Strategic Network Formation with Attack and Immunization

Sanjeev Goyal[1], Shahin Jabbari[2(✉)], Michael Kearns[2], Sanjeev Khanna[2], and Jamie Morgenstern[2]

[1] Faculty of Economics, University of Cambridge, Cambridge, UK
sg472@cam.ac.uk
[2] Department of Computer and Information Sciences,
University of Pennsylvania, Philadelphia, USA
{jabbari,mkearns,sanjeev,jamiemor}@cis.upenn.edu

Abstract. Strategic network formation arises in settings where agents receive some benefit from their connectedness to other agents, but also incur costs for forming these links. We consider a new network formation game that incorporates an adversarial attack, as well as *immunization* or protection against the attack. An agent's network benefit is the expected size of her connected component post-attack, and agents may also choose to immunize themselves from attack at some additional cost. Our framework can be viewed as a stylized model of settings where *reachability* rather than centrality is the primary interest (as in many technological networks such as the Internet), and vertices may be vulnerable to attacks (such as viruses), but may also reduce risk via potentially costly measures (such as an anti-virus software).

Our main theoretical contributions include a strong bound on the edge density at equilibrium. In particular, we show that under a very mild assumption on the adversary's attack model, every equilibrium network contains at most only $2n-4$ edges for $n \geq 4$, where n denotes the number of agents and this upper bound is tight. We also show that social welfare does not significantly erode: every non-trivial equilibrium with respect to several adversarial attack models asymptotically has social welfare at least as that of any equilibrium in the original attack-free model.

We complement our sharp theoretical results by a behavioral experiment on our game with over 100 participants, where despite the complexity of the game, the resulting network was surprisingly close to equilibrium.

1 Introduction

In network formation games, distributed and strategic agents receive benefit from their connectedness to others, but also incur some cost for forming these links. Much research in this area [4,6,9] studies the structure of equilibrium networks

The full version of this paper with all the omitted details is available at https://arxiv.org/abs/1511.05196.

Y. Cai and A. Vetta (Eds.): WINE 2016, LNCS 10123, pp. 429–443, 2016.
DOI: 10.1007/978-3-662-54110-4_30

formed as the result of various choices for the network benefit function, as well as the social welfare in equilibria. In many such games, the costs incurred from forming links are direct: each edge costs $C_E > 0$ for an agent to purchase. Recently, motivated by scenarios as diverse as financial crises, terrorism and technological vulnerability, games with indirect connectivity costs have been considered: an agent's connections expose her to negative, contagious shocks.

We begin with the well-studied *reachability* network formation game [4], in which players purchase links to each other, and enjoy a network benefit equal to the size of their connected component in the formed graph. We modify this model by introducing an adversary who is allowed to examine the network, and choose a single vertex or player to attack. This attack then spreads throughout the entire connected component of the originally attacked vertex, destroying all of these vertices. Crucially however, players also have the option of purchasing *immunization* against attack. Thus the attack spreads only to those non-immunized (or *vulnerable*) vertices reachable from the originally attacked vertex. We examine several natural adversarial attacks such as an adversary that seeks to maximize destruction, an adversary that randomly selects a vertex for the start of infection and an adversary that seeks to minimize the social welfare of the network post-attack to name a few. A player's overall payoff is thus the expected size of her post-attack component, minus her edge and immunization expenditures.[1]

Our game can be viewed as a stylized model for settings where reachability rather than centrality is the primary interest in joining a network vulnerable to adversarial attack. Examples include technological networks such as the Internet, where packet transmission times are sufficiently low that being "central" [9] or a "hub" [6] is less of a concern, but in the presence of attacks such as viruses or DDoS, mere reachability may be compromised. Parties may reduce risks via costly measures such as anti-virus. In a financial setting, vertices might represent banks and edges credit/debt agreements. The introduction of an attractive but extremely risky asset is a threat or attack on the network that naturally seeks its largest accessible market, but can be mitigated by individual institutions adopting balance sheet requirements or leverage restrictions. In a biological setting, vertices could represent humans, and edges physical proximity or contact. The attack could be an actual biological virus that randomly infects an individual and spreads by physical contact through the network; again, individuals may have the option of immunization. While our simplified model is obviously not directly applicable to any of these examples in detail, we do believe our results provide some high-level insights about the strategic tensions in such scenarios.

[1] The spread of the initial attack to reachable non-immunized vertices is deterministic in our model, and the protection of immunized vertices is absolute. It is also natural to consider relaxations such as probabilistic attack spreading and imperfect immunization, as well as generalizations such as multiple initial attack vertices. However, as we shall see, even the basic model we study here exhibits substantial complexity. We refer the reader to the full version for a discussion on possible extensions/relaxations.

Immunization against attack has recently been studied in games played on a network where risk of contagious shocks are present [7] but only in the setting in which the network is first designed by a centralized party, after which agents make individual immunization decisions. We endogenize both these aspects, which leads to a model incomparable to this earlier work.

The original reachability game [4] permitted a sharp and simple characterization of the equilibria: any tree as well as the empty graph. We demonstrate that once attack and immunization are introduced, the set of possible equilibria becomes considerably more complex, including networks that contain multiple cycles, as well as others which are disconnected but nonempty. This diversity leads to our primary questions of interest: How dense can equilibria become? In particular, does the presence of the attacker encourage the creation of massive redundancy of connectivity? Also does the introduction of attack and immunization result in dramatically lower social welfare compared to the original game?

Our Results and Techniques. The main theoretical contributions of this work are to show that our game still exhibits edge sparsity at equilibrium, and has high social welfare properties despite the presence of attacks. First we show that under a mild assumption on the adversary's attack model, the equilibrium networks with $n \geq 4$ players have at most $2n-4$ edges, fewer than twice as many edges as any nonempty equilibria of the original game without attack. We prove this by introducing an abstract representation of the network and use tools from graph theory to upper bound the resources globally invested by the players to mitigate connectivity disruptions due to any attack.

We then show that with respect to several attack models, in any equilibrium with at least one edge and one immunized vertex, the resulting network is connected. This implies that any *new* equilibrium network (i.e. one which was not an equilibrium of the original reachability game) is either a sparse but connected graph, or is a forest of unimmunized vertices. The latter occurs only in the rather unnatural case where the cost of immunization or edges grows with the population size, and in the former case we further show the social welfare is at least $n^2 - O(n^{5/3})$ – asymptotically the maximum possible with a polynomial rate of convergence. These results provide us with a complete picture of welfare in our model. We prove the welfare lower bound by showing that there cannot be many targeted vertices who are *critical* for global connectivity, where critical is defined formally in terms of both the vertex's probability of attack and the size of the components remaining after the attack. Thus players myopically optimizing their own utility create highly resilient networks in presence of attack.

We conclude by reporting on a behavioral experiment on our network formation game with over 100 participants, where despite the complexity of the game, the resulting network was surprisingly close to equilibrium.

Organization. We formally present our model and review some related work in Sect. 2. In Sect. 3 we briefly describe some interesting topologies that arise as equilibria and then prove our sparsity result. We present our lower bound on welfare in Sect. 4. Section 5 describe our behavioral experiment.

In the full version, we provide simulations demonstrating fast and general convergence of *swapstable* best response, a type of limited best response which generalizes linkstable best response but is more powerful in our game. The computational complexity of full best response dynamics was unknown to us at the time of conducting our simulations but this question has been recently studied by Ihde et al. [13]. The simulations illustrate a number of interesting further features of equilibria e.g. heavy-tailed degree distributions. Whether swapstable best response provably converges (as seen empirically) is an open question.

2 Model

We assume the n vertices of a graph (network) correspond to individual players. Each player has the choice to purchase edges to other players at a cost of $C_{\mathrm{E}} > 0$ per edge. Each player additionally decides whether to immunize herself at a cost of $C_{\mathrm{I}} > 0$ or remain *vulnerable*.

A (pure) *strategy* for player i (denoted by s_i) is a pair consisting of the subset of players i purchased an edge to and her immunization choice. Formally, we denote the subset of edges which i buys an edge to as $x_i \subseteq \{1, \ldots, n\}$, and the binary variable $y_i \in \{0, 1\}$ as her immunization choice ($y_i = 1$ when i immunizes). Then $s_i = (x_i, y_i)$. *We assume that edge purchases are unilateral i.e. players do not need approval in order to purchase an edge to another but that the connectivity benefits and risks are bilateral.* We restrict our attention to pure strategy equilibria and our results show they exist and are structurally diverse.

Let $\mathbf{s} = (s_1, \ldots, s_n)$ denote the strategy profile for all the players. Fixing $\mathbf{s}$, the set of edges purchased by all the players induces an undirected graph and the set of immunization decisions forms a bipartition of the vertices. We denote a game *state* as a pair $(G, \mathcal{I})$, where $G = (V, E)$ is the undirected graph induced by the edges purchased by the players and $\mathcal{I} \subseteq V$ is the set of players who decide to immunize. We use $\mathcal{U} = V \setminus \mathcal{I}$ to denote the vulnerable vertices i.e. the players who decide not to immunize. We refer to a subset of vertices of $\mathcal{U}$ as a *vulnerable region* if they form a maximally connected component. We denote the set of vulnerable regions by $\mathcal{V} = \{\mathcal{V}_1, \ldots, V_k\}$ where each $\mathcal{V}_i$ is a vulnerable region.

Fixing a game state $(G, \mathcal{I})$, the adversary inspects the formed network and the immunization pattern and chooses to attack some vertex. If the adversary attacks a vulnerable vertex $v \in \mathcal{U}$, then the attack starts at v and spreads, killing v and any other vulnerable vertices reachable from v. Immunized vertices act as "firewalls" through which the attack cannot spread. *We point out that in this work we restrict the adversary to only pick one seed to start the attack.*

More precisely, the adversary is specified by a function that defines a probability distribution over vulnerable regions. We refer to a vulnerable region with non-zero probability of attack as a *targeted region* and the vulnerable vertices inside of a targeted region as *targeted vertices*. We denote the targeted regions by $\mathcal{T} = \{\mathcal{T}_1, \ldots, \mathcal{T}_{k'}\}$ where each $\mathcal{T}' \in \mathcal{T}$ denotes a targeted region.[2]

[2] The index k' in the definition of $\mathcal{T}$ satisfies $k' \leq k$ (see k in the definition of $\mathcal{V}$).

$\mathcal{T} = \emptyset$ corresponds to the adversary making no attack, so player i's *utility* (or *payoff*) is equal to the size of her connected component minus her expenses (edge purchases and immunization). When $|\mathcal{T}| > 0$, player's i expected utility (fixing a game state) is equal to the expected size of her connected component[3] less her expenditures, where the expectation is taken over the adversary's choice of attack (a distribution on $\mathcal{T}$). Formally, let $\Pr[\mathcal{T}']$ denote the probability of attack to targeted region $\mathcal{T}'$ and $CC_i(\mathcal{T}')$ the size of the connected component of player i post-attack to $\mathcal{T}'$. Then the expected utility of i in strategy profile s denoted by $u_i(s)$ is precisely

$$u_i(\mathbf{s}) = \sum_{\mathcal{T}' \in \mathcal{T}} \Big(\Pr\left[\mathcal{T}'\right] CC_i\left(\mathcal{T}'\right) \Big) - |x_i| C_{\mathrm{E}} - y_i C_{\mathrm{I}}.$$

We refer to the sum of expected utilities of all the players playing $\mathbf{s}$ as the *(social) welfare* of $\mathbf{s}$.

Examples of Adversaries. We highlight several natural adversaries that fit into our framework. We begin with a natural adversary whose goal is to maximize the number of agents killed.

Definition 1. *The* maximum carnage *adversary attacks the vulnerable region of maximum size. If there are multiple such regions, the adversary picks one of them uniformly at random. Once a targeted region is selected, the adversary selects a vertex inside of that region uniformly at random to start the attack.*

So a targeted region with respect to a maximum carnage adversary is a vulnerable region of maximum size and the adversary defines a uniform distribution over such regions (see Fig. 1). Another natural but less sophisticated adversary starts an attack by picking a vulnerable vertex at random.

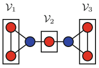


Fig. 1. Blue and red vertices denote $\mathcal{I}$ and $\mathcal{U}$, respectively. The probability of attack to the vulnerable regions denoted by $\mathcal{V}_1, \mathcal{V}_2$ and $\mathcal{V}_3$ (in that order) for each adversary are as follows. maximum carnage: 0.5, 0, 0.5; random attack: 0.4, 0.2, 0.4; maximum disruption: 0, 1, 0. (Color figure online)

Definition 2. *The* random attack *adversary attacks a vulnerable vertex uniformly at random.*

So every vulnerable vertex is targeted with respect to the random attack adversary and the adversary induces a distribution over targeted regions such that the probability of attack to a targeted region is proportional to its size (see Fig. 1). Lastly, we define another natural adversary whose goal is to minimize the post-attack welfare.

[3] If a vertex is killed, the size of her connected component is defined to be 0.

Definition 3. *The* maximum disruption *adversary attacks the vulnerable region which minimizes the post-attack social welfare. If there are multiple such regions, the adversary picks one of them uniformly at random. Once a targeted region is selected for the attack, the adversary selects a vertex inside of that region uniformly at random to start the attack.*

Thus the maximum disruption adversary only attacks those vulnerable regions which minimize the post-attack welfare and the adversary defines a uniform distribution over such regions (see Fig. 1).

Equilibrium Concepts. We analyze the networks formed in our game under two types of equilibria. We model each of the n players as strategic agents who choose deterministically which edges to purchase and whether or not to immunize, knowing the exogenous behavior of the adversary defined as above. We say a strategy profile $\mathbf{s}$ is a *pure strategy Nash equilibrium* (Nash equilibrium for short) if, for any player i, fixing the behavior of the other players to be $\mathbf{s}_{-i}$, the expected utility for i cannot strictly increase playing any action $\mathbf{s}'_i$ over $\mathbf{s}_i$.

In addition to Nash, we study another equilibrium concept that is closely related to linkstable equilibrium [5], a bounded-rationality generalization of Nash. We call this concept *swapstable equilibrium.*[4] A strategy profile is a swapstable equilibrium if no agent's expected utility (fixing other agents' strategies) can strictly improve under any of the following *swap deviations:* (1) dropping any single purchased edge, (2) purchasing any single unpurchased edge, (3) dropping any single purchased edge and purchasing any single unpurchased edge, (4) any one of the deviations above and also changing the immunization status.

The first two deviations correspond to the standard linkstability. The third permits the more powerful *swapping* of one purchased edge for another. The last additionally allows reversing immunization status. Our interest in swapstable networks derives from the fact that while they only consider "simple" or "local" deviation rules, they share several properties with Nash networks that linkstable networks do not. Hence, swapstability is a bounded rationality concept that moves us closer to full Nash. Intuitively, in our game (and in many of our proofs), we exploit the fact that if a player is connected to some other set of vertices via an edge to a targeted vertex, and that set also contains an immune vertex, the player would prefer to connect to the immune vertex instead. This deviation involves a swap not just a single addition or deletion. It is worth mentioning explicitly that by definition every Nash equilibrium is a swapstable equilibrium and every swapstable equilibrium is a linkstable equilibrium. The reverse of none of these statements are true in our game. Also the set of equilibrium networks with respect to adversaries defined in Definitions 1, 2 and 3 are disjoint.

2.1 Related Work

Our paper is a contribution to the study of strategic network design and defense. The problem has been extensively studied in economics, electrical engineering,

[4] Lenzner [17] introduced this equilibrium concept under the name *greedy equilibrium.*

and computer science (see e.g. [1,2,11,18]). Most of the existing work takes the network as given and examines optimal security choices (see e.g. [3,8,12,14,16]). To the best of our knowledge, our paper offers the first model in which both links and defense (immunization) are chosen by the players.

Combining linking and immunization within a common framework yields new insights. We start with a discussion of the network formation literature. In a setting with no attack, our model with respect to the maximum carnage adversary reduces to the original model of one-sided reachability network formation of Goyal [4]. They showed that a Nash equilibrium network is either a tree or an empty network. By contrast, we show that in the presence of a security threat, Nash networks exhibit very different properties: both networks containing cycles and partially connected networks can emerge in equilibrium. We also show that while networks may contain cycles, they are sparse (we provide a tight upper bound on the number of links in any equilibrium network of our game).

Regarding security, a recent paper by Cerdeiro et al. [7] studies optimal design of networks in a setting where players make immunization choices against a maximum carnage adversary but the network design is given. They show that an optimal network is either a hub-spoke or a network containing k-critical vertices[5] or a partially connected network (a k-critical vertex can secure $n - k$ vertices by immunization). We extend this work by showing that there is a pressure toward the emergence of k-critical vertices even when linking is decentralized. We also contribute to the study of welfare costs of decentralization. Cerdeiro et al. [7] show that the Price of Anarchy (PoA) is bounded, when the network is centrally designed while immunization is decentralized (their welfare measure includes the edge expenditures of the planner). By contrast, we show that the PoA is unbounded when both decisions are decentralized. Although we also show that non-trivial equilibrium networks with respect to various adversaries have a PoA very near 1. This highlights the key role of linking and resonates with the original results on the PoA of pure network formation games [10].

Recently Blume et al. [6] study network formation where new links generate direct (but not reachability) benefits, infection can flow through paths of connections and immunization is not a choice. They demonstrate a fundamental tension between socially optimal and stable networks: the former lie just below a linking threshold that keeps contagion under check, while the latter admit linking just above this threshold, leading to extensive contagion and very low payoffs.

Finally, Kliemann [15] introduced a reachability game with attacks but without defense. In their model, the attack happens after the network is formed and the adversary destroys exactly one *link* (with no spread) according to a probability distribution over links that can depend on the structure of the network. They show equilibrium networks are chord-free and hence sparse. We also show an abstract representation of equilibrium networks in our model corresponds to chord-free graphs and then use this observation to prove sparsity. While both models lead to chord-free graphs in equilibria, the analysis of *why* these graphs

[5] Vertex v is k-critical in a connected network if the size of the largest connected component after removing v is k.

are chord-free is quite different. In their model, the deletion of a single link destroys at most one path between any pair of vertices. So if there were two edge-disjoint paths between any pairs of vertices, they will certainly remain connected after any attack. In our model the adversary attacks a vertex and the attack can spread and delete many links. This leads to a more delicate analysis. The welfare analysis is also quite different, since the deletion of an edge can cause a network to have at most two connected components, while the deletion of vertices might lead to many connected components.

3 Sparsity

In contrast to the original game [4], our game exhibits equilibrium networks with cycles, as well as disconnected but non-empty graphs. Figure 2 gives several examples of Nash networks with respect to the maximum carnage adversary for small populations, each of which is representative of a broad family of equilibria for large populations and a range of values for C_{E} and C_{I}.[6] So the tight characterization of the original game, where equilibrium networks are either empty graph or trees, fails to hold for our game. However, we show that an approximate version of this characterization continues to hold for several adversaries.

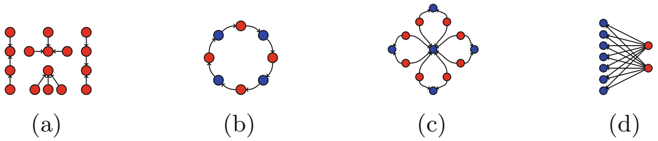


Fig. 2. Examples of equilibria with respect to the maximum carnage adversary: (a) Forest equilibrium, $C_{\mathrm{E}} = 1$ and $C_{\mathrm{I}} = 9$; (b) cycle equilibrium, $C_{\mathrm{E}} = 1.5$ and $C_{\mathrm{I}} = 3$; (c) 4-petal flower equilibrium, $C_{\mathrm{E}} = 0.1$ and $C_{\mathrm{I}} = 3$, (d) Complete bipartite equilibrium, $C_{\mathrm{E}} = 0.1$ and $C_{\mathrm{I}} = 4$. (Color figure online)

We show that despite the existence of equilibria containing cycles as shown in Fig. 2, under a very mild restriction on the adversary, *any* Nash, swapstable or linkstable equilibrium network of our game has at most $2n-4$ edges and is thus quite sparse. Moreover, this upper bound is tight as the generalized complete bipartite graph in Fig. 2d has exactly $2n-4$ edges.

The rest of this section is organized as follows. We start by defining a natural restriction on the adversary. We then propose an abstract view of the networks in our game and proceed to show that the abstract network is chord-free in equilibria with respect to the restricted adversary. We finally derive the edge density of the original network by connecting its edge density to the density of the abstract network. We start by defining equivalence classes for networks.

[6] We represent immunized and vulnerable vertices as blue and red, respectively. Although we treat the networks as undirected (the benefits and risks are bilateral), we use directed edges in some figures to denote which player purchased the edge.

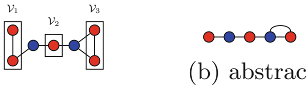
Definition 4. *Let $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ be two networks. G_1 and G_2 are* equivalent *if for all vertices $v \in V$, the connected component of v is the same in both G_1 and G_2 for every possible choice of initial attack vertex in V.*

Based on equivalence, we make the following natural restriction on the adversary.

Assumption 1. *An adversary is* well-behaved *if on any pair of equivalent networks $G_1 = (V, E_1)$ and $G_2 = (V, E_1)$, the probability that a vertex $v \in V$ is chosen for attack, is the same.*

The adversaries in Definitions 1–3 are all well-behaved. Next, we abstract the network formed by the agents and analyze the edge density in the abstraction.

Let $G = (V, E)$ be any network, $\mathcal{I} \subseteq V$ the immunized vertices and $\mathcal{V}_1, \ldots, \mathcal{V}_k$ the vulnerable regions in G. In the abstract network every vulnerable region in G is contracted to a single vertex. Formally, let $G' = (V', E')$ be the abstract network. Define $V' = \mathcal{I} \cup \{u_1, \ldots u_k\}$ where each u_i represents a contracted vulnerable region of G. E' is constructed as follows. For any edge $(v_1, v_2) \in E$ such that $v_1, v_2 \in \mathcal{I}$ there is an edge $(v_1, v_2) \in E'$. For any edge $(v_1, v_2) \in E$ such that $v_1 \in \mathcal{V}_i$ for some i and $v_2 \in \mathcal{I}$ there is an edge $(u_i, v_2) \in E'$ where u_i denotes the contracted vulnerable region of G that v_1 belongs to. For any edge (v_1, v_2) such that $v_1, v_2 \in \mathcal{V}_i$ for some i there is no edge in G' (see Fig. 3).



(a) original

(b) abstract

Fig. 3. Example of original and abstract network. Blue: immunized vertices in both networks. Red: the vulnerable vertices and regions in the original and abstract network, respectively. (Color figure online)

We next show that if G is an equilibrium network then G' is a chord-free graph. We defer all the omitted proofs to the full version.

Lemma 1. *Let $G = (V, E)$ be a Nash, swapstable or linkstable equilibrium network and $G' = (V', E')$ the abstraction of G. Then G' is a chord-free graph if the adversary is well-behaved.*

As the next step we bound the edge density of chord-free networks in Theorem 1 using tools from the graph theory literature.

Theorem 1. *Let $G = (V, E)$ be a chord-free graph on $n \geq 4$ vertices. Then $|E| \leq 2n - 4$.*[7]

[7] Kliemann [15] proved Theorem 1 with a different technique for a density bound of $2n - 1$ for all n.

Theorem 1 implies the edge density of the abstract network $G' = (V', E')$ is at most $2|V'| - 4$. To derive the edge density of the original network, we first show that any vulnerable region in G is a tree when G is an equilibrium network.

Lemma 2. *Let $G = (V, E)$ be a Nash, swapstable or linkstable equilibrium network. Then any vulnerable region in G is a tree if the adversary is well-behaved.*

We use Lemmas 1, 2 and Theorem 1 to prove our sparsity result.

Theorem 2. *Let $G = (V, E)$ be a Nash, swapstable or linkstable equilibrium network on $n \geq 4$ vertices. Then $|E| \leq 2n - 4$ for any well-behaved adversary.*

4 Connectivity and Social Welfare in Equilibria

The results of Sect. 3 show that despite the potential presence of cycles at equilibrium, there are still sharp limits on collective expenditure on edges. However, they do not directly lower bound the welfare, due to connectivity concerns: if the graph could become highly fragmented after the attack, or is sufficiently fragmented prior to the attack, the reachability benefits to players could be sharply lower than in the attack-free reachability game. We now show that when $C_{\rm I}$ and $C_{\rm E} > 1$ are both constants with respect to n,[8] none of these concerns are realized in any "interesting" equilibrium network, described precisely below.

In the original reachability game [4], the *maximum* welfare achievable in any equilibrium is $n^2 - O(n)$. Here we will show that the welfare achievable in any "non-trivial" equilibrium is $n^2 - O(n^{5/3})$. Obviously with no restrictions on the adversary and the parameters this cannot be true. Just as in the original game, for $C_{\rm E} > 1$, the empty graph with a social welfare of only $O(n)$ remains an equilibrium in our game with respect to all the natural adversaries in Sect. 2. We thus assume the equilibrium network contains at least *one* edge and at least *one* immunized vertex. We refer to all equilibrium networks that satisfy the above assumption as *non-trivial* equilibria. They capture the equilibria that are new to our game compared to the original attack-free setting — the network is not empty, and at least one player has chosen immunization.

Limiting attention to non-trivial equilibria is *necessary* if we hope to guarantee that the welfare at equilibrium is $\Omega(n^2)$ when $C_{\rm E} > 1$. As already noted, without the edge assumption, the empty graph is an equilibrium with respect to several natural adversaries. Furthermore, without the immunization assumption, $n/3$ disjoint components where each component consists of 3 vulnerable vertices is an equilibrium (for carefully chosen $C_{\rm E}$ and $C_{\rm I}$) with respect to e.g. the maximum carnage adversary. In both cases, the social welfare is only $O(n)$.

Similar to Sect. 3, to get any meaningful results for the welfare we need to restrict the adversary. To simplify presentation, we state and analyze our results for the maximum carnage adversary. We later show how these results (or their slight modifications) can be extended to several other adversaries.

[8] We view this condition as the most interesting regime, since in natural circumstances we do not expect the edge or immunization costs to grow with the population size.

Consider any connected component that contains an immunized vertex and an edge in a non-trivial equilibrium network with respect to the maximum carnage adversary. We first show that any targeted region in such component (if exists) has size 1 when $C_{\mathrm{E}} > 1$.

Lemma 3. *Let G be a non-trivial Nash or swapstable equilibrium network with respect to the maximum carnage adversary. Then in any component of G with at least one immunized vertex and at least one edge, the targeted regions (if they exist) are singletons when $C_E > 1$.*

We then show that non-trivial equilibrium networks with respect to the maximum carnage adversary are connected when $C_{\mathrm{E}} > 1$.

Theorem 3. *Let G be a non-trivial Nash, swapstable or linkstable equilibrium network with respect to the maximum carnage adversary. Then, G is a connected graph when $C_E > 1$.*

So any non-trivial equilibrium network with respect to maximum carnage adversary is a connected network with targeted regions of size 1. Finally, we state our main result regarding the welfare in such non-trivial equilibria.

Theorem 4. *Let G be a non-trivial Nash or swapstable equilibrium network on n vertices with respect to the maximum carnage adversary. If C_E and C_I are constants (independent of n) and $C_E > 1$ then the welfare of G is $n^2 - O(n^{5/3})$.*

Our proof techniques for Theorem 4 might not extend to non-trivial linkstable equilibrium networks with respect to the maximum carnage adversary since such networks can have targeted regions of size bigger than 1 when $C_{\mathrm{E}} > 1$.

Remarks. We proved our sparsity result with a rather mild restriction on the adversary. However, we presented our welfare results only with respect to the maximum carnage adversary. Our proofs in this section only rely on the following two properties of the maximum carnage adversary: (1) Adding an edge between any 2 vertices (at least 1 of which is immunized) does not change the distribution of the attack and (2) Breaking a link inside of a targeted region does not increase the probability of attack to the targeted region while at the same time does not decrease the probability of attack to any other vulnerable region. The same properties hold for the random attack adversary and other adversaries that set the probability of attack to a vulnerable region directly proportional to an increasing function of the size of the region. Thus our welfare results extend to random attack adversary and other such adversaries without any modifications.

However, some natural adversaries (e.g. the maximum disruption adversary) might not satisfy these properties. While the same techniques might not be directly applicable to such adversaries, it is still possible to reason about the welfare using different methods e.g. we can still show that in any non-trivial and *connected* equilibrium with respect to the maximum disruption adversary, when C_{E} and C_{I} are constants and $C_{\mathrm{E}} > 1$, then the welfare is $n^2 - O(n^{5/3})$. See the full version for more details.

5 A Behavioral Experiment

To complement our theory, we conducted a behavioral experiment on our game with 118 participants. The participants were students in an undergraduate survey course on network science at the University of Pennsylvania. As training, participants were given a detailed document and lecture on the game, with simple examples of payoffs for players on small graphs under various edge purchase and immunization decisions. (See http://www.cis.upenn.edu/~mkearns/teaching/NetworkedLife/NetworkFormationExperiment2015.pdf for the training document provided to participants.) Participation was a course requirement, and students were instructed that their grade on the assignment would be exactly equal to their payoffs according to the rules of the game.

The payoffs used the maximum carnage adversary, with costs of $C_{\mathrm{E}} = 5$ and $C_{\mathrm{I}} = 20$. With $n = 118$ participants (so a maximum connectivity benefit of 118 points), it felt that these values made edge purchases and immunization significant expenses and thus worth careful deliberation. Second, running swapstable best response simulations using these values generally resulted in non-trivial equilibria with high welfare, whereas raising C_{E} and C_{I} significantly generally resulted in empty or fragmented graphs with low welfare.

In a game of such complexity, with so many participants, it is unreasonable and uninteresting to formulate the experiment as a one-shot simultaneous move game. Rather, some form of communication must be allowed. We chose to conduct the experiment in a courtyard with the single ground rule that *all conversations be quiet and local* i.e. in order to hear what a participant was saying to others, one should have to stand next to them.

Other than the quiet rule, there were no restrictions on the nature of conversations: participants were free to enter agreements, make promises or threats and move freely. However, it was made clear that any agreements or bargains struck would *not* be enforced by the rules of the experiment (thus were non-binding). Each subject was given a handout that required them to indicate which other subjects they chose to purchase edges to (if any), and whether or not they chose to purchase immunization. The handout contained a list of subject names, along with a unique identification number for each subject used to indicate edge purchases. Thus subjects knew the names of the others as well as their assigned ID numbers. An entire class session was devoted to the experiment, but subjects were free to (irrevocably) turn in their handout at any time and leave. Subjects committed and exited sequentially, and the entire duration was approximately 30 min. During the experiment, subjects tended to gather quickly in small discussion groups that reformed frequently, with subjects moving freely from group to group. It is clear from the outcome that despite adherence to the quiet rule, the subjects engaged in widespread coordination via this rapid mixing.

In the left panel of Fig. 4, we show the final undirected network formed by the edge purchases and immunization decisions. The graph is clearly anchored by two main immunized hub vertices, each with many spokes who purchased their single edge to the respective hub. These two large hubs are both directly connected, as well as by a longer "bridge" of three vulnerable vertices. There is also a smaller hub with just a handful of spokes, again connected to one of the larger hubs via a chain of two vulnerable vertices.

For the payoffs, inspection of the network reveals that there are 2 groups of 3 vertices that are the largest vulnerable connected components, and thus are the targets of the attack. These 6 players are each killed with probability 1/2 for a payoff that is only half that of the wealthiest players (the vulnerable spokes of degree 1). In between are the players who purchased immunization including the 3 hubs and 2 immunized spokes. The immunized spoke of the upper hub is unnecessarily so, while the immunized spoke in the lower hub is best responding — had they not purchased immunization, they would have formed a unique largest vulnerable component of size 4 and thus been killed with certainty.

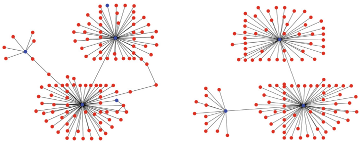


Fig. 4. Left: the final undirected network formed by the edge purchases and immunization decisions (blue for immunized, red for vulnerable). Right: a "nearby" Nash network. (Color figure online)

It is striking how many properties the behavioral network shares with the theory: multiple hub-spoke structures with sparse connecting bridges, resulting in high welfare and a heavy-tailed degree distribution; a couple of cycles. To quantify how far the behavioral network is from equilibrium we use it as the starting point for swapstable best response dynamics and run it until convergence. In the right panel of Fig. 4, we show the resulting Nash network reached from the behavioral network in only 4 rounds of swapstable dynamics, and with only 15 of 118 vertices updating their choices. The dynamics simply *clean up* some suboptimal behavioral decisions — the vulnerable bridges between hubs are replaced by direct edges, the other targeted group of three spokes drops theirs fatal edges, and immunizing spokes no longer do so.

Participants were required to complete a survey after the experiment: they were asked to comment on any strategies they contemplated prior to the experiment; whether and how those strategies changed during the experiment; and what strategies or behaviors they observed in other participants.

Many subjects reported entering the experiment with not just a strategy for themselves, but also a "master plan" they hoped to convince others to join. One frequently reported plan involved variations on cycles. Though little thought seems to have been given to how to coordinate a global ordering in a cycle via only the quiet rule. Another frequently cited plan involved the hub-spoke. Although most strategies are based on abstractions, others reported planning to use real-world social relationships e.g. connecting to students they know.

Of course, of particular interest are the surveys of the hubs. One seems to report an altruistic motivation for purchasing immunization, hoping to maximize welfare. In contrast, the other displays a more Machiavellian attitude and was willing to immunize in the hopes of creating 3 distinct groups of participants: the "winners" who would connect to the hub; the hub with slightly lower payoff; a large group of "losers" deliberately left out of the hub-spoke structure.

It is clear from the surveys that the word quickly spread during the experiment to connect to hubs and that many participants joined though not without some reported mistrust and hesitation.

Acknowledgments. We thank Chandra Chekuri, Yang Li and Aaron Roth for useful suggestions. We also thank anonymous reviewers for detailed suggestions regarding some of the proofs. Sanjeev Khanna is supported in part by National Science Foundation grants CCF-1552909, CCF-1617851, and IIS-1447470.

References

1. Alpcan, T., Baar, T.: Network security: a decision and game-theoretic approach, 1st edn. Cambridge University Press, Cambridge (2010)
2. Anderson, R.: Security engineering: a guide to building dependable distributed systems, 2nd edn. Wiley Publishing (2008)
3. Aspnes, J., Chang, K., Yampolskiy, A.: Inoculation strategies for victims of viruses and the sum of squares partition problem. J. Comput. Syst. Sci. **72**(6), 1077–1093 (2006)
4. Bala, V., Goyal, S.: A noncooperative model of network formation. Econometrica **68**(5), 1181–1230 (2000)
5. Bloch, F., Jackson, M.: Definitions of equilibrium in network formation games. Int. J. Game Theo. **34**(3), 305–318 (2006)
6. Blume, L., Easley, D., Kleinberg, J., Kleinberg, R., Tardos, É.: Network formation in the presence of contagious risk. In: EC, pp. 1–10 (2011)
7. Cerdeiro, D., Dziubinski, M., Goyal, S.: Contagion risk and network design. Working Paper (2014)
8. Cunningham, W.: Optimal attack and reinforcement of a network. J. ACM **32**(3), 549–561 (1985)
9. Fabrikant, A., Luthra, A., Maneva, E., Papadimitriou, C., Shenker, S.: On a network creation game. In: PODC, pp. 347–351 (2003)
10. Goyal, S.: Connections: an introduction to the economics of networks. Princeton University Press, Princeton (2007)
11. Goyal, S.: Conflicts and Networks. The Oxford Handbook on the Economics of Networks (2015)
12. Gueye, A., Walrand, J., Anantharam, V.: A network topology design game, how to choose communication links in an adversarial environment. In: GameNets (2011)
13. Ihde, S., Keßler, C., Neubert, S., Schumann, D., Lenzner, P., Friedrich, T.: Efficient best-response computation for strategic network formation under attack. CoRR abs/1610.01861 (2016)
14. Kearns, M., Ortiz, L.: Algorithms for interdependent security games. In: NIPS, pp. 561–568 (2003)
15. Kliemann, L.: The price of anarchy for network formation in an adversary model. Games **2**(3), 302–332 (2011)

16. Laszka, A., Szeszlér, D., Buttyán, L.: Linear loss function for the network blocking game: an efficient model for measuring network robustness and link criticality. In: Grossklags, J., Walrand, J. (eds.) GameSec 2012. LNCS, vol. 7638, pp. 152–170. Springer Berlin Heidelberg, Berlin, Heidelberg (2012). doi:10.1007/978-3-642-34266-0_9
17. Lenzner, P.: Greedy selfish network creation. In: Goldberg, P.W. (ed.) WINE 2012. LNCS, vol. 7695, pp. 142–155. Springer, Heidelberg (2012). doi:10.1007/978-3-642-35311-6_11
18. Roy, S., Ellis, C., Shiva, S., Dasgupta, D., Shandilya, V., Wu, Q.: A survey of game theory as applied to network security. In: HICSS, pp. 1–10 (2010)