Subjects receive meticulous explanations of the rules and payoff calculations through instructions supplemented with examples. These experimental instructions

[14] In our continuous-time framework, we can consider at least three alternative approaches for incentivizing participants: 1) basing the payoff consequences of players' decisions on a randomly chosen moment during the round rather than based on the end-of-the-round moment; 2) tying the payment to the length of time that a stable structure is maintained (e.g., implement payment corresponding to a specific configuration remains unchanged for at least 15 seconds); 3) associating each decision with immediate benefits or non-refundable costs. Our analysis indicates that participants typically established their final network configurations in a fairly short amount of time (i.e., 10 seconds). Thus, the first two alternative payment scheme approaches would likely have little impact on our results. Using a setup with instantaneous payoffs could hinder coordination on specific equilibria, since it may lead to more cautious participant behavior when extending costly negative links. A rigorous analysis of players' actions in such a setting would demand a considerably different theoretical model than the one we present here.

[15] We choose not to add information about rival relationships because a participant can be either a receiver or an initiator of a red link, which entails different payoff consequences. In the design, we need to balance the potential benefits for coordination of presenting more information and the potential costs of confusion due to too much information.

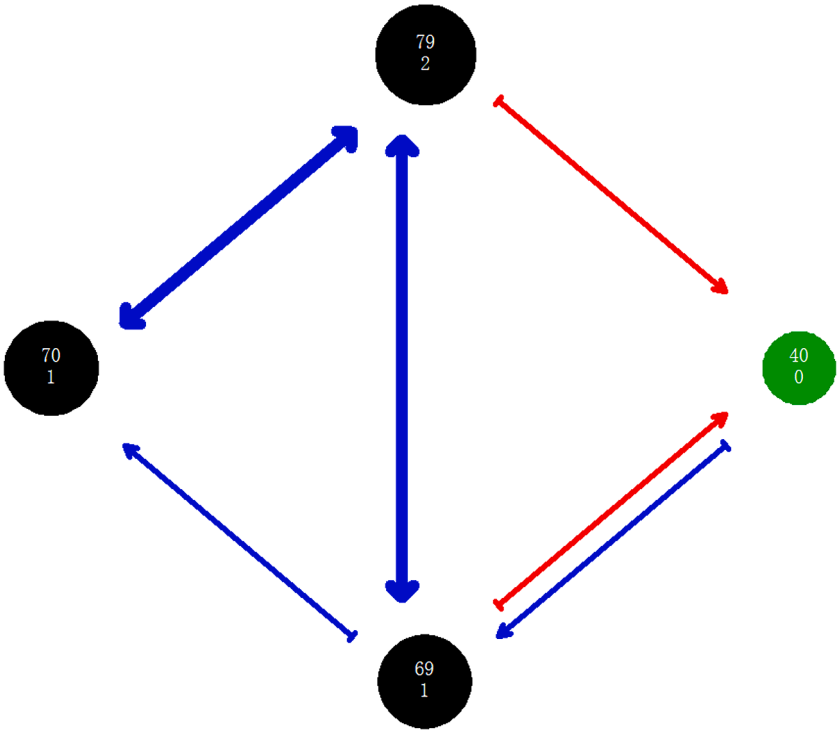

Fig. 3. Screenshot Example of Decision Screen, Cost = 11.

can be found in Online Appendix B.

Each session consists of 20 rounds of the network formation game. At the beginning of each round, participants are randomly matched into 4-person groups and assigned a position in the network (upper, lower, left, or right) displayed on their screens. We divide the session into five blocks, with each block consisting of four rounds. Within each block, the cost to extend a rival link remains constant for all group members. However, between blocks, the cost changes, taking one of five different values without repetition: 3, 5, 7, 9, or 11. We conduct 10 experimental sessions using five distinct ordering sequences of costs, generated by a Latin square,[16] to ensure each participant experiences all cost levels while minimizing order effects. Participants receive a 70-point endowment in each round, which will increase or decrease based on their link formation choices and the game's outcome. At the end of the experiment, one block is randomly selected for payment, with participants receiving the accumulated earnings from that block's four rounds.

Our choice of specific cost levels in the main experiment is based on an earlier pilot study with two sessions. These sessions shared an almost identical setup as the main experiment, maintaining a constant cost for rival link extension. The only difference is that group matching and participants' spatial positions were randomized across blocks. The two sessions had costs of attack of 3 and 11. Results indicated that almost all groups with a cost of 11 reached Peace in the end. Consequently, we inferred that costs above 11 would be of limited interest in terms of the relative occurrence of Bully and Peace. By contrast, Bully is reached in almost all groups in the session with a cost of 3. Therefore, we designed the main experiment with attack cost levels of between 3 and 11, anticipating that this cost range would be most relevant for observing treatment effects of cost on network outcomes and dynamics.

The rationale behind employing a within-subject design is primarily based on its suitability for testing our hypotheses (which will be discussed in greater detail in Section 4.4). Crucially, one hypothesis (i.e., Hypothesis 3) emphasizes investigating the transition between Bully and Peace in relation to cost, necessitating that participants have adequate exposure to both situations, which is achievable through a within-subject design. This decision is further supported by our between-subject pilot study results, which demonstrated an overstated cost effect when either cost of 11 or 3 was employed, as participants may not fully experience the consequences of both Bully and Peace.

The purpose of using a Latin square design in our within-subject design is to minimize potential order effects. Aware of the possibility that session effects may still confound our treatment comparisons, we conducted a between-subject version of the experiment as a robustness measure. This version involved three cost levels (3, 7, and 11) and four sessions per level, covering the same range of costs as the within-subject design. We ensured consistency between the within- and between-subject designs, except that the cost did not vary from block to block in the between-subject version, allowing us to identify any potential differences in the robustness of our

[16] The specific ordering is (3, 5, 11, 7, 9) in sessions 1 and 6, (5, 7, 3, 9, 11) in sessions 2 and 7, (7, 9, 5, 11, 3) in sessions 3 and 8, (9, 11, 7, 3, 5) in sessions 4 and 9, and (11, 3, 9, 5, 7) in sessions 5 and 10.

findings across experimental designs.

4.3. Implementation procedure

The experiments were conducted at the Nanjing Audit University Economics Experimental Lab with a total of 356 university students, using the software z-Tree (Fischbacher, 2007).[17] 164 students participated in the within-subject experiment, while 192 students participated in the between-subject experiment. Each subject could only participate in one treatment of either the within-subject or the between-subject experiment, so that no subject had prior experience with our specific experiment. Either 16 or 20 students participated in each session of 20 rounds of the network formation game, with partners randomly re-matched in each round.

Participants were randomly seated at a partitioned computer terminal upon arrival. The experimental instructions were provided to participants in written form and were also read aloud by the experimenter at the start of each session. Participants then completed a comprehension quiz before proceeding, which was designed to ensure that every participant understood the instructions. An average of 25 min per session was dedicated to ensuring comprehension. At the end of the experiment, participants completed a questionnaire eliciting their social preferences and numerical sophistication (implemented for half of all sessions of the within-subject experiment and for all sessions of the between-subject experiment) and finally a short survey concerning their demographics, their attitudes toward risk and competitiveness and strategies they used in the game.

For every 5 points earned in the subject's randomly selected block, participants earned 1 RMB. At the end of the session, participants were paid privately in cash and instructed to leave the laboratory one at a time. A typical session lasted about 75 min with average earnings of 69.7 RMB, including a show-up fee of 15 RMB.[18]

4.4. Hypotheses

We derive two empirically testable hypotheses regarding *Peace* and *Bully* outcomes from our theoretical results summarized in Table 1, in the context of our experimental setup as follows.

Hypothesis 1. *The observed relative frequency of Bully to Peace networks is higher when $c < 10$ than when $c = 11$.*

Hypothesis 2. *The observed frequency of Bully networks is higher than that of Peace networks both when $c < 10$ and when $c = 11$.*

Hypothesis 3. *The likelihood of reversion from Bully to Peace networks is lower than the likelihood of reversion from Peace to Bully networks during the course of continuous-time play within a round.*

Table 1 and the discussion thereof directly imply all of the above Hypotheses. The intuition of Hypothesis 1 is that when $c < 10$, a person only needs to make one ally to be profitable in a fight against another person. However, he needs to make two allies to be profitable when $c = 11$. Both Hypotheses 2 and 3 follow from the fact that the Bully equilibrium is generally more robust than the Peace equilibrium. A "more robust" network based on the theoretical criteria discussed earlier is expected to occur more frequently in the final configurations of each round (Hypothesis 2) and, once a "more robust" network is formed, it is expected to be more dynamically stable within-round (Hypothesis 3). Therefore, if a group temporarily coordinates on the *Peace* network, it is still likely that some players might jointly find it more profitable to deviate to the *Bully* network. However, there is no similar incentive for them to revert to the *Peace* network once they settle upon the *Bully* network.

The dynamics of the continuous-time decision environment of the game are complex, and aside from Hypothesis 3, our existing theory does not provide specific guidance as to the precise dynamics of alliance formation. Hence, for the analysis of dynamic network formation, we adopt an explorative approach to examining how players coordinate and attack, especially how a common enemy in a *Bully* network emerges and how an alliance emerges and grows. In Online Appendix G, however, we will discuss a simple quasi-dynamic model to account for some of the observed dynamic patterns.

5. Experimental results

We divide our results section into two sections. In the first section, we present basic results on the final networks formed, and evaluate the results with respect to the Hypotheses of the previous section separately for the within-subject and between-subject experiments. Then, in the second section, we analyze the dynamics of how the final networks were reached for the within-subject experiment, while a similar set of analyses for the between-subject experiments is reported in Online Appendix F.

[17] The experimental procedure was reviewed and approved for ethics considerations by Survey and Behavioral Research Ethics Committee, The Chinese University of Hong Kong.

[18] The average per-round earnings of 68.4 points are close to 70 points due to two factors: a substantial number of peaceful outcomes and a preference for establishing negative links at low costs. Taking into account the 15 RMB show-up fee and a conversion rate of 5 points to 1 RMB, participants' average earnings amount to 69.7 RMB. Based on exchange rates during the experiment schedule, the average earnings per subject was equivalent to around $10 USD, which is well-within the standard experimental payment range in mainland China.

5.1. Final network formation

5.1.1. Within-subject experiment

Fig. 4 shows the frequency of the main final network types under each cost level. In total, we have 820 network observations and find two major categories of final networks corresponding to the Peace and Bully equilibrium networks depicted in Fig. 1. Peace networks represent 38.9% (319/820) of all cases. Bully networks represent 49.8% (408/820) of all cases. All other final networks account for the remaining 11.3% of all cases. Table 2 shows a more detailed categorization of final network types, including the frequency data for other equilibrium networks and non-equilibrium networks. In general, equilibrium network types other than the Peace and Bully equilibria were rarely observed in our data.

In terms of the speed with which the final network was reached, on average it took groups 37.6 s (s.d.: 30.4, min: 2.5, max: 105.8; median: 29.2) to settle upon a final network. By each final network category, it took an average of 22.9 s to settle on Peace networks, 43.2 s for Bully networks, and 68.1 s for all other networks. Given that the average duration of each round is approximately 90 s, most groups were able to stabilize on a specific network about halfway through the round.

Turning to test of our hypotheses, our data support Hypothesis 1, which predicts that the relative frequency of Bully to Peace networks is higher for costs below 10 than the cost of 11.[19] Table 3 reports estimates from a Probit model that regresses a binary dependent variable for whether the final network is Bully (=1) or Peace (=0) on cost level dummy variables and round-fixed effects. The estimates confirm that the relative frequency of Bully over Peace networks is significantly higher for every cost level below 10 than for the cost of 11.[20] While the theory only predicts difference in relative frequencies of the two main equilibria between costs above and below 10, the theory is technically silent about comparisons among cost levels below 10. One intuitive conjecture is that a higher cost should lead to more Peace relative to Bully. Fig. 4 appears consistent with this conjecture. However, as reported in the bottom part of Table 3, the pairwise comparisons for costs below 10 indicate that only the comparisons between the cost of 3 vs. 9 and 5 vs. 9 are statistically significant.

Our data also partially support Hypothesis 2 which predicts that the frequency of Bully is consistently higher than that of Peace at any cost level. It is clear, however, from Fig. 4 that this only appears to be true for the cost levels of 3, 5 and 7. The frequency of Bully and Peace is almost the same for the cost of 9 and the direction is reversed for the cost of 11. The Wilcoxon sign-ranked test (at the session level) shows that the frequency of Bully is significantly higher than that of Peace for the cost of 3 ($p = 0.047$), but that there is no significant difference for other cost levels. One interpretation is that the possible cognitive approximations, which lead to an increase in the ratio of Peace to Bully as attack costs increase (even those below 10), dominate the tendency of the more theoretically robust equilibrium (i.e., Bully) to occur more frequently.

Finally, Hypothesis 3 proposes that Bully networks are more stable than Peace ones in the sense that once a group is in a Peace network it is more likely to converge to a Bully situation, compared to the other way around. This prediction is well-reflected in the data. The data show that Peace networks arose in 408 groups at some point within a round; however, 19.1% (78/408) of these groups ended up in Bully networks as their final network structure. By contrast, Bully networks arose in 317 groups at some point during the round; but only 0.6% (2/317) of them reverted to Peace networks as their final network structure. This is consistent with our intuitive interpretation of Hypothesis 3 that the three allies in a Bully network can receive the highest possible payoff and, therefore, if they can successfully coordinate, they would like to deviate from a Peace network to a Bully one, but not the other way around.

A further piece of evidence is that in the regression of Table 3, the coefficients on the round-fixed effects (not displayed for space concerns) show that Bully networks are generally more likely as the number of rounds played increases, and in particular, significantly so among the last played rounds.[21] This suggests, as a further cross-round consequence of Hypothesis 3, that Bully outcomes are enhanced by learning and experience.

We summarize our findings about final networks as follows in Result 1:

Result 1. *i) Around 90% of the time, groups reached either Bully networks or Peace networks; ii) the relative frequency of Bully over Peace networks tended to decline as a function of the cost of attack; iii) Groups were more likely to transition from a Peace network to a Bully network than to transition from a Bully network to a Peace network.*

5.1.2. Between-subject experiment

In our within-subject experiment described in the previous section, we utilized the Latin square design to mitigate potential order effects of treatments. However, despite this, the ordering may still introduce confounding variables to treatment comparisons due to possible session-level effects. To address this concern, in this section we present results from a between-subject version of our

[19] We conduct a series of Wilcoxon signed-rank tests comparing the frequency of Bully under each cost level below 10 with that under the cost of 11 (at the session level). The frequency of Bully is significantly different in 3 vs. 11 ($p = 0.028$) and 5 vs. 11 ($p = 0.052$), but not in 7 vs. 11 ($p = 0.153$) and 9 vs. 11 ($p = 0.183$). We also conduct the same tests comparing the frequency of Peace. The results show that the frequency is significantly different in 3 vs. 11 ($p = 0.012$), 5 vs. 11 ($p = 0.025$) and 7 vs. 11 ($p = 0.041$), but not in 9 vs. 11 ($p = 0.103$).

[20] Recall that we used a Latin square design to sequence treatments across sessions, reducing potential order effects. To verify the robustness of our results, we separately plotted Bully and Peace frequencies, similar to Fig. 4, for each ordering. Figure C2 in Online Appendix C demonstrates that the pattern of higher costs leading to increased Peace occurrences is consistent in 4 out of 5 orderings; the exception occurs when the cost of 11 is introduced last. A potential explanation is that the early low-cost Bully experience causes spillover effects by predisposing participants to conflict even when costs increase later.

[21] Out of space considerations, the round-fixed effect estimates are omitted here, but are available upon request.

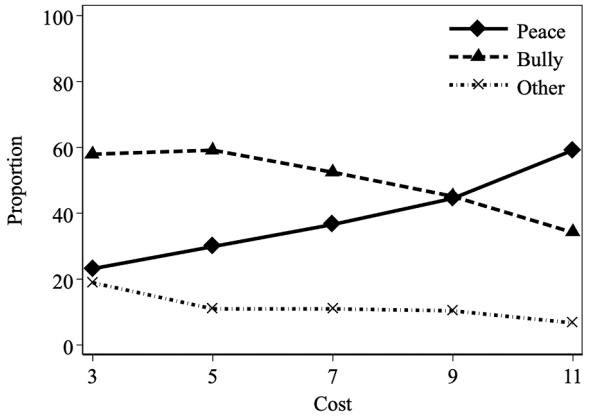

Fig. 4. Proportions of Final Network Types – Within-Subject Experiment.
Note: This figure shows the final network configuration of all 820 groups of all cost levels. "Other" includes all equilibrium networks other than the Peace and Bully equilibria, together with non-equilibrium networks.

Table 2
Frequency (in%) of Final Network Types – Within-Subject Experiment.

	Bully	Peace	Other equilibrium networks	Non-equilibrium networks
Cost = 3	57.9	23.8	0.6	17.7
Cost = 5	59.1	29.9	1.2	9.8
Cost = 7	52.4	37.2	0	10.4
Cost = 9	45.1	44.5	1.2	9.2
Cost = 11	34.1	59.1	3.0	3.8

Note: The sets of equilibrium networks are different for costs below 10 and the cost of 11 as shown in Fig. 1. In our data, equilibrium network types 5 to 8 (quasi-peace), 10 (quasi-bully) and 12 were never formed. Fig. C1 in Online Appendix C shows all non-equilibrium network types formed at least once. Most of these networks were rarely formed except for types 30 and 39 which were formed relatively more frequently. Table C1 shows an even finer categorization of final network types for each cost level.

Table 3
Probit Model: Treatment Effects on Final Networks – Within-Subject Experiment.

	Average marginal effects	Standard error
Cost = 3	0.337***	0.077
Cost = 5	0.291***	0.068
Cost = 7	0.220***	0.060
Cost = 9	0.131**	0.059
H0: 3 vs. 5	$p = 0.515$	
H0: 3 vs. 7	$p = 0.112$	
H0: 3 vs. 9	$p = 0.003$	
H0: 5 vs. 7	$p = 0.433$	
H0: 5 vs. 9	$p = 0.027$	
H0: 7 vs. 9	$p = 0.165$	
N	725	

Note: The dependent variable is whether the final network is Bully (=1) or Peace (=0). The cost of 11 serves as the benchmark. The regression also includes round fixed effects which show trends toward more Bully over time. Standard errors are clustered at the session level.
** $p < 0.05$.
*** $p < 0.01$.

experiment. In this design, we implemented three cost levels: 3, 7, and 11, representing low (below cost of 10), medium (below cost of 10) and high cost (above cost of 10), respectively. The objective of these between-subject treatments is to validate the robustness of our primary findings from the within-subject experiment when participants do not have experience across a range of different attack cost levels.

Fig. 5 depicts the frequency of final network types for each cost level, with a total of 320 network observations per level. Table 4 offers a more detailed classification of final network types, incorporating frequency data for other equilibrium networks and non-equilibrium networks. Consistent with our observations in the within-subject experiment, groups reached either Bully or Peace networks approximately 90% of the time. The relative frequency of Bully to Peace networks is higher for costs below 10 compared to the

cost of 11, supporting Hypothesis 1. As the cost of attack increases, the frequency of Bully networks decreases, while that of Peace networks increases, in a nearly linear fashion. Table 5 presents estimates derived from a Probit model, similar to that in Table 3. These estimates confirm that the relative frequency of Bully networks is significantly greater for all cost levels below 10 as compared to the cost of 11.

We note that there are some differences between the within-subject and between-subject experiments. For example, under the cost of 7 in the between-subject experiment, Bully and Peace networks occur with almost equal frequency, while the within-subject experiment shows that Bully networks are slightly more frequent than Peace networks (though not significantly so) at the cost of 7. Additionally, the frequency of Bully networks is somewhat higher under the cost of 3, and lower under the cost of 11 in the between-subject experiment, compared to the earlier described within-subject experiment. These differences could be explained by the lack of experience with a variety of cost levels in the between-subject experiment, leading participants to behave more extremely in response to nominal cost levels.

The between-subject experiment results also partly support Hypothesis 2, similarly to those of the within-subject experiment. Specifically, the frequency of Bully networks was significantly higher than that of Peace networks at a cost of 3 ($p = 0.068$, Wilcoxon signed-rank test at the session level). However, the frequency of Bully as compared to Peace was not significant at the cost of 7 ($p = 1.000$), while it was significantly reversed at the cost of 11 ($p = 0.068$).

Furthermore, our findings in the between-subject experiment are consistent with Hypothesis 3, which posits that Bully networks are more stable than Peace networks. The data shows that Peace networks occurred in 433 groups at some point within a round, and among them, 88 out of 433 groups (20.3%) transitioned to Bully networks. By contrast, Bully networks emerged in 418 groups throughout the round, but only eight out of 410 groups (1.9%) reverted to Peace networks.

Overall, the between-subject experiment shows a similar set of results as those found in the within-subject experiment, and helps rule out that the effects found in the within-subject experiment are driven primarily by participants' prior experiences in each treatment.

5.2. Dynamics of network formation

In this section, we analyze the dynamic processes of groups within a round of the game as they transition towards Bully and Peace networks. Our objective is to identify the critical factors influencing network formation which determine whether a Bully or Peace network is established. We concentrate our analysis on the within-subject experiment, pooling data across all cost levels. We note that groups reaching a particular outcome (Peace or Bully) exhibit no significant differences in their dynamic network formation patterns by different cost levels once the respective dynamics are underway. In Online Appendix D, we present separate analyses for each cost level, confirming the consistency of observed patterns across varying costs.

The benefit of focusing on the within-subject experiment is that participants have adequate exposure to the range of different cost levels, so that on average, the patterns observed should be reflective of experienced participants who know the game well under different cost parameter values. In Online Appendix F, we also report the same set of analyses for the between-subject experiment, demonstrating that the primary dynamic patterns are strikingly similar.

We begin by examining the overall linking activity of participants in the game. Fig. 6 illustrates the activity levels per second of extending any type of link (friendly or rival) to other group members within the continuous-time decision environment. The figure reveals high participant activity at the onset, particularly in forming friendly connections. The overall activity rate drops dramatically after the tenth second, indicating that groups successfully coordinate on a network early in each round.

5.2.1. Differences in dynamic formation between groups converging to Bully vs Peace equilibria

Next, we examine whether the patterns of extending friendly and rival links differ between Bully and Peace groups. In fact, within the first five seconds, the divergence in alliance formation between the two sets of groups is apparent. Table 6 shows that while the percentage of three-member alliances notably increases over time in Bully groups, it is relatively low and declining in Peace groups starting from the 3rd second. On the contrary, the percentage of fully connected networks notably increases over time in Peace groups but remains relatively low and stable in Bully groups also starting from the 3rd second.

Given these patterns, a key question is why some groups converge to Bully networks while others manage to reach Peace networks? One factor is group members' success in coordinating on a common rival. Fig. C3 in Online Appendix C shows the maximum number of rival links received by any player in a group over time within a round, averaged across all groups. By this measure, Bully and Peace groups diverge almost immediately in their pattern of making rivals. In Bully groups, at the 5th second, the maximum number of attacks any player in a group receives is 1.5 on average, and this number increases to 2.1 attacks by the 10th second. By contrast, in Peace groups a player rarely ever receives more than one attack throughout the entire round.

We later show that in fact, first victims account for the majority of the final victims (i.e., the player who is the final target in a Bully group). Thus, Peace groups who have ever had an attack in the network, effectively pass on the opportunity to coordinate on this salient target. Taken together, these patterns show that Bully groups quickly coordinate on a common rival whereas Peace groups fail to (or perhaps do not attempt to) coordinate on a common rival, particularly the natural coordination target of the first victim.

To provide statistical evidence on factors that can explain the divergent paths of the Bully and Peace final networks, we also implement a series of group-level Probit regression analyses with a binary dependent variable for whether a group eventually converges to Bully or Peace networks. We define several temporary network structure states empirically observed in the first few seconds and utilize them as independent variables. We note that this particular analysis should be interpreted with caution since all of these temporary network patterns may be endogenous to some unobserved factors within the game. Nonetheless, the findings can be

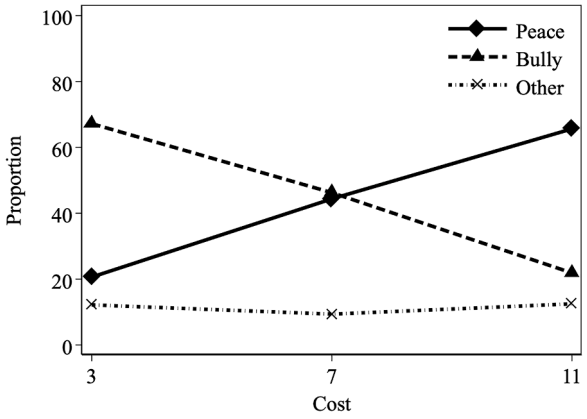

Fig. 5. Proportions of Final Network Types – Between-Subject Experiment.
Note: This figure shows the final network configuration of all 960 groups of all cost levels. "Other" includes all equilibrium networks other than the Peace and Bully equilibria, together with non-equilibrium networks.

Table 4
Frequency (in%) of Final Network Types – Between-Subject Experiment.

	Bully	Peace	Other equilibrium networks	Non-equilibrium networks
Cost = 3	67.2	20.6	1.3	10.9
Cost = 7	46.6	44.7	0	8.7
Cost = 11	22.2	65.3	6.3	6.2

Note: The sets of equilibrium networks are different for costs below 10 and the cost of 11 as shown in Fig. 1. In our data, equilibrium network types 5 to 8 (quasi-peace), 10 (quasi-bully) and 12 were never formed. Fig. C1 in Online Appendix C shows all non-equilibrium network types formed at least once. Most of these networks were rarely formed except for types 30 and 39 which were formed relatively more frequently. Table C1 shows an even finer categorization of final network types for each cost level.

Table 5
Probit Model: Treatment Effects on Final Networks – Between-Subject Experiment.

	Average marginal effect	Standard error
Cost = 3	0.514***	0.069
Cost = 7	0.261***	0.092
H0: 3 vs. 7	$p < 0.001$	
N	851	

Note: The dependent variable is whether the final network is Bully (=1) or Peace (=0). The cost of 11 serves as the benchmark. The regression also includes round dummies which show trends toward more Bully over time. Standard errors are clustered at the session level.
*** $p < 0.01$.

informative as to how early network states lead to the formation of the final network structure from a predictive standpoint.

The detailed tables and discussions are relegated to Online Appendix E, but are summarized here for convenience. First, the analysis shows that the occurrence of a single three-member alliance in the first few seconds strongly predicts the Bully network, whereas the occurrence of players being fully connected via friendly links (i.e., a universal alliance) in the first few seconds strongly predicts the Peace network. Second, the incidence of at least one player receiving one or two attacks in the very first seconds strongly predicts the Bully network. Finally, while variables related to attacking are more influential on the final network in earlier seconds of each round, variables related to alliances become more dominant in terms of predictive power in the later seconds. Overall, these results confirm our earlier descriptive observations that the network patterns occurring in the first few seconds readily predict the type of final network eventually formed.

The analysis in this section leads us to the following set of findings about network formation dynamics:

Result 2. *In terms of the type of alliance formed and the tendency to coordinate on a common rival, Bully groups and Peace groups rapidly diverge within the first few seconds. While Bully groups quickly formed a three-member alliance, coordinating to attack a first victim, Peace groups quickly became fully connected via friendly links while avoiding attacking each other.*

5.2.2. Coordinating on a final victim

We now focus on the cases of Bully equilibria formed, to explore in further detail the dynamic process of reaching Bully networks.

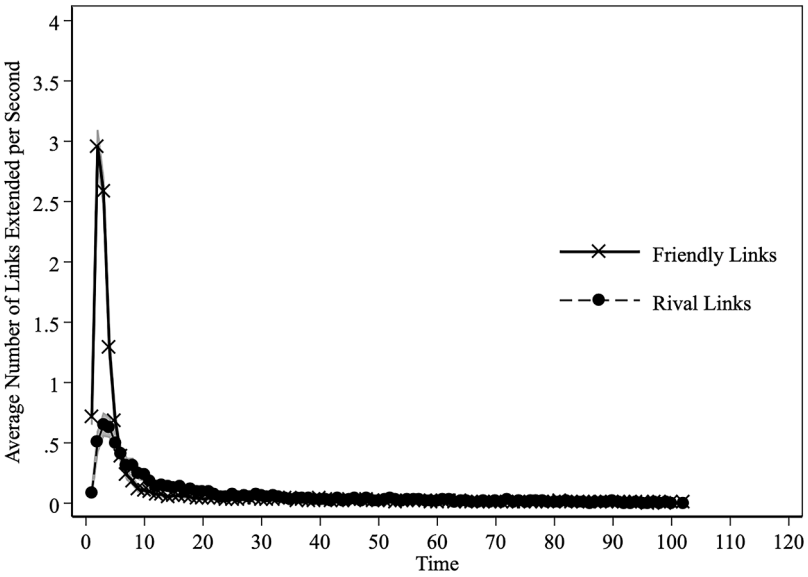

Fig. 6. Extension of Links per Group, by Second – Within-Subject Experiment.
Note: The grey shaded area indicates 95% confidence intervals.

Table 6
Percentages of 3-member Alliances and Fully Connected Networks, First 10 s – Within-Subject Experiment.

Time (seconds)	Peace		Bully	
	3-member alliance	Fully connected	3-member alliance	Fully connected
1	0.6	0	1.2	0
2	18.3	1.3	18.6	2.2
3	23.0	15.8	34.6	5.6
4	22.1	33.1	45.6	7.1
5	18.9	46.1	50.0	7.4
6	14.8	53.6	56.1	6.6
7	11.4	58.4	62.5	6.4
8	19.7	64.0	65.9	5.9
9	9.1	66.6	68.9	5.4
10	8.5	67.5	72.3	4.7

Note: "3-member alliance" is the situation in which three out of four players are mutual friends and the other player is a lone player; this is a necessary condition for Bully group. "Fully connected" is the situation in which all four players in the group are mutual friends.

We jointly consider two related processes which comprise the Bully network equilibrium: the process of coordinating on a final victim and the process of forming a three-member alliance.

In Fig. 7, we plot a time-series figure showing both the number of rival links the final victim receives and the number of effective friendships among the other three players in Bully groups. The Figure shows three clear features. First, the final victim receives attacks very early in a round: on average they receive 1.4 enemy links by the 5th second and 2.0 enemy links by the 10th second. Second, the other three players form an alliance even faster: 50.0% of the groups form a three-member alliance within 5 s, while 72.3% of groups have done so within 10 s. The third feature is that the formation of a three-member alliance generally precedes the emergence of the final victim. The alliance forms before the final victim receives the first attack in 49.5% (202/408) of the groups; the proportion of three-player alliances increases to 72.8% (297/408) and 88.0% (359/408) before the final victim receives the second and third attacks, respectively. Fig. C4 in Online Appendix C shows a similar pattern using the median number of attacks instead of the average. Overall, in the groups that reach Bully networks in the final configuration, alliance formation is swift and precedes targeting the final victim.

We now turn to a more detailed analysis to understand *how* the final victim is coordinated upon by the other three players. We conjecture that the first player who receives an attack (the first victim) becomes salient and is subsequently more likely to be coordinated upon than others (Schelling, 1960). It is also possible that the player who initiates the first attack (referred to as the initiator) could be another target for coordination for similar reasons of salience, although it may require additional steps of coordination to successfully adjust the collective target on the attacker.

We analyze the likelihood that each type of player receives one, two, three attacks and becomes the final victim, respectively. For this analysis, we include data on all of the types of networks to obtain a broad picture of how final victims are coordinated upon. Fig. 8 shows that among these 630 first victims, 66.3% (418/630) of them receive two attacks subsequently; 51.7% (326/630) receive three

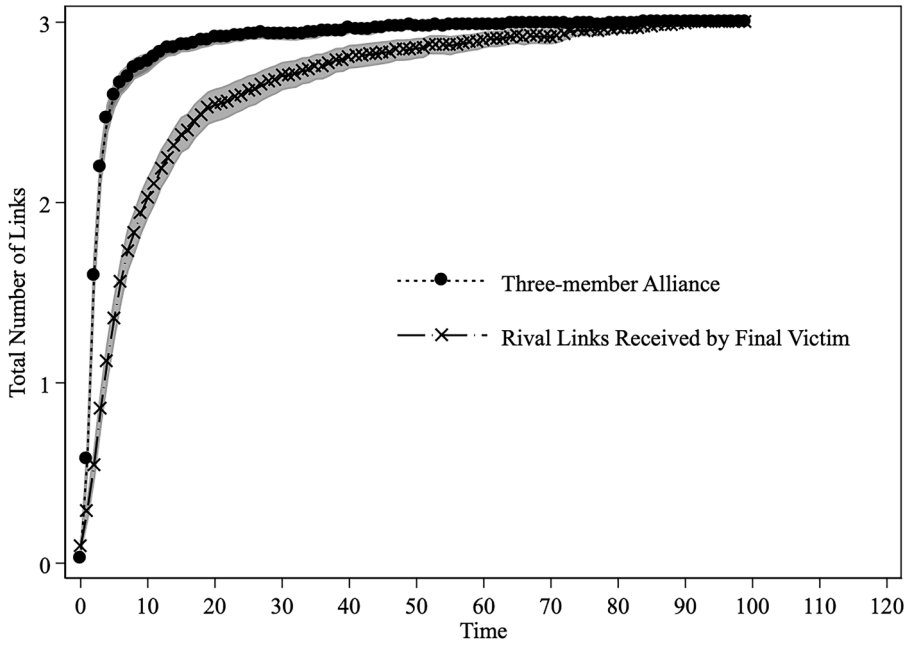

Fig. 7. Evolution of Average Attacks Received by the Final Victim and Average Effective Friendships Among the Other Three Players, Bully Groups – Within-Subject Experiment.

attacks subsequently; and finally 48.7% (307/630) become the final victims. Overall, 48.7% of the first victims are also the final victims, despite the fact that the choices of the first victims are seemingly random. Overall, first victims account for 75.4% (307/407) of all final victims, implying that the coordination on final victims is mostly path-dependent.[22] In terms of attack initiators, among 630 of them, 12.7% (80/630) are the final victims. Thus, the proportion of first attackers who eventually become final victims is still notable and non-negligible. Among the 2020 other players (who are neither first victims nor initiators), the proportion of final victims is negligible 1.0% (20/2020).

5.2.3. Escape from victimhood?

Given that our previous analysis suggests that final victimhood is predictable based on the early conditions of the network structure, in this section we more thoroughly examine the statistical determinants of final victimhood, as well as whether any actions taken by the initial victim can reduce that player's chance of becoming the final victim.

We implement individual-level random effects Probit regressions with a binary variable for final victimhood as the dependent variable. As explanatory variables, we include whether the player is a victim of a first attack, or an initiator of a first attack, with other player types as the comparison group. Table 7 reports the average marginal effect estimates for all groups and for Bully equilibrium groups, separately. Columns (1) and (5) show that both being a first victim and being an initiator (as compared to other players who were not in either category) strongly predict becoming the final victim. Furthermore, F-tests indicate that a first victim is significantly more likely to be the final victim compared to an initiator across all specifications ($p < 0.001$). These regression results are consistent with our previous descriptive statistics.

There are some behaviorally intuitive actions that victims tend to take in order to escape from being bullied, and our empirical analysis can inform us about whether these approaches are actually effective or not. One frequently observed approach by first victims is to quickly attack the initiator back. One possible motive is that other players might then have difficulty in coordinating on a victim. Overall, we find 176 (out of 630) cases of first victims counter-attacking within five seconds of the initial attack, resulting in pairs of a first victim and an initiator with mutual rival links.[23] However, columns (2) and (6) in Table 7 indicate that this strategy is ineffective in reducing a first victim's likelihood of being a final victim.

We also examine whether first victims can escape final victimhood by befriending other players. Estimates from columns (3) and (7) in Table 7 indicate that such efforts at that point in time, do not help in avoiding final victimhood: above-median level of befriending activity (measured by the number of extensions and retractions of friendly links to any other player in a group) after being attacked is

[22] The observed coordination on a final victim can potentially be interpreted as focal behavior. Focality has been shown to influence behavior in other types of conflict such as Colonel Blotto games (Chowdhury et al., 2021).

[23] There are 151 other cases where pairs of a first victim and an initiator have an attack on each other in place in the same instance during the round. However, for these cases, first victims attacked the initiator back at least 5 seconds later. Our results are robust to including all these cases in the regression.

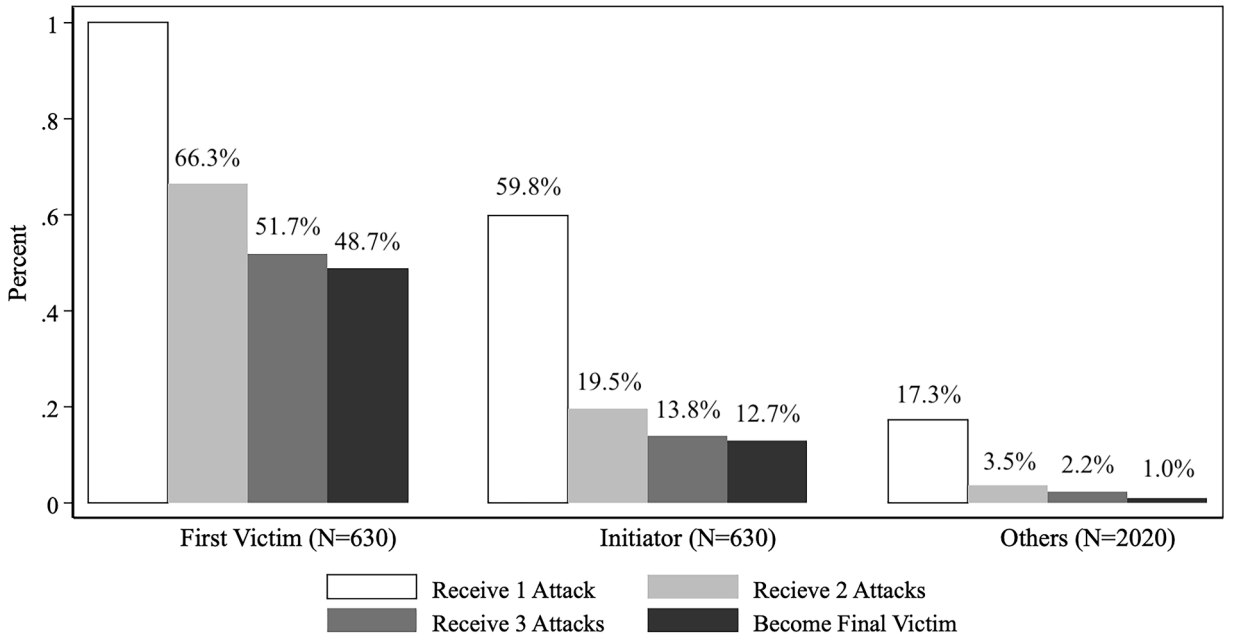

Fig. 8. Transition to Final Victimhood – Within-Subject Experiment.
Notes: This figure includes all 820 groups with a total of 3280 players. In 630 groups, there is ever a first victim. Correspondingly, there are 630 initiators of these first victims. "% Receive 2 (3) attacks" means whether the percentage of players who ever receive 2 (3) attacks during the whole round. "% final victim" means the percentage of players who become final victims.

Table 7
Random Effects Probit model: Determinants of Final Victimhood – Within-Subject Experiment.

	All groups				Bully groups			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
First victim	0.292***	0.286***	0.249***	0.294***	0.455***	0.469***	0.384***	0.464***
	(0.017)	(0.018)	(0.020)	(0.017)	(0.006)	(0.012)	(0.020)	(0.007)
Initiator	0.131***	0.131***	0.129***	0.131***	0.167***	0.167***	0.164***	0.166***
	(0.011)	(0.011)	(0.011)	(0.011)	(0.021)	(0.021)	(0.021)	(0.021)
First victim (attack back)		0.022				−0.042		
		(0.017)				(0.028)		
First victim (befriending activity, above median)			0.075***				0.129***	
			(0.015)				(0.032)	
First victim (more friends than initiator)				−0.023				−0.097***
				(0.020)				(0.029)
N	3280	3280	3280	3280	1632	1632	1632	1632

Note: The dependent variable is whether a player is a final victim (=1) or not (=0). The table reports average marginal effect estimates with standard errors clustered at the session level. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

in fact positively related to the likelihood of becoming final victims. It is possible that first victims' statuses in the network at the point of being attacked are already disadvantaged, such that attempts to further form friend links are rendered ineffective.

However, those first victims who at the moment of being attacked have more friends than initiators appear to be less vulnerable to final victimhood than those first victims who have fewer friends. Estimates from column (8) in Table 7 support this conjecture: having more friends as compared to the initiator significantly reduces the likelihood of a first victim becoming a final victim by 9.7% in Bully groups. The fact that having more friends helps first victims escape victimhood is also consistent with our previous finding that alliance formation generally precedes the coordination on a final victim.

In summary, first victims are far more likely to be final victims than any other players, despite any of the intuitive actions taken by first victims examined above. First victims' efforts to escape from being bullied are mostly futile, at least from a statistical standpoint. The only factor that appears to help is from an ex-ante standpoint, having more allies in the first place prior to being attacked. This in turn, can explain the urgency with which players tend to form alliances in the very beginning of each round.

The above analyses lead us to the following findings regarding final victimhood:

Result 3. *i) In Bully groups, the formation of a three-member alliance generally precedes the emergence of the final victim; ii) The final victim*

is most likely the first-attacked victim, followed by the initiator of the first attack; iii) First victims' efforts to escape from being bullied, whether through counter-attack or post-attack alliance formation, are mostly futile.

5.2.4. Why initiate an attack?

The puzzling scenario in our data that attack initiators themselves face a considerable likelihood of eventually becoming victims prompts us to investigate the underlying motivations for initiating an attack. One may question why individuals choose to initiate bullying, given the potential for it to backfire, when acting instead as a follower could lead to a similar outcome but with notably lower risk of being ultimately targeted. To address this question, our analysis begins by illustrating this dilemma, highlighting that initiators do not seem to be prioritizing their responses for either short-term gains or long-term benefits. As a follow-up, we conduct a detailed examination of the decision to initiate an attack at the individual level, exploring the connection between a player's prior status and behavioral patterns in relation to their inclination to engage in bullying behavior.

In our study, initiating an attack is theoretically optimal if the attacker possesses strictly more friends than the targeted first victim for attack cost below 10, and for attack cost of 11, at least two more friends than the targeted first victim. Surprisingly, our data show that in only 25.1% of cases does the initiator have strictly more friends than the first victim; the initiator has fewer friends than the targeted victim in 5.8% of cases, and an equal number of friends in 69.1% of cases. This suggests that launching an attack is not primarily motived by immediately apparent payoff considerations, but more likely for coordination purposes. Furthermore, most attacks (74.0%) from any player at any time ($N = 3533$) do not yield a hypothetical increase in payoff associated with the attack action. Interestingly, however, an attack retraction ($N = 2125$) frequently corresponds to a hypothetical payoff increase (85.9%), although over 50% of all attack-related actions are not immediate optimal responses in this manner.

To understand the optimal responses over extended time horizons, we employ a method akin to fictitious play to assess whether initiating an attack is the best action based on the empirical distribution of final victims at the end of the round. By closely studying the empirical distribution of bullying cases (illustrated in Fig. 8), we find that players who initiate the first attack are subjected to bullying in 12.7% of cases, whereas those not engaging in the initial provocation only experience bullying in 11.8% of cases.[24] Thus, factoring in the precise empirical distribution of bullying incidents, it becomes apparent that initiating an attack is not the best response. Moreover, the average payoff for initiators is 69.5 points, while for followers, who we define as those attacking the first victim after the first attack has already occurred ($N = 779$), the average payoff stands at 79.8 points.[25] This difference is statistically significant at the 5% level when applying the Wilcoxon signed-rank test using the session average as the unit of observation. As a result, becoming a first attacker carries a significantly higher risk compared to assuming the role of an (unvictimized) follower.[26]

The above analysis suggests that the decision to initiate an attack is not empirically optimal, either in the short-term or long-term. This prompts questions regarding the factors influencing participants who instigate bullying, whether they are reacting to previous experiences of being bullied themselves, or how personal characteristics such as selfishness, competitiveness, or lack of strategic sophistication may influence their decisions. To investigate these aspects, in half of all sessions we conducted a post-experimental survey featuring the Berlin Numeracy Task (BNT), which measures individual risk literacy and numerical sophistication (Cokely et al., 2012), and the Social Value Orientation (SVO) task to assess individual social preferences (Sonnemans et al., 2006).[27] We also ask subjects to indicate their general risk-taking attitude and competitiveness.

We implement individual-level random effects Probit regressions, using the binary dependent variable of being an initiator. We included explanatory variables such as whether the player was an initiator in the previous round, a first victim in the previous round, or

[24] There are two scenarios in which a player may become victimized if she refrains from initiating the first attack. First, she may be the target of the initial attack, which carries a 48.7% chance of resulting in her becoming the final victim. Second, she may be on the receiving end of a non-initial attack, presenting a 1.0% chance of becoming the final victim. Thus, as illustrated in Fig. 8, the probability of such a player emerging as the final victim can be calculated as follows: $(48.7\% \times 630/2650) + (1.0\% \times 630/2650) = 11.8\%$.

[25] We also study a specific subset of followers who have established all three friendships before the initial attack and target the first victim afterward. With an average payoff of 80.4 points, this group shows no substantial advantage over followers in general. Comparatively, we assess peaceful players, who consistently avoid attacking during the entire round. Their average payoff of 66.7 points suggests no notable benefit from their peace-maintaining behavior.

[26] The choice between strategies of initiating conflict or waiting for another to assume the aggressor role bears similarity to the fight-or-flight model examined by van Leeuwen et al. (2022). Their model depicts a dynamic interaction between two players, with options to fight, flee, or wait. Employing a waiting tactic aims to secure the prize without engaging in a costly battle. This observed behavior also resembles a volunteer's dilemma, wherein a player faces the choice between bearing the cost of volunteering or waiting to reap the benefit without any expense.

[27] The Berlin Numeracy task result takes a value from 0 to 4 with a higher number indicating a higher level of numerical sophistication. Social Value Orientation is measured as an angle, where 0 degrees corresponds to a dictator giving everything to herself, 45 degrees corresponds to splitting equally between a recipient and herself, and 90 degrees corresponds to giving everything to the recipient. Both the numeracy task and the social value orientation task were incentivized.

a final victim in the previous round, as well as various measures of individual characteristics. Table 8 presents the average marginal effect estimates for all groups and for Bully groups separately. Consistently across all specifications, being an initiator in the previous round is a substantial predictor of initiating an attack again in the current round. At the same time, those who were first victims in the previous round are also significantly more likely to become initiators in the current round, presumably to avoid being bullied again..[28] Furthermore, none of the measures of individual characteristics are significant predictors of a subject's tendency to be an initiator.[29] Our overall interpretation of these findings is that choosing to be an initiator is primarily a path-dependent decision and a reaction to prevent oneself from being bullied. Compared to adopting the role of a follower or a peaceful player, the cost incurred from being an initiator may not be subjects' primary consideration in these cases.

Result 4. *Being a first attacker does not payoff empirically in expectation because first attackers bear a substantial risk of ultimately being bullied themselves. The choice of being a first attacker is highly path-dependent upon whether the same individual was the initiator or the first victim in the previous round, and thus is interpreted as being a reactive choice in order to avoid being bullied again.*

6. Conclusion

While the origins of real-world conflicts can be complex and multifaceted, we study the propensity for conflict in a network formation game with the potential for costly capture of peer resources. In our experiment, which is implemented as a continuous-time decision environment in the laboratory, the absence of any rivalrous links in the network yields the greatest social surplus, with equal division of surplus across the four players. Hence, in our setting, a peaceful network is both efficient and fair, two of the most typically prized social welfare objectives.

However, players can choose to disrupt peace by directing a rival link at another player, incurring a cost. The player with more alliance links obtains a portion of the rival's surplus based on the final network formation of a round. Our theoretical analysis, modifying Hiller's (2017) signed network formation game, demonstrates that while Peace is an equilibrium, a 3-against-1 Bully configuration is also a highly robust equilibrium. Our experimental data shows that groups reach these two equilibria about 90% of the time, and their relative frequencies depend on attack costs, as predicted by our theory-generated hypotheses.

Our experiments address the question of conflict emergence among homogeneous decision-makers with equal initial endowments. In a flexible network environment, even with homogenous players, substantial socially costly conflicts arise. Bully outcomes occur more frequently than Peace when attack costs are low, but this frequency decreases relative to Peace as costs increase. In the highest cost treatment, Peace outcomes become more prevalent, occurring 59% and 65% of the time in within-subject and between-subject experiments, respectively. Consistent with our hypothesis, based on the higher stability of the Bully equilibrium, transitions from Peace to Bully occur more frequently than from Bully to Peace. This indicates that once a Bully outcome emerges within a round, Peace proves challenging to reinstate.

We delve deeper to study the rich dynamics of alliance formation and conflict emergence. Examining the dynamics of the network formations, the data reveal that despite the complex structure of the game, the two main equilibrium networks are reached with remarkable speed. Most of the active linking activity occurs within the first few seconds of each round, and the early network configurations which hint at eventual outcomes, in fact strongly predict final network configurations. Thus, there is a substantial path dependency in the network formation over the course of a round, and furthermore, early on Bully and Peace networks diverge sharply in their linking patterns.

When examining Bully situations specifically, alliance formation generally precedes coordination on a common rival, showing participants' tendency to gather a circle of friends before making an attack. In terms of the determination of the final victim, the most likely candidate is the player who receives the first attack from any other player in the group, while the initial attacker also faces a non-trivial likelihood of being left in the role of final victim. For first victims, there is also a heavy path dependency in that it is subsequently very difficult for first victims to escape from being bullied. Intuitive ex-post tactics such as counter-attacking or extending friend links are largely ineffective.

Finally, given the importance of the first attack launched, combined with the non-trivial chance of a first attacker becoming the final victim in the data, we examine whether some path-dependent states or individual characteristics correlate with the decision of being first attackers.[30] The data show that being first victims in the previous round is strongly correlated with being first attackers in the current round, who presumably attempt to escape from being bullied again. By contrast, none of the individual characteristics such as numerical sophistication, social preferences, risk or competitive attitudes is a significant predictor of initiating a first attack. Thus, choosing to be first attackers is mostly a path-dependent, reactive choice to avoid being bullied.

There are many potentially interesting extensions to the signed network formation games used in the current study. Although our

[28] Interestingly, being a final victim in the previous round does not significantly correlate with initiating an attack in the next round for most specifications, likely due to the high correlation between the two explanatory variables of being the first victim and being the final victim.

[29] We also examine the predictive capabilities of these individual characteristics in determining a subject's tendency to act as a follower or a peaceful player. Specifically, we define a follower as an individual who attacks the first victim after the onset of initial aggression initiated by another player, while a peaceful player is defined as someone who refrains from attacking others during the entire round. In unreported regression analyses akin to those presented in Table 8, we consistently observe that none of the individual characteristics measured in our post-experiment elicitations can significantly predict characterization of either followers or peaceful players.

[30] In Online Appendix G, we attempt to explain a player's decision to initiate the first attack using a quasi-dynamic model, highlighting the role of beliefs in the coordination process of reaching a bullying outcome. The model is consistent with a number of stylized facts in our experimental data.

Table 8
Random Effects Probit model: Determinants of Initiating First Attack – Within-Subject Experiment.

	All groups				Bully groups			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
L1.Initiator	0.160***	0.173***		0.174***	0.229***	0.252***		0.255***
	(0.022)	(0.025)		(0.027)	(0.028)	(0.028)		(0.029)
L1.First victim	0.089***	0.083***		0.082***	0.080***	0.096**		0.096**
	(0.010)	(0.011)		(0.010)	(0.020)	(0.040)		(0.037)
L1.Final victim	0.015	0.029*		0.030*	0.010	0.028		0.031
	(0.016)	(0.017)		(0.016)	(0.028)	(0.050)		(0.049)
BNT score			−0.000	0.003			0.002	0.004
			(0.007)	(0.007)			(0.005)	(0.004)
SVO angle			−0.007	−0.021			−0.016	−0.047
			(0.042)	(0.046)			(0.023)	(0.034)
Risk-taking			−0.002	−0.006			−0.003	−0.006
			(0.004)	(0.004)			(0.004)	(0.005)
Competitive			−0.001	−0.002			−0.003	−0.009
			(0.006)	(0.007)			(0.006)	(0.007)
N	3116	1520	1600	1520	1596	860	876	1632

Note: The dependent variable is whether a player is an initiator (=1) or not (=0). L1.First Victim, L1.Final Victim and L1.Initiator denote being a first victim, a final victim, and an initiator in the previous round, respectively. BNT score takes a value from 0 to 4 with a higher number indicating a higher level of numerical sophistication. SVO angle takes a value from 0 to 90 with a higher degree indicating a higher level of prosociality. "Risk-taking" is the self-reported general attitude toward risk-taking in daily life on the scale from 1 (not risk-taking at all) to 7 (extremely risk-taking). "Competitive" is the self-reported general attitude toward competitiveness in daily life on the scale from 1 (not competitive at all) to 7 (extremely competitive). The table reports average marginal effect estimates with standard errors clustered at the session level. Column (1) includes data from all sessions, whereas column (2) only includes the sessions in which those measures about individual characteristics are also elicited.

* $p < 0.1$;
** $p < 0.05$;
*** $p < 0.01$.

current study is focused on a homogeneous player setting with equal endowments and no existing bilateral links, our study can also serve as a benchmark to understand situations with additional layers of complexity. For example, heterogeneity in players' characteristics may substantially alter the dynamics in alliance formation and conflict. A natural direction is to manipulate heterogeneity in fighting strength by making one player stronger than others. Player heterogeneity creates a tension between bandwagoning, i.e., siding with the stronger player, and balancing, i.e., targeting at this uniquely salient player to make everyone's payoff more equal. Importantly, both situations create incentives to disrupt peace as the equilibrium and thus may lead to more chaotic coordination.

Accompanying the idea of asymmetric players, another potential direction is to have an initial non-empty network structure and observe how linking decisions evolve from the initial state. This line of investigation will also provide *causal* insights about the dynamics that we explore in the current study. Yet, another possible extension, which may involve more substantial changes to the current game, is to allow players to make commitments or promises about linking decisions or possibly resource transfers in pre-game negotiations. In addition, making adjustments to the nature of alliances is a further direction for extension of this work. In our current setup, the alliance is effective for both offensive and defensive purposes. However, an alternative setup is to vary the effectiveness of the alliance based on use for a defensive purpose versus for an attacking purpose.[31]

Another promising direction for future research involves examining large-scale network formation. In our current setting, a group consisting of more than four players can sustain various alliance formations, allowing for the emergence of multiple alliances with at least two members in equilibrium, in which larger alliances attack the smaller ones (Hiller, 2017). It will be interesting to investigate whether specific networks, such as the Bully network, maintain prominence among all other equilibria or if alternative robust network patterns emerge. Although conducting large-scale network formation games in a laboratory setting poses technical challenges, the innovative experimental platform introduced by Choi et al. (2020) provides a practical toolkit for carrying out network experiments on

[31] We have already pursued this direction by studying the theoretical properties of different alliance types and implementing the corresponding experiments, however these experiments may be outside the scope of the current paper. In case of reader interest, we summarize the basic findings here. In one treatment, a player's friends only help when the player is an initiator of a rival link, but do not come into the rescue when he is a receiver. This reflects a type of *offensive alliance* in which agents agree to fight together but do not commit to intervene when one alliance member is attacked. In the other treatment, a player's friends only come to the rescue when the player receives a rival link but do not help when the player is an initiator. This reflects a type of *defensive alliance* in which agents agree to defend together but not commit to intervene when one ally initiates an attack. These two alliance treaties do have real world counterparts and are studied by political scientists who are mostly concerned about how different treaties affect the risk of war. For example, Siverson and King (1980) analyzed the Correlates of War data and found that the formation of offensive (defensive) alliances increases (decreases) the occurrence of war. The theory predicts that peace is not impossible in the offensive alliance treatment and conflict is not impossible in the defensive alliance treatment, and this is confirmed in our data. Furthermore, 75% of the groups in the offensive alliance treatment reach the same bullying situation and their dynamic patterns are also similar to those in the main experiment. On the contrary, almost all groups retain peace in the defensive alliance treatment. More details about design and results are available on request.

an expanded scale. This development paves the way for probing questions similar to those in our research but on a larger scale, thereby deepening our understanding of alliance and conflict network dynamics.

Declaration of competing interest

None.

Acknowledgement

Financial support from the National Natural Science Foundation of China (Grant 72073080, 72192842, 72203099, 72250710170, 72373083, 72394394), Hong Kong Research Grants Council (Grant 14502922), and Shandong University Direct Grants is gratefully acknowledged. For helpful comments we thank the associate editor, two anonymous reviewers and participants in the 2020 ESA World Meetings (online), 2020 ECNU Industrial Organization and Behavioral Economics Workshop, Virtual East Asia Experimental and Behavioral Economics Seminar Series, HKUST Workshop on Industrial Organization, Global Seminar on Contests & Conflict, and Duke Kunshan Workshop on Social Networks in Economics. All authors are co-first authors.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.geb.2024.05.009.

References

Abbink, K., Doğan, G., 2019. How to choose your victim. Games Econ. Behav. 113 (January), 482–496.
Acemoglu, D., Egorov, G., Sonin, K., 2008. Coalition formation in non-democracies. Rev. Econ. Stud. 75 (4), 987–1009.
Antal, T., Krapivsky, P.L., Redner, S., 2006. Social balance on networks: the dynamics of friendship and enmity. Phys. D Nonlinear Phenom. 224 (1–2), 130–136.
Bala, V., Goyal, S., 2000. A noncooperative model of network formation. Econometrica 68 (5), 1181–1229.
Berninghaus, S.K., Ehrhart, K.M., Ott, M., Vogt, B., 2007. Evolution of networks–an experimental analysis. J. Evol. Econ. 17 (3), 317–347.
Berninghaus, S.K., Ehrhart, K.M., Ott, M., 2006. A network experiment in continuous time: the influence of link costs. Exp. Econ. 9 (3), 237–251.
Bloch, F., 2012. Endogenous formation of alliances in conflicts. The Oxford Handbook of the Economics of Peace and Conflict. Oxford University Press, Oxford, pp. 473–502 edited by Michelle R Garfinkel and Stergios Skaperdas.
Bloch, F., Sánchez-Pagés, S., Soubeyran, R., 2006. When does universal peace prevail? Secession and group formation in conflict. Econ. Gov. 7 (1), 3–29.
Burger, M.J., Buskens, V., 2009. Social context and network formation: an experimental study. Soc. Netw. 31 (1), 63–75.
Callander, S., Plott, C.R., 2005. Principles of network development and evolution: an experimental study. J. Public Econ. 89 (8), 1469–1495.
Cartwright, D., Harary, F., 1956. Structural balance: a generalization of Heider's theory. Psychol. Rev. 63 (5), 277–293.
Choi, S., S. Goyal, and F. Moisan. 2020. "Large scale experiments on networks: a new platform with applications." Cambridge-INET Working Paper Series No: 2020/29.
Chowdhury, S.M., Kovenock, D., Arjona, D.R., Wilcox, N.T., 2021. Focality and asymmetry in multi-battle contests. Econ. J. 131 (636), 1593–1619.
Cokely, E.T., Galesic, M., Schulz, E., Ghazal, S., Garcia-Retamero, R., 2012. Measuring risk literacy: the Berlin numeracy test. Judgm. Decis. Mak. 7 (1), 25–47.
Cortes-Corrales, S., and P.M. Gorny. 2018. "Generalising conflict networks." MPRA Paper No. 90001.
Easley, D., Kleinberg, J., 2010. Networks, Crowds, and Markets: Reasoning About a Highly Connected World. Cambridge University Press.
Falk, A., Kosfeld, M., 2012. It's All about connections: evidence on network formation. Rev. Netw. Econ. 11 (3), 2.
Fischbacher, U., 2007. Z-Tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178.
Franke, J., Öztürk, T., 2015. Conflict Networks. J. Public Econ. 126 (June), 104–113.
Galeotti, A., Goyal, S., 2010. The law of the few. Am. Econ. Rev. 100 (4), 1468–1492.
Garfinkel, M.R., 2004. Stable alliance formation in distributional conflict. Eur. J. Political Econ. 20 (4), 829–852.
Goeree, J.K., Riedl, A., Ule, A., 2009. In search of stars: network formation among heterogeneous agents. Games Econ. Behav. 67 (2), 445–466.
Goyal, S., Rosenkranz, S., Weitzel, U., Buskens, V., 2017. Information acquisition and exchange in social networks. Econ. J. 127 (606), 2302–2331.
Goyal, S., Vigier, A., Dziubinski, M., 2016. Conflict and networks. The Oxford Handbook of the Economics of Networks. Oxford University Press, Oxford, pp. 214–243 edited by Y Bramoulle, A Galeotti, and B Rogers.
Guido, A., Robbett, A., Romaniuc, R., 2019. Group formation and cooperation in social dilemmas: a survey and meta-analytic evidence. J. Econ. Behav. Organ. 159 (March), 192–209.
Haller, H., Hoyer, B., 2019. The common enemy effect under strategic network formation and disruption. J. Econ. Behav. Organ. 162 (June), 146–163.
Harary, F., Norman, R.Z., Cartwright, D., 1965. Structural Models: An Introduction to the Theory of Directed Graphs. John Wiley & Sons, New York.
Heider, F., 1958. The Psychology of Interpersonal Relations. John Wiley & Sons, New York.
Herbst, L., Konrad, K.A., Morath, F., 2015. Endogenous group formation in experimental contests. Eur. Econ. Rev. 74 (February), 163–189.
Hiller, T., 2017. Friends and enemies: a model of signed network formation. Theor. Econ. 12 (3), 1057–1087.
Huitsing, G., Duijn, M.A.J., Snijders, T.A.B., Wang, P., Sainio, M., Salmivalli, C., Veenstra, R., 2012. Univariate and multivariate models of positive and negative networks: liking, disliking, and bully–victim relationships. Soc. Netw. 34 (4), 645–657.
Jackson, M.O., Wolinsky, A., 1996. A strategic model of social and economic networks. J. Econ. Theory 71 (1), 44–74.
Jackson, M.O., Nei, S., 2015. Networks of military alliances, wars, and international trade. Proc. Natl. Acad. Sci. 112 (50), 15277–15284.
De Jaegher, K., 2021. Common-enemy effects: multidisplinary antecedents and economic perspectives. J. Econ. Surv. 35 (1), 3–33.
Ke, C., Konrad, K.A., Morath, F., 2013. Brothers in arms–an experiment on the alliance puzzle. Games Econ. Behav. 77 (1), 61–76.
Ke, C., Konrad, K.A., Morath, F., 2015. Alliances in the shadow of conflict. Econ. Inq. 53 (2), 854–871.
Kirchsteiger, G., Mantovani, M., Mauleon, A., Vannetelbosch, V., 2016. Limited farsightedness in network formation. J. Econ. Behav. Organ. 128 (August), 97–120.
König, M.D., Rohner, D., Thoenig, M., Zilibotti, F., 2017. Networks in conflict: theory and evidence from the great war of Africa. Econometrica 85 (4), 1093–1132.
Konrad, K.A., 2014. Strategic aspects of fighting in alliances. The Economics of Conflict. MIT Press, Cambridge & London, pp. 1–22 edited by K Wärneryd.
Kosfeld, M., 2004. Economic networks in the laboratory: a survey. Rev. Netw. Econ. 3 (1), 20–41.
Leeuwen, B., Offerman, T., Schram, A., 2020. Competition for status creates superstars: an experiment on public good provision and network formation. J. Eur. Econ. Assoc. 18 (2), 666–707.

Leeuwen, B., Offerman, T., van de Ven, J., 2022. Fight or flight: endogenous timing in conflicts. Rev. Econ. Stat. 104 (2), 217–231. https://doi.org/10.1162/rest_a_00961.
Macfarlan, S.J., Walker, R.S., Flinn, M.V., Chagnon, N.A., 2014. Lethal coalitionary aggression and long-term alliance formation among Yanomamö Men. Proc. Natl. Acad. Sci. 111 (47), 16662–16669.
O'Connell, P., Pepler, D., Craig, W., 1999. Peer involvement in bullying: insights and challenges for intervention. J. Adolesc. 22 (4), 437–452.
Pruetz, J.D., Boyer Ontl, K., Cleaveland, E., Lindshield, S., Marshack, J., Wessling, E.G., 2017. Intragroup lethal aggression in West African Chimpanzees (Pan Troglodytes Verus): inferred killing of a former alpha male at Fongoli, Senegal. Int. J. Primatol. 38 (1), 31–57.
Ray, D., 2007. A Game-Theoretic Perspective on Coalition Formation. Oxford University Press, Oxford.
Ray, D., Vohra, R., 1999. A theory of endogenous coalition structures. Games Econ. Behav. 26 (2), 286–336.
Rezaei, S., Rosenkranz, S., Weitzel, U., Westbrock, B., 2024. Social preferences on networks. J. Public Econ. 234, 105113.
Rong, R., Houser, D., 2015. Growing stars: a laboratory analysis of network formation. J. Econ. Behav. Organ. 117, 380–394.
Rosenkranz, S., Weitzel, U., 2012. Network structure and strategic investments: an experimental analysis. Games Econ. Behav. 75 (2), 898–920.
Roser, M., 2016. War and Peace. Our World in Data.
Salmivalli, C., Huttunen, A., Lagerspetz, K.M.J., 1997. Peer networks and bullying in schools. Scand. J. Psychol. 38 (4), 305–312.
Sánchez-Pagés, S., 2007. Endogenous coalition formation in contests. Rev. Econ. Des. 11 (2), 139–163.
Schelling, T.C., 1960. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA.
Siverson, R.M., King, J., 1980. Attributes of national alliance membership and war participation, 1815-1965. Am. J. Political Sci. 24 (1), 1–15.
Smith, A.C., Skarbek, D.B., Wilson, B.J., 2012. Anarchy, groups, and conflict: an experiment on the emergence of protective associations. Soc. Choice Welf. 38 (2), 325–353.
Snyder, G.H., 1997. Alliance Politics. Cornell University Press, Ithaca, NY.
Sonnemans, J., van Dijk, F., van Winden, F., 2006. On the dynamics of social ties structures in groups. J. Econ. Psychol. 27 (2), 187–204.
Teteryatnikova, M., Tremewan, J., 2020. Myopic and farsighted stability in network formation games: an experimental study. Econ. Theory 69 (4), 987–1021.
Tremewan, J., Vanberg, C., 2016. The dynamics of coalition formation – a multilateral bargaining experiment with free timing of moves. J. Econ. Behav. Organ. 130 (October), 33–46.
Wolke, D., Lereya, S.T., 2015. Long-term effects of bullying. Arch. Dis. Child. 100 (9), 879–885.
Xu, J., Zenou, Y., Zhou, J., 2022. Equilibrium characterization and shock propagation in conflict networks. J. Econ. Theory 206, 105571.