Optimal Welfare in Noncooperative Network Formation Under Attack

Natan Doubez[1], Pascal Lenzner[2], Marcus Wunderlich[2]

[1]École Polytechnique, Palaiseau, France
[2]University of Augsburg, Augsburg, Germany
natan.doubez@polytechnique.org, {pascal.lenzner,marcus.wunderlich}@uni-a.de

Abstract

Communication networks are essential for our economy and our everyday lives. This makes them lucrative targets for attacks. Today, we see an ongoing battle between criminals that try to disrupt our key communication networks and security professionals that try to mitigate these attacks. However, today's networks, like the Internet or peer-to-peer networks among smart devices, are not controlled by a single authority, but instead consist of many independently administrated entities that are interconnected. Thus, both the decisions of how to interconnect and how to secure against potential attacks are taken in a decentralized way by selfish agents.

This strategic setting, with agents that want to interconnect and potential attackers that want to disrupt the network, was captured via an influential game-theoretic model by Goyal, Jabbari, Kearns, Khanna, and Morgenstern (WINE 2016). We revisit this model and show improved tight bounds on the achieved robustness of networks created by selfish agents. As our main result, we show that such networks can resist attacks of a large class of potential attackers, i.e., these networks maintain asymptotically optimal welfare post attack. This improves several bounds and resolves an open problem. Along the way, we show the counter-intuitive result, that attackers that aim at minimizing the social welfare post attack do not actually inflict the greatest possible damage.

Extended version — https://arxiv.org/abs/2511.10845

Introduction

Given the importance of today's networks, they have always been prone to attack and network operators have always tried to secure their networks. A prominent example is the use of firewalls in routers to prevent a computer virus from infecting other parts. Also, social distancing and vaccination measures can be understood as security measures in a (social) network to prevent the spread of a virus.

However, both examples show that in today's networks there is no central authority that could enforce certain security measures or a certain structure of the network. The use of a firewall or to vaccinate is an individual decision of the participants of the networks. The same holds for the decision of which connections to establish: while a direct connection yields benefit in terms of low latency, it also poses a risk since a possible attack could spread along the created link. Thus, today's networks can be better modeled and understood as a complex multi-agent system consisting of strategic smart agents. These agents can be people, routers, smart devices, or simply interacting components of an AI system. Each agent strategically decides on its security measures and on its links it wants to establish towards other agents.

This viewpoint of networks as multi-agent systems of strategic agents has sparked the multifaceted research on game-theoretic network formation models in Economics, Computer Science, and Artificial Intelligence in the last three decades, see e.g. (Papadimitriou 2001; Jackson et al. 2008). In these models, agents that correspond to nodes of a network strategically decide which connections to other nodes to establish. The agents are selfish and act according to some given utility function that encodes the agents' objectives. In this paper, we focus on the objective of network robustness. Modern communication and infrastructure networks have to cope with hardware failures or even deliberate attacks while still providing a reliable service. For incorporating this, researchers have studied agent-based network formation models where the agents prepare for single-link failures (Bala and Goyal 2003; Meirom, Mannor, and Orda 2015; Chauhan et al. 2016) or try to maximize the obtained min-cut in the created network (Echzell et al. 2020). Also the scenario of fighting a virus that spreads in the created network was considered and this is our main reference point.

In this work we revisit the elegant strategic network formation model with attack and immunization by Goyal, Jabbari, Kearns, Khanna, and Morgenstern (Goyal et al. 2016) that features a smart adversary. Each agent selfishly decides which costly links to form and if to acquire protection from attack. In such an attack, a single node of the network is targeted for infection and then this infection spreads along the subgraph consisting of unprotected nodes. The authors consider three natural types of attack:

- **maximum carnage:** the attacker targets a node to infect as many nodes as possible,
- **random attack:** the attacker targets an unprotected node uniformly at random, and
- **maximum disruption:** the attacker targets a node to minimize the social welfare post attack.

The utility of uninfected agents is the (expected) number of reachable uninfected nodes post attack, while infected agents have utility zero. The social welfare is the total utility.

Goyal et al. (2016) give non-trivial bounds on the social welfare of equilibrium networks for the maximum carnage and the random adversary and they pose the analysis of the maximum disruption attacker as open problem. In this paper, we completely resolve the question of the social welfare by providing tight optimal bounds for all three attackers. Even more, we show that the optimal bound holds for more general class of attackers that subsumes the maximum disruption attacker. In fact, we show that the created equilibrium networks have asymptotically the same social welfare post attack as in a setting without adversary. This highlights that networks that are created in a decentralized way by strategic agents are highly robust against various types of attackers.

Model and Preliminaries

Sets & Functions: We use $\mathbb{N}$ to denote the set of natural numbers and $\mathbb{R}^+$ to denote the set of non-negative real numbers. We will refer to the derivative of a discrete function $f : \mathbb{N} \to \mathbb{R}^+$ as the function $f' : n \mapsto f(n+1) - f(n)$.

The Game: For our network formation game, we mostly use the original notation by Goyal et al. (2016). A game instance is defined by the tuple $(n, C_E, C_I, \mathcal{A})$, where n is the number of agents (or nodes of the network), the value $C_E > 1$ is the cost at which agents can buy an edge to any other node, the cost for a player to *immunize* itself is $C_I > 0$, and $\mathcal{A}$ describes the attacker (or opponent) targeting the created network. We use $[n] \coloneqq \{1, \dots, n\}$ as the set of agents and we use the terms agents and nodes interchangeably.

Strategies: Every agent can (1) buy arbitrarily many undirected incident edges to other agents, at a cost of C_E per edge, and (2) immunize itself at a cost of C_I. The set of nodes to which an agent $i \in [n]$ buys an edge is $X_i \subseteq [n]$, where $j \in X_i$ indicates that agent i buys the edge $\{i, j\}$. Whether agent i immunizes itself is represented via Boolean variable $y_i \in \{0, 1\}$, i.e., $y_i = 1$ if and only if agent i is immunized. If $y_i = 0$, agent i is called *vulnerable*. The pair $s_i \coloneqq (X_i, y_i)$ is the *strategy* of agent i, and a vector $\mathbf{s} = (s_1, \dots, s_n)$ of strategies of all agents is called a *strategy profile*. Every strategy profile $\mathbf{s}$ induces an undirected network $G(\mathbf{s}) = (V, E(\mathbf{s}))$, with $V \coloneqq [n]$ being the set of agents, partitioned into immunized agents $\mathcal{I}(\mathbf{s})$ and vulnerable agents $\mathcal{U}(\mathbf{s})$. The set $E(\mathbf{s}) = \{\{i, j\} \mid i \in X_j \vee j \in X_i\}$ is the set of bought edges. If it is clear from the context, we will omit the reference to the strategy profile $\mathbf{s}$.

Attacks: Given a strategy profile $\mathbf{s}$ and its induced graph $G(\mathbf{s}) = (\mathcal{I}(\mathbf{s}) \cup \mathcal{U}(\mathbf{s}), E(\mathbf{s}))$, the attacker $\mathcal{A}$ will choose a single node as target. Attacking node $v \in \mathcal{U}(\mathbf{s})$ *infects* its connected component in the network $G[\mathcal{U}(\mathbf{s})]$ induced by the vulnerable agents (while attacking an immunized node has no impact). Infected nodes will be completely removed from the network. We call the connected components of $G[\mathcal{U}(\mathbf{s})]$ *vulnerable regions* and define $\mathcal{V}(\mathbf{s})$ to be the set of all vulnerable regions of $G(\mathbf{s})$. We similarly call a connected component of $G[\mathcal{I}(\mathbf{s})]$ an *immunized region*. Formally, an *attacker* $\mathcal{A}$ is defined by a probability distribution over the vulnerable regions $\mathcal{V}(\mathbf{s})$ that states how likely it is for $\mathcal{A}$ to attack each of them. This distribution heavily depends on the type of attacker, that we will describe in detail later. Given a fixed attacker $\mathcal{A}$, we refer to the vulnerable regions of $G(\mathbf{s})$ that $\mathcal{A}$ attacks with non-zero probability as *targeted regions* and define $\mathcal{T}(\mathbf{s}, \mathcal{A})$ as the set of all regions targeted by $\mathcal{A}$. All nodes inside a targeted region are called *targeted nodes*.

Utilities: The utility of an agent i in profile $\mathbf{s}$ is a combination of its connectivity in network $G(\mathbf{s})$ post attack minus the agent's costs for edges and immunization. Formally, the post attack connectivity of agent i is the expected size of i's connected component after the attacker infected (and thus destroyed) a targeted region. To be precise, let $CC_i(T, \mathbf{s})$, for some vulnerable region T, be the size of the connected component of $G[V \setminus T]$ that contains agent i. Then, the *connectivity* of agent i in strategy profile $\mathbf{s}$ with attacker $\mathcal{A}$ is

$$\mathbb{E}_{\mathcal{T}(\mathbf{s},\mathcal{A})}[CC_i(\mathbf{s})] \coloneqq \textstyle\sum_{T \in \mathcal{T}(\mathbf{s},\mathcal{A})} \left(\mathbb{P}_{\mathcal{A}}[T, \mathbf{s}] \cdot CC_i(T, \mathbf{s})\right),$$

where $\mathbb{P}_{\mathcal{A}}[T, \mathbf{s}]$ is the probability at which $\mathcal{A}$ attacks T in profile $\mathbf{s}$. Further, the *utility* of agent i in this setting then is

$$u_i(\mathbf{s}) \coloneqq \mathbb{E}_{\mathcal{T}(\mathbf{s},\mathcal{A})}[CC_i(\mathbf{s})] - |X_i| C_E - y_i C_I,$$

where $|X_i| C_E + y_i C_I$ is the *cost* of agent u in strategy profile $\mathbf{s}$. The sum over the utilities of all agents is called the *social welfare* of profile $\mathbf{s}$. We sometimes write $CC_i(v, \mathbf{s})$ instead of $CC_i(T, \mathbf{s})$, where T is the targeted region containing $v \in \mathcal{U}$. And similarly $U_f(v, \mathbf{s})$ for $U_f(T, \mathbf{s})$.

Equilibria: Given strategy profile $\mathbf{s}$, we say that agent i *plays best response*, if given the strategies of all other agents, agent i cannot deviate to a different strategy that yields strictly higher utility. If all agents play best response with respect to strategy profile $\mathbf{s}$, then we say that $\mathbf{s}$ is a *Nash equilibrium*, and we say that $G(\mathbf{s})$ is an *equilibrium network*. We say that profile $\mathbf{s}$ is a *non-trivial* Nash equilibrium, if $G(\mathbf{s})$ has at least one edge and if $|\mathcal{I}(\mathbf{s})| > 0$. These equilibria do not exist in the attack-free setting (Bala and Goyal 2000).

Different Types of Attackers: The simplest attacker is the one that attacks every vulnerable node uniformly at random. We refer to this one as *random attack*. This attacker, and the two other attackers, *maximum carnage* and *maximum disruption*, introduced by Goyal et al. (2016), belong to the class of *well-behaved opponents* that have the same attack distribution for equivalent networks (see Definition 4 in (Goyal et al. 2016) for details). We will consider a subclass of this, called f-opponents, that subsumes the maximum carnage and the maximum disruption opponents. For f-opponents, every attack is defined by a function $f\colon \{0, \dots, n\} \to \mathbb{R}^+$ that maps every size of a possible connected component of the network to some non-negative value, with[1] $f(0) = 0$. With this, an f-opponent $\mathcal{A}$ aims to minimize $U_f(T, \mathbf{s}) \coloneqq \sum_{K \in \mathcal{K}(T)} f(|K|)$, where $\mathcal{K}(T)$ is the set of connected components of the network after $\mathcal{A}$ attacked some vulnerable region $T \in \mathcal{T}(\mathbf{s}, \mathcal{A})$. Formally, the

[1] Setting $f(0) = 0$ is not necessary but useful. E.g. when the opponent targets a region inside a connected component Z that does not create a new component. The change in utility is the difference of the images of f of the sizes of Z before and after the attack. Setting $f(0) = 0$ allows it to still hold if the attack completely deletes Z. It is also useful when using the convexity of f.

f-opponent attacks only vulnerable regions $T \in \mathcal{T}(\mathbf{s}, \mathcal{A})$ of $G(\mathbf{s})$ that minimize $U_f(\cdot, \mathbf{s})$ and chooses uniformly at random among them. In this framework, the *maximum carnage* attacker and the *maximum disruption* can be defined to be x^1- and x^2-opponents respectively, where $x^r \colon n \mapsto n^r$ is the monomial of degree r for $r \in \mathbb{N}$.

Example 1. *To better understand the difference between the maximum carnage and maximum disruption attackers, consider the instance depicted in Figure 1. The maximum carnage attacker would treat both vulnerable regions A and B of size 7 the same. In contrast, the maximum disruption attacker would uniquely attack node k in profile $\mathbf{s}$, since $U_{x^2}(C, \mathbf{s}) = 7^2 + 2^2 + 8^2 + 7^2 = 166$, while $U_{x^2}(A, \mathbf{s}) = 11^2 + 7^2 = 170$. So, targeting region C minimizes it. However, in $\mathbf{s}'$ we have $U_{x^2}(C, \mathbf{s}') = 9^2 + 8^2 + 7^2 = 194$, while $U_{x^2}(A, \mathbf{s}') = 11^2 + 7^2 = 170$. Hence, only A is targeted.*

This shows the counter-intuitive behavior that agent i prefers to buy the edge e even if it costs $C_E = 9 - \varepsilon$, for $\varepsilon > 0$, to prevent the destruction of node k. In fact, if, as indicated in Figure 1, agent i buys no other edges, we have $u_i(\mathbf{s}') = 11 - C_E - C_I$ and $u_i(\mathbf{s}) = 2 - C_I$. Such behavior makes the maximum disruption attacker difficult to analyze, as tiny strategy changes may completely shift the targeted regions. Also, a best response strategy may create an edge to a node, like node j, that for sure gets destroyed. ◁

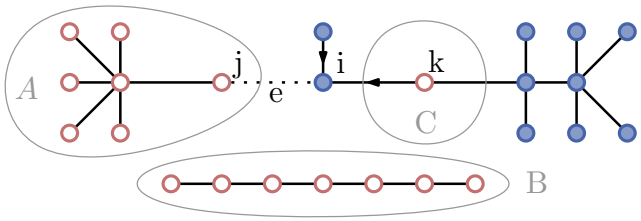

Figure 1: Strategy profile $\mathbf{s}$ without edge $e = \{i, j\}$ and $\mathbf{s}'$ where e is bought by i. Red (blue) nodes are vulnerable (immunized). A, B and C are vulnerable regions. Without edge e the maximum carnage attack randomizes between infecting a node in A or B, while the maximum disruption attacker targets node k. In $\mathbf{s}'$, the latter targets a node in A. Arrows indicate edge ownership, directed away from the owner.

However, besides the counter-intuitive behavior in Example 1, the maximum disruption opponent still respects some properties. For instance it is *edge-averse*, i.e., removing edges can only reduce the size of the connected components of each agent, and therefore their utility. We will see that this property is a direct consequence of the strict convexity of the function x^2. Also, this opponent is biased towards targeting the largest components as they contribute more to the social welfare. For our proofs, any function f that grows faster than x^2 will have the same bias. This leads us to naturally encapsulate these properties into a class of attackers, called *super-quadratic-disruptor (SQD)*, that we will use for the remainder of this paper to prove statements for maximum disruption opponent and similar attackers. A SQD is an f-opponent such that the function f satisfies: (1) strict convexity, and (2) the function $f(x)/x^2$ is non-decreasing.

Related Work

The study of network formation games has a long tradition in Economics, Computer Science and AI. We restrict our discussion to models and results that are closely related to our work and that of Goyal et al. (2016).

The objective of ensuring reachability of all other agents was proposed by Bala and Goyal (2000) in a setting that is identical to our model but without attack or immunization. They show that equilibrium networks are either empty or trees, depending on the edge price C_E. For non-empty equilibria the social welfare is $n^2 - \mathcal{O}(n)$, while the empty network has welfare in $\mathcal{O}(n)$. Moreover, they investigate convergence dynamics for finding equilibria. Also, versions with directed edges or connection benefit decay are studied. Later, the authors augmented the model with single edge failures (Bala and Goyal 2003) and find that agents build minimally more edges to ensure post failure connectivity. Haller and Sarangi (2005) extended the model by allowing different failure probabilities per edge and Kliemann (2011, 2013) studied the price of anarchy of several versions.

Our reference point, the model by Goyal et al. (2016), is also a direct extension of the reachability model by Bala and Goyal (2000). As one of their main results, they show for the very broad class of well-behaved opponents (and also for multiple stability concepts other than Nash equilibria) that obtained equilibria have at most $2n - 4$ edges and that vulnerable regions are trees. Since our f-opponents are well-behaved, these results carry over to our analysis. Moreover, the authors show for $C_E, C_I > 1$ that the social welfare of non-trivial equilibria with respect to the maximum carnage and random attack opponent is $n^2 - \mathcal{O}(n^{5/3})$. The same holds for maximum disruption, but only for connected equilibria. For $C_E > 1$, the empty network is an equilibrium with welfare $\mathcal{O}(n)$ and also for $C_I \leq 1$ such bad equilibria exist. Most important for us, they leave the general analysis of the maximum disruption opponent as an open problem.

The complexity of computing a best response strategy in the model by Goyal et al. (2016) was analyzed by Friedrich et al. (2017). They give a polynomial time algorithm for the maximum carnage and the random attack opponents. Later, Àlvarez and Messegué (2023) also achieved this for the maximum disruption attacker. Another extension by Chen et al. (2019) is the natural variant of the model where the infection spreads with probability $p > 0$ over each link. For this, the authors only consider the random attack opponent and show that equilibria have at most $\mathcal{O}(n \log n)$ many edges and that equilibria with $\Omega(n)$ edges exist. Also, they show that the expected social welfare of equilibria with $\mathcal{O}(n)$ edges is in $\Theta(n^2)$, i.e., asymptotically optimal and comparable to the setting without attacker.

Also models with other objectives exist. Close to us is the model by Echzell et al. (2020), where agents buy edges to maximize the min-cut in the network. A model with an intermediary, and where edges can have different connection strength was studied by Anshelevich, Bhardwaj, and Kar (2015). Another large class of related models are network formation games where the objective of an agent is centrality in the created network. Starting from the semi-

nal works of Jackson and Wolinsky (1996) and Fabrikant et al. (2003), many variants have been studied. Among them, there are also versions focusing on the robustness of the created network (Meirom, Mannor, and Orda 2015; Chauhan et al. 2016), versions that focus on social networks (Bilò et al. 2021) or social distancing (Friedrich et al. 2022), variants with temporal edges (Bilò et al. 2023, 2025), homophilic agents (Bullinger, Lenzner, and Melnichenko 2022), or greedy routing (Berger et al. 2025).

Our Contribution

We revisit the elegant noncooperative network formation model with attack and immunization by Goyal et al. (2016). Their main goal was to study the impact of the attacker on the obtained equilibrium networks. They find that non-trivial equilibria with respect to the maximum carnage or random attack opponents have a social welfare of $n^2 - \mathcal{O}(n^{5/3})$, whereas without attacker the social welfare is $n^2 - \mathcal{O}(n)$. Due to the n^2-term, this is asymptotically optimal. However, note that the n^2-term is a result of using a "benefit minus cost" utility function instead of using a pure cost function. For every agent, every reachable node yields a connection benefit of 1 whereas unreachable nodes contribute value 0. This essentially adds the n^2-term to the social welfare. In contrast, with a pure cost function, where a reachable node contributes cost 0 and unreachable nodes yield cost 1, the n^2-term would vanish and the obtained bounds would be far from being tight. We resolve this by proving robust tight bounds on the social welfare. See Table 1 for an overview.

	Goyal et al. (2016)	Our Results
Max. Carnage	$n^2 - \mathcal{O}(n^{5/3})$	$n^2 - \mathcal{O}(n)$
Random Attack	$n^2 - \mathcal{O}(n^{5/3})$	$n^2 - \mathcal{O}(n)$
Max. Disruption	open problem	$n^2 - \mathcal{O}(n)$
Tailored	not considered	$\mathcal{O}(n)$

Table 1: Comparison of our results on the social welfare of non-trivial Nash equilibria with the previous results. "Tailored" means a specifically designed opponent.

As our main result, we show the optimal social welfare bound of $n^2 - \mathcal{O}(n)$ for super-quadratic-disruptor opponents, which subsume the maximum disruption opponent. This shows that even with attackers that try to minimize the post attack social welfare, the obtained welfare is asymptotically the same as in the setting without attacker. Thus, smart strategic agents can cope with strong attackers and still build networks with optimal welfare. This solves an open problem, since for the maximum disruption opponent only a bound of $n^2 - \mathcal{O}(n^{5/3})$ for *connected* equilibrium networks was known (Goyal et al. 2015). Thus, we give a general optimal bound for a larger class of attackers.

Finally, we show that the optimal welfare bound does not hold for all attackers. For this, we prove the counter-intuitive result that opponents exist, that enforce asymptotically lower social welfare post attack than the attacker that aims at minimizing this value. In fact, we present a tailored attacker that achieves the low social welfare of empty networks.

Improved Social Welfare for Maximum Carnage and Random Attack

We study the (expected) social welfare post attack of equilibria with respect to the maximum carnage and the random attack opponent. For this, we focus on the block-cut decomposition of any equilibrium network $G = (V, E)$:

Definition 1. *Consider a network $G = (V, E)$. A node $v \in V$ is called a* cut-vertex *if $G[V \setminus v]$ has more connected components than G. A network is* non-separable *if it is connected and has no cut-vertices. Every subnetwork of G that is non-separable and maximal in that property is called a* block. *With that, consider the set of blocks $\mathcal{B}(G)$ and the set of cut-vertices $\mathcal{S}(G)$ of G. Connecting all cut-vertices to all blocks they are part of yields a tree $(\mathcal{B}(G) \cup \mathcal{C}(G), E')$ that is called the* block-cut decomposition *of G.*

First, we bound the length of specific paths in the block-cut decomposition.

Lemma 1. *Take the block-cut decomposition $(B(G) \cup C(G), E')$ of an equilibrium network $G = (V, E)$ with respect to the maximum carnage or the random attack opponent. If a path $b_1, c_1, b_2, c_2, ...b_r$, that goes through p vulnerable cut-vertices exists, then $p \leq 2C_E + 1$.*

Proof Sketch. Let $G(\mathbf{s})$ be the equilibrium network and consider the vulnerable cut-vertices on the path in the block-cut decomposition of $G(\mathbf{s})$. Moreover, let $x_1 \in b_1$ and $x_r \in b_r$ be two immunized agents. See Figure 2 for an illustration.

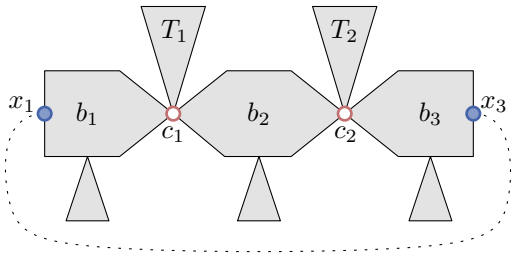

Figure 2: The block-cut decomposition of $G(\mathbf{s})$ with two vulnerable cut-vertices c_1 and c_2. The dotted edge $\{x_1, x_r\}$ is bought in both strategy profiles $\mathbf{s}'_1$ and $\mathbf{s}'_2$.

Now consider the strategy profiles $\mathbf{s}'_1$ and $\mathbf{s}'_2$ that result from either agent x_1 buying an edge to agent x_r or vice versa in strategy profile $\mathbf{s}$. Note, that the connectivity of agent x_1 in profile $\mathbf{s}'_1$ and of agent x_r in profile $\mathbf{s}'_2$ is the same.

By computing the obtained utilities, we show that the desired result is a consequence of the fact that neither agent x_1 nor agent x_r can improve from this deviation. □

Next, we use the concept of a centroid of a tree, extend it to connected components and show that it always exists.

Definition 2. *Let Z be a connected component in $G(\mathbf{s})$. A node $c \in Z$ is called* centroid *of Z if it respects the property that if it immunizes (or stays immunized if it already was), then after any vulnerable region in Z is destroyed, the size of the connected component containing c is at least $|Z|/2$.*

Lemma 2. *Every connected component of G has a centroid.*

From the proof of Lemma 2, we get the following:

Corollary 1. *In any connected component Z of a network $G(\mathbf{s})$, one of the two following properties must hold:*

- *there exists an immunized centroid of Z,*
- *there exists a vulnerable region containing a centroid that if removed, every remaining connected component has size $< |Z|/2$.*

In the following, we use a result from Goyal et al. (2016) that holds for all well-behaved opponents.

Lemma 3. *(Lemma 2 & Theorem 2 of Goyal et al. (2016)) Let $G = (V, E)$ be an equilibrium network with a well-behaved opponent. Then*

1. *all vulnerable regions in G are trees, and*
2. *$|E| \leq 2n - 4$, for all $n \geq 4$.*

Now, we can prove the main result of this section.

Theorem 1. *The social welfare of non-trivial Nash equilibria with respect to the maximum carnage or the random attack opponent is $n^2 - O(n)$.*

Proof Sketch. Consider the immunized agent x and the block-cut decomposition of the network rooted in the block containing node x. For every $0 \leq i \leq 2C_E + 1$, define H_i the set of vulnerable cut-vertices of rank i. The rank is defined as follows. Let H_0 is the set of vulnerable cut-vertices such that no other vulnerable cut-vertices are in the subtree of this node. Then H_{i+1} is the set of cut-vertices whose subtree contains a different node in H_i. Finally, let $H = \cup_i H_i$ the set of all vulnerable cut-vertices. See Figure 3.

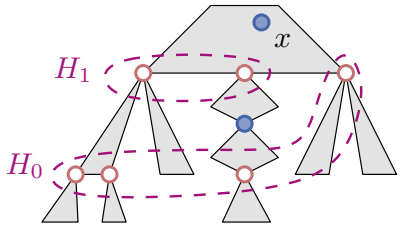

Figure 3: The layers H_i from the proof of Theorem 1.

Note that for a given $0 \leq i \leq 2C_E + 1$ and a node y, there is at most one node in H_i that disconnects nodes x and y when removed. Therefore, $\sum_{z \in H_i} CC_x(z, \mathbf{s}) \geq n(|H_i| - 1)$, since every node is in all of these components except for at most one. Also, all targeted regions are singletons. Therefore,

$$\mathbb{E}_{\mathcal{U}(\mathbf{s})}[CC_x(\mathbf{s})] \geq n - 2 - 4C_I(C_E + 1).$$

Finally the social welfare is the expected value of the sum of squares of connected components minus the cost. This is higher or equal to the expected value of the square of the connected component of x minus the cost. We therefore have, using Jensen's inequality and Lemma 3, that the social welfare of Nash equilibrium $\mathbf{s}$ is at least

$$\begin{aligned}
&\mathbb{E}_{\mathcal{U}(\mathbf{s})}[CC_x(\mathbf{s})^2] - (2n-4)C_E - nC_I \\
&\geq \mathbb{E}_{\mathcal{U}(\mathbf{s})}[CC_x(\mathbf{s})]^2 - (2n-4)C_E - nC_I \\
&\geq (n - 2 - 4C_I(C_E + 1))^2 - (2n-4)C_E - nC_I \\
&= n^2 - O(n).
\end{aligned}$$

□

Optimal Welfare Bounds for SQD Opponents

In this section, all our statements are with respect to a fixed game $(n, C_E, C_I, \mathcal{A})$ with $C_E, C_I > 1$, where $\mathcal{A}$ is an SQD opponent with function f. Since $\mathcal{A}$ is fixed for this section, we will omit it as a parameter and in subscripts.

We show the intuitive statement, that the SQD opponent favors splitting components into small components. Also, selling edges in targeted regions makes the region less attractive for the SQD opponent, i.e., it is edge-averse. Both follows from **??** (see Supp. Material) which states that SQD opponents favor attacking large components.

Corollary 2. *Let $\mathbf{s}$ be a strategy profile such that a component K in $G(\mathbf{s})$ contains at least two vulnerable regions T_{cut}, T_{leaf} of equal size, such that the remaining nodes in component K stay connected when removing T_{leaf} but not when removing T_{cut}. Then $U_f(T_{cut}, \mathbf{s}) < U_f(T_{leaf}, \mathbf{s})$, which means that T_{leaf} is not targeted by the SQD opponent.*

Corollary 3. *Let $\mathbf{s}$ be a strategy profile and consider a vulnerable node u inside a vulnerable region T in $G(\mathbf{s})$. Further, let $\mathbf{s}'$ be the strategy profile derived from $\mathbf{s}$, where*

- *some edges are sold that are incident to nodes in T, and/or*
- *some agents in T got immunized.*

Then, if T' is the vulnerable region containing node u in network $G(\mathbf{s}')$, it holds that $U_f(T', \mathbf{s}') \geq U_f(T, \mathbf{s})$.

The next corollary is one of the main tools for our other results. It follows from **??** (see Supp. Material).

Corollary 4. *Let u be a cut-vertex in an equilibrium network w.r.t. an SQD opponent $\mathcal{A}$, and Z be a remaining connected component after node u is removed. If agent u bought all of the k edges that it shares with nodes in Z, and if there are either targeted regions outside of Z or no targeted regions included in Z, then it holds that $|Z| \geq kC_E$. Also, if node u was targeted, then $|Z| > kC_E$.*

Next, we reprove and generalize a result from the full-version of (Goyal et al. 2016) (see (Goyal et al. 2015)).

Lemma 4. *(generalizes Theorem 6 in (Goyal et al. 2015)) In any connected component with at least one immunized node, the targeted regions are singletons for equilibrium networks w.r.t. an SQD opponent.*

The next results consider what happens if the SQD opponent targets cut-vertices of the network.

Lemma 5. *In a Nash equilibrium with respect to an SQD opponent $\mathcal{A}$, if a connected component contains an immunized node and a targeted cut-vertex, then,*

i) when removing any targeted node, the size of all remaining connected components is strictly higher than C_E, and

ii) there are at least two targeted regions outside of this component.

Corollary 5. *In a Nash equilibrium $\mathbf{s}$ with respect to an SQD opponent $\mathcal{A}$, if a connected component contains an immunized node and a targeted cut-vertex, then there is no isolated node in $G(\mathbf{s})$.*

Next, we bound the sizes of connected components.

Lemma 6. *In a Nash equilibrium* $\mathbf{s}$ *with respect to an SQD opponent* $\mathcal{A}$*, the size of a connected component is either* 1 *or at least* $C_E + 1$*.*

The next lemma deals with the implications of the existence of a targeted region that is a cut-vertex inside a component with an immunized node.

Lemma 7. *In a Nash equilibrium* $\mathbf{s}$ *with respect to an SQD opponent* $\mathcal{A}$*, if two different connected components each contain an immunized node and a targeted region, then none of the targeted regions are cut-vertices.*

Later will be bound the number of agents in some cases of Nash equilibria. The following statement is useful for this.

Lemma 8. *In a Nash equilibrium* $\mathbf{s}$ *with respect to an SQD opponent* $\mathcal{A}$*, the number of connected components is bounded by* $C_E + C_I + 2$ *if there are no isolated nodes.*

The next lemma allows us to eliminate the possibility of targeted cut-vertices as they only exist in small networks and thus do not impact the asymptotic bounds.

Lemma 9. *If there is a Nash equilibrium* $\mathbf{s}$ *with respect to an SQD opponent* $\mathcal{A}$ *such that* $G(\mathbf{s})$ *has a component with a targeted cut-vertex and an immunized node, then the number of agents in this game instance is at most*

$$(C_I + C_E + 2)(2C_I + 3C_E).$$

Now we consider a general result that guarantees the existence of a specific connected component.

Lemma 10. *In a non-trivial Nash equilibrium w.r.t. an SQD opponent, if there is a targeted region, at least one such region is in a connected component with an immunized node.*

We continue with the setting without targeted cut-vertices and show the existence of at least two vulnerable nodes.

Lemma 11. *In a non-trivial Nash equilibrium* $\mathbf{s}$ *with respect to an SQD opponent, if there are no targeted cut-vertices in connected components with an immunized node, every connected component with an immunized node and a targeted region contains at least two vulnerable nodes.*

Finally, we can show that most one connected component with a targeted regions exists.

Lemma 12. *In a non-trivial Nash equilibrium with respect to an SQD opponent, if there are no targeted cut-vertices in connected components with an immunized node, then there is at most one connected component with an immunized agent and a targeted region.*

Now, we derive the important statement, that without targeted cut-vertices no isolated agents exist. This is helpful, since such agents contribute to low social welfare.

Lemma 13. *In a non-trivial Nash equilibrium w.r.t. an SQD opponent, if no targeted cut-vertices exist in connected components with an immunized node, no agent is isolated.*

Proof Sketch. Let $\mathbf{s}$ be a non-trivial Nash equilibrium. By Lemma 10, there is a component K with at an immunized agent and, if there are vulnerable nodes in $G(\mathbf{s})$, a targeted region. By Lemma 4, targeted regions must be singletons and by assumption there are no targeted cut-vertices. Hence, every vulnerable agent in K is targeted and not a cut-vertex.

Now, towards a contradiction, assume there exists an isolated agent x. If agent x was already immunized, then buying an edge to any immunized node in K would be a strict utility increase, as there are at least $C_E + 1$ nodes in K and no vulnerable cut-vertex exists in K. Therefore, agent x is vulnerable, which means that there exists a targeted region, and thus targeted non-cut-vertices in component K.

Now there will be several cases depending on the structure of component K, see Figure 4 for an illustration.

1. Some agent in K bought two edges.
2. There is only one immunized agent in K.
3. An immunized agent bought an edge that disconnects K when removed and no agent bought more than one edge.
4. Every edge bought by an immunized agent in K is part of a cycle, no agent bought more than one edge, and some immunized agent bought an edge.

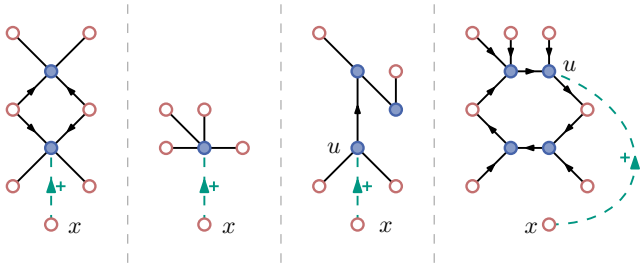

Figure 4: The four cases of the proof. The dashed edge is the proposed deviation.

We find that in all cases buying an edge to an immunized node of K will be a strictly improving deviation for agent x. Without buying immunity in cases $1, 2$ and 4 and sometimes buying immunity in Case 3. □

Finally, we combine all our structural observations, to show that large enough Nash equilibria must be connected.

Lemma 14. *All non-trivial Nash equilibria with respect to an SQD opponent that are not connected have a bounded number of agents.*

Proof. First of all, Lemma 9 ensures that if the number of nodes is greater than $(C_E + C_I + 2)(2C_I + 3C_E)$, then there are no targeted cut-vertices in connected components with immunized nodes. Assume this is the case, i.e., that we have more than $(C_E + C_I + 2)(2C_I + 3C_E)$ many agents.

Now if there are no vulnerable agents, then non-trivial Nash equilibria are connected. This was already proven by Bala and Goyal (2000), since without vulnerable agents we are in the setting without attacker.

Hence, we assume that at least one vulnerable agent exists. For this, by Lemma 10, we know that there is a connected component K with an immunized agent and a targeted region. There are also no isolated nodes, according to

Lemma 13, and therefore, only a bounded number of connected components, as guaranteed by Lemma 8.

By Lemma 4, component K has targeted singletons that are not cut-vertices. Now we show, that K is the largest component with vulnerable regions. Indeed, assume another component M is strictly larger than K. If M contains vulnerable nodes, then with the same reasoning as above, the targeted regions in M are singletons and not cut-vertices. Thus, if $|M| > |K|$ the SQD opponent would attack only regions in M, which contradicts that component K has targeted regions. If, on the other hand, component M only contains immunized nodes, then there is a strictly improving deviation for the immunized agents of K. This consists of buying an edge to any node in M. Since, by Lemma 13, component M contains at least two nodes, we get, by Lemma 6, that $|M| \geq C_E + 1$. Thus, buying the edge is profitable.

Therefore, we can assume that K is the largest connected component. Let L be the second largest component. We now find an agent in L and consider a specific strategy change, which then yields a bound on the size of K. For this, take a centroid c of component L, and consider the deviation of agent c immunizing (if it was not already immunized) and buying an edge to an immunized node of component K. The expected size of the connected component of agent c after the deviation is at least $|K| + \frac{|L|}{2} \geq |L| + \frac{|K|}{2}$. This holds, since if a vulnerable region in K is attacked, then agent c's connected component has size $|K| - 1 + |L| \geq |K| + \frac{|L|}{2}$, since $|L| \geq 2$. If a vulnerable region in L is attacked, then K is not affected and, since agent c is an immunized centroid of component L, the size of its connected component is at least $|K| + \frac{|L|}{2}$. If a vulnerable region outside of K and L is attacked, then the size of agent u's connected component is $|K| + |L| > |K| + \frac{|L|}{2}$.

Hence, for the original strategy profile to be a Nash equilibrium, the inequality $|K| \leq 2(C_I + C_E)$ must hold, since the additional cost of the deviation of agent c is $C_I + C_E$, while the expected additional connectivity is at least $\frac{|K|}{2}$.

Hence, the number of agents is bounded by $(C_E + C_I + 2)(2C_I + 2C_E)$, if there are no targeted cut-vertices. By Lemma 9, the latter holds. The stated upper bound on the number of agents is true, since by Lemma 8 and Lemma 13, we have $C_E + C_I + 2$ many connected components, each of size at most $|K| \leq 2C_I + 2C_E$.

This bound of $(C_E + C_I + 2)(2C_I + 2C_E)$ is lower than our assumed minimum number of agents of $(C_E + C_I + 2)(2C_I + 3C_E)$, which is a contradiction. Thus, every non-trivial Nash equilibrium with more than $(C_E + C_I + 2)(2C_I + 3C_E)$ nodes must be connected. □

Now we are ready, to prove our main result of this paper.

Theorem 2. *The social welfare of any non-trivial Nash equilibrium with respect to an SQD opponent is* $n^2 - \mathcal{O}(n)$.

Proof. Because this is an asymptotic result, we can ignore the social welfare of instances with bounded size. Thus, for a growing number of agents n, Lemma 14 ensures that every non-trivial Nash equilibrium is connected. With Lemma 4, we know that every targeted region is a singleton. Finally, Lemma 9 ensures that no targeted regions are cut-vertices.

Therefore, if some agents are vulnerable, the size of the connected component after the attack is $n - 1$. And because the total cost for all edges and immunizations are linear in n, as seen in Lemma 3, the social welfare is $n^2 - \mathcal{O}(n)$. □

Low Welfare for a Tailored Opponent

As contrast for our positive results on the maximum carnage, random attack, and the SQD opponents, we show that there exist opponents, such that Nash equilibria can have social welfare of $\Theta(n)$, i.e., the lowest possible welfare. This shows the counter-intuitive result, that opponents exist, that achieve a lower social welfare than the opponent that actually aims for minimizing the social welfare.

Theorem 3. *There are non-trivial Nash equilibria with respect to some attacker $\mathcal{A}$ with a social welfare of $\Theta(n)$.*

Proof Sketch. We consider a family of instances with n agents and $C_E = C_I = 6$ and a tailored f-opponent $\mathcal{A}$, where f is defined as follows:

i	1	7	8	9	10	otherwise
$f(i)$	2	3	5	4	7	0

For every $n \geq 10$, we consider the strategy profile $\mathbf{s}_{\text{bad}}$ illustrated in Figure 5. There, all but nine agents are isolated. There are therefore $n - 5$ vulnerable regions, and $n - 9$ targeted nodes. Note, that one of the isolated nodes will be targeted. Thus, the social welfare of profile $\mathbf{s}_{\text{bad}}$ is

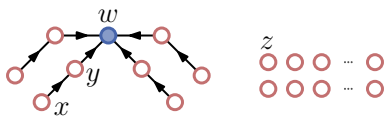

Figure 5: The Nash equilibrium $\mathbf{s}_{\text{bad}}$ with social welfare in $\Theta(n)$, for our tailored opponent, with $C_E = C_I = 6$.

$$(n-9) \cdot \frac{n-10}{n-9} \cdot 1 + 9 \cdot 9 - 8C_E - C_I \in \Theta(n),$$

since each of the $n - 9$ isolated nodes survives with probability $\frac{n-10}{n-9}$ and since the nine nodes in the large connected component are not targeted, each of them has connectivity 9.

We verify that profile $\mathbf{s}_{\text{bad}}$ is a Nash equilibrium. □

Conclusion

We have revisited the elegant noncooperative network formation model with attack and immunization from Goyal et al. (2016). We prove optimal social welfare of Nash equilibria under various strong types of attackers. This solves an open problem and highlights, that decentralized strategic agents can create very robust networks even under attack.

Our last result, the tailored opponent that yields Nash equilibria with low social welfare, indicates that there is still much to explore even for this very basic model. For example, characterizing what kind of attackers yield suboptimal social welfare, e.g., which properties of the f-function are crucial. Also, other variants of the model, for example with cooperation among the agents, could be studied.

References

Àlvarez, C.; and Messegué, A. 2023. Computing a Best Response against a Maximum Disruption Attack. *CoRR*, abs/2302.05348.

Anshelevich, E.; Bhardwaj, O.; and Kar, K. 2015. Strategic Network Formation through an Intermediary. In *IJCAI 2015*, 447–453.

Bala, V.; and Goyal, S. 2000. A noncooperative model of network formation. *Econometrica*, 68(5): 1181–1229.

Bala, V.; and Goyal, S. 2003. A strategic analysis of network reliability. In *Networks and Groups*, 313–336. Springer.

Berger, J.; Friedrich, T.; Lenzner, P.; Machaira, P.; and Ruff, J. 2025. Strategic Network Creation for Enabling Greedy Routing. In *AAAI 2025*, 13622–13630.

Bilò, D.; Cohen, S.; Friedrich, T.; Gawendowicz, H.; Klodt, N.; Lenzner, P.; and Skretas, G. 2023. Temporal Network Creation Games. In *IJCAI 2023*, 2511–2519.

Bilò, D.; Cohen, S.; Friedrich, T.; Gawendowicz, H.; Klodt, N.; Lenzner, P.; and Skretas, G. 2025. Temporal Network Creation Games: The Impact of Non-Locality and Terminals. In *AAMAS 2025*, 334–342.

Bilò, D.; Friedrich, T.; Lenzner, P.; Lowski, S.; and Melnichenko, A. 2021. Selfish Creation of Social Networks. In *AAAI 2021*, 5185–5193.

Bullinger, M.; Lenzner, P.; and Melnichenko, A. 2022. Network Creation with Homophilic Agents. In *IJCAI 2022*, 151–157.

Chauhan, A.; Lenzner, P.; Melnichenko, A.; and Münn, M. 2016. On Selfish Creation of Robust Networks. In *SAGT 2016*, 141–152.

Chen, Y.; Jabbari, S.; Kearns, M. J.; Khanna, S.; and Morgenstern, J. 2019. Network Formation under Random Attack and Probabilistic Spread. In *IJCAI 2019*, 180–186.

Echzell, H.; Friedrich, T.; Lenzner, P.; and Melnichenko, A. 2020. Flow-Based Network Creation Games. In *IJCAI 2020*, 139–145.

Fabrikant, A.; Luthra, A.; Maneva, E. N.; Papadimitriou, C. H.; and Shenker, S. 2003. On a network creation game. In *PODC 2003*, 347–351.

Friedrich, T.; Gawendowicz, H.; Lenzner, P.; and Melnichenko, A. 2022. Social Distancing Network Creation. In *ICALP 2022*, 62:1–62:21.

Friedrich, T.; Ihde, S.; Keßler, C.; Lenzner, P.; Neubert, S.; and Schumann, D. 2017. Efficient Best Response Computation for Strategic Network Formation Under Attack. In *SAGT 2017*, 199–211.

Goyal, S.; Jabbari, S.; Kearns, M. J.; Khanna, S.; and Morgenstern, J. 2015. Strategic Network Formation with Attack and Immunization. *CoRR*, abs/1511.05196.

Goyal, S.; Jabbari, S.; Kearns, M. J.; Khanna, S.; and Morgenstern, J. 2016. Strategic Network Formation with Attack and Immunization. In *WINE 2016*, 429–443.

Haller, H.; and Sarangi, S. 2005. Nash networks with heterogeneous links. *Mathematical Social Sciences*, 50(2): 181–201.

Jackson, M. O.; and Wolinsky, A. 1996. A strategic model of social and economic networks. *Journal of economic theory*, 71(1): 44–74.

Jackson, M. O.; et al. 2008. *Social and economic networks*, volume 3. Princeton university press Princeton.

Kliemann, L. 2011. The Price of Anarchy for Network Formation in an Adversary Model. *Games*, 2(3): 302–332.

Kliemann, L. 2013. The price of anarchy in bilateral network formation in an adversary model. *arXiv preprint arXiv:1308.1832*.

Meirom, E. A.; Mannor, S.; and Orda, A. 2015. Formation games of reliable networks. In *INFOCOM 2015*, 1760–1768.

Papadimitriou, C. H. 2001. Algorithms, games, and the internet. In *STOC 2001*, 749–753.