Dong, Lu; Huang, Lingbo; Lien, Jaimie W.; Zheng, Jie

Working Paper

How alliances form and conflict ensues

CeDEx Discussion Paper Series, No. 2021-04

Provided in Cooperation with:
The University of Nottingham, Centre for Decision Research and Experimental Economics (CeDEx)

Suggested Citation: Dong, Lu; Huang, Lingbo; Lien, Jaimie W.; Zheng, Jie (2021) : How alliances form and conflict ensues, CeDEx Discussion Paper Series, No. 2021-04, The University of Nottingham, Centre for Decision Research and Experimental Economics (CeDEx), Nottingham

This Version is available at:
https://hdl.handle.net/10419/248295

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

University of Nottingham
UK | CHINA | MALAYSIA

Discussion Paper No. 2021-04

Lu Dong, Lingbo Huang,
Jaimie W Lien, & Jie Zheng

How Alliances Form and Conflict Ensues

August 2021

CeDEx Discussion Paper Series
ISSN 1749 - 3293

CENTRE FOR DECISION RESEARCH & EXPERIMENTAL ECONOMICS

The Centre for Decision Research and Experimental Economics was founded in 2000, and is based in the School of Economics at the University of Nottingham.

The focus for the Centre is research into individual and strategic decision-making using a combination of theoretical and experimental methods. On the theory side, members of the Centre investigate individual choice under uncertainty, cooperative and non-cooperative game theory, as well as theories of psychology, bounded rationality and evolutionary game theory. Members of the Centre have applied experimental methods in the fields of public economics, individual choice under risk and uncertainty, strategic interaction, and the performance of auctions, markets and other economic institutions. Much of the Centre's research involves collaborative projects with researchers from other departments in the UK and overseas.

Please visit http://www.nottingham.ac.uk/cedex for more information about the Centre or contact

Suzanne Robey
Centre for Decision Research and Experimental Economics
School of Economics
University of Nottingham
University Park
Nottingham
NG7 2RD
Tel: +44 (0)115 95 14763
suzanne.robey@nottingham.ac.uk

The full list of CeDEx Discussion Papers is available at

http://www.nottingham.ac.uk/cedex/publications/discussion-papers/index.aspx

How Alliances Form and Conflict Ensues

Lu Dong	Lingbo Huang	Jaimie W. Lien	Jie Zheng
Nanjing Audit University	Nanjing Audit University	The Chinese University of Hong Kong	Tsinghua University

Current draft: July, 2021
Initial draft: April 2020

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. Consistently with the theoretical model predictions, peaceful and bullying outcomes are prevalent among the randomly re-matched experimental groups, based on 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 by initiating the coordination. These findings provide insights for understanding social dynamics in group coordination.

Keywords: network formation, conflict, alliance, bully, peace

JEL: C72, C92, D74, D85, F51

Dong: Economics Experimental Laboratory, Nanjing Audit University, Nanjing 211815, China, lu.dong@outlook.com. Huang: Economics Experimental Laboratory, Nanjing Audit University, Nanjing 211815, China, lingbo.huang@outlook.com. Lien: Department of Decision Sciences and Managerial Economics, CUHK Business School, Chinese University of Hong Kong, jaimie.academic@gmail.com. Zheng: School of Economics and Management, Tsinghua University, jie.academic@gmail.com.

Financial support from the National Natural Science Foundation of China (Grant 72073080, 71873074 and 71903092), Hong Kong Research Grants Council (Grant 14501919 and 14500516), and Chinese University of Hong Kong Direct Grants is gratefully acknowledged. For helpful comments we thank 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, and Global Seminar on Contests & Conflict. The authors are named in alphabetical order. All authors contributed equally.

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 common 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, Huttunen, and Lagerspetz 1997; O'Connell, Pepler, and Craig 1999; Huitsing et al. 2012).[2] 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 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, Krapivsky, and Redner 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. 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

[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.

[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 data 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.

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 a laboratory experiment 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: Thus, two players form an alliance if each has a friendly link extended to the other; they become rivals if at least one of the players sends a rival link to the other. 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 subjects can freely make links to each other while the network structure and hypothetical momentary payoffs are updated in real-time. Since our interest is in the process of alliance and conflict formation, 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 seconds. Only the final network configuration determines the experimental subjects' 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: pairwise stability, no pairwise profitable deviation and no 3-person profitable deviation. 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

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

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, with potential insights for social and political phenomena, including social bullying and international relations. 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 (73.1% of the time) to become the final victim in Bully scenarios. Analysis also shows that coordinating on which player to attack is highly path dependent, in that players who have become salient in the link formation process via a negative link are the most vulnerable. The dynamic results are consistent with the notion that aggressors are often backed by peers, while peers are much less likely to intervene on behalf of victims (O'Connell, Pepler, and Craig 1999).

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. In order to draw further insights on the motives to attack first, we provide a quasi-dynamic theoretical analysis focusing on players' incentive to initiate an attack from an initially peaceful state. Given that it is significantly riskier to take on the role of first attacker compared to being a follower, why would a player initiate the first attack? First attacks pose a puzzle to some extent, because our experimental data imply that being the first attacker does not pay off empirically. Thus, it is likely that the initiation of a first attack serves as a costly coordination device on behalf of the other two players. The success of coordinating an attack among alliance members hinges upon a player's stronger belief about other players' propensity to follow suit in attacking the first victim rather than attacking the first attacker. The model thus also provides a rationale for why first victims in the experiment sometimes attempt to counter-attack quickly.

By examining the formation of alliances and conflicts in a controlled environment, our experiment shows that even in a four-player network game with homogeneous players, no obvious leaders, and equal initial resource allocations, subjects often rapidly and successfully coordinate a group attack on a lone arbitrary player to capture and distribute that player's resources amongst themselves. A higher cost of attacks facilitates a higher frequency of peace based on both theoretical and empirical results, suggesting one potential channel for public policies promoting peaceful outcomes. However, both our theoretical and empirical results caution that it is easier to for peace to descend into conflict, than for peace to be restored from conflict.

The remainder of the paper proceeds as follows. The next section discusses the related literature. Section 3 presents the theoretical framework and hypothesis about final network structures. Section 4 describes the experimental design and Section 5 reports the results. Section 6

presents a quasi-dynamic theoretical analysis, shedding light on the coordination process reaching the bullying situation. Section 7 concludes with remarks on future work.

2. Related Literature

Our study is related to both the theoretical and experimental literature on alliance formation and conflict. We begin with the review of the related theoretical literature. Our work is built directly on Hiller (2017) which applies a novel network approach to endogenous alliance formation, establishes and characterizes equilibrium results.[4] We are the first to put this framework to an experimental test, and to further examine the dynamics of alliance formation.

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, Zenou, and Zhou 2019). For example, Franke & 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, Vigier, & Dziubinski (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.

Outside of network games, alliance formation and conflict has also been investigated using more traditional approaches. A class of theoretical work studies the stability and structure of coalition formation in contests (Ray and Vohra 1999; Garfinkel 2004; Bloch, Sánchez-Pagés, and Soubeyran 2006; Sánchez-Pagés 2007; Ray 2007; Acemoglu, Egorov, and Sonin 2008). For example, Bloch, Sánchez-Pagés, and Soubeyran (2006) study a model in which players, following an exogenous rule of order, propose to form coalition with others. Once a coalition is formed, its members expend effort for the group to win a contest against all outsiders. They find that the grand coalition is the unique stationary perfect equilibrium. Acemoglu, Egorov, and Sonin (2008) investigate a different coalition formation model under the assumption that players have no commitment power, so that any stable coalition must be self-enforcing. They find that such coalitions always exist theoretically, and that they are (generically) unique.

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). A typical setting involves three agents with two rounds of conflict: in the first round, two agents choose to form an alliance and fight against the third agent. The alliance, if having won the first round, then engages in a second round of conflict to determine which member obtains the prize. The key insight of this model is the so-called "paradox of alliance formation" which describes the failure of the first two agents to form an alliance in the first round (Konrad 2009). Two forces may

[4] 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.

hinder the formation of alliances: one is the lack of synergy between the two agents' efforts; the other is free-riding during the conflict which lowers the benefits of joining an alliance, particularly for strong partners. Along similar lines, Baik (2016) examines alliance formation in contest games when the number of agents is more than three.

The main focus of our study is the dynamic process of alliance formation and conflict, that is, choices regarding whom to befriend and whom to fight against. We abstract away from considerations of within-alliance tensions regarding resource distribution emphasized by political scientists (Waltz 1979; Snyder 1997; Morrow 2000). This allows us to focus sharply on the dynamic process of alliance formation and conflict when faced with basic economic incentives in intergroup conflict.

Turning to experimental evidence, our study is related to a growing literature about alliances in contests.[5] Ke, Konrad, and Morath (2013, 2015) study how alliances tend to ruin their chances against outsiders due to in-group conflict. Herbst, Konrad, and Morath (2015) explore whether an intrinsic motivation to form alliances affects players' choices of exerting effort in inter-group contests by comparing exogenously and endogenously formed alliances. Benenson et al. (2009) find that subjects systematically use an intuitive strategy to choose allies based on their own and others' relative strength, rather than on payoff calculations. Jandoc and Juarez (2019) also focus on heterogeneous power and experimentally test a sequential model of coalition formation with farsighted agents as proposed by Acemoglu, Egorov, and Sonin (2008). Our study differs from these by focusing on dynamics of alliance formation and conflict in network games using a continuous decision-time experimental design. In addition, our focus is on homogenous players and in understanding whether and how uneven conflict may nevertheless occur.

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.

In terms of experimental work, two studies on alliance formation and conflict are most closely related to the research questions proposed in our work, albeit not in a network setting. The first is Smith, Skarbek, and Wilson (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. Comparing to a baseline treatment in which

[5] 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, Robbett, and Romaniuc (2019) for a review).

forming groups is not allowed, they find that allowing autonomy does not lead to more cooperation and peace. 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 study closely related to our work in topic, but does not examine a network setting, 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 real-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, our study is also related to the experimental literature about network formation games in which players can essentially link with each other as "friends" (Kosfeld 2004; Callander and Plott 2005; Berninghaus, Ehrhart, and Ott 2006; Berninghaus et al. 2007; Burger and Buskens 2009; Goeree, Riedl, and Ule 2009; Falk and Kosfeld 2012; Rong and Houser 2015; Goyal et al. 2017; van Leeuwen, Offerman, and Schram 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).

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.

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 difference between our model and Hiller (2017) is that we allow for the possibility of no particular relationship between any two agents, whereas in Hiller (2017), each

pair of agents has either a positive or a negative relationship. This proves to be useful in our study for two main reasons: First, 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; Second, we later utilize the expanded set of equilibria along with equilibrium refinement criteria to generate testable predictions of the model.

An agent's strategy is defined as a row vector $\boldsymbol{g}_i = (g_{i,1}, g_{i,2}, g_{i,3}, g_{i,4})$, 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 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.[6] 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, $N_i^+(\boldsymbol{g}) = \{j \in N \mid \bar{g}_{i,j} = 1\}$, 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 $f_i(\boldsymbol{g}) = |N_i^+(\boldsymbol{g})|$. Similarly, we define $N_i^-(\boldsymbol{g}) = \{j \in N \mid \bar{g}_{i,j} = -1\}$ as the set of agents with whom agent i forms rival relationships. A subset of $N_i^-(\boldsymbol{g})$, defined as $N_i^{e-}(\boldsymbol{g}) = \{j \in N \mid g_{i,j} = -1\}$, is the set of agents to whom agent i *extends* a negative link, and we denote by $e_i(\boldsymbol{g}) = |N_i^{e-}(\boldsymbol{g})|$ the number of negative links agent i extends.

An agent draws strength from his friends when engaging in a conflict. Denote an agent's intrinsic strength as $\lambda > 0$. The potential fighting strength of agent i is then given by $s_i(\boldsymbol{g}) = \lambda(f_i(\boldsymbol{g}) + 1)$.

Extending a positive link is assumed to be costless. Extending a negative link, i.e., attacking, however, is assumed to incur a cost of $c > 0$. The payoff of agent i from extending an attack to agent j are determined by the attack payoff function $h_i^A(s_i(\boldsymbol{g}), s_j(\boldsymbol{g}))$, and the payoff of agent i from receiving an attack from agent j is determined by defense payoff function $h_i^D(s_j(\boldsymbol{g}), s_i(\boldsymbol{g}))$. We make three reasonable assumptions regarding the payoff functions $h_i^A(s_i(\boldsymbol{g}), s_j(\boldsymbol{g}))$ and $h_i^D(s_j(\boldsymbol{g}), s_i(\boldsymbol{g}))$.

Assumption 1 (Monotonicity): Both $h_i^A(s_i(\boldsymbol{g}), s_j(\boldsymbol{g}))$ and $h_i^D(s_j(\boldsymbol{g}), s_i(\boldsymbol{g}))$ are weakly increasing in $s_i(\boldsymbol{g})$ and weakly decreasing in $s_j(\boldsymbol{g})$.

Assumption 2 (Balance): $h_i^A\left(s_i(\boldsymbol{g}), s_j(\boldsymbol{g})\right) + h_j^D\left(s_i(\boldsymbol{g}), s_j(\boldsymbol{g})\right) = 0$.

Assumption 3 (Neutrality): $h_i^A\left(s_i(\boldsymbol{g}), s_j(\boldsymbol{g})\right) = h_i^D\left(s_j(\boldsymbol{g}), s_i(\boldsymbol{g})\right)$.

Basically, Assumption 1 implies that the higher an agent's own fighting strength or the lower his opponent fighting strength, the higher that both the agent's attack payoff and defense payoff are. Assumption 2 simply means that the payoffs from an attack are internalized and what

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

the winning side receives is exactly equal to what the other side loses, except that the attacking side has to incur a cost of c. Assumption 3 means that an agent's payoff in a fight does not depend on whether an agent is a receiver or an initiator of the negative link.

There is no direct payoff 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

$$u_i(\boldsymbol{g}) = \sum_{j \in N_i^-(\boldsymbol{g}) \setminus N_i^{e-}(\boldsymbol{g})} h_i^D(s_j(\boldsymbol{g}), s_i(\boldsymbol{g})) + \sum_{j \in N_i^{e-}(\boldsymbol{g})} h_i^A(s_i(\boldsymbol{g}), s_j(\boldsymbol{g})) - e_i(\boldsymbol{g})c.$$

Note that by Assumption 3, the above expression simplifies to

$$u_i(\boldsymbol{g}) = \sum_{j \in N_i^-(\boldsymbol{g})} h_i^A(s_i(\boldsymbol{g}), s_j(\boldsymbol{g})) - e_i(\boldsymbol{g})c.$$

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.

For simplicity as well as consistency with our experimental implementation, we additionally assume that the attack payoff function is linear in the difference between two sides' strengths derived from their friendships. We also normalize each player's intrinsic strength by setting $\lambda = 1$, without loss of generality. k thus represents the gross payoff benefit of implementing an attack, from having one unit of additional strength compared to the target of the attack.

Assumption 4 (Linearity): $h_i^A(s_i(\boldsymbol{g}), s_j(\boldsymbol{g})) = k\left(s_i(\boldsymbol{g}) - s_j(\boldsymbol{g})\right)$, where $k > 0$.

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 be either friends or enemies. One consequence resulting from this more flexible setup, however, is that the set of Nash equilibria is much larger compared to that in Hiller (2017). Therefore, in our analysis we also consider three equilibrium refinement criteria: pairwise stability, no pairwise profitable deviation condition, and no 3-person profitable deviation condition, 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 $\bar{g}_{i,j} = 1$, either $u_i(\boldsymbol{g}^*) \geq u_i(\boldsymbol{g}^* + \boldsymbol{ij})$ or $u_j(\boldsymbol{g}^*) \geq u_j(\boldsymbol{g}^* + \boldsymbol{ij})$ or both.

Condition 2 (No Pairwise Profitable Deviation): Equilibrium $\boldsymbol{g}^*$ is robust to pairwise profitable deviation if for any agent pair (i, j), for any $(\boldsymbol{g}_i', \boldsymbol{g}_j') \neq (\boldsymbol{g}_i^*, \boldsymbol{g}_j^*)$, either $u_i\left(\boldsymbol{g}_i', \boldsymbol{g}_j', \boldsymbol{g}_{-(i,j)}^*\right) \leq u_i(\boldsymbol{g}^*)$ or $u_j\left(\boldsymbol{g}_i', \boldsymbol{g}_j', \boldsymbol{g}_{-(i,j)}^*\right) \leq u_j(\boldsymbol{g}^*)$ or both, 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 $(\boldsymbol{g}_i', \boldsymbol{g}_j', \boldsymbol{g}_k') \neq (\boldsymbol{g}_i^*, \boldsymbol{g}_j^*, \boldsymbol{g}_k^*)$, either $u_i\left(\boldsymbol{g}_i', \boldsymbol{g}_j', \boldsymbol{g}_k', \boldsymbol{g}_{-(i,j,k)}^*\right) \leq u_i(\boldsymbol{g}^*)$, or $u_j\left(\boldsymbol{g}_i', \boldsymbol{g}_j', \boldsymbol{g}_k', \boldsymbol{g}_{-(i,j,k)}^*\right) \leq u_j(\boldsymbol{g}^*)$ or $u_k\left(\boldsymbol{g}_i', \boldsymbol{g}_j', \boldsymbol{g}_k', \boldsymbol{g}_{-(i,j,k)}^*\right) \leq u_k(\boldsymbol{g}^*)$ or any combination of the above three inequalities holds, where $\boldsymbol{g}_{-(i,j,k)}^*$ represents the agent other than i, j and k's equilibrium strategy profile.

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 regarding the robustness of an equilibrium with regard to subsets of individual players' entire network strategies. 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.[7]

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 $k < c < 2k$. Appendix A provides a detailed depiction of all Nash equilibria, and equilibria that survive our equilibrium refinements for attack cost levels $c < k$ and $k < c < 2k$, respectively. Here we summarize the main findings.

In general, attacking another player may be profitable as long as one has successfully formed an alliance with at least one other player. However, one must balance the potential benefits

[7] 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, as shown in Appendix A, 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.

of attacking another player with the benefit of forming an alliance with them, which will serve as a mutual aid in attacking another player. Among the set of equilibria, two emerge as the most prevalent empirically. In the *Peace* equilibrium, all possible links are positive. In the *Bully* equilibrium, three agents reciprocate friendly links with each other and each extends a rival link to the fourth agent. These two equilibria are illustrated in Figure 1.[8]

In the current setup, the Peace and Bully equilibria coexist. For $c < k$, the Peace equilibrium is pairwise stable but not robust to pairwise profitable deviation; the Bully equilibrium is robust to both refinements. For $k < c < 2k$, both equilibria are pairwise stable and robust to pairwise profitable deviation.[9]

(a) Peace equilibrium (b) Bully equilibrium

Figure 1: 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. The above depicted network structures account for over 95% of *Peaceful* (absence of rival links) groups and *Bullying* (3 in alliance against 1) groups in our study. Hence, for clarity, we focus our discussions primarily on these two structures, and refer the reader to the Appendix for the full set of equilibrium conflict/non-conflict outcomes.

From the equilibrium refinement Conditions above, we obtain the following Propositions, which serve as the basis for our testable Hypotheses in the experiment.

Proposition 1 (Prevalence of Peace and Bully)*: For $c < k$, any pairwise stable equilibrium is either a Bully equilibrium or an equilibrium without any rival link. For $k < c < 2k$, any pairwise stable equilibrium robust to pairwise profitable deviation is either a Bully equilibrium or an equilibrium without any rival link.*

Note that as the analysis in Appendix A shows, there are several other equilibria besides the Peace equilibrium and Bully equilibrium specifically illustrated in Figure 1. Among them, some of the equilibria are also neither classifiable as the bullying "outcome" (exactly the same as the Bully equilibrium among the set of equilibria), nor a peaceful "outcome" (which for clarity,

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

[9] Note that the Peace equilibrium is not robust to 3-person profitable deviation. In that sense, the Bully equilibrium is more robust than the Peace equilibrium. There exist other "peaceful" equilibria in which no negative link is made while some pairs have no relationship. However, none of the peaceful equilibria are robust to 3-person profitable deviation, while some peaceful equilibria are robust to pairwise profitable deviation at higher cost levels of attack. See Appendix A for details.

we refer to in the above proposition as any equilibrium without a rival link). Proposition 1 informs us that among all the equilibria of this game, those that are classifiable as either bullying or peaceful in outcome are more robust. Furthermore, for the higher range of cost levels ($k < c < 2k$), the equilibrium selection criteria to deliver such outcomes are more stringent (additionally robust to pairwise profitable deviation) than for the lower cost levels ($c < k$).

Proposition 2 (Across-cost Comparison)*: The Peace equilibrium is more robust for* $k < c < 2k$ *than for* $c < k$*. The Bully equilibrium is equally robust for* $k < c < 2k$ *and for* $c < k$*.*

Proposition 2 presents the comparative statics analysis result on equilibrium robustness with respect to cost level. When the cost is relatively low (c < k), Table A2 in Appendix A shows that the Peace equilibrium is pairwise stable, neither robust to pairwise profitable deviation, nor to 3-person profitable deviation. When the cost is relatively high ($k < c < 2k$), Table A1 in Appendix A shows that the Peace equilibrium is pairwise stable, robust to pairwise profitable deviation, but not robust to 3-person profitable deviation. This means that the Peace equilibrium becomes more robust in the sense that Condition 2 is satisfied as the cost increases from below k to above k. In contrast, the Bully equilibrium satisfies all three robustness conditions for both low-cost level and high-cost level.

Proposition 3 (Across-equilibria Comparison)*: The Bully equilibrium is in general more robust than the Peace equilibrium, both for* $c < k$ *and for* $k < c < 2k$*.*

Proposition 3 provides a comparison of robustness level between the Bully equilibrium and the Peace equilibrium across different cost levels. Based on the results in Appendix A and the discussion in the previous paragraph, we can immediately arrive at the conclusion that the Bully equilibrium is more robust than the Peace equilibrium at both low-cost levels and high-cost levels, while the difference in robustness between these two equilibria decreases as the cost increases from below k to above k.

4. Experimental Design and Implementation

4.1 Basic setup

We implement a real-time experimental design in which subjects can freely make links to each other, and both the network structure and momentary payoff implications are updated in real-time (Khavas et al. 2018; Goyal et al. 2017).[10] Within each round, participants can freely adjust their linking decisions for a period lasting between 75 and 105 seconds and ending at an unknown moment, and this random termination feature is known by all subjects. The random termination

[10] 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 real-time design can those help to facilitate coordination.

design is used to minimize potential end-game effects so that participants are less likely to change decisions at the last few seconds.[11]

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 real-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 subjects.

Figure 2 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).

[11] In fact, as our later analysis shows, participants tended to settled down into their final network structure relatively quickly and rarely utilized the entire allotted time.

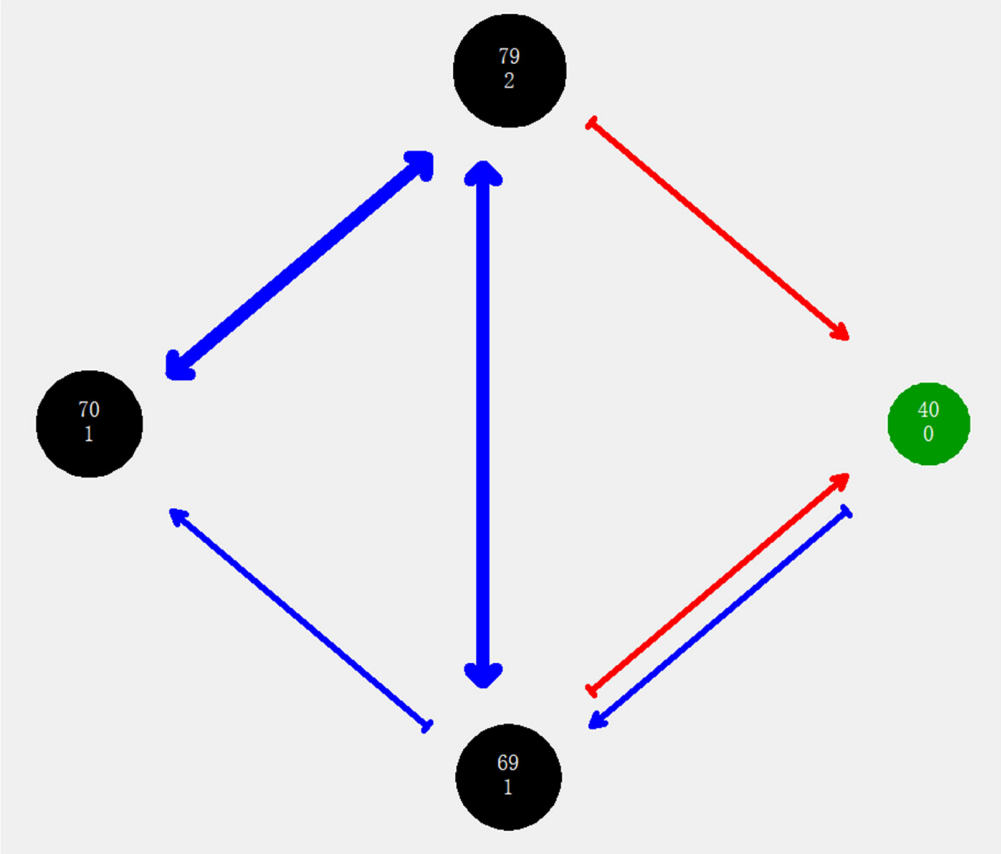


Figure 2: Screenshot Example of Decision Screen, Cost = 11

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 participant's screens. The number on top of each circle indicates that participant's current points. A larger circle indicates that that participant has more points. The

bottom number in each circle indicates the number of effective pairwise alliances that a participant currently has.[12]

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, the parameter k from the model is 10, and the strength function $s_i(\boldsymbol{g})$ is solely determined by the number of effective friendships n_i that player i has. 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

We use a within-subject design with five treatments, which in accordance with our theoretical hypotheses, vary in terms of the cost of forming negative links, the specific cost levels motivated by an earlier pilot experiment we conducted.[13] In the experiment, the parameter k in the payoff function is set to 10 and the parameter c takes one of the five values $\{3, 5, 7, 9, 11\}$. Thus, for four out of five cost values, $c < k = 10$ and for one value, $c = 11 > k = 10$. The instructions carefully explain the rules and payoff calculations to subjects with the help of an example. The experimental instructions are provided in Appendix B.

Each session has 20 rounds of the network formation game. At the beginning of *each* round, participants are randomly matched into 4-person groups and are randomly assigned a position in the network depicted in the software, either upper, lower, left or right, which is shown on their screen. We label every four rounds as a block, making a total of five blocks in a session. Within each block, the cost to extend a rival link is the same for all group members. However, from block to block, the cost changes and takes one out of five different values without repetition: 3, 5, 7, 9, 11. To ensure every participant receives all the different costs/treatments while minimizing any order effects, we conduct five experimental sessions using five different ordering sequences of the cost generated by a Latin square. Participants receive an endowment worth 70 points in each round, from which their link formation choices will be added or subtracted, depending on the outcome of

[12] 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.

[13] The choice of cost levels was partly motivated by a pilot experiment consisting of two sessions. In the pilot, we implemented a similar identical setup as the main experiment but with the cost of extending a rival link constant throughout a session. Another difference was that the group matching and subjects' spatial position were only randomized across blocks but unchanged within each block. One session has a cost of 3 and the other session has a cost of 11. The results show that almost all groups in the session with the cost of 11 reached Peace. We thus conjectured that for any cost above 11 there would be little scope for observing interesting dynamics leading separately to Bully and Peace. On the other hand, the session with the cost of 3 shows that Bully was reached most of the time. Thus, the main design was implemented under the suggestion that cost levels between 3 and 11 are more likely to provide a good amount of scope for observing treatment effects of cost levels on final networks or dynamics.

the game. At the end of the experiment, one block is randomly selected for each participant who receives the accumulated payment across the four rounds in that block.

4.3 Implementation procedure

The experiment was conducted in November and December of 2019 at the Nanjing Audit University Economics Experimental Lab with a total of 84 university students, using the software z-Tree (Fischbacher 2007).[14] Either 16 or 20 students participated in each session of 20 rounds of the network formation game, with randomly re-matched partners in each round.

Participants were randomly seated at a partitioned computer terminal upon arrival. The experimental instructions were provided to subjects 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 minutes per session was dedicated to ensuring comprehension. At the end of the experiment, participants completed a short survey concerning their demographics and strategies they used in the game.

For every 5 points earned in the subject's randomly selected block, subjects 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 minutes with average earnings of 69.7 RMB, including a show-up fee of 15 RMB.[15]

4.4 Hypotheses

As suggested earlier, throughout the empirical analysis, we define a group as being *Peaceful* if there is no rival relationship in the network. Being *Peaceful* not only includes the situation described specifically by the Peace equilibrium in which all group members are mutual friends, but also situations in which some of the group members extend friendly links while others do not. Note that the empirically defined set of *Peaceful* networks includes equilibria without any rival link (as mentioned in Proposition 1) as well as non-equilibrium outcomes without any rival link. In this sense, *Peaceful* empirically is defined more generally as the absence of any conflict. However, note that there is a very high correspondence (95.2%) between empirically Peaceful networks in the data, and the formal Peace equilibrium.

We define a group as *Bullying*, which includes the specific situation dictated in the Bully equilibrium, that is, three members form an alliance and all three attack the fourth member (who will be referred to as the "final victim"). *Bullying* networks also include a situation in which the targeted subject counter-attacks against one of the allied members, which, as shown in the next section, represents a very small number of cases but nevertheless is consistent in essence with a bully situation. The situation of the exact Bully equilibrium corresponds to our empirical definition of *Bullying* 96.9% of the time.

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

[15] 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.

We can derive three empirically testable hypotheses regarding *Peaceful* and *Bullying* outcomes from our theoretical Propositions 1, 2 and 3, in the context of our experimental setup as follows.

Hypothesis 1*: The Bullying or Peaceful networks are prevalent, given the robustness of these two equilibrium configurations as established in Proposition 1.*

Hypothesis 2*: The relative frequency of Bullying to Peaceful networks is higher when $c < 10$ than when $c = 11$, following from Proposition 2.*

Hypothesis 3*: The likelihood of reversion from Bullying to Peaceful networks is lower than the likelihood of reversion from Peaceful to Bullying networks during the continuous-time play, based on the relative robustness result in Proposition 3.*

Propositions 1 and 2 directly imply Hypotheses 1 and 2, respectively. The intuition of Hypothesis 2 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 $c = 11$. Hypothesis 3 is derived from Proposition 3 which predicts that the Bully equilibrium is generally more robust than the Peace equilibrium. Therefore, if a group temporarily coordinates on a *Peaceful* network, it is still likely that some players might jointly find it more profitable to deviate to a *Bullying* network. However, there is no similar incentive for them to revert to a *Peaceful* network once they settle upon a *Bullying* network.

The dynamics of the real-time decision environment of the game are complex, and our existing theory does not provide specific guidance as to the dynamics of alliance formation. Hence, for the analysis of dynamic network formation, we take an explorative approach to examining how players coordinate and attack, especially how a common enemy in a *Bullying* network emerges and how an alliance emerges and grows. In Section 6, 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 subsections. First, we present the final networks formed, and evaluate the network results with respect to the Hypotheses of the previous section. Then, in the second subsection, we analyze the dynamics of how the final networks were reached.

5.1 Final Network Formation

Figure 3 shows the frequency of the main final network types under each cost level. In total, we have 420 network observations and find two major categories of final networks corresponding to the Peaceful and Bullying networks defined in Section 4.4. Peaceful networks (networks without any rival links) represent 44.5% (187/420) of all cases, among which 95.2% correspond exactly to the Peace equilibrium (all players are mutual friends). Bullying networks (three in alliance against one) represent 46.4% (195/420) of all cases, among which 96.9% correspond exactly to the Bully equilibrium (see Figure 1).

All other final networks account for the remaining 9.1% of all cases.[16] Thus, consistently with Hypothesis 1, Peaceful and Bullying networks which largely correspond to the specific Peace and Bully equilibria, emerge as the two dominant final networks.

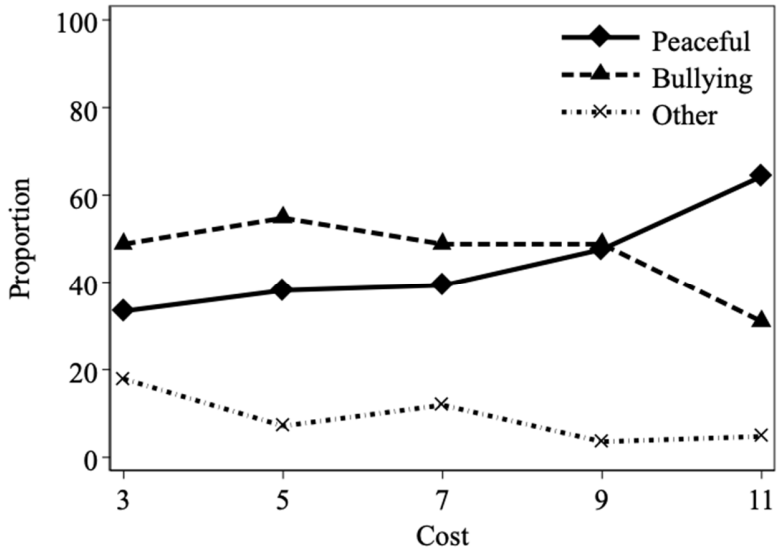


Figure 3: Proportions of Final Network Types

Note: This figure shows the final network configuration of all 420 groups of all cost levels.

In terms of the speed with which the final network was reached, on average it took groups 35.8 seconds (s.d.: 30.5, min: 2.6, max: 104.6; median: 26.4) to settle upon a final network. Within each final network category, it took an average of 21.4 seconds to settle on Peaceful networks, 43.5 seconds for Bullying networks, and 67.7 seconds for Other. Given that the average duration of each round is around 90 seconds, most groups were able to stabilize on a specific network about halfway through a round.

Our data also support Hypothesis 2, which predicts that the relative frequency of Peaceful and Bullying networks is similar for costs below 10 and has a spike at the cost of 11. Using Figure 3, we calculate that the ratio of the numbers of Peaceful and Bullying networks is 0.68, 0.70, 0.80 and 0.98 for the costs of 3, 5, 7 and 9, respectively, and the difference in the joint distribution of these four treatments is marginally significant ($\chi_2(3) = 7.139$, $p = 0.068$). However, the ratio

[16] These other networks include a specific configuration we call Kite (due to the shape formed by the network structure), representing 5.7% of all cases. Kite differs from Bully only in that one ally befriends both two other allies and the fourth member, although the fourth member is still a rival to the other two allies. 21 out of 24 cases are exactly Kite as described. In one out of 24 cases, the fourth member also extends a rival link to one of the allied members. In two out of 24 cases, the fourth member only receives one rival link from one of the allied members and no other member receives any rival link. Figure C1 in Appendix C shows the Kite specification and the rest of 14 uncategorized networks.

increases to 2.08 for the cost of 11, and the difference in the joint distributions of all five treatments is significant at the 0.1% level ($\chi_2(3) = 132.412, p < 0.001$).[17]

Table 1 also reports estimates from a Probit model that regresses a binary dependent variable for whether the final network is Peaceful (=1) or Bullying (=0) on cost level dummies and round-fixed effects. The estimates confirm that the relative frequency of Peaceful over Bullying networks is not significantly different for costs below 10, but is significantly higher for the cost of 11.[18]

Table 1: Probit Model: Treatment Effects on Final Networks

	Average marginal effects	Standard error
Cost = 5	0.003	0.067
Cost = 7	0.051	0.089
Cost = 9	0.068	0.065
Cost = 11	0.256***	0.093
N	382	

Note: The dependent variable is whether the final network is Peaceful (=1) or Bullying (=0). The cost of 3 serves as the benchmark. The regression also includes round dummies which show trends toward more Bullying outcomes over time. Standard errors are clustered at the session level. *** p < 0.01.

Finally, Hypothesis 3 proposes that Bullying networks are more robust than Peaceful ones. In other words, once a group is Peaceful it is still likely to converge to a Bullying situation, but not the other way around. This is well-reflected in the data. For the ease of interpretation, among the Peaceful and Bullying networks characterized earlier, for the analysis of Hypothesis 3, we only counted *exact* Peace and *exact* Bully as depicted in Figure 1, which arose in some intermediate stage.[19]

The data show that Peace networks arose in 206 groups at some point within a round; however, 14.1% (29/206) of these groups ended up in Bully networks as their final network structure. By contrast, Bully networks arose in 207 groups at some point during the round; but only 2.4% (5/207) 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

[17] The results from a pilot experiment show an even starker difference in the Peaceful/Bullying ratio. The session with the cost of 3 has the ratio of 0.14 (=9/66) and the session with the cost of 11 has the ratio of 9.9 (=69/7). Compared to the results from the main experiment, the lack of experience with different cost levels probably explains the more extreme results in the pilot.

[18] Technically speaking, the regression shows that costs of 5, 7 and 9 are not significantly different from the cost of 3, while the cost of 11 is significantly different from the cost of 3. We also ran the same regression using different cost levels than 3 as the benchmark and obtained similar results.

[19] We focus on exact Peace in testing Hypothesis 3 to help preclude the default peaceful state which is merely each group's starting point.

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 1, the coefficients on the round-fixed effects show that Peaceful outcomes are generally less likely as the number of rounds played increases, and in particular, significantly so among the last played rounds.[20] This suggests, as a cross-round consequence of Hypothesis 3, that Bullying is enhanced by learning and experience.

The following results regarding final network structures are consistent with Hypotheses 1, 2 and 3:

Result 1: *i) Over 90% of the time, groups reached Peaceful networks or Bullying networks; ii) the relative frequency of Peaceful and Bullying networks was similar for costs below 10 and increased significantly for the cost of 11; iii) Groups were more likely to revert from a Peaceful network to a Bullying network than from a Bullying network to a Peaceful network.*

5.2 Dynamics of Network Formation

We now turn to analyzing the dynamic processes of groups in reaching Bullying and Peaceful networks. In particular, whether there are factors at any critical juncture in the network formation process that could be influential in terms of whether a Bullying or Peaceful network is ultimately reached. In the following analyses, we pool the data together from all cost levels, under the premise that groups reaching each type of final outcome do not differ in their dynamic network formation based on the cost level. In Appendix D, we report analyses separately for each cost level, showing that the patterns described are in fact highly consistent across cost levels.

First, we examine the overall linking activity of participants playing the game. Figure 4 plots the activity levels of extending any type of link (friendly or rival) to each other per group in each second of the continuous real-time decision environment. It shows that the participants are highly active at the very beginning, especially in making friends. The overall activity rate plummets passing the tenth second. Figure C2 in Appendix C plots the activity for finalized Bullying and Peaceful groups separately, and shows a similar pattern in terms of concentrated link formation in the first few seconds. This pattern indicates that groups successfully coordinate on a particular network very early on in a round.

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

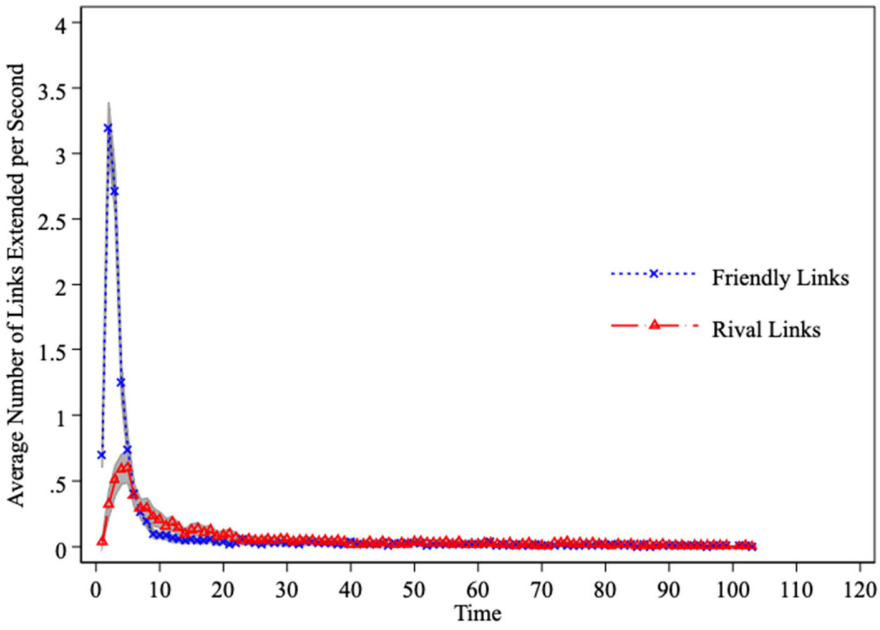


Figure 4: Extension of Links per Group, by Second

Note: The grey shaded area indicates 95% confidence intervals.

5.2.1 How do the dynamics in Bullying and Peaceful groups differ?

Next, we examine whether the patterns of extending friendly and rival links differ between Bullying and Peaceful groups. Figure 5 plots the average number of effective friend links formed over time. It shows that the speed of befriending looks remarkably similar for Bullying and Peaceful groups within the first five seconds.

However, passing that moment, the two sets of groups diverge and move decisively toward very different network patterns. While players in Peaceful groups quickly tend to become fully connected with each other, those in Bullying groups steadily form one and only one three-member alliance. In fact, within the first five seconds, the divergence in alliance formation is apparent. Table 2 shows that while the percentage of three-member alliances notably increases over time in Bullying groups, it is relatively low and declining in Peaceful groups starting from the 3rd second.[21] On the contrary, the percentage of fully connected networks notably increases over time in Peaceful groups but remains relatively low and stable in Bullying groups also starting from the 3rd second.

Given these patterns, our key question is why some groups converge to Bullying networks while others manage to reach Peaceful networks? The answer lies at least partly in group members' success in coordinating on a common rival. Figure 6 shows the maximum number of rival links

[21] Figure C4 in Appendix C plots the percentage of exactly one three-member alliance separately for Peaceful and Bullying groups over all seconds, confirming that the stable frequency of the three-member alliance over time.

received by any player in a group, averaged across all groups. Bullying and Peaceful groups diverge almost immediately in their pattern of making rivals by this measure. In Bullying 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 2.2 by the 10th second. It means that potential victims have already received more than 2 attacks on average by the 10th second, they were in turn very likely to be the final target in Bullying groups. By contrast, in Peaceful groups a player rarely ever receives more than one attack throughout the entire round.

Similarly, when only considering groups in which attacks ever occur, in Bullying groups a first victim (the player who receives the first ever rival link from another player) on average receives the second attack within 10 seconds, whereas in Peaceful groups a first victim rarely receives the second attack (see Figure C3 in Appendix C). Later we show that first victims account for the majority of the final victims (the player who is the final target in a Bullying group). Thus, Peaceful groups who have ever had an attack effectively pass on the opportunity to coordinate on this salient target. Taken together, these patterns show that Bullying groups can quickly coordinate on a common rival whereas Peaceful groups fail to (or perhaps do not attempt to) coordinate on a common rival, especially on the first victim.

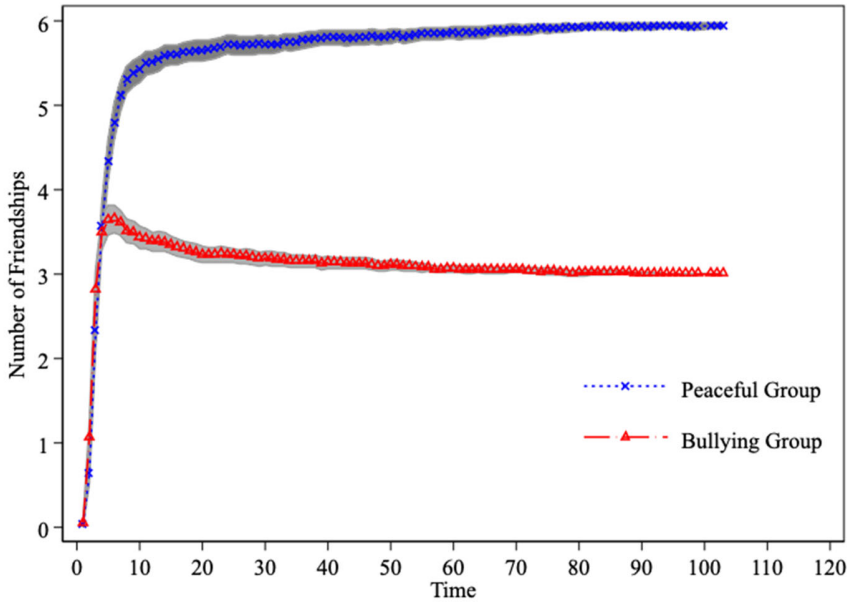


Figure 5: Evolution of Effective Friendships per Group

Note: This Figure shows the average effective friendship links formed in each second over time. If four players are all mutual friends, there are six effective friendship links in the group (Peaceful groups); if three out of the four players are mutual friends and the other player is a lone player, there are three effective friendship links in the group (a necessary condition for Bullying groups). The grey shaded area indicates 95% confidence intervals.

Table 2: Percentages of 3-member Alliances and Fully Connected Networks, First 10 Seconds

Time (seconds)	Peaceful		Bullying	
	3-member alliance	Fully connected	3-member alliance	Fully connected
1	0	0	1.5	0
2	17.1	2.1	22.6	2.6
3	23.5	16.6	34.9	4.6
4	19.8	33.7	46.9	6.7
5	17.1	48.1	53.3	6.2
6	11.8	54.5	59.0	5.1
7	6.4	61.0	64.6	5.1
8	9.1	67.4	67.2	5.6
9	7.0	70.1	71.8	5.1
10	7.0	72.2	74.4	4.1

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.

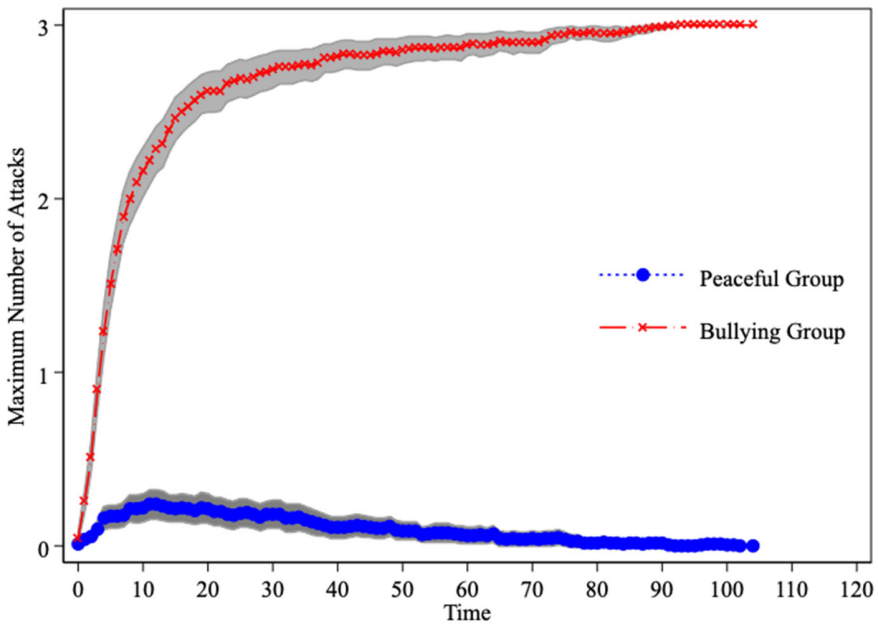


Figure 6: Evolution of Maximum Number of Attacks Received by Any Player per Group

Note: This Figure shows the maximum number of attacks received by any player in a group. By definition, no player receives attacks at the end in Peaceful groups, and a player (final victim) receives 3 attacks at the end in Bullying groups. The grey shaded area indicates 95% confidence intervals.

To provide some statistical evidence on factors that might explain the divergent paths of Bullying and Peaceful networks, we also implement a series of group-level Probit regression analyses with a binary dependent variable of whether a group eventually converges to Bullying or Peaceful networks. We define several temporary network structure states observed in the first few seconds and utilize them as the independent variables. We note that this particular analysis should be taken with caution as all network temporary patterns may be endogenous to some unobserved behavior within the game. Nonetheless, the findings can be informative as to how early network states lead to the formation of the final type of network from a predictive standpoint.

The detailed Tables and discussions are relegated to Appendix E, but summarized here for convenience. First, the analysis shows that the occurrence of a single three-member alliances in the first few seconds strongly predicts the Bullying network, whereas the occurrence of being fully connected via friendly links (or a universal alliance) in the first few seconds strongly predicts the Peaceful network. Second, the incidence of at least one player receiving one or two attacks in the very first seconds strongly predicts the Bullying 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. These results confirm our earlier descriptive observations that the network patterns occurring in the first few seconds already predict the type of final network formed.

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

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

5.2.2 Coordinating on a final victim

We now explore in further detail the dynamic process of reaching Bullying networks. We jointly consider two related processes: the process of coordinating on a final victim and the process of forming a three-member alliance. To this end, in Figure 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 Bullying groups. The Figure shows three clear patterns. First, the final victim receives attacks very early in a round: on average they receive 1.2 enemy links by the 5^{th} second and 2.1 enemy links by the 10^{th} second. Second, the other three players form an alliance even faster: 46.9% of the groups form a three-member alliance within 5 seconds, while 71.1% of groups have done so within 10 seconds.

The third pattern is that the formation of a three-member alliance generally precedes the emergence of the final victim, evidenced by Figure 7. The alliance forms before the final victim receives the first attack in 54.8% (107/195) of the groups; the proportion of three-player alliances

increases to 82.6% (161/195) and 93.8% (183/195) before the final victim receives the second and third attacks, respectively.[22] Figure C5 in Appendix C shows a similar pattern using the median instead of average. Overall, in these groups that reach Bullying networks in the final configuration, alliance formation is swift and precedes targeting the final victim.

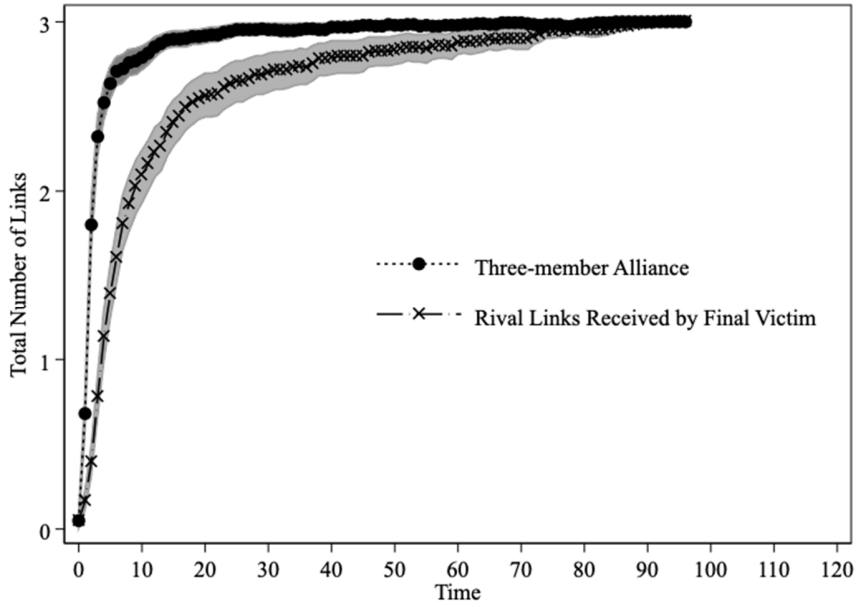


Figure 7: Evolution of Average Attacks Received by the Final Victim and Average Effective Friendships Among the Other Three Players, Bullying Groups

Note: This Figure shows dynamics in Bullying groups (195 out of 420 groups). It plots the average number of enemy links received by the final victim and average number of effective friendships formed by the other three players (except for the final victim). The grey shaded area indicates 95% confidence intervals.

Now we 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 possible that the player who initiates the first attack (referred to as the initiator) could also be a target for coordination for similar reasons of salience, although it could require additional steps of coordination to successfully adjust the collective target.

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 types of networks to obtain a broad picture of how final victims are coordinated upon. Figure 8 shows that among

[22] There are in total 197 final victims. Two cases do not correspond to the exact Bully: the three members excluding the final victim were not fully connected. These two cases are not included in the analysis.

these 303 first victims, 65.0% (197/303) of them receive two attacks subsequently; then 50.8% (154/303) receive three attacks subsequently; and finally 47.5% (144/303) become the final victims. Overall, 47.5% of the first victims are also the final victims, despite the fact that the choices of the first victims are almost at random.[23] By contrast, among the 303 attack initiators, 11.9% (36/303) of them are the final victims. Among 1074 other players, the proportion of final victims is negligible 0.7% (7/1074). Thus, the proportion of first attackers who eventually become final victims is still notable and non-negligible. Overall, first victims account for 73.1% (144/197) of all final victims (see footnote 22 for the explanation of two cases of final victims that do not belong to Bullying groups), implying that the coordination on final victims is mostly path-dependent.[24]

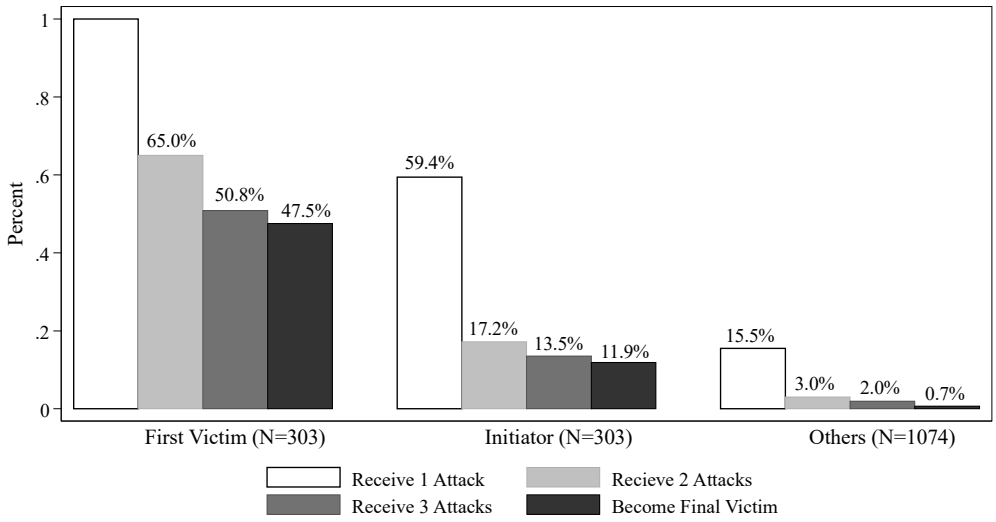


Figure 8: Transition to Final Victimhood

Notes: This Figure includes all 420 groups with a total of 1680 players. In 303 groups, there is ever a first victim. Correspondingly, there are 303 initiators of these first victims. "% Receive 2 (3) attacks" means whether the

[23] In our experiment, group members were spatially located on the screen, which may naturally lead to the question of whether any member was more likely to be the first victim based on spatial position? The player on the right corner was slightly more likely to be chosen as the first victim. 21.5%, 22.1%, 23.8% and 32.7% of first victims were located in the upper, lower, left and right positions, respectively, and the distribution is significantly different from a uniform distribution ($\chi^2(3) = 9.858$, $p = 0.020$). We do not have a theoretically-grounded explanation for this statistically significant result. One possibility is that the green circle which represents self is positioned on the right in the screenshot used in the instructions. This could have primed participants to attack the player on the right despite the fact that a player' position is randomized across rounds. Another possibility is that clicking on the right is more convenient for right-handed subjects. An initiator also seems to be more likely to attack the player diagonally to his own position rather than one of his neighbors: Diagonal players were attacked in 40.6% (=123/303) of the time, significantly higher than the expected 33.3% ($\chi^2(1) = 7.188$, $p = 0.007$). We also checked whether the spatial position of the first victim affected the probability of becoming the final victim. First victims positioned in the upper, lower, left and right positions had 41.5%, 50.7%, 44.4% and 51.5% of becoming final victims, respectively. However, the percentage of final victims are not significantly different across the four positions ($\chi^2(3) = 2.119$, $p = 0.548$).

[24] 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).

percentage of players who ever receive 2 (3) attacks during the whole round. "% final victim" means the percentage of players who become final victims.

In the next section we examine the determinants of final victimhood and find that both being a first victim and being an initiator of the attack significantly predict final victimhood.

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 by the initial victim can reduce the chance of being the final victim.

We implemented individual-level random effects Probit regressions with final victimhood as the binary dependent variable. As explanatory variables, we include whether the player is a first victim, an initiator, with other player types as a comparison group. Table 3 reports the average marginal estimates for all groups and for Bullying groups, separately. Columns (1) and (5) show that both being a first victim and being an initiator (compared to other types) strongly predict being the final victim. Furthermore, F-tests indicate that a first victim is significantly more likely than an initiator to be the final victim in all specifications ($p < 0.001$). These regression results are consistent with our previous descriptive statistics.[25]

There are some possible intuitive strategies that victims can use to escape from being bullied, and our analysis can inform us about whether these approaches are effective or not. One frequently observed strategy in our data is that a first victim quickly attacks back to the initiator so that other players might have difficulty in coordinating on a victim. Overall, we find 90 (out of 303) cases of first victims counter-attacking within five seconds, resulting in pairs of a first victim and an initiator with mutual rival links.[26] However, columns (2) and (6) in Table 3 show that this strategy is ineffective in reducing the chance of being a final victim in Bullying groups: column (2) on the full data shows that those first victims who attacked back within 5 seconds were actually *more* likely to be final victims overall, while column (6) shows that this strategy has no significant influence when focusing exclusively on Bullying groups.

We also examine whether first victims making more effort to befriend other players can help them to escape victimhood. Estimates from columns (3) and (7) in Table 3 do not support this conjecture: above-median level of befriending activity (measured by the number of extending and retracting friendly links to any other player in a group) after being attacked is in fact positively

[25] Further regression analyses also show that those who are first victims in the previous round are more likely to be initiators, presumably attempting to escape from being bullied again (see Table C5 in Appendix C). Interestingly, being final victims in the previous round is not significantly correlated with being initiators in the next round, presumably because being final victims and first victims are highly correlated.

[26] There are 66 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.

related to the likelihood of becoming final victims. However, it is also possible that at the point of being attacked, the victim's effort to develop further friend links were ineffective. However, it is also possible that first victims' typical statuses in the network at the point of being attacked are already such that attempts to develop further friend links are rendered less effective.

Table 3: Random Effects Probit model: Determinants of Final Victims

	All groups				Bullying groups			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
First victim	0.270***	0.252***	0.211***	0.272***	0.453***	0.456***	0.339***	0.464***
	(0.031)	(0.031)	(0.021)	(0.032)	(0.010)	(0.023)	(0.009)	(0.011)
Initiator	0.121***	0.121***	0.115***	0.120***	0.158***	0.158***	0.148***	0.156***
	(0.018)	(0.018)	(0.016)	(0.018)	(0.035)	(0.035)	(0.034)	(0.035)
First victim (attack back)		0.054***				-0.008		
		(0.019)				(0.035)		
First victim (befriending activity, above median)			0.098***				0.203***	
			(0.020)				(0.014)	
First victim (more friends than initiator)				-0.031***				-0.138***
				(0.011)				(0.021)
N	1680	1680	1680	1680	780	780	780	780

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

On the other hand, those first victims who at the moment of being attacked have more friends than initiators appear to be less vulnerable than those first victims who have fewer friends. Estimates from columns (4) and (8) in Table 3 support this conjecture: having more friends than the initiator significantly reduces the likelihood of a first victim becoming a final victim by 11% in all groups and reduces the likelihood of final victimhood by 30% in Bullying groups. By additionally interacting the dummy variable of whether first victims launched a counter-attack within five seconds, with the dummy variable indicating first victims having more friends, the coefficient reveals that it is precisely when first victims attacked back that having more friends increases their likelihood of escaping from being bullied. This effect is marginally significant only when we include all groups, but not when we exclusively examine Bullying groups. Thus, first victims' intuitive strategy of fighting back quickly may only work towards supporting peaceful outcomes when there are enough friends backing them up.[27] 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.

[27] The regression with this interaction term is omitted due to space considerations but is available on request.

In summary, despite any of the examined intuitive strategies to avoid being bullied, first victims are far more likely to be final victims than any other players. Their efforts to escape from being bullied at least from a statistical standpoint, are mostly futile. The only factor that appears to help is 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 Bullying 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.*

6. A Quasi-dynamic analysis of network dynamics

In the preceding section, we established several important dynamic patterns regarding coordinating on a final victim. However, it remains unclear why a player would initiate a first attack. In particular, based on empirical considerations, is it an optimal action for a player to initiate a rival link?

In our setup, an attack is theoretically an optimal action if the attacker has more friends than the first victim for costs lower than 10, and if the attacker had at least two more friends than the first victim for the cost of 11. However, we find that this is not the case most of the time in the data: the initiator of the attack has more friends than the first victim in only 28.1% of all cases; fewer friends in 8.6% of all cases; and same number of friends in an overwhelming 63.4% of all cases. This suggests that an attack was not initially motivated by a temporary payoff increase but more likely for coordination purposes. More broadly, for any attack from any player at any moment (N=1656), most of them (73.6%), are not for a temporary payoff increase either. Interestingly, however, the retraction of an attack (N=989) often increases a player's temporary payoff (86.5%). Still, over 50% of all actions related to attacking are not immediate best responses.[28]

To further examine the best response over a longer time horizon, we apply a logic similar to fictitious play and ask whether an initiator's attack is an optimal action given the empirical distribution of final victims at the end of the game. From the empirical distribution of being bullied (see Figure 8), players who attack first end up being bullied in 11.9% of the cases, while players who do not initiate the first attack end up being bullied in 11.0% of the cases.[29] Thus, an initiator's attack is still not a best response even if he can anticipate the true empirical distribution of being bullied. This is again in line with the coordination purpose of initiating an attack, which may not pay off for the initiator in the end. Finally, it should be clear that once the first victim receives the

[28] We do not find much evidence of treatment differences in this regard. If anything, when the cost is 11, the rate of best responses is slightly lower compared to the other cost levels.

[29] There are two ways to victimize a player if she does not initiate a first attack. First, she is a target of the first attack, who has 47.5% chance of being the final victim. Second, she is a target of non-first attack, who has 0.7% chance of being the final victim. Therefore, according to Figure 8, the probability for such a player to be the final victim is 47.5%*303/1377 + 0.7%*1074/1377 = 11.0%.

attack, other players have a strong incentive to follow the initiator to attack the first victim: while other players' chance of being bullied is almost negligible, the chance of the first victim to be the final victim is as high as 73.1% (47.5%/65.0%) once he receives the second attack. Therefore, it is significantly riskier to take on the role of first attacker if one can instead become an (unvictimized) follower. Thus, it is plausible that if subjects are sophisticated enough to realize this, the initiation of a first attack is likely intended as a coordination device for the other two non-victims.

The above discussion suggests that the initiator's attacking decision is not an empirically optimal action in either the short-run or the long-run. The observed pattern bears similarity to a volunteer's dilemma in which someone needs to pay a cost to volunteer while others can receive the benefit without paying any cost. However, the situation here differs from the volunteer's dilemma in that being the initiator carries a reduced risk of being bullied compared to being the first attacked.

In the remaining part of this section, we attempt to investigate the non-equilibrium dynamic interactions among players, with a particular focus on players' incentive to initiate an attack in an initial peaceful state. Given that we do not explore all possible dynamic patterns nor cover all possible initial states, we consider our analysis quasi-dynamic. The main purpose of this simple model is not to capture every player's strategy in real time accurately, but rather to highlight the coordination nature of the dynamics and hopefully inspire future theoretical research.

Given an initially peaceful state with 4 players, there are three possible final consequences, each of which can be supported as a reasonable equilibrium once an attack is initiated from player i to player j, with the other two players being bystanders n_1 and n_2. In Case 1, bystanders n_1 and n_2 successfully coordinate on supporting player i, resulting in a bullying outcome with player j as the final victim. In Case 2, bystanders n_1 and n_2 successfully coordinate on supporting player j, resulting in a bullying outcome with player i as the final victim. In Case 3, bystanders n_1 and n_2 cannot successfully coordinate and the attempt to initiate an attack fails, resulting in a peaceful outcome.

From the bystanders' perspective, since a successful coordination will lead to a payoff of $2k - c$ while a failed coordination provides a payoff of 0, players n_1 and n_2 have strong incentive to coordinate. However, given that decisions are made in a real-time setting, bystanders may not be able to achieve coordination in a timely manner. For simplicity, we assume that the two-bystanders make their decisions independently and simultaneously.[30]

Note that from a payoff perspective, the bystanders are indifferent between Case 1 and Case 2, both of which deliver a payoff of $2k - c$ to players n_1 and n_2. Suppose that the initiator i believes that each bystander attacks player j with probability μ (Case 1) and attacks player i with probability $1 - \mu$ (Case 2). For simplicity, not attacking is considered a dominated strategy and thus occurs with probability zero. Thus, the initiator i's expected payoff will be

[30] We have also considered an alternative setting where the two-bystanders make decisions sequentially. The main results remain the same and are available upon request.

$$u_i = \mu^2(2k - c) + (1 - \mu)^2(-6k) + \mu(1 - \mu)(0) = \mu^2(2k - c) + (1 - \mu)^2(-6k),$$

where the terms $2k - c$, $-6k$, and 0 stand for the initiator's payoffs in Case 1, Case 2 and Case 3, respectively.

The *individual rationality condition* for initiator i requires that $u_i \geq 0$, that is, the expected payoff by initiating an attack should be at least no less than staying in the initial peaceful state. This condition is equivalent to the following inequality based on the payoff function parameters in our game:

$$\mu \geq \mu^*(\tfrac{c}{k}) \equiv \tfrac{6-\sqrt{12-6\frac{c}{k}}}{4+\frac{c}{k}}.$$

A player may choose not to be an initiator even if initiating an attack is profitable. This can happen when letting someone else initiate the attack can potentially bring a higher payoff. By not initiating an attack, one can become a bystander, with 2/3 chance enjoying a higher payoff than being the initiator, while with 1/3 chance one may suffer from becoming the victim. Thus, the expected payoff for not initiating an attack will be

$$u_i^{'} = \frac{2}{3}[\mu^2(2k - c) + (1 - \mu)^2(2k - c)] + \frac{1}{3}[\mu^2(-6k) + (1 - \mu)^2(2k - c)].$$

The *incentive compatibility condition* for initiator i requires that $u_i \geq u_i^{'}$, which is equivalent to the following inequality:

$$\mu \geq \mu^{**}(\tfrac{c}{k}) \equiv \tfrac{3\left(8-\frac{c}{k}\right)-\sqrt{3\left(8-\frac{c}{k}\right)\left(16-\frac{c}{k}\right)}}{8-2\frac{c}{k}}.$$

We can show that $\mu^{*'} > 0$, $\mu^{*''} > 0, \mu^{**'} < 0$ and $\mu^{**''} < 0$. Also note that $\mu^*(0) = \min \mu^* > \max \mu^{**} = \mu^{**}(0)$, which means the incentive compatibility condition is always satisfied as long as the individual rationality condition is satisfied. We more formally state these results in the following proposition:

Proposition 4: *Given an initial peaceful state, the initiator's threshold belief on the bystanders supporting the initiated attack, denoted by μ^*, is an increasing and convex function of $\frac{c}{k}$.*

Proposition 4 implies that an increase in the cost of initiating an attack leads to an increase in the threshold belief level regarding bystanders following the initiator, which all else equal, makes the initiation of an attack less likely. In addition, such an effect strengthens as the cost of initiating an attack rises. When $c = 0$, $\mu^* = \frac{3-\sqrt{3}}{2} \approx 0.634$; when $c = k$, $\mu^* = \frac{6-\sqrt{6}}{5} \approx 0.710$; when $c = 2k$, $\mu^* = 1$, which implies that initiation of an attack is not an optimal action under any possible belief if $c = 2k$.

In summary, by linking initiators' decisions to their belief about bystanders' behavior, our quasi-dynamic model provides a simple account of why some players are motivated to initiate an attack. The convex relationship between the initiator's threshold belief and the attacking cost

(Proposition 4) is qualitatively consistent with the final networks observed in Figure 3, in that the higher the attacking cost, (the increasingly higher the threshold belief, and) the increasingly lower likelihood of a bullying outcome. It also appears to be consistent with our data showing that both the frequency of initiations of an attack (i.e., the proportion of initiators) and the likelihood of first victims becoming final victims tend to decrease with the attacking cost, especially when the cost increases from 9 to 11 (see Table D1 in Appendix D).

The quasi-dynamic model also provides an explanation for why first victims are most likely to be the final victim: since the value of threshold μ^* is always higher than 0.5 regardless of the attacking cost, an attack indicates that the initiator holds the belief that each bystander will join in attacking the first victim with more than 50% chance.

Furthermore, although we do not explicitly model the first victim's strategy and how that might influence the initiator's belief, the analysis provides a rationale for why first victims often fight back against initiators (probably as a strategy to escape from victimhood as discussed in Section 5.2.3). If the first victim fights back, the positions of the initiator and the first victim will become more symmetric. Since there is no payoff difference for bystanders between following the initiator and supporting the first victim, a more symmetric position should make the strategy of following the initiator less salient.

Finally, since the model assumes that the initiator will be the target of the bystanders with some probability (which is consistent with our earlier observation in Section 5.2.2), the expected payoff of a bystander is always higher than the expected payoff of the initiator. This prediction is also borne out in the data: among all groups in which attacking ever happened, initiators received on average 69.7, while bystanders earned on average 76.6. The difference is statistically significant at the 5% level using the Wilcoxon signed-rank test with session average as the unit of observation.

Overall, we view our simple quasi-dynamic model as a first step toward a better understanding of the rich dynamics observed in our network formation game. It would be valuable to explore further how to model players' coordination behavior more fully. For example, the real-time setup naturally calls for a model allowing for endogenous timing decisions by bystanders, whereby bystanders decide both when and whom to attack. An even more complete version should also endogenize the timing decision by the initiator given the observation that the initiator expects to earn less than bystanders.

We also do not necessarily consider coordination failure as the sole reason for not reaching a bullying outcome as the current model implies. The frequently observed peaceful outcome could be a result of players' other-regarding preferences such as inequality aversion. These preferences can lead players to prefer a Peaceful network in which everyone earns the same over a Bullying network in which a substantial payoff gap arises between the alliance members and the final victim.[31]

[31] Researchers have developed some dynamic models, albeit not micro-founded, on signed networks to explore how networks evolve to be structurally balanced. One typical dynamic is as follows (Antal, Krapivsky, and Redner 2006): (1) pick a random triad; if it is balanced do nothing; (2) if the triad has one rival link and thus is imbalanced, then the

7. Conclusion

While the origins of real-world conflicts can be complex and multifaceted, this paper studies 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 real-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, player(s) can choose to disrupt the peace, incurring a cost to direct a rival link at another player, such that the player with the greater number of alliance links obtains a portion of the rival's surplus based on the final network formation at the end of a round. Our theoretical analysis, which modifies the signed network formation game of Hiller (2017), shows that while Peace is an equilibrium, a 3-against-1 Bully configuration is also a highly robust equilibrium. These two equilibria together are reached over 90% of the time in the experimental data, and their relative frequencies depend on the cost of making an attack, as predicted by our theory-generated hypotheses.

Our experiment thus allows us to answer the question of to what extent conflicts arise among ex-ante homogeneous decision-makers with initially equal endowments. The results show that in a flexible network decision environment, even with ex-ante identical players, the tendency to generate socially costly conflict is substantial. Bullying was more common than Peace in all cost treatments for which our model predicted a higher Bullying to Peaceful ratio, occurring approximately 50% of the time. In the highest cost treatment, Peaceful outcomes were relatively more prevalent, occurring 64% of the time.

Our results also showed, consistent with our hypothesis, generated from the relatively higher level of robustness of the 3-against-1 Bully equilibrium, that the transition from Peaceful to Bullying was more prevalent than the transition from Bullying to Peaceful. The implication is that once Bullying emerges in a round, the Peaceful setting is disrupted and is subsequently difficult to restore.

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 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

---

rival link changes to a friendly link with probability p and the friendly link changes to a rival link with probability $1 - p$; (3) if the triad has three rival links and thus is imbalanced, then change a rival link to a friendly link. In our setting, since groups are able to reach the Bully or Peace equilibrium, both of which are structurally balanced, the forces that are present in the non-game-theoretical dynamics may also operate here. Our intuition, however, is that these dynamic forces are simplistic and unlikely to fully explain our dynamic data since they do not consider, for example, incentives related to the cost of attacking, which we have found to affect the final outcome.

course of a round, and furthermore, Bullying and Peaceful networks diverge sharply in their linking patterns early.

When examining Bullying situations specifically, alliance formation generally precedes coordination on a common rival, showing subjects' tendency to gather their 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. The 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 in the data of a first attacker becoming the final victim, we examine the trade-offs faced by potential first attackers based on the observations in the data. We then 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.

There are many potential interesting extensions to the signed network formation games used in the current study. Although our 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. 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.[32]

[32] 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

Another interesting direction for future research is to study large scale network formation. In our setting, a group with more than four players can sustain other forms of alliances such that more than one alliance with at least two members can emerge in equilibrium and larger alliances attack smaller ones (Hiller 2017). It will be interesting to see whether a certain network such as the Bully network can still stand out among all other equilibria, or whether other robust network patterns will emerge. Running large scale network formation games in the lab presents significant technical difficulties, but a new experimental platform recently developed by Choi, Goyal, and Moisan (2020) provides a promising toolkit to conduct network experiments to answer questions such as the ones we present here on a substantially larger scale.

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.

References:

Abbink, Klaus, and Gönül Doğan. 2019. "How to Choose Your Victim." *Games and Economic Behavior* 113 (January): 482–96.

Acemoglu, Daron, Georgy Egorov, and Konstantin Sonin. 2008. "Coalition Formation in Non-Democracies." *Review of Economic Studies* 75 (4): 987–1009.

Antal, T, P L Krapivsky, and S Redner. 2006. "Social Balance on Networks: The Dynamics of Friendship and Enmity." *Physica D: Nonlinear Phenomena* 224 (1–2): 130–36.

Baik, Kyung Hwan. 2016. "Endogenous Group Formation in Contests: Unobservable Sharing Rules." *Journal of Economics & Management Strategy* 25 (2): 400–419.

Bala, Venkatesh, and Sanjeev Goyal. 2000. "A Noncooperative Model of Network Formation." *Econometrica* 68 (5): 1181–1229.

Benenson, Joyce F, Henry Markovits, Melissa Emery Thompson, and Richard W Wrangham. 2009. "Strength Determines Coalitional Strategies in Humans." *Proceedings of the Royal Society B: Biological Sciences* 276 (1667): 2589–95.

Berninghaus, Siegfried K, Karl-Martin Ehrhart, Marion Ott, and Bodo Vogt. 2007. "Evolution of Networks—an Experimental Analysis." *Journal of Evolutionary Economics* 17 (3): 317–47.

Berninghaus, Siegfried K, Karl Martin Ehrhart, and Marion Ott. 2006. "A Network Experiment in Continuous Time: The Influence of Link Costs." *Experimental Economics* 9 (3): 237–51.

Bloch, Francis. 2012. "Endogenous Formation of Alliances in Conflicts." In *The Oxford Handbook of the Economics of Peace and Conflict*, edited by Michelle R Garfinkel and Stergios Skaperdas, 473–502. Oxford: Oxford University Press.

Bloch, Francis, Santiago Sánchez-Pagés, and Raphaël Soubeyran. 2006. "When Does Universal Peace Prevail? Secession and Group Formation in Conflict." *Economics of Governance* 7 (1): 3–29.

Burger, Martijn J, and Vincent Buskens. 2009. "Social Context and Network Formation: An Experimental Study." *Social Networks* 31 (1): 63–75.

Callander, Steven, and Charles R Plott. 2005. "Principles of Network Development and Evolution: An Experimental Study." *Journal of Public Economics* 89 (8): 1469–95.

Cartwright, Dorwin, and Frank Harary. 1956. "Structural Balance: A Generalization of Heider's Theory." *Psychological Review* 63 (5): 277–93.

Choi, Syngjoo, Sanjeev Goyal, and Frederic Moisan. 2020. "Large Scale Experiments on Networks: A New Platform with Applications." Cambridge-INET Working Paper Series No: 2020/29.

Chowdhury, Subhasish M, Dan Kovenock, David Rojo Arjona, and Nathaniel T Wilcox. 2021. "Focality and Asymmetry in Multi-Battle Contests." *The Economic Journal* 131 (636): 1593–1619.

Cortes-Corrales, Sebastián, and Paul M. Gorny. 2018. “Generalising Conflict Networks.” MPRA Paper No. 90001.

Falk, Armin, and Michael Kosfeld. 2012. “It’s All about Connections: Evidence on Network Formation.” *Review of Network Economics* 11 (3): Article 2.

Fischbacher, Urs. 2007. “Z-Tree: Zurich Toolbox for Ready-Made Economic Experiments.” *Experimental Economics* 10 (2): 171–78.

Franke, Jörg, and Tahir Öztürk. 2015. “Conflict Networks.” *Journal of Public Economics* 126 (June): 104–13.

Galeotti, Andrea, and Sanjeev Goyal. 2010. “The Law of the Few.” *American Economic Review* 100 (4): 1468–92.

Garfinkel, Michelle R. 2004. “Stable Alliance Formation in Distributional Conflict.” *European Journal of Political Economy* 20 (4): 829–52.

Goeree, Jacob K, Arno Riedl, and Aljaž Ule. 2009. “In Search of Stars: Network Formation among Heterogeneous Agents.” *Games and Economic Behavior* 67 (2): 445–66.

Goyal, Sanjeev, Stephanie Rosenkranz, Utz Weitzel, and Vincent Buskens. 2017. “Information Acquisition and Exchange In Social Networks.” *Economic Journal* 127 (606): 2302–31.

Goyal, Sanjeev, Adrien Vigier, and Marcin Dziubinski. 2016. “Conflict and Networks.” In *The Oxford Handbook of the Economics of Networks*, edited by Yann Bramoulle, Andrea Galeotti, and Brian Rogers, 214–43. Oxford: Oxford University Press.

Guido, Andrea, Andrea Robbett, and Rustam Romaniuc. 2019. “Group Formation and Cooperation in Social Dilemmas: A Survey and Meta-Analytic Evidence.” *Journal of Economic Behavior & Organization* 159 (March): 192–209.

Herbst, Luisa, Kai A Konrad, and Florian Morath. 2015. “Endogenous Group Formation in Experimental Contests.” *European Economic Review* 74 (February): 163–89.

Hiller, Timo. 2017. “Friends and Enemies: A Model of Signed Network Formation.” *Theoretical Economics* 12 (3): 1057–87.

Huitsing, Gijs, Marijtje A J van Duijn, Tom A B Snijders, Peng Wang, Miia Sainio, Christina Salmivalli, and René Veenstra. 2012. “Univariate and Multivariate Models of Positive and Negative Networks: Liking, Disliking, and Bully–Victim Relationships.” *Social Networks* 34 (4): 645–57.

Jackson, Matthew O, and Stephen Nei. 2015. “Networks of Military Alliances, Wars, and International Trade.” *Proceedings of the National Academy of Sciences* 112 (50): 15277–84.

Jandoc, Karl, and Ruben Juarez. 2019. “An Experimental Study of Self-Enforcing Coalitions.” *Games* 10 (3): 31.

Ke, Changxia, Kai A Konrad, and Florian Morath. 2013. “Brothers in Arms – An Experiment on the Alliance Puzzle.” *Games and Economic Behavior* 77 (1): 61–76.

———. 2015. “Alliances in the Shadow of Conflict.” *Economic Inquiry* 53 (2): 854–71.

Khavas, Sarah Rezaei, Stephanie Rosenkranz, Utz Weitzel, and Bastian Westbrock. 2018. "Fairness Concerns on Networks." *SSRN Electronic Journal*.

König, Michael D, Dominic Rohner, Mathias Thoenig, and Fabrizio Zilibotti. 2017. "Networks in Conflict: Theory and Evidence From the Great War of Africa." *Econometrica* 85 (4): 1093–1132.

Konrad, Kai A. 2009. *Strategy and Dynamics in Contests*. New York, NY: Oxford University Press.

———. 2014. "Strategic Aspects of Fighting in Alliances." In *The Economics of Conflict*, edited by Karl Wärneryd, 1–22. Cambridge & London: MIT Press.

Kosfeld, Michael. 2004. "Economic Networks in the Laboratory: A Survey." *Review of Network Economics* 3 (1): 20–41.

Leeuwen, Boris van, Theo Offerman, and Arthur Schram. 2020. "Competition for Status Creates Superstars: An Experiment on Public Good Provision and Network Formation." *Journal of the European Economic Association* 18 (2): 666–707.

Macfarlan, Shane J., Robert S. Walker, Mark V. Flinn, and Napoleon A. Chagnon. 2014. "Lethal Coalitionary Aggression and Long-Term Alliance Formation among Yanomamö Men." *Proceedings of the National Academy of Sciences* 111 (47): 16662–69.

Morrow, James D. 2000. "Alliances: Why Write Them Down?" *Annual Review of Political Science* 3 (1): 63–83.

O'Connell, PAUL, DEBRA Pepler, and WENDY Craig. 1999. "Peer Involvement in Bullying: Insights and Challenges for Intervention." *Journal of Adolescence* 22 (4): 437–52.

Pruetz, Jill D., Kelly Boyer Ontl, Elizabeth Cleaveland, Stacy Lindshield, Joshua Marshack, and Erin G. Wessling. 2017. "Intragroup Lethal Aggression in West African Chimpanzees (Pan Troglodytes Verus): Inferred Killing of a Former Alpha Male at Fongoli, Senegal." *International Journal of Primatology* 38 (1): 31–57.

Ray, Debraj. 2007. *A Game-Theoretic Perspective on Coalition Formation*. Oxford: Oxford University Press.

Ray, Debraj, and Rajiv Vohra. 1999. "A Theory of Endogenous Coalition Structures." *Games and Economic Behavior* 26 (2): 286–336.

Rong, Rong, and Daniel Houser. 2015. "Growing Stars: A Laboratory Analysis of Network Formation." *Journal of Economic Behavior and Organization* 117: 380–94.

Rosenkranz, Stephanie, and Utz Weitzel. 2012. "Network Structure and Strategic Investments: An Experimental Analysis." *Games and Economic Behavior* 75 (2): 898–920.

Roser, Max. 2016. "War and Peace." *Our World in Data*.

Salmivalli, Christina, Arja Huttunen, and Kirsti M J Lagerspetz. 1997. "Peer Networks and Bullying in Schools." *Scandinavian Journal of Psychology* 38 (4): 305–12.

Sánchez-Pagés, Santiago. 2007. “Endogenous Coalition Formation in Contests.” *Review of Economic Design* 11 (2): 139–63.

Siverson, Randolph M, and Joel King. 1980. “Attributes of National Alliance Membership and War Participation, 1815-1965.” *American Journal of Political Science* 24 (1): 1–15.

Smith, Adam C, David B Skarbek, and Bart J Wilson. 2012. “Anarchy, Groups, and Conflict: An Experiment on the Emergence of Protective Associations.” *Social Choice and Welfare* 38 (2): 325–53.

Snyder, Glenn H. 1997. *Alliance Politics*. Ithaca, NY: Cornell University Press.

Waltz, Kenneth N. 1979. *Theory of International Politics*. New York, NY: Random House.

Xu, Jin, Yves Zenou, and Junjie Zhou. 2019. “Networks in Conflict: A Variational Inequality Approach.” *SSRN Electronic Journal*.

Online Appendix:

Appendix A. Full theoretical analysis of equilibria

We assume the payoff function is linear in the difference between two players' strengths, $h_i^A(s_i, s_j) = k(s_i - s_j), k > 0$. Each player's intrinsic strength is set at $\lambda = 1$. The cost of extending a negative link is $c > 0$. In the main text, we only consider two equilibria, Peace and Bully, in the case of $k < c < 2k$. Here, we fully characterize Nash equilibria and show how they survive the 3 refinement conditions, both when $k < c < 2k$ and when $c < k$ for each type of alliance.

First, we consider $k < c < 2k$. All Nash equilibrium outcomes are illustrated in Figures A1 and refinements are summarized in Table A1.

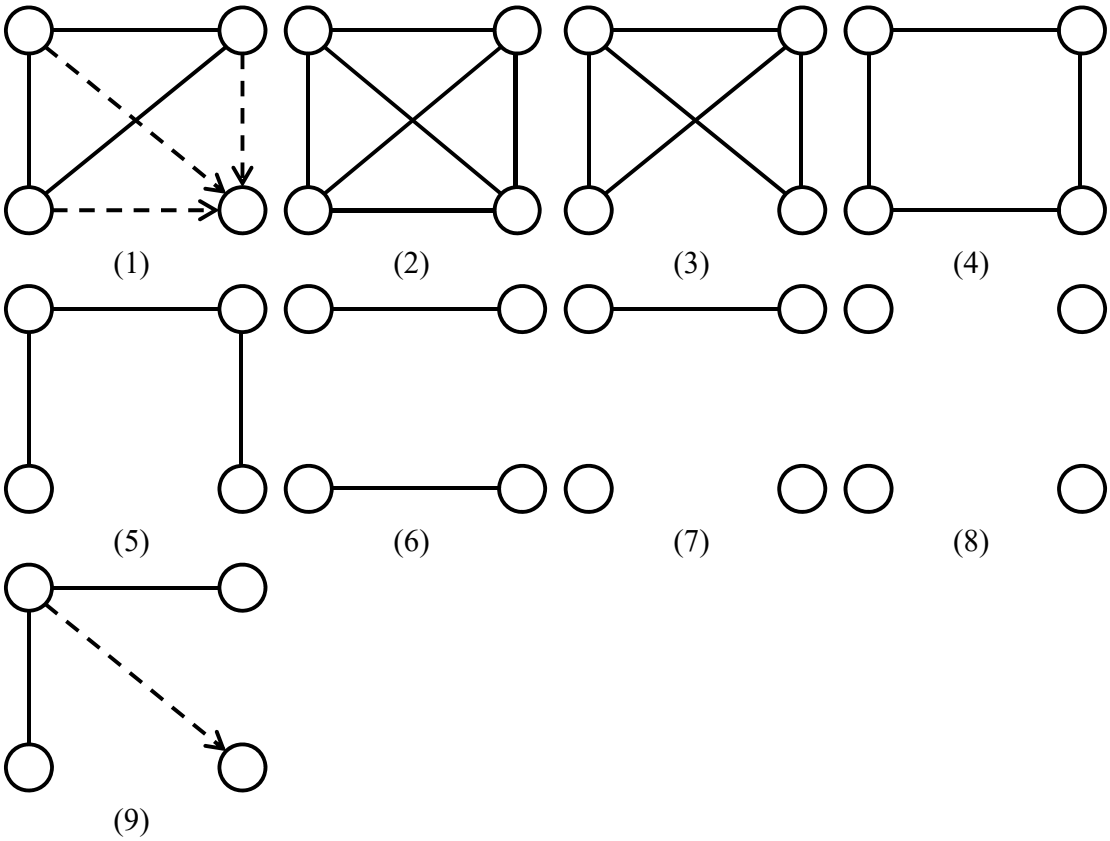


Figure A1. All Nash equilibria ($k < c < 2k$)

Table A1. Equilibrium refinements ($k < c < 2k$)

Equilibrium refinement	**Outcome**
Nash (single profitable deviation)	1,2,3,4,5,6,7,8,9
Pairwise stability	1,2,3,4,5,6,7,8,9
No Pairwise profitable deviation	1,2,6,7,8
No 3-person profitable deviation	1

Next, we consider $c < k$. All Nash equilibrium outcomes are illustrated in Figures A2 and refinements are summarized in Table A2.

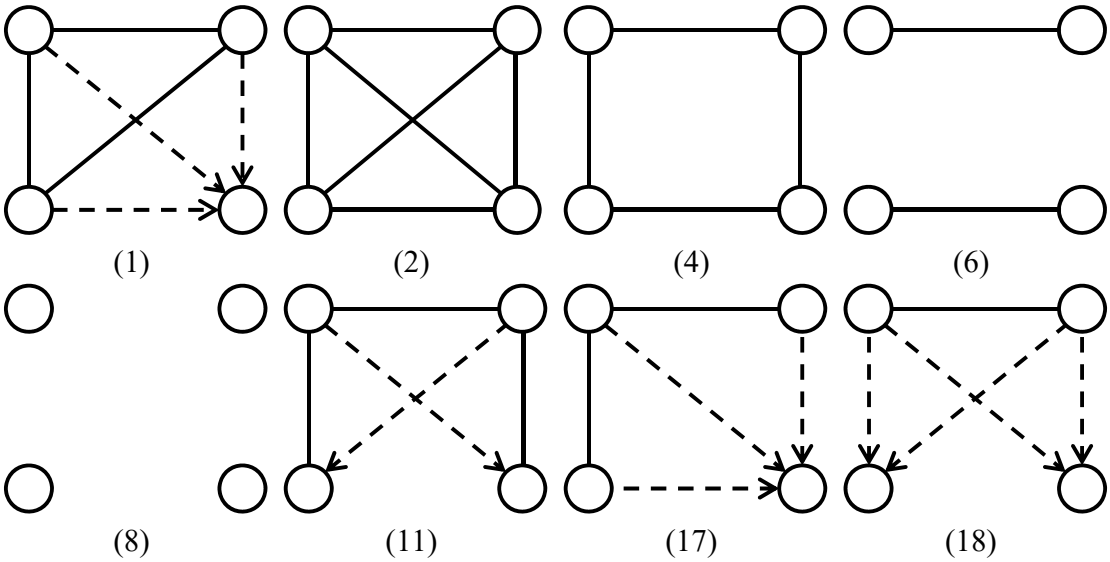


Figure A2. All Nash equilibria ($c < k$)

Table A2. Equilibrium refinements ($c < k$)

Equilibrium refinement	**Outcome**
Nash (single profitable deviation)	1,2,4,6,8,11,17,18
Pairwise stability	1,2,4,6,8
No Pairwise profitable deviation	1
No 3-person profitable deviation	1

Appendix B. Experimental instructions (English translation)

General Information:

You are participating in a decision-making study. Please read the following instructions carefully. These instructions are the same for all the participants. During the experiment, you are not allowed to communicate with other participants. Turn-off your mobile phone and put it in the envelope on your desk. If you have any questions, please raise your hand. One of the experimenters will approach you to answer your question.

You have earned 15 RMB for showing up on time. You can earn additional money by means of earning points during the experiment. The number of points that you earn depends on your own choices and the choices of other participants. At the end of the experiment, the total number of points that you earn during the experiment will be exchanged at the rate of:

5 points = 1 RMB

The money you earn will be paid out in cash via WeChat. Your decisions in this experiment will be anonymous, meaning no one can associate your name with your action throughout this study, and no other participants will be able to see how much you earn.

Overview of the experiment:

The experiment consists of 5 sections, each of which has 4 rounds. There are 20 rounds in total. At the beginning of each section, you will be randomly matched with three other participants. Each participant's position in your group will be shown as a circle on a specific position of the screen (either upper, lower, left or right, see screenshot below). The green circle represents yourself, while the black circles represent other three participants in your group. These participants are all currently in this room, but everyone's identity will be anonymous. The groups and the positions within a group will remain unchanged across all 4 rounds within a section, but will change from section to section.

Your decisions:

During each round, you may connect to one or more of the other participants in your group via two different means (you can also choose not to connect):

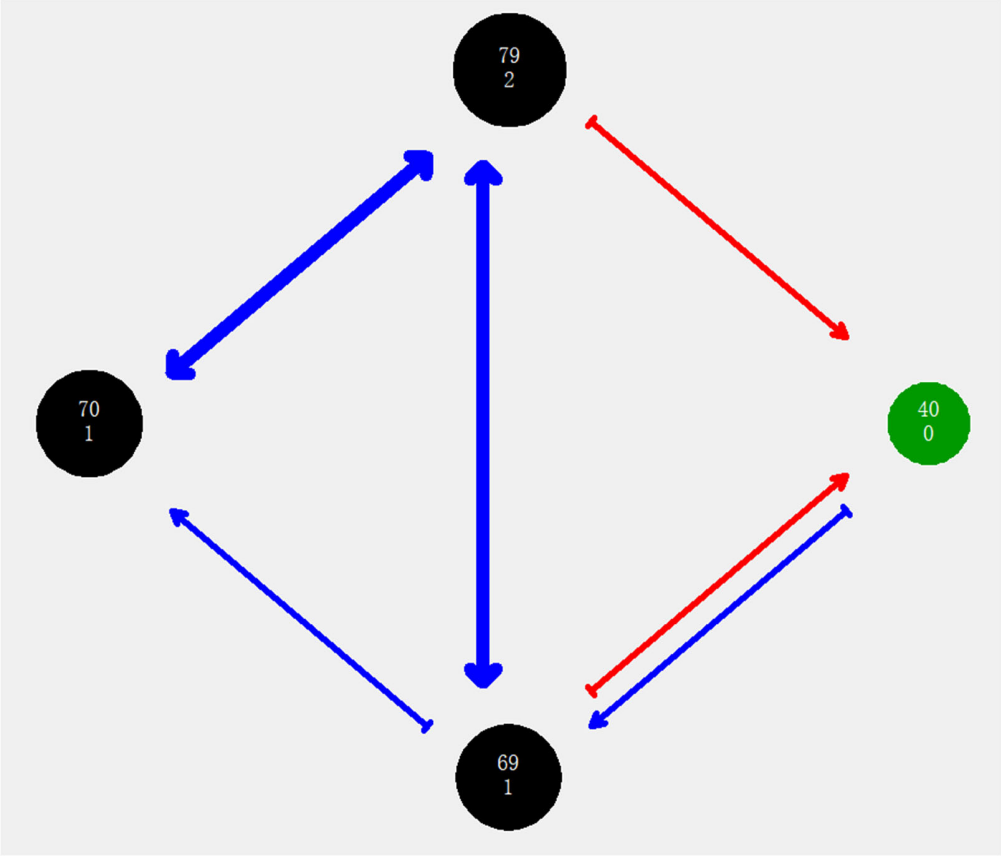
1. If you LEFT-MOUSE-CLICK on one of the black circles representing another participant, a blue link with an arrow pointing to that participant will appear. Left-clicking again on that participant, the blue link will be removed.

- Initiating a blue link represents an attempt to establish a partnership with another participant. Importantly, a partnership is only effective if ***both*** participants have initiated blue links, otherwise the partnership is ineffective. On the screen, if both participants initiate a blue link to each other, the blue link then becomes a bold double-headed arrow link (see screenshot below, the upper & lower players, and the upper & left players). Only in this case, the partnership is effective. Note that as long as one of the participants removes the link, the partnership will be ineffective.

2. If you RIGHT-MOUSE-CLICK on another participant, a red link with an arrow pointing to that participant will appear. Right-clicking again on that participant, the red link will be removed.

- Initiating a red link represents establishing a competitive relationship with another participant. Any unilateral initiation of a red link is effective (see screenshot below, both the upper and lower players initiate a red link to the right player). It means that a competitive relationship is effective as long as ***at least one side*** initiates a red link.

When you are making your linking decisions, you will be able to see your group members making and removing links simultaneously in real time. Likewise, other members in your group can see your making and removing decisions simultaneously in real time.



The number on top of the green circle indicates your current points, and the number on top of the black circle indicates other's current points. The size of a circle changes with the points that a player will receive: a larger circle means that that participant receives more points. The bottom number in each circle indicates the number of effective blue (partnership) links a player currently has.

Remarks:

- Note that to change a blue link to a red link, or vice versa, there is no need to 'unlink' the previous choice. You can simply directly left-click for blue or right-click for red. You cannot initiate a blue link and a red link to the same other participant at the same time.
- Note that there may be a slight time-lag between your click and the changes of the numbers on the screen. One click is enough to change a link successfully. A subsequent click will not be effective until the previous click is successfully in place. Therefore, be patient until a link is changed in order to make subsequent changes.

Your earnings:

Below we explain how to calculate your points for each round. Points depend on the links you and other participants make. Read this carefully. Do not worry if you find it difficult to grasp immediately—recall that the concurrent point values will be shown as the top number in each circle representing a player. We present an example with calculations below.

At the beginning of each round, each of the players will receive an endowment of 70 points. Note you start with 70 points every round. Formation of blue links is costless, while initiating each red link costs some points, which are either 3, 5, 7, 9 and 11. Before a section begins, you will know the cost of a red link for the 4 rounds in this section. You also know that the cost is the same for all members in your group.

If a player neither initiates nor receives a red link, her points remain as 70.

In the presence of a red link between say player A and player B, the point change depends on the difference between A's effective blue (partnership) links and B's effective blue (partnership) links. Take player A as an example, if A initiated a red link to B, A's additional points from the competitive relationship with player B are:

10*(A's effective blue links – B's effective blue links) – costs of red links

If A did not initiate a red link (thus it must be B who initiated a red link to A), A's additional points from the competitive relationship with player B are:

10*(A's effective blue links – B's effective blue links)

Point changes are calculated separately for each red link that you initiate or receive. Therefore, the total points, which are shown as the top number in each circle, are the sum of the endowment and point changes across all of your existing red links.

In the example shown in the above figure, the cost of each red link is 7 points:

The upper player

- initiates a red link to the right player;
- has two effective blue links with the lower and left players respectively, while the right player has no effective blue link;
- has payoff = 70 + 10(2-0) – 7 = 83.

The right player

- receives two red links from the upper and lower players respectively;
- has no effective blue link, while the upper player has two effective blue links and the lower player has one effective blue link;
- has payoff = 70 + 10(0-2) + 10(0-1) = 40.

The lower player

- initiates a red link to the left player;
- has one effective blue link, while the left player has no effective blue link;
- has payoff = 70 + 10*(1-0) – 7 = 73.

The left player

- neither initiates nor receives a red link;
- has one effective blue link;
- has payoff = 70.

Each player's final points in that round are determined at the end of that round. You can make as many adjustments of links as you like during a round; these adjustments are free. Both links and points in the circles are updated in real time. However, once that round ends, your points are determined by whatever the situation is in terms of your links at that point in time. Each round lasts somewhere between 75 and 105 seconds. The end will be at an unknown and random moment within this time interval. Please note that different rounds will not last equally long.

The computer will randomly choose one section (4 rounds) to calculate the total points as your final earnings. To give yourself the best chance of earning the most, you should decide carefully about every single round.

Questionnaire:

After the 20 rounds, you will be asked to fill in a brief questionnaire. Please take your time to fill in this questionnaire accurately. After you finish the questionnaire, the total amount you have earned from this experiment will be shown on the computer screen. Please remain seated until being instructed to leave.

This concludes the instructions. To make sure that everyone understands the instructions, you will now be asked to answer some comprehension questions. Please raise your hand if you need help. We will start the experiment once every participant has correctly answered all the comprehension questions.

Comprehension Quiz:

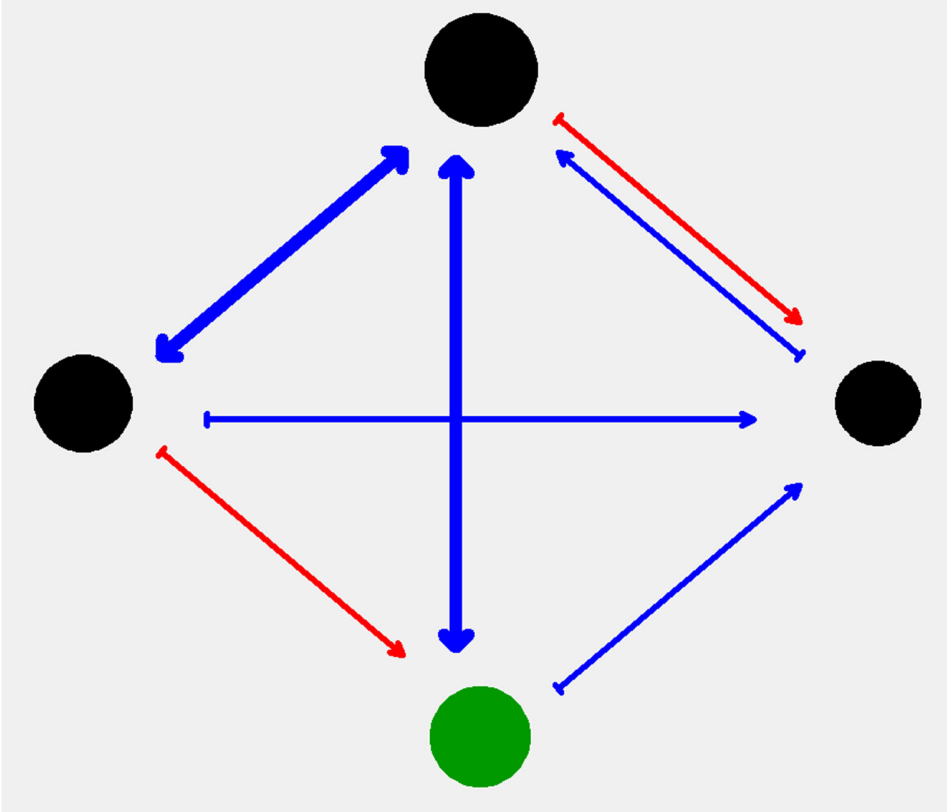
(on participants' computer screens)

General quiz (True or False):

- Your actions are anonymous in this study.
- You will receive money in cash via WeChat.
- In each section, you meet the same three other participants, but you don't know who they are. And their positions on the screen are unchanged within a section.
- You will meet the same three other participants from section to section.
- Every participant gets 70 points and starts with 70 points at the beginning of each round.
- Each round will end between 75 and 105 seconds, and the ending time for each round is different.
- If a round ends at the 105-th second, then your points in that round are determined by the nature of all links at the 105-th second.

Quiz on calculating the payoff:

Suppose a round ends at the 120th second, and at that moment, players have the following link configuration:



Upper player:

- Has initiated __ red links and received __ red links.
- Has __ effective blue links.
- The final payoff is __ points.

Bottom player:

- Has initiated __ red links and received __ red links.
- Has __ effective blue links.
- The final payoff is __ points.

Left player:

- Has initiated __ red links and received __ red links.
- Has __ effective blue links.
- The final payoff is __ points.

Right player:

- Has initiated __ red links and received __ red links.
- Has __ effective blue links.
- The final payoff is __ points.

Appendix C. Additional figures and tables

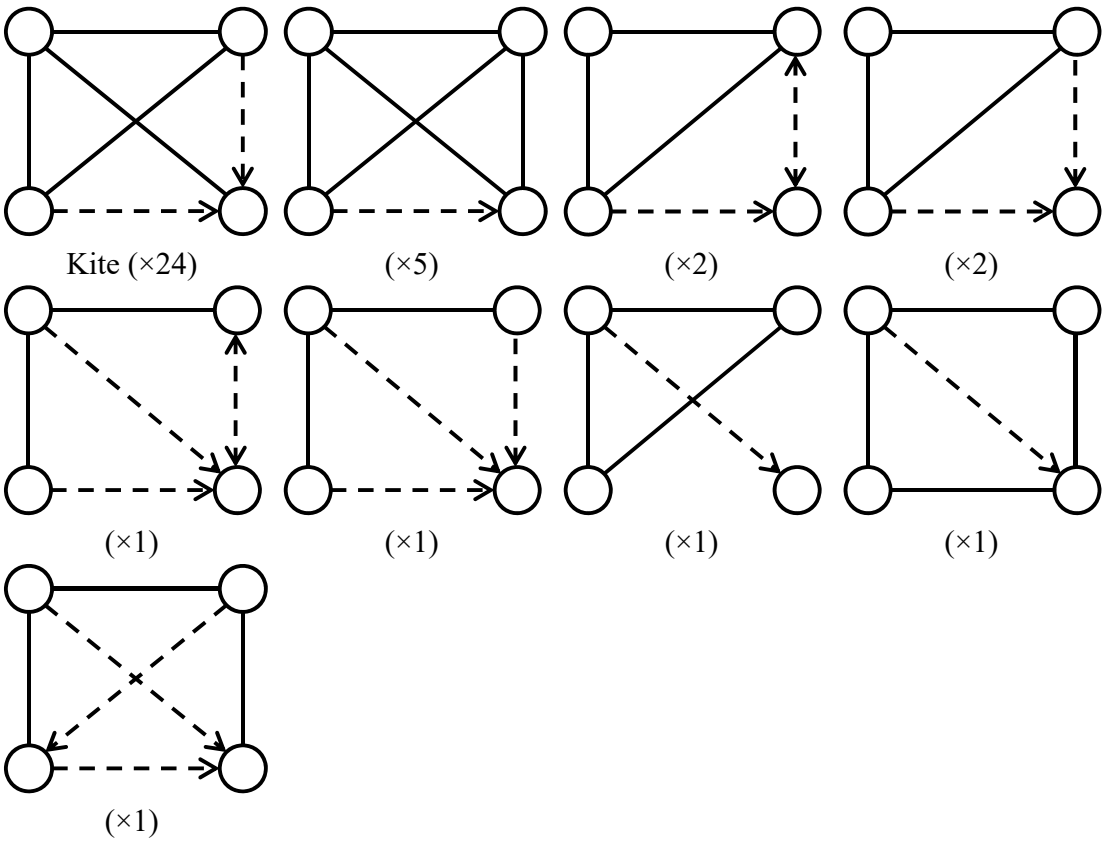


Figure C1. The Kite and other uncategorized networks (number of cases in bracket)

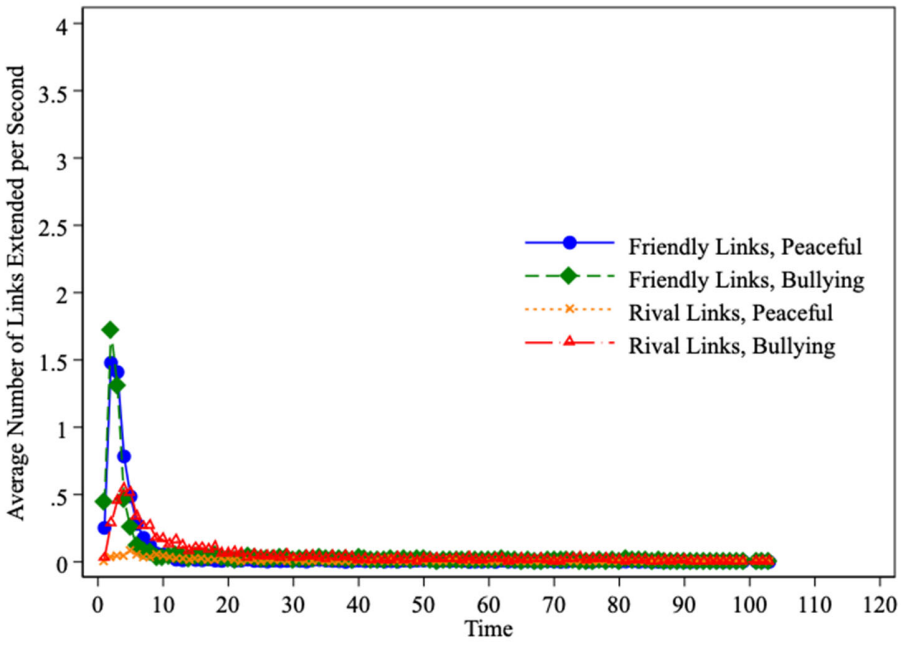


Figure C2: Extension of links per group per second for Peaceful and Bullying groups

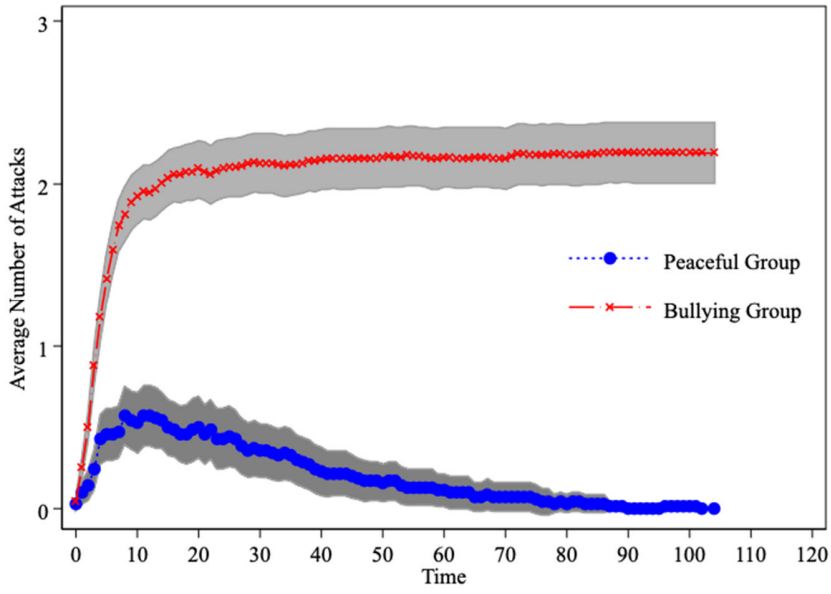


Figure C3: The evolution of the number of attacks received by first victims

Note: This Figure shows the average attacks received by first victims, in groups where at least one attack link was extended. In total, there are 303 out of 420 groups in which there ever exist a first victim. 195 out of 303 groups end up as Bullying groups and 142 first victims are final victims in these groups. 108 out of 303 groups resolve as Peaceful groups. The grey shaded area indicates 95% confidence intervals.

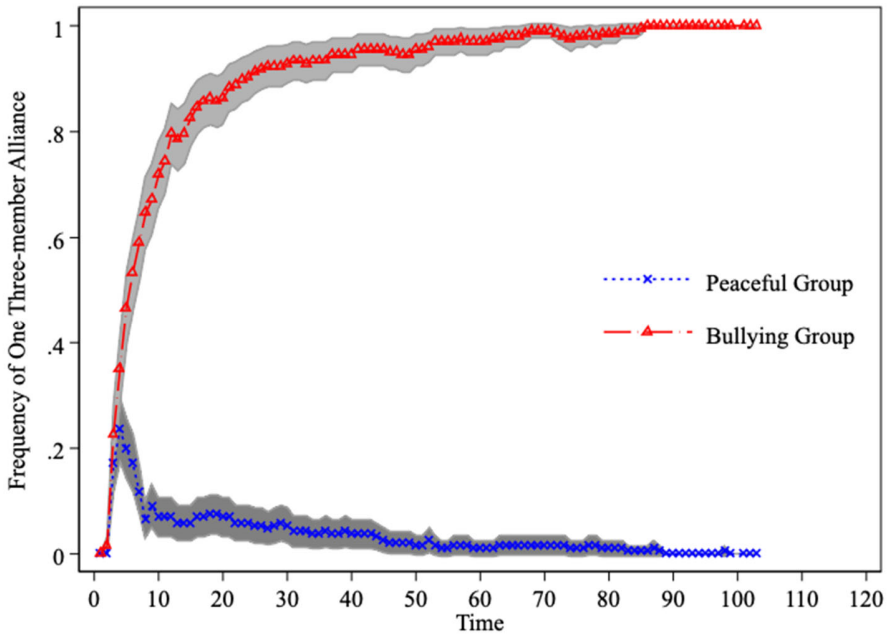


Figure C4: The percentage of exactly one three-member alliance for Peaceful and Bullying groups

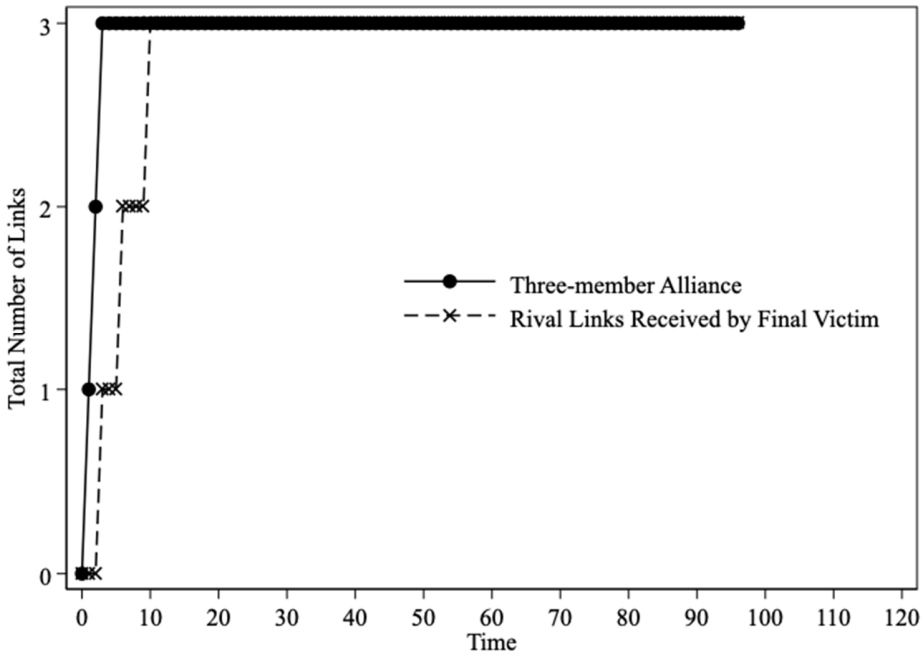


Figure C5: The evolution of median attacks received by the final victim and median effective friendships among the other three players in Bullying groups

Table C1: Probit model estimates of determinants of initiators

	All groups	Bullying groups
	(1)	(3)
L1.First Victim	0.104***	0.071***
	(0.012)	(0.018)
L1.Final Victim	0.031	0.005
	(0.031)	(0.030)
L1.Initiator	0.250***	0.265***
	(0.014)	(0.023)
N	1596	760

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. The table reports average marginal effect estimates with standard errors clustered at the session level. The dependent variable is whether a player is a final victim. * $p<0.1$; ** $p<0.05$; *** $p<0.01$

Appendix D. Additional figures and tables separately for different treatments/cost levels

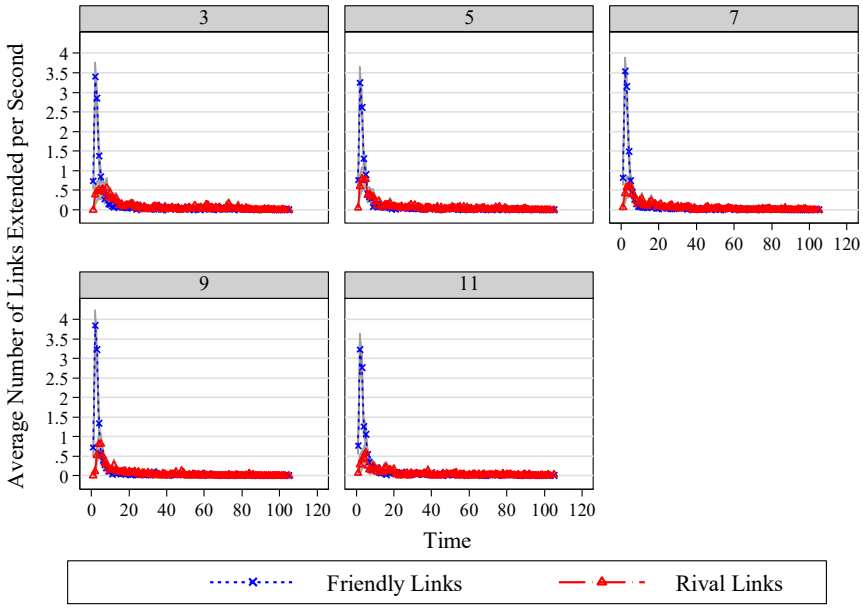


Figure D1: Extension of links per group per second by cost level

Note: The grey shaded area indicates 95% confidence intervals.

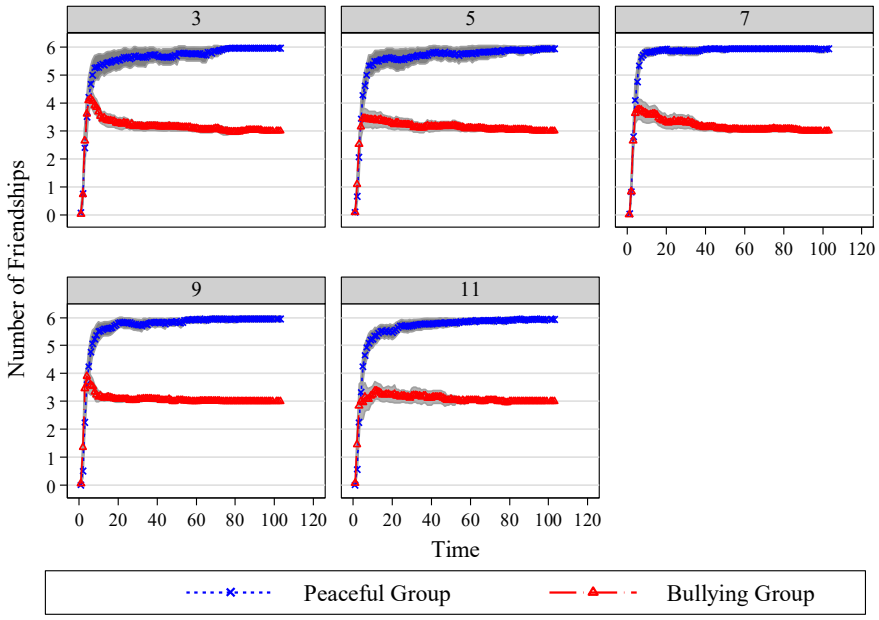


Figure D2: The evolution of effective friendships per group by cost level

Note: The grey shaded area indicates 95% confidence intervals.

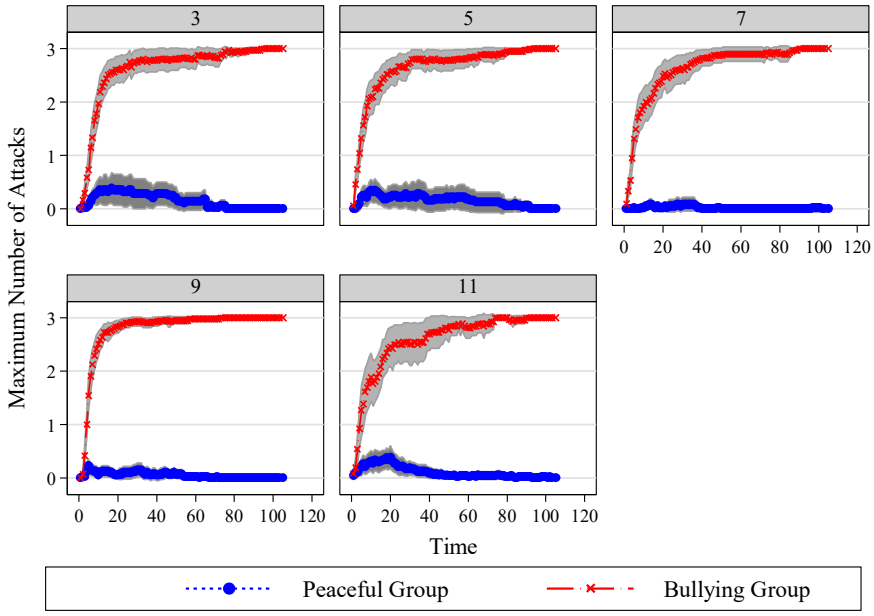


Figure D3: The evolution of the maximum number of attacks received by any player per group by cost level

Note: The grey shaded area indicates 95% confidence intervals.

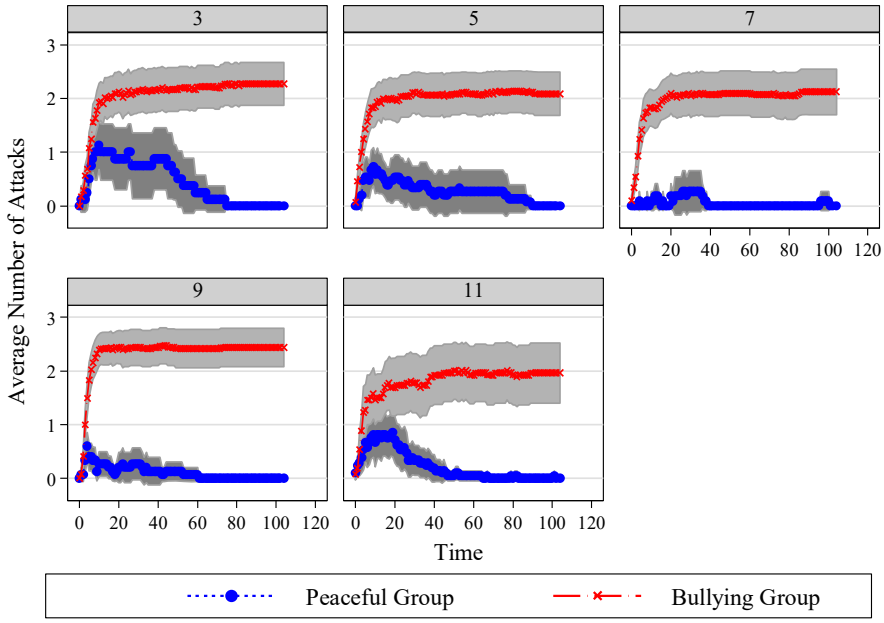


Figure D4: The evolution of the number of attacks received by first victims by cost level

Note: The grey shaded area indicates 95% confidence intervals.

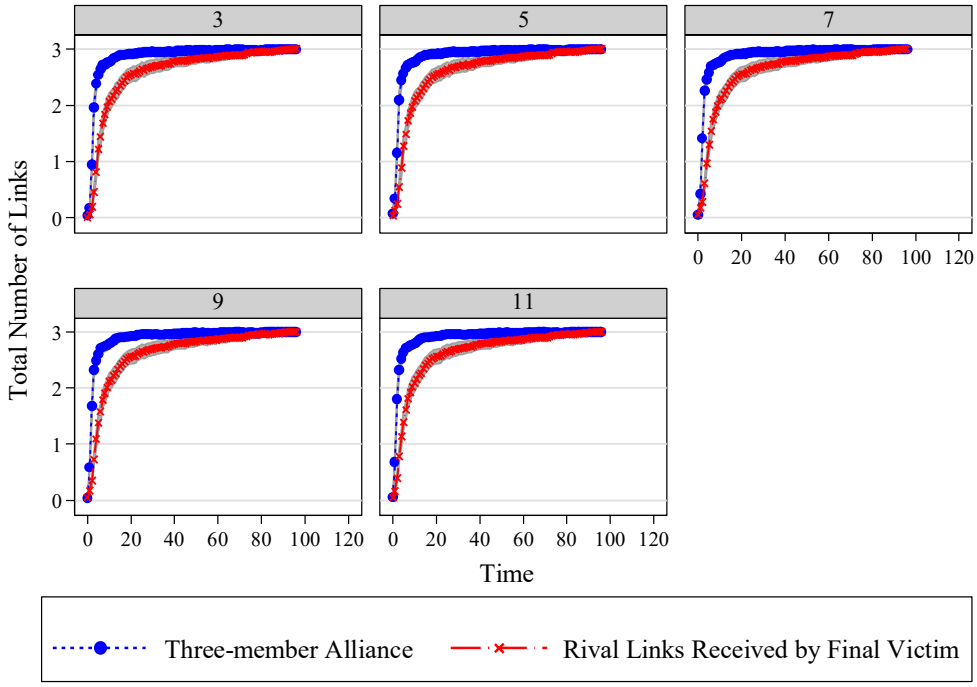


Figure D5: The evolution of average attacks received by the final victim and average effective friendships among the other three players in Bullying groups by cost level

Note: The grey shaded area indicates 95% confidence intervals.

Table D1: The pattern of transition to final victims by cost level

Type	N	% Receive 1 attack	% Receive 2 attacks	% Receive 3 attacks	% Final victim
Cost=3					
First victim	64	100%	70.3%	53.1%	51.6%
Initiator	64	68.8%	21.9%	14.1%	12.5%
Others	208	22.6%	3.4%	1.9%	0.5%
Cost=5					
First victim	67	100%	64.2%	50.7%	47.8%
Initiator	67	65.7%	16.4%	13.4%	10.4%
Others	202	16.3%	4.5%	4.0%	2.0%
Cost=7					
First victim	62	100%	66.1%	50.0%	46.8%
Initiator	62	56.5%	19.4%	17.7%	16.1%
Others	212	12.7%	2.4%	1.4%	0%
Cost=9					
First victim	59	100%	67.8%	57.6%	55.9%
Initiator	59	49.2%	8.5%	6.8%	6.8%
Others	218	15.1%	3.2%	1.8%	0.9%
Cost=11					
First victim	51	100%	54.9%	41.2%	33.3%
Initiator	51	54.9%	19.6%	15.7%	13.7%
Others	234	11.1%	1.7%	0.9%	0%

Appendix E. The relationship between network patterns in the first few seconds and the final outcome

Given our findings of early divergence between Peaceful and Bullying groups discussed in the previous subsection, we now consider whether particular categories of network formations are predictive of eventual convergence to Peaceful or Bullying outcomes. To provide statistical evidence on factors that can explain the divergent paths of Bullying and Peaceful networks, we turn to a group-level regression analysis with a binary dependent variable of whether a group eventually converges to Bullying or Peaceful networks.

We define five key explanatory variables for this analysis. The first two binary variables relate to the pattern of forming alliance: *OneAlliance* indicates whether there is one and only one three-member alliance; *FullConnect* indicates whether all group members are mutual friends. The justification for these variables is that the formation of an alliance that is exclusive to the fourth member is might be a precondition to Bullying networks, whereas the formation of four fully connected members is conducive to Peaceful networks. We thus hypothesize that *OneAlliance* predicts whether a group converges to Bullying networks whereas *FullConnect* predicts whether a group reaches Peaceful networks.

The next two binary variables are related to the pattern of making rivals. *maxAttack1* indicates whether the maximum number of rival links received by any player in a group is equal to 1; and *maxAttack2* indicates whether the maximum number of rival links received by any player in a group is equal to 2. Since both variables measure different degrees of progress in coordinating on a common rival, we hypothesize that both *maxAttack1* and *maxAttack2* are predictive of Bullying networks while *maxAttack2* has a stronger impact than *maxAttack1*.

We test how these state variables of the network status at time t (3~10 seconds) predict the final network, second by second. Table E1 reports the mean for each of these explanatory variables at each second from the third to the tenth second (the variables at the first two seconds are not included because there are very few observations. For ease of interpretation, we first consider only variables related to the pattern of alliance, then separately consider only variables related to the pattern of making rivals, and finally consider all variables together to see the relative importance of these two subsets of variables. The dependent variable for all Probit regressions below is whether the final network is Bullying (as opposed to Peaceful).

Table E1: The means of all explanatory variables at each second

	3 sec	4 sec	5 sec	6 sec	7 sec	8 sec	9 sec	10 sec
OneAlliance	0.199	0.291	0.332	0.356	0.359	0.361	0.387	0.401
TwoAlliance	0.107	0.199	0.199	0.181	0.204	0.186	0.136	0.123
FullConnect	0.024	0.105	0.199	0.264	0.291	0.322	0.356	0.366
maxAttack1	0.209	0.249	0.217	0.181	0.183	0.152	0.134	0.120
maxAttack2	0.034	0.094	0.128	0.160	0.139	0.160	0.154	0.149
N	382	382	382	382	382	382	382	382

Table E2 reports estimates from regressions that include *OneAlliance* and *FullConnect*. It is striking how quickly the final outcome is resolved: *FullConnect* starts to negatively predict Bullying networks by the 4th second, and continues to do so with generally increasing strength as the seconds proceed. For *OneAlliance*, the significant positive prediction of Bullying networks starts in the 5th second, and generally strengthens as the seconds proceed, mostly up to the 8th second.

Table E2: Probit model estimates of network state variables (alliance variables)

	3 sec	4 sec	5 sec	6 sec	7 sec	8 sec	9 sec	10 sec
OneAlliance	0.086	0.073	0.144**	0.144***	0.226***	0.302***	0.212***	0.279***
	(0.077)	(0.082)	(0.047)	(0.019)	(0.051)	(0.042)	(0.059)	(0.046)
FullConnect	0.011	-0.350**	-0.371***	-0.430***	-0.402***	-0.336***	-0.372***	-0.305***
	(0.274)	(0.137)	(0.070)	(0.061)	(0.022)	(0.039)	(0.048)	(0.038)
N	382	382	382	382	382	382	382	382

Note: The dependent variable is 1 if the final status of the group is Bullying, and 0 if Peaceful. We only include groups in which their final status is either Bullying or Peaceful (382 out of 420 groups). The table reports average marginal effect estimates with standard errors clustered at the session level. All regressions include period fixed effects. * $p<0.1$; ** $p<0.05$; *** $p<0.01$

Table E3 reports estimates from regressions that include *maxAttack1* and *maxAttack2*. In general, 1 or 2 maximum attacks predicts Bullying networks from the very beginning, and for the case of 1 maximum attack, the predictive power is not necessarily increasing in strength over time, while for 2 maximum attacks, the estimate is stable, and more influential than 1 maximum attack as the seconds go by.

Table E3: Probit model estimates of network state variables (attacking variables)

	3 sec	4 sec	5 sec	6 sec	7 sec	8 sec	9 sec	10 sec
maxAttack1	0.494***	0.488***	0.355***	0.309***	0.316***	0.224**	0.162*	0.182
	(0.506)	(0.049)	(0.062)	(0.064)	(0.078)	(0.081)	(0.091)	(0.118)
maxAttack2	0.506*	0.522***	0.615***	0.587***	0.546***	0.682***	0.580***	0.528***
	(0.261)	(0.062)	(0.080)	(0.094)	(0.122)	(0.116)	(0.129)	(0.133)
N	382	382	382	382	382	382	382	382

Note: The dependent variable is 1 if the final status of the group is Bullying, and 0 if Peaceful. The table reports average marginal effect estimates with standard errors clustered at the session level. All regressions include period fixed effects. * $p<0.1$; ** $p<0.05$; *** $p<0.01$

For completeness, we also include another variable, named *maxAttack3*, meaning that the maximum number of rival links received by any player in a group is 3. Thus, there is already a common rival in the group if *maxAttack3* = 1. This only happens starting from the fifth second. By definition, this variable is strongly correlated with *OneAlliance* as these two state variables often imply the realization of the Bullying networks. Table E4 reports Probit estimates by including all

three state variables related to attacking. Not surprisingly, *maxAttack3* strongly predicts Bullying networks and its strength also tends to be the largest.

Table E4: Probit estimates of network state variables (all three attacking variables)

	5 sec	6 sec	7 sec	8 sec	9 sec	10 sec
maxAttack1	0.375***	0.326***	0.328***	0.263***	0.234***	0.254***
	(0.039)	(0.021)	(0.025)	(0.031)	(0.042)	(0.048)
maxAttack2	0.619***	0.575***	0.525***	0.586***	0.502***	0.452***
	(0.035)	(0.027)	(0.039)	(0.035)	(0.025)	(0.022)
maxAttack3	0.679***	0.701***	0.705***	0.618***	0.640***	0.627***
	(0.128)	(0.127)	(0.098)	(0.086)	(0.082)	(0.066)
N	382	382	382	382	382	382

Note: The dependent variable is 1 if the final status of the group is Bullying, and 0 if Peaceful. The table reports average marginal effect estimates with standard errors clustered at the session level. All regressions include period fixed effects. * p<0.1; ** p<0.05; *** p<0.01

Next, we include *OneAlliance*, *FullConnect*, *maxAttack1* and *maxAttack2* in the same regression to investigate the relative importance of these state variables for each second. Table E5 reports the estimates, showing that while *maxAttack1* and *maxAttack2* tend to be more influential in earlier seconds, *OneAlliance* and *FullConnect* tend to take over the predictive power in later seconds. While some of these variables might overlap to an extent that prohibits a very precise interpretation, the overall result suggests that while intermediate state variables such as *maxAttack1* and *maxAttack2* are good predictors of Bullying networks, it is eventually the stabilized pattern of alliance that absorbs their predictive power and determines the final outcome. We explore the dual dynamic process of attacking and alliance formation in more detail in the next subsection.

Table E5: Probit model estimates of network state variables (alliance and attacking variables)

	3 sec	4 sec	5 sec	6 sec	7 sec	8 sec	9 sec	10 sec
OneAlliance	0.031	0.013	0.152***	0.147***	0.248***	0.279***	0.164***	0.258***
	(0.092)	(0.093)	(0.028)	(0.027)	(0.054)	(0.052)	(0.053)	(0.049)
FullConnect	0.128	-0.099	-0.156	-0.273**	-0.280***	-0.273***	-0.386***	-0.320***
	(0.222)	(0.143)	(0.103)	(0.090)	(0.078)	(0.071)	(0.047)	(0.041)
maxAttack1	0.493***	0.464***	0.285***	0.179**	0.166*	0.050	-0.079*	-0.060
	(0.093)	(0.080)	(0.093)	(0.088)	(0.092)	(0.068)	(0.042)	(0.075)
maxAttack2	0.515*	0.494***	0.504***	0.357***	0.229**	0.221***	0.086*	0.022
	(0.255)	(0.095)	(0.110)	(0.138)	(0.107)	(0.064)	(0.046)	(0.045)
N	382	382	382	382	382	382	382	382

Note: The dependent variable is 1 if the final status of the group is Bullying, and 0 if Peaceful. The table reports average marginal effect estimates with standard errors clustered at the session level. All regressions include period fixed effects. * p<0.1; ** p<0.05; *** p<0.01

Finally, we also examine an additional explanatory variable, *TwoAlliance*, indicating

whether there are exactly two three-member alliances. This is the case in which all but one pair of members are friends. We are agnostic about the predictive power of this variable but would like to know whether it has the same predictive direction as *OneAlliance* or *FullConnect:* it is possible that any pattern of alliances falling short of being fully connected would eventually lead to Bullying networks; but it is also possible that *TwoAlliance* serves as an intermediate step toward Peaceful networks. To investigate, we first ran Probit regressions on its own for each second. The estimates are reported in Table E6.

On its own, *TwoAlliance* does not seem to have much regular significant predictive power on the final outcome. However, when we estimate its coefficient together with those of *OneAlliance* and *FullConnect*, *TwoAlliance* significantly negatively predicts Bullying networks starting from the 5th second or so, with regularity. The estimates are reported in Table E7. It is also interesting to observe that *TwoAlliance* tends to soak up part of the previous explanatory power of *OneAlliance*, previously a very significant predictor of Bullying networks, although the estimate of *OneAlliance* never becomes negative. It is probably because there is no longer any variable that is a strong period by period predictor of Peaceful networks when *TwoAlliance* is included (that is, unobserved variables tend to predict Bullying networks). It thus becomes easier to predict Peaceful networks than Bullying networks. These results suggest that the groups with almost mutual friends are likely to find a way to eventually keep the peace, whereas the ones with one and only one three-member alliance consistently lead up to a bullying situation.

Table E6: Probit model estimates of network state variables (two alliances)

	3 sec	4 sec	5 sec	6 sec	7 sec	8 sec	9 sec	10 sec
TwoAlliance	0.067	-0.031	-0.086***	0.004	-0.063	-0.116**	-0.021	-0.100
	(0.201)	(0.104)	(0.028)	(0.050)	(0.056)	(0.055)	(0.087)	(0.103)
N	382	382	382	382	382	382	382	382

Note: The dependent variable is 1 if the final status of the group is Bullying, and 0 if Peaceful. The table reports average marginal effect estimates with standard errors clustered at the session level. All regressions include period fixed effects. * p<0.1; ** p<0.05; *** p<0.01

Table E7: Probit model estimates of network state variables (all three alliance variables)

	3 sec	4 sec	5 sec	6 sec	7 sec	8 sec	9 sec	10 sec
OneAlliance	0.100	0.048	0.053	0.064	0.095	0.166**	0.109	0.181**
	(0.098)	(0.111)	(0.057)	(0.054)	(0.058)	(0.060)	(0.068)	(0.065)
TwoAlliance	0.093	-0.063	-0.166***	-0.140***	-0.193**	-0.187***	-0.150**	-0.143***
	(0.230)	(0.159)	(0.034)	(0.064)	(0.092)	(0.065)	(0.073)	(0.054)
FullConnect	0.030	-0.373*	-0.452***	-0.504***	-0.517***	-0.444***	-0.455***	-0.379***
	(0.316)	(0.190)	(0.070)	(0.052)	(0.044)	(0.047)	(0.252)	(0.050)
N	382	382	382	382	382	382	382	382

Note: The dependent variable is 1 if the final status of the group is Bullying, and 0 if Peaceful. The table reports average marginal effect estimates with standard errors clustered at the session level. All regressions include period fixed effects. * p<0.1; ** p<0.05; *** p<0.01