# The Complexity of Estimating Systematic Risk in Networks

Benjamin Johnson
University of California, Berkeley
johnsonb@ischool.berkeley.edu

Aron Laszka
Budapest University of Technology and Economics
aron@laszka.hu

Jens Grossklags
Pennsylvania State University
jensg@ist.psu.edu

***Abstract*—**

**This risk of catastrophe from an attack is a consequence of a network's structure formed by the connected individuals, businesses and computer systems. Understanding the likelihood of extreme events, or, more generally, the probability distribution of the number of compromised nodes is an essential requirement to provide risk-mitigation or cyber-insurance. However, previous network security research has not considered the higher moments of this distribution, while previous cyber-insurance research has not considered the effect of topologies on the supply side.**

**We provide a mathematical basis for bridging this gap: we study the complexity of computing these loss-number distributions, both generally and for special cases of common real-world networks. In the case of scale-free networks, we demonstrate that expected loss alone cannot determine the riskiness of a network, and that this riskiness cannot be naively estimated from smaller samples, which highlights the lack/importance of topological data in security incident reporting.**

## I. Introduction

Computer systems, businesses, and individuals often form networks. Computers, for example, are connected by physical and logical links; businesses provide services to one another; and individuals make friends and acquaintances encompassing various implicit levels of trust.

While these networks can be very beneficial, they can also increase risks, as attackers are often able to exploit the access and trust relationships that network connections entail. For example, in 2011, RSA, a major security company, was compromised; and information on about 40 million SecureID tokens were stolen. This successful compromise was later used to attack Lockheed Martin, one of the world's largest defense contractors [1]. More recently, hackers calling themselves the Syrian Electronic Army sent e-mails to Financial Times employees containing phishing links, which were used to gain access to FT.com corporate e-mail accounts. These accounts were then used to propagate the social engineering attack to a larger number of FT.com users, eventually compromising the organization's website and Twitter account [2].

These examples serve to illustrate that implicit trust from network connections can be used to compromise trusting neighbors through attacks on their peers. From the attacker's perspective, the network structure gives rise to what we might term systematic opportunity, because the opportunity for an attacker to strike a large payoff is a consequence of the system itself. Correspondingly, the users of such systems become subject to *systematic risks*, arising from the structure of their connections.

These systematic artifacts can have consequential effects on the motivations of users of such systems, as they recognize that their security is dependent on the investments of their peers. The resulting environment gives rise to well-documented problems such as under-investment or free-riding [3]; and it may also motivate users to consider alternative risk-mitigation strategies such as purchasing insurance.

Insurance is a promising remedy to many risk-related problems because it facilitates risk diversification; however, structural consequences of networked systems can also affect insurers. Traditionally, insurance is based on the diversifiability of risks: if an insurance provider has enough clients, the variability in individual risks cancel, and the aggregate risk is predictable. But if individual risks are correlated, then even for a large number of clients, there may be a non-negligible probability of a catastrophic event in which many clients are compromised at the same time.

This risk of catastrophe is a consequence of the network structure formed by the connected individuals, businesses and computer systems; and this causal relationship warrants our attention. However, to the best of our knowledge, the effects of a network's connective structure on general risk-mitigation concerns, such as those relevant to a cyber-insurance provider, have not been researched. For example, Lelarge and Bolot model interdependent security with insurance, but assume that there is an insurance provider with an *exogenously* priced premium [4], thus sidestepping the question of whether an insurance provider would be willing to offer such a contract. Many elements for understanding the relationship between a network's structure and the resulting risk to its components can be found in related work addressing cyber-insurance, models of interdependent security, or properties of scale-free networks. But a persistent research gap remains.

In this paper, we provide a mathematical basis for studying the distribution of the number of losses from a set of interconnected nodes, after individual risk propagates through a network structure. We illustrate and explain why network-wide risk-mitigation solutions, such as cyber-insurance, must consider the variability in the number of compromised nodes; and that in contrast to its expected value, the variability of this number cannot be naively estimated from sampling a small part of the network. This failure is especially interesting from a practical point of view, as many real-world business and social networks are resilient against comprehensive data collection, so that the only viable prediction mechanism for determining the risk portfolio of these networks relies on extrapolation from smaller samples. Our previous work emphasized the

importance of loss distributions to risk mitigation [5], and exhibited simulations of these distributions for some real-world networks [6]; while the current work for the first time proves the NP-hardness of computing the loss-number distribution, and more generally focuses on important aspects of computational complexity.

The rest of the paper is organized as follows. In Section II, we discuss related work from the areas of risk mitigation, interdependent security, and network structure. Section III introduces our network risk propagation model, which is derived from the literature on interdependent security games. In Section IV, we address the computational complexity of computing the distribution of the number of compromised nodes for this model. Section V addresses the question of systematic risk for scale-free networks. Finally, Section VI concludes the paper.

## II. Related work

First, we present current challenges in risk mitigation and risk-transfer mechanisms, such as cyber-insurance. Then, we summarize previous work on interdependent security models, which model how risk is propagated between connected nodes, the main concern of our study. Finally, we discuss scale-free networks, which realistically model many real-world networks and which form the basis of our simulation-based analysis.

### A. Risk Mitigation and Risk Transfer

Markets for risk-mitigation and risk-transfer mechanisms, such as cyberinsurance, suffer from the difficulty to predict systematic risk in networks. A functioning market for cyber-insurance and a good understanding of the insurability of networked resources both matter, because they signal that stakeholders are able to manage modern threats which cause widespread damage across many systems [7], [8]. However, the cyber-insurance market is developing at a slow pace due to a number of factors and is still not fully understood from an economic modeling perspective (see, in particular, the survey by Böhme and Schwartz [9]).

A primary difficulty for insurance providers is the correlation of risk. A group of defenders might appear as a particularly appealing target to an attacker because of a high correlation in the risk profiles of the defended resources. For example, even though systems may be independently owned and administrated, they may exhibit similar software configurations leading to so-called monoculture risks [10], [11]. Böhme and Kataria as well as Chen et al. study the impact of correlations that are due to such monoculture risks [12], [13].

Our research is complementary to the studies cited above: they investigate (the effect of) correlations arising from nodes having the same software configurations, while we study how correlations arise from nodes being connected to each other.

### B. Interdependent Security

The notion of correlation of risks can be extended to include a better understanding of the underlying interdependent nature of networks. That is, the mere vulnerability of a large number of systems to a particular attack is less significant if an attacker cannot easily execute a sufficiently broad attack and/or propagation is limited. Interdependence has been considered in different ways in the academic literature [3].

Varian, for example, studies security compromises that result from the failure of independently-owned systems to contribute to an overall prevention objective (i.e., security is a public good) [14]. In this model, security compromises are often the result of misaligned incentives which manifest as coordination failures, such as free-riding on others' prevention investments. Grossklags et al. extend this work to allow for investments in system recovery (i.e., self-insurance) and find that it can serve as a viable investment strategy to sidestep such coordination failures [15]. However, the availability of system recovery will further undermine incentives for collective security investments. Johnson et al. add the availability of cyber-insurance to this modeling framework, and identify solution spaces in which these different investment approaches may be used as bundled security strategies [16]. However, due to the fact that those models capture primarily two security outcomes (i.e., everybody is compromised, or nobody is compromised), they can only serve as approximate guidance for realistic insurance models.

A second group of economic models derives equilibrium strategies for containing the propagation of a virus or an attack in a network. For example, the models by Aspnes et al. as well as Moscibroda et al. would be applicable to the study of loss distributions, however, several simplifying assumptions in those models limit the generality of the results [17], [18]. Those limitations include the assumption that every infected node deterministically infects all unprotected neighbors.

A third class of propagation models that has been widely studied is the class of epidemic models, which describe how a virus spreads or extinguishes in a network. In the literature on epidemic models, the results of Kephart and White [19] are the closest to our analysis. Kephart and White study one of the simplest of the standard epidemic models, the susceptible-infected-susceptible (SIS) model, using various classes of networks. For Erdős-Rényi random graphs, they approximate both the expected value and the variance of the number of infected nodes using formulas. For the more realistic hierarchical network model, they show that the expected number of infected nodes does not increase with the size of the graph. This indicates that, even though variance is typically very high in this case, catastrophic events are unlikely as the magnitude of losses is low. Kephart and White also present two new, more realistic epidemiological models in [20]. In the first model, called "kill signals", they find large oscillations in the number of affected nodes using simulation, but only in two-dimensional square lattices, which are too regular to model a number of practical networks. In the second model, called "organizations", they show that the incident size distribution follows approximately an exponential distribution, which implies that catastrophic events are unlikely. Pastor-Satorras and Vespignani analyze real data from computer virus infections in order to define a dynamical SIS model for epidemic spreading in scale-free networks [21]. They find that, in scale-free networks, the epidemic threshold (the minimum ratio of the virus's birth rate to death rate such that an epidemic occurs) and its associated critical behavior do not exist (in other words, the threshold is zero). Eguíluz and Klemm study

the spreading of viruses in scale-free networks with large clustering coefficient and degree correlation, which they model as highly clustered scale-free graphs [22]. They show that, in contrast to randomly-wired scale-free networks, there exists a finite epidemic threshold for highly clustered scale-free networks, even if they are infinite in size. Pastor-Satorras and Vespignani study epidemic dynamics in finite-size scale-free networks, and show that, even for relatively small networks, the epidemic threshold is much smaller than that of homogeneous systems [23]. Wang et al. propose a general epidemic threshold condition, which applies to arbitrary graphs, based on the largest eigenvalue of the adjacency matrix [24], [25]. To show that the model yields precise results, they conduct simulations. They also prove that, when the network is below the epidemic threshold, the number of infected nodes decays exponentially over time. In a follow-up, Ganesh et al. obtain the same epidemic threshold result (along with other results) using another approach [26]: Wang et al. use point estimate, while Ganesh et al. derive exact bounds on the propagation equations.

Finally, a popular approach to model interdependent risk is taken by Kunreuther and Heal, and forms the basis for our formal analysis [27], [28]. The basic premise of this work is to separately consider the impact of direct attacks and propagated attacks. We explain the details of the model in Section III. The model has been generalized to consider distributions of attack probabilities [29] and strategic attackers [30]. Similarly, Ogut et al. proposed a related model that allows for continuous (rather than binary) security investments [31]. Our analysis setup draws from these extensions by implicitly considering a continuum of risk parameters to study the distribution of outcomes.

### C. Scale-Free Networks

Many real-world networks are believed to be scale-free, including social, financial, and biological networks [32]. A scale-free network's degree distribution is a scale-free power law distribution, which is generally attributed to robust self-organizing phenomena. Recent interest in scale-free networks started with [33], in which the Barabási-Albert (BA) model is introduced for generating random scale-free networks. The BA model is based on two concepts: network growth and preferential node attachment. We discuss this model in detail in Section V. Li et al. introduce a new, mathematically more precise, and structural definition of "scale-free" graphs [34], which promises to offer a more rigorous and quantitative alternative. The networks discussed in our paper satisfy this definition as well.

One of the important questions addressed by our paper is whether small samples can be used to predict systematic risks in scale-free networks. Stump et al. show that the degree distributions of randomly sampled subnets of scale-free networks are not scale-free [35]; thus, subnet data cannot be naively extrapolated to every property of the entire network. However, random samples are unbiased estimators of some properties (e.g., average degree). In Section V, we investigate whether they are unbiased estimators of systematic risk.

## III. Model overview

Our modeling framework builds on the network security game introduced by Kunreuther and Heal [27], [28]. This model has been studied and extended by many authors (e.g., [29], [30], [36], [37]), with a common focus on understanding how individuals in a networked system make individualized choices in response to probabilistic threats, along with how these choices affect other individuals.

Although we use this model's risk propagation structure, our focus is different from prior work. We concentrate exclusively on properties of the network's loss distribution as a function of each node's direct risk, and the probabilities of propagation from one node to another.

This propagation structure yields a simple mechanism for studying network risk, and it captures the core dynamics of risk transfer from a node-centric perspective. For example, consider node 1 in Figure 1. Any risk to this node may be categorized as either originating outside the network or originating within the network. If the risk originates outside the network we may categorize it in terms of its expected magnitude; while if the risk originates from within the network, from one of its connected nodes, we may quantify the magnitude of the risk derived from that connection.

For a list of symbols used in the paper, see Table I.

TABLE I: List of Symbols

| Sym. | Description |
|---|---|
| $N$ | number of nodes |
| $p$ | probability of direct compromise (when it is uniform over the nodes) |
| $p_i$ | probability of node $i$ being directly compromised |
| $q$ | probability of compromise propagation (when it is uniform over the links) |
| $q_{ij}$ | probability of compromise propagation from node $i$ to node $j$ (given that node $i$ is directly compromised) |
| $q_{in}$ | probability of compromise propagation from an outer node to the internal node in star topologies |
| $q_{out}$ | probability of compromise propagation from the internal node to an outer node in star topologies |
| $NL$ | random variable measuring the number of compromised nodes |
| $\rho$ | edge inclusion probability in ER random graph model |
| $m_0$ | size of the initial clique in the BA random graph model |
| $m$ | number of connections per additional node in the BA random graph model |

### A. Network Risk Propagation Model

Consider a network of $N$ nodes. Each node has two types of connections: one type which connects to other nodes in the same network, and another type which connects to a system outside the network. For an illustration, see Figure 1. Threats originate outside the network, and subject each node to some risk of compromise. If an outside threat successfully reaches a node, that node is compromised. This outcome is binary so that each node is either compromised or not.

If a node is compromised, the risk may propagate within the network to that node's direct neighbors. In our interpretation of the model, the risk does not propagate further than one hop, so that each node's risk exposure is bounded by the aggregate probability of direct external risks to itself and its immediate neighbors. While this model does not encompass all conceivable multiple-hop propagation structures, it strikes

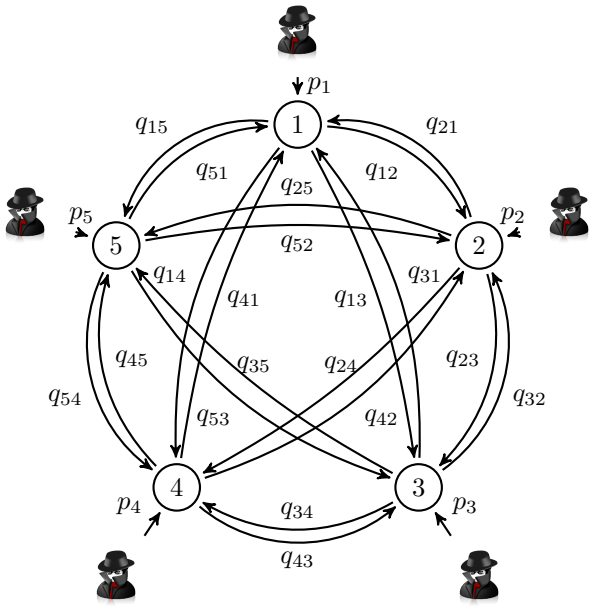



Fig. 1: Network Risk Arrival and Propagation

a good balance between realistic risk transfer properties and conceptual simplicity.

Risk of direct compromise threatens each node $i$ with probability $p_i$, and for the analysis we assume that direct compromises for different nodes are independent events. Our framework is agnostic about the origin of direct risk, although it could be motivated in an active attacker model by assuming that each node has a different attacker.

If a node is directly compromised, it transfers risk to each neighboring node $j$ with probability $q_{ij}$. If a node is not directly compromised, it cannot transfer risk to any other node. Notice that we can use the matrix $[q_{ij}]$ to directly specify network topology alongside risk propagation simply by requiring $q_{ij} = 0$ whenever node $i$ is not connected to node $j$.

### B. Game-Theoretic Actors

Many studies have used this model to understand interdependent security by considering a game in which each individual can reduce the risk of her own node by making a security investment. In this case, actors with various motivations make choices whose consequences either increase or decrease the node-centric risk parameters $p_i$ and $q_{ij}$. A game-theoretic analysis informs us about the set of configurations in which the model might be likely to end up after some time, but once each actor has made her choice, the system rests in a fixed configuration. In this paper, we build on the previous results on this model, and assume that this configuration is given. We then study the challenging problem of ascertaining the probability distribution on loss outcomes from the perspective of the entire network.

### C. Loss Distribution

A loss outcome is an event in which some nodes are compromised and others are not. To completely specify a loss outcome requires listing the set of compromised nodes. So a complete distribution on loss outcomes is a probability distribution on all subsets of nodes. This distribution is not tractable to analyze since the number of subsets of nodes is exponential in the number of nodes. However, if we consider only the *number* of compromised nodes, then its distribution is tractable to analyze. Moreover, the information obtained from studying this distribution remains highly relevant to network security and insurability. Let $NL$ be the random variable that counts the number of compromised nodes in an outcome of the model. Then, the *loss distribution* is a set of $N+1$ numbers giving $Pr[NL = k]$ for $k = 0, \ldots, N$.

## IV. Computability of the Loss Distribution

*Notational Conventions*

Whenever necessary for convenience throughout this paper, we adopt the following common mathematical conventions:

$$0^0 = 1, \qquad \sum_{\varnothing} = 0, \qquad \prod_{\varnothing} = 1$$

$$\binom{n}{m} = 0 \quad \text{whenever } m < 0 \text{ or } m > n.$$

### A. General Formula

We start by giving a general formula for the $N+1$ terms of the loss distribution on $NL$.

**Theorem 1.** *For each $k = 0, \ldots, N$,*

$$Pr[NL = k] = \sum_{\substack{C,D:\\ C\subseteq\{1,\ldots,N\}\\ D\subseteq C\\ |C|=k}} \left[\prod_{i\in D} p_i \cdot \prod_{i\in C\setminus D}\Big((1-p_i)(1-\prod_{j\in D}(1-q_{ji}))\Big) \cdot \prod_{i\notin C}\Big((1-p_i)\prod_{j\in D}(1-q_{ji})\Big)\right].$$

*Proof:* We compute the probability of the event $NL = k$ by enumerating all events in which $k$ nodes are compromised and summing their probabilities.

Let us first subdivide outcomes meeting the criteria $NL = k$ into disjoint classes according to which nodes were compromised directly, indirectly, or neither. Let $C$ be the set of all compromised nodes, and let $D$ be the set of directly compromised nodes. Then $D \subseteq C$ and, for outcomes in the class specified by this pair $(C, D)$, we know that:

1) every node in $D$ is directly compromised, and
2) every node in $C \setminus D$ is not directly compromised but is indirectly compromised by at least one of the nodes in $D$, and
3) every node not in $C$ is neither directly compromised nor indirectly compromised by a node in $D$.

Denoting these events with their numbers, respectively, we

have

$$\Pr[1] = \prod_{i \in D} p_i$$
$$\Pr[2|1] = \prod_{i \in C \setminus D} \Big( (1-p_i)\Big(1 - \prod_{j \in D}(1-q_{ji})\Big)\Big)$$
$$\Pr[3|1] = \prod_{i \notin C} \Big( (1-p_i) \prod_{j \in D}(1-q_{ji})\Big) \ .$$

In any outcome where 1 happens, events 2 and 3 are independent, which implies $\Pr[2 \wedge 3|1] = \Pr[2|1] \cdot \Pr[3|1]$.

The probability of an event in the class $(C, D)$ happening is then

$$\begin{aligned} \Pr[1 \wedge 2 \wedge 3] &= \Pr[1] \cdot \Pr[2 \wedge 3|1] \\ &= \Pr[1] \cdot \Pr[2|1] \cdot \Pr[3|1] \ . \end{aligned} \tag{1}$$

The probability that any event satisfying $NL = k$ happens can now be computed by taking the sum of Equation (1) over all pairs $C, D$ with $D \subseteq C \subseteq \{1, \ldots, N\}$ and $|C| = k$. ■

Notice that the number of terms in the theorem's formula is exponential in the number of nodes $N$. Consequently, the running time of a straightforward algorithm computing the value of the formula is also exponential. Even for relatively small networks, the number of terms can be considerably large; for example, the number of 150-element-subsets of the set $\{1, \ldots, 300\}$ is approximately $10^{88}$, which is greater than the number of atoms in the observable universe. In practice, this prevents us from using the above formula for networks that are not very small.

### *B. NP-Hardness*

The question naturally arises: is this exponential running time a defect of our formula or an inherent property of the problem? In this subsection, we show that, unfortunately, the general problem is indeed NP-hard; thus, assuming that $P \neq NP$[1], no polynomial-time algorithm can exist that computes the exact value of $\Pr[NL = k]$ for each $k$. However, in the subsequent subsections, we also show that the distribution can be computed efficiently for certain classes of networks.

Our hardness proof is based on reduction from a well-known NP-complete problem, the Minimum Set Cover problem.

**Definition 1. Minimum Set Cover:** Given a universe $U$, a family $F$ of subsets of $U$, and an integer $m$, is there a collection of at most $m$ subsets in $F$ whose union is $U$?

To perform the reduction, we first define a decision problem that can be easily reduced to computing the distribution of $NL$.

**Definition 2. Total Loss Probability:** Given an integer $N$, probabilities $p_i, q_{ij}$ for $i, j = 1, \ldots N$, and a real number $\delta$, does the network of $N$ nodes having direct compromise probabilities $p_i$ and indirect risk transfer probabilities $q_{ij}$ satisfy $\Pr[NL = N] \geq \delta$?

[1] $P \neq NP$ is a widely accepted conjecture; if it were not true, we would be able to solve all NP-hard problems in polynomial time.

**Theorem 2.** *Set Cover reduces to Total Loss Probability in polynomial time.*

*Proof:* Given an instance $(U, F, m)$ of Set Cover, we construct an instance $(N, \{p_i\}, \{q_{ij}\}, \delta\})$ of Total Loss Probability as follows.

- Let $N = 1 + |F| + |U|$.
- Let $r$ be a network node with direct compromise probability $p_r = 1$.
- For each subset $S \in F$, let $S$ also be a network node with direct compromise probability $p_S = \frac{1}{|F|!}$.
- For each element $u \in U$, let $u$ also be a network node with direct compromise probability $p_u = 0$.
- For each $S \in F$, add an edge from $r$ to $S$ with risk propagation probability $q_{rS} = 1$.
- For each pair $(u, S)$ from $U \times F$ with $u \in S$, add an edge from $S$ to $u$ with risk propagation probability $q_{Su} = 1$.
- Let $\delta = \frac{1}{|F|!^m}$.

This reduction can be carried out in time and space that is polynomial (quadratic) in the size of the Minimum Set Cover problem instance. To see this, note that the size of the proscribed Total Loss Probability instance is at most quadratic in the size of the Minimum Set Cover instance. (It is potentially quadratic because there are potentially a quadratic number of propagation probabilities $q_{Su}$.) There is no computation involved in the reduction except for computing the factorial of $|F|$, which can done naively in $|F|$ multiplications/divisions and using at most $\log_2 |F|! < |F| \log_2 |F| < |F|^2$ bits.

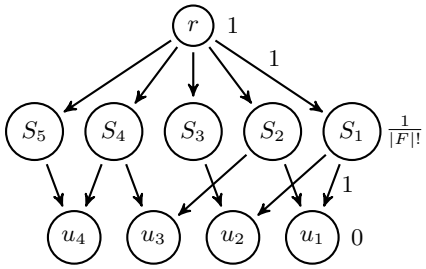



Fig. 2: NP-hardness reduction.

See Figure 2 for an example instance of Total Loss Probability generated by the reduction. Observe that in every outcome for this network, the node $r$ and the nodes $S \in F$ are always compromised, because $r$ is directly compromised, and $r$ propagates its risk to each node $S \in F$ with probability 1. Thus all nodes are compromised and $NL = N$ if and only if each node $u \in U$ is compromised.

We claim that there exists a collection of at most $m$ subsets in $F$ whose union is $U$ if and only if $\Pr[NL = N] \geq \delta$.

For the forward direction, assume that there exists a collection $C$ of at most $m$ subsets in $F$ whose union is $U$. Then,

$$\begin{aligned} \Pr[NL = N] &\geq \Pr[\text{every subset in } C \text{ is compromised}] \\ &= \left(\frac{1}{|F|!}\right)^m = \delta \ . \end{aligned}$$

Conversely, assume that there does not exist an $m$-cover of $U$, so that every collection of sets in $F$ that covers $U$ has size at least $m+1$. Then

$$\begin{aligned}
&\Pr[NL = N] \\
&= \Pr\left[\begin{matrix}\text{some collection } C \subseteq F \\ \text{that covers } U \text{ is compromised}\end{matrix}\right] \\
&\leq \Pr\left[\begin{matrix}\text{some collection } C \subseteq F \text{ having} \\ m+1 \text{ subsets is compromised}\end{matrix}\right] \\
&\leq \binom{|F|}{m+1}\left(\frac{1}{|F|!}\right)^{m+1} \\
&< |F|!\left(\frac{1}{|F|!}\right)^{m+1} = \left(\frac{1}{|F|!}\right)^{m} = \delta \ .
\end{aligned}$$

The equivalence shows that if we had an efficient algorithm to solve the Total Loss Probability problem, we could apply the above reduction to an arbitrary instance of the Minimum Set Cover problem, and use the reduction to solve that instance efficiently. However, since Minimum Set Cover is NP-hard, assuming $P \neq NP$, no algorithm solves arbitrary instances of that problem efficiently. Thus there is no efficient means to compute Total Loss Probability unless $P = NP$. ∎

### C. Special Case Topologies

Since the problem of computing the exact distribution is NP-hard, we have two viable options for larger networks. First, we can focus on restricted classes of networks. We give efficient formulas for three such classes in the following subsubsections. A second option is to use heuristic algorithms to approximate the general case. We take this second approach in Subsection IV-D.

*1) Homogeneous Topologies:* For a homogeneous network, the topology of the network is a complete graph; each node has a direct compromise probability of $p$; and each edge has a propagation probability of $q$ (in both directions). See Figure 3a for an illustration.

**Lemma 1.** *The probability of $k$ nodes being compromised in a homogeneous network is*

$$\begin{aligned}
&Pr[NL = k] \\
&= \binom{N}{k}\sum_{d=0}^{k}\Big[\binom{k}{d}p^d(1-p)^{(N-d)} \\
&\qquad \cdot \left(1-(1-q)^d\right)^{k-d} \cdot \left((1-q)^d\right)^{(N-k)}\Big] \ .
\end{aligned}$$

*Proof:* See Appendix A. ∎

*2) Star Topologies:* A star graph is a tree with one internal node and $N-1$ outer nodes. See Figure 3b for an illustration. We let $p_0$ denote the direct compromise probability of the internal node, and assume that the outer nodes have a uniform direct compromise probability, denoted by $p_1$. Furthermore, we assume that the probability of propagation is uniform from the internal node to the outer nodes, denoted by $q_{out}$, and from the outer nodes to the internal node, denoted by $q_{in}$.

This can model, for example, a network that consists of a single server and $N-1$ clients. We can assume that each client communicates directly only with the server; e.g., there are strict firewalls or no physical connections between the clients. Hence, there is no propagation between the clients.

**Lemma 2.** *The probability of $k$ nodes being compromised in the star network is*

$$\begin{aligned}
&Pr[NL = k] \\
&= \binom{N-1}{k-1}\sum_{d=0}^{k-1}\Big[\binom{k-1}{d}\cdot p_0 p_1^d(1-p_1)^{N-1-d} \\
&\qquad \cdot q_{out}^{(k-1)-d}\cdot(1-q_{out})^{N-k}\Big] \\
&+ \binom{N-1}{k-1}p_1^{k-1}(1-p_0)(1-p_1)^{N-k}\cdot(1-(1-q_{in})^{k-1}) \\
&+ \binom{N-1}{k}p_1^{k}(1-p_0)(1-p_1)^{N-1-k}\cdot(1-q_{in})^{k} \ .
\end{aligned}$$

*Proof:* See Appendix A. ∎

*3) E-R Random Topologies:* In the Erdős-Rényi (E-R) random graph model, undirected edges are set between each pair of nodes with equal probability $\rho$, independently of other edges [38].

Assume that the propagation probability of every edge is $q$. Then, the probability that a directly compromised node $i$ propagates compromise to any given node $j$ is

$$\Pr[i \text{ and } j \text{ are connected}] \cdot q = \rho q \ . \tag{2}$$

Consequently, the probability of any given node $i$ being compromised in an E-R random graph with a propagation probability of $q$ and an edge probability of $\rho$ is equal to the probability of $i$ being compromised in a homogeneous network with a propagation probability of $\rho q$. Therefore, the distribution of $NL$ is the same for a random network with parameters $p$, $q$, and $\rho$ and for a homogeneous network with parameters $p$ and $\rho q$. See Figure 3c for an illustration.

**Lemma 3.** *The probability of $k$ nodes being compromised in an E-R random network is*

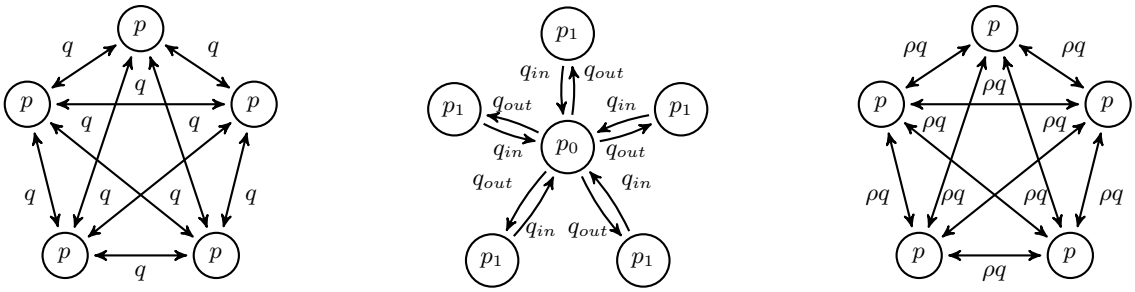
$$\begin{aligned}
&Pr[NL = k] \\
&= \binom{N}{k}\sum_{d=0}^{k}\Big[\ \binom{k}{d}p^d(1-p)^{(N-d)} \\
&\qquad \cdot \left(1-(1-\rho q)^d\right)^{k-d} \cdot \left((1-\rho q)^d\right)^{(N-k)}\Big] \ .
\end{aligned}$$

*Proof:* It follows immediately from Lemma 1 and the structure of the E-R random network. ∎

Notice that for each of these network topologies, the formula giving the distribution on the number of losses is at most quadratic in $N$. Thus we can compute the distribution efficiently for networks with these topologies.

### D. Simulation

For more general network topologies, we use simulation to obtain an approximate distribution. The simulation computes an empirical distribution by repeatedly choosing outcomes that




(a) Homogeneous network (b) Star network (c) E-R random network

Fig. 3: Special case topologies.

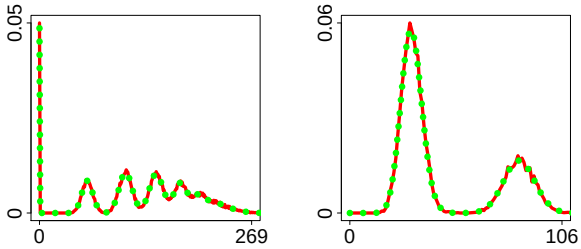
result from a simulated attack following the direct compromise and propagation probabilities, as follows:

- In each iteration, choose an outcome randomly in the following way:
  - First, for each node $i$, decide whether node $i$ is directly compromised (or not) at random according to $p_i$.
  - Second, for each directly compromised node $i$, iterate over all of its non-compromised neighbors. For each non-compromised neighbor node $j$, decide whether node $i$ propagates compromise to node $j$ (or not) at random according to $q_{ij}$.
- Count the nodes that have been compromised and add 1 to the number of occurrences of this outcome.
- After a fixed number of iterations, terminate the simulation and, for each outcome, output the number of occurrences over the number of iterations as the empirical probability of that outcome.

The running time of the above algorithm is polynomial in the size of the network, given a constant number of iterations. Furthermore, we have from the strong law of large numbers that the empirical distribution function converges to the actual function almost surely. To show that this convergence is indeed fast enough in practice, we ran the simulation for a number of homogeneous and star graph networks and compared the approximate distributions to the exact ones. In the following, we present two of these results.

Figure 4 compares distributions obtained from simulations to the exact distributions for a homogeneous and a star network, respectively. The homogeneous network consists of 300 nodes with $p = 0.01$ and $q = 0.2$. For this network, the simulation ran for 50,000 iterations, which took less than 14 seconds on an average desktop computer. The star network consists of 300 nodes with $p_i = 0.3$, $p_1 = 0.1$, and $q_{in} = q_{out} = 0.2$. For this network, the simulation ran for 20,000 iterations, which took 19 seconds. As can be seen in the figures, the distributions obtained from the simulations, which consisted of only relatively small numbers of iterations, are very good approximations to the exact distributions.

Notice that these distribution have multiple humps which distinguish them substantially from the common bell shape




(a) Homogeneous (b) Star

Fig. 4: Comparison of distributions obtained using simulation (solid red) to the exact distributions obtained from the formulas (dotted green).

of a binomial distribution. To explain this phenomenon, in the homogeneous network, the tall line at the beginning represents the event in which no nodes are directly compromised, while each consecutive bump primarily contains events in which one additional node is directly compromised. In the star network, the first bump primarily contains events in which the center node is not compromised, and the second bump consists primarily of events in which the center node is compromised.

## V. Systematic Risk in Scale-Free Networks

To study how systematic risk is affected by the network topology, we ran a large number of simulations on scale-free networks. The networks were generated according to one of the most prevalent models, the Barabási-Albert (BA) model [33]. The BA model is based on the concept of preferential attachment, which means that the more connected a node is, the more likely it is to receive new connections.

The BA model generates scale-free graphs as follows. First, a clique of $m_0$ initial nodes is constructed. Then, the remaining $N - m_0$ nodes are added to the network one by one. Each new node is randomly connected to $m$ existing nodes with probabilities proportional to the degrees of the existing nodes.

TABLE II: Comparison of the actual loss distribution to the binomial distribution for various direct compromise and propagation probabilities, and constant network size $N = 600$.

| $p$ | $q$ | $E[NL]$ | Variance $Var(NL)$ | | Quantile $Q(NL, 0.999)$ | | Safety loading $Q(NL, 0.999) - E[NL]$ | |
|---|---|---|---|---|---|---|---|---|
| | | | actual | binomial | actual | binomial | actual | binomial |
| 0.005 | 0.05 | 4.22 | 7.76 | 4.19 | 17 | 12 | 12.78 | 7.78 |
| | 0.1 | 5.41 | 14.82 | 5.36 | 24 | 14 | 18.59 | 8.59 |
| | 0.5 | 14.87 | 145.51 | 14.50 | 77 | 28 | 62.13 | 13.13 |
| 0.01 | 0.05 | 8.40 | 15.31 | 8.29 | 24 | 19 | 15.60 | 10.60 |
| | 0.1 | 10.80 | 28.83 | 10.60 | 34 | 22 | 23.20 | 11.20 |
| | 0.5 | 29.11 | 264.77 | 27.70 | 101 | 47 | 71.89 | 17.89 |
| 0.05 | 0.05 | 41.31 | 67.72 | 38.47 | 69 | 62 | 27.69 | 20.69 |
| | 0.1 | 52.19 | 118.52 | 47.65 | 90 | 75 | 37.81 | 22.81 |
| | 0.5 | 125.56 | 728.81 | 99.28 | 218 | 157 | 92.44 | 31.44 |

### A. *Measuring Systematic Risk*

In this subsection, we compute the mean, the variance, the 99.9% quantile, and the safety loading[2] requirement at probability 99.9% for the loss distributions of several scale-free networks. We also compare these quantities to those of the binomial distributions having the same mean. Note that binomial distributions are of special interest to us, because if the node compromise events were independent, then the loss outcomes would follow a binomial distribution. Consequently, the binomial distribution with the same mean is the distribution with the same overall risk, but with minimal systematic risk. We use the comparison to illustrate the systematic risk of scale-free networks.

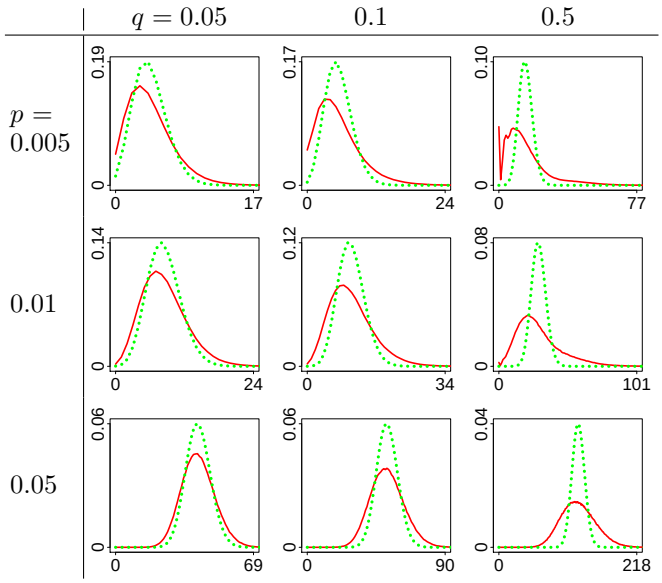



Fig. 5: Comparison of the actual loss distribution (solid red) to the binomial distribution (dotted green) for various direct compromise and propagation probabilities, and constant network size $N = 600$. (Note that the slightly irregular subfigure for $p = 0.005$ and $q = 0.5$ is correctly drawn.)

Figure 5 and Table II compare binomial distributions to the actual loss distributions resulting from various direct compromise and propagation probabilities. The network consists of $N = 600$ nodes, and it was generated using the parameters $m_0 = 15$, and $m = 4$. As expected, increasing the propagation probability increases the difference between the actual loss distribution and the binomial distribution through increasing the interdependence between the nodes of the network. Increasing the direct node compromise probability has a less pronounced effect in the same direction.

Figure 6 and Table III compare binomial distributions to the actual loss distributions for varying network sizes. The direct compromise and propagation probabilities are $p = 0.01$ and $q = 0.1$. As can be seen in both the figure and the table, the difference between the actual loss distribution and the binomial distribution does not diminish as the size of the network increases.

### B. *Sampling Scale-Free Graphs*

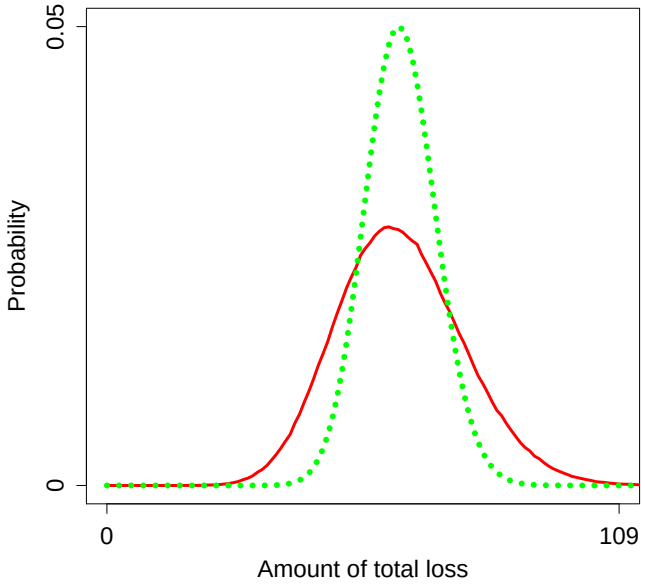



Fig. 7: The loss distribution (solid red) of a scale-free network of $N = 600$ nodes with parameters $m_0 = 15$, $m = 4$, $p = 0.05$, $q = 0.15$, compared to the binomial distribution (dotted green).

In the previous subsection, we illustrated the extent to which systematic risk is present in scale-free networks. In this subsection, we investigate the properties of random subnets of scale-free networks. The network from which the samples

[2]Safety loading is the excess premium required to ensure that the probability of ruin for an insurer is at most the given probability.

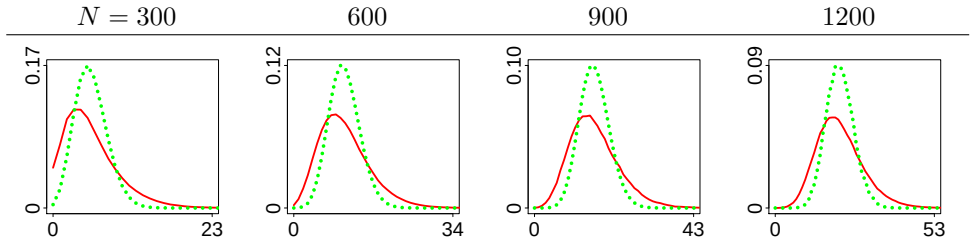



Fig. 6: Comparison of the actual loss distribution (solid red) to the binomial distribution (dotted green) for various network sizes and constant $p = 0.01$, $q = 0.1$.

TABLE III: Comparison of the actual loss distribution to the binomial distribution for various network sizes and constant $p = 0.01$, $q = 0.1$.

| $N$ | $E[NL]$ | Variance $Var(NL)$ | | Quantile $Q(NL, 0.999)$ | | Safety loading $Q(NL, 0.999) - E[NL]$ | |
|---|---|---|---|---|---|---|---|
| | | actual | binomial | actual | binomial | actual | binomial |
| 300 | 5.44 | 14.81 | 5.35 | 23 | 14 | 17.56 | 8.56 |
| 600 | 10.80 | 28.83 | 10.60 | 34 | 22 | 23.20 | 11.20 |
| 900 | 16.17 | 43.42 | 15.88 | 43 | 30 | 26.83 | 13.83 |
| 1200 | 21.51 | 58.81 | 20.69 | 53 | 36 | 31.49 | 14.49 |

are drawn is a scale-free network with parameters $N = 600$, $m_0 = 15$, and $m = 4$, and with compromise probabilities $p = 0.05$ and $q = 0.15$. Figure 7 shows the loss distribution for the entire network, compared to the binomial distribution with the same expected number of losses.

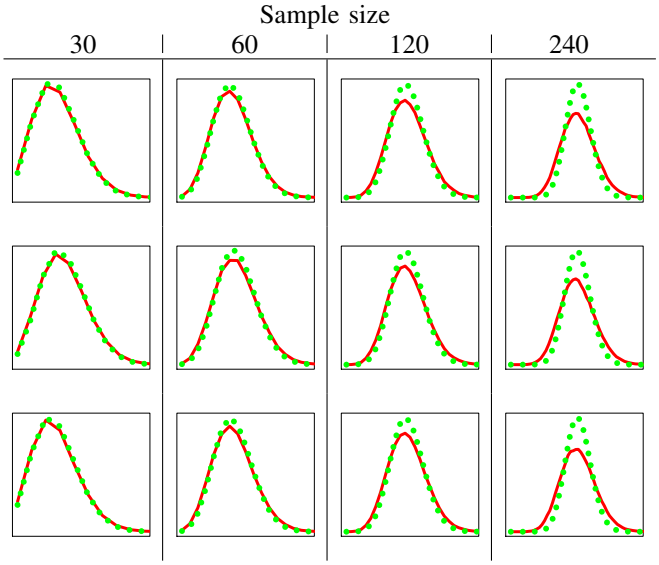



Fig. 8: Loss distributions of randomly drawn samples (solid red) compared to binomial distributions (dotted green). For each size, three randomly chosen samples are used to compare the distributions.

Figure 8 and Table IV compare the actual loss distributions of randomly drawn samples to binomial distributions. We study four different sample sizes: 30, 60, 120, and 240. For each samples size, three samples of that size were drawn uniformly at random from the set of all nodes. For each sample, its loss distribution was computed by running the simulation for the *entire network*, but counting only the compromised nodes belonging to the sample. This models the real-world scenario where incident reports are collected from only a sample, but the security of this sample is affected by the rest of the world through external connections. The number of iterations was 200,000 for each sample. As before, the binomial distributions, with which the samples are compared, have exactly the same expected losses $E[NL]$ as the corresponding sample distributions.

Figure 8 shows the three random samples for each size, together with the corresponding binomial distributions, while Table IV gives a more detailed comparison in terms of the metrics we are considering. As can be seen in the figure, the loss distributions of the samples are almost indistinguishable from the binomial distributions for sample sizes of 30 and 60 nodes. Consequently, by observing only a sample of the entire network, one might arrive at the wrong conclusion that individual node compromises are independent events. As the sizes of the samples increase, the loss distributions become more distinguishable from the binomial distribution, eventually approaching the distribution of losses for the full network.

### C. Application to Cyber-Insurance

As a motivating example, consider an insurer who provides insurance coverage to the entities that form a network with parameters $N = 600$, $p = 0.01$, and $q = 0.1$. Suppose that the insurer uses a smaller sample of incident reports to estimate the risk associated with these insurance policies. Since even small samples are unbiased estimators of the average probability that a given node is compromised[3], it can correctly estimate the average risk as $E[NL]/N = 1.79\%$ based on the individual incident reports. In order to keep its probability of ruin below a given level $0.001$, the insurer wants to compute the necessary safety loading $Q(NL, 0.999) - E[NL]$ using the quantile premium principle. Since the insurer assumes that risks are very close to independent, it estimates the necessary safety loading based on a binomial distribution, which gives a value of $11.1$. However, its safety loading should

[3]This result can be seen in Table IV by comparing $\frac{E[NL]}{N}$ across the rows.

TABLE IV: Comparison of the actual loss distribution of randomly drawn samples to the binomial distribution for varying sample sizes.

| $N$ | $E[NL]$ | Variance $Var(NL)$ | | Quantile $Q(NL, 0.999)$ | | Safety loading $Q(NL, 0.999) - E[NL]$ | |
|---|---|---|---|---|---|---|---|
| | | actual | binomial | actual | binomial | actual | binomial |
| 30 | 3.05 | 2.96 | 2.74 | 9.33 | 9.00 | 6.28 | 5.95 |
| 60 | 6.58 | 6.92 | 5.86 | 16.00 | 15.00 | 9.42 | 8.42 |
| 120 | 12.50 | 15.24 | 11.20 | 26.00 | 24.00 | 13.50 | 11.50 |
| 240 | 25.59 | 42.80 | 22.86 | 48.33 | 41.33 | 22.74 | 15.74 |

be in fact 23.2 (see Table II). This mistake has rather harsh consequences for the insurance provider: the probability that the total loss exceeds the erroneously computed insurance premium is $\Pr[NL > E[NL] + 11.1] = 3.1\%$. In other words, the probability of ruin is $3.1\%$ instead of $0.1\%$.

## VI. Conclusions

The systematic risk of a networked system depends jointly on the topology of the network and the security levels of individual nodes. In this paper, we studied a well-known risk propagation model which concretely specifies this connection.

We found that the distribution of the number of compromised nodes has a number of interesting properties. It is expressible as a simple closed formula; it is NP-hard to compute in general; it is efficiently computable for several interesting special cases; and it can be efficiently approximated using simulation for other, more general cases.

By applying our methodology to scale-free networks, we found that the full network possesses systematic risks, which may require large amounts of safety capital to properly insure. Yet we found much lower systematic risk in random samples of the same networks. This observation yields two contrasting applications to cyber-insurance. On the one hand, it may be possible to insure random subsamples of a network with scale-free properties while bearing only a modest loading cost. On the other hand, an insurer cannot readily deduce the systematic risk of a networked system by taking random samples.

## References


[1] C. Drew, "Stolen data is tracked to hacking at Lockheed," New York Times, http://www.nytimes.com/2011/06/04/technology/04security.html, June 3, 2011.

[2] A. Betts, "A sobering day," Financial Times Labs, http://labs.ft.com/2013/05/a-sobering-day/, May 29, 2013.

[3] A. Laszka, M. Felegyhazi, and L. Buttyán, "A survey of interdependent security games," CrySyS Lab, Budapest University of Technology and Economics, Tech. Rep. CRYSYS-TR-2012-11-15, Nov 2012.

[4] M. Lelarge and J. Bolot, "Economic incentives to increase security in the internet: The case for insurance," in *Proceedings of the 33rd IEEE International Conference on Computer Communications (INFOCOM)*, 2009, pp. 1494–1502.

[5] B. Johnson, A. Laszka, and J. Grossklags, "How many down? Toward understanding systematic risk in networks," to appear in *Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security (ASIACCS)*, 2014.

[6] A. Laszka, B. Johnson, J. Grossklags, and M. Felegyhazi, "Estimating systematic risk in real-world networks," to appear in *Proceedings of the 18th International Conference on Financial Cryptography and Data Security (FC)*, 2014.

[7] R. Anderson, "Liability and computer security: Nine principles," in *Proceedings of the Third European Symposium on Research in Computer Security (ESORICS)*, Brighton, UK, Nov. 1994, pp. 231–245.

[8] R. Böhme, "Towards insurable network architectures," *it - Information Technology*, vol. 52, no. 5, pp. 290–293, 2010.

[9] R. Böhme and G. Schwartz, "Modeling cyber-insurance: Towards a unifying framework," in *Workshop on the Economics of Information Security*, 2010.

[10] K. Birman and F. Schneider, "The monoculture risk put into context," *IEEE Security and Privacy*, vol. 7, no. 1, pp. 14–17, Jan. 2009.

[11] D. Geer, C. Pfleeger, B. Schneier, J. Quarterman, P. Metzger, R. Bace, and P. Gutmann, "Cyberinsecurity: The cost of monopoly. How the dominance of Microsoft's products poses a risk to society," 2003.

[12] R. Böhme and G. Kataria, "Models and measures for correlation in cyber-insurance," in *Workshop on the Economics of Information Security*, 2006.

[13] P.-Y. Chen, G. Kataria, and R. Krishnan, "Correlated failures, diversification, and information security risk management," *MIS Quarterly*, vol. 35, no. 2, pp. 397–422, Jun. 2011.

[14] H. Varian, "System reliability and free riding," in *Economics of Information Security*, J. Camp and S. Lewis, Eds. Dordrecht, The Netherlands: Kluwer Academic Publishers, 2004, pp. 1–15.

[15] J. Grossklags, N. Christin, and J. Chuang, "Secure or insure?: A game-theoretic analysis of information security games," in *Proceedings of the 17th International World Wide Web Conference*, 2008, pp. 209–218.

[16] B. Johnson, R. Böhme, and J. Grossklags, "Security games with market insurance," *Decision and Game Theory for Security*, pp. 117–130, 2011.

[17] J. Aspnes, K. Chang, and A. Yampolskiy, "Inoculation strategies for victims of viruses and the sum-of-squares partition problem," *J. Computer and System Sciences*, vol. 72, no. 6, pp. 1077–1093, Sep. 2006.

[18] T. Moscibroda, S. Schmid, and R. Wattenhofer, "When selfish meets evil: Byzantine players in a virus inoculation game," in *Proceedings of the ACM Symposium on Principles of Distributed Computing*, 2006, pp. 35–44.

[19] J. Kephart and S. White, "Directed-graph epidemiological models of computer viruses," in *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy*, 1991, pp. 343–359.

[20] ——, "Measuring and modeling computer virus prevalence," in *Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy*, 1993, pp. 2–15.

[21] R. Pastor-Satorras and A. Vespignani, "Epidemic spreading in scale-free networks," *Physical Review Letters*, vol. 86, no. 14, pp. 3200–3203, 2001.

[22] V. Eguiluz and K. Klemm, "Epidemic threshold in structured scale-free networks," *Physical Review Letters*, vol. 89, no. 10, p. 108701, 2002.

[23] R. Pastor-Satorras and A. Vespignani, "Epidemic dynamics in finite size scale-free networks," *Physical Review E*, vol. 65, no. 3, p. 035108, 2002.

[24] Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos, "Epidemic spreading in real networks: An eigenvalue viewpoint," in *Proceedings of the 22nd International Symposium on Reliable Distributed Systems*, 2003, pp. 25–34.

[25] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos, "Epidemic thresholds in real networks," *ACM Transactions on Information and System Security*, vol. 10, no. 4, p. 1, 2008.

[26] A. Ganesh, L. Massoulié, and D. Towsley, "The effect of network topology on the spread of epidemics," in *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, Mar. 2005, pp. 1455–1466.

[27] H. Kunreuther and G. Heal, "Interdependent security," *Journal of Risk and Uncertainty*, vol. 26, no. 2, pp. 231–249, 2003.

[28] G. Heal and H. Kunreuther, "Interdependent security: A general model," NBER Working Paper No. 10706, August 2004.

[29] B. Johnson, J. Grossklags, N. Christin, and J. Chuang, "Uncertainty in interdependent security games," *Decision and Game Theory for Security*, pp. 234–244, 2010.

[30] H. Chan, M. Ceyko, and L. Ortiz, "Interdependent defense games: Modeling interdependent security under deliberate attacks," in *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence (UAI)*, Catalina Island, CA, Aug. 2012, pp. 152–162.

[31] H. Ogut, N. Menon, and S. Raghunathan, "Cyber insurance and IT security investment: Impact of interdependent risk," in *Workshop on the Economics of Information Security*, 2005.

[32] A.-L. Barabási, "Scale-free networks: A decade and beyond," *Science*, vol. 325, no. 5939, pp. 412–413, Jul. 2009.

[33] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.

[34] L. Li, D. Alderson, J. Doyle, and W. Willinger, "Towards a theory of scale-free graphs: Definition, properties, and implications," *Internet Mathematics*, vol. 2, no. 4, pp. 431–523, 2005.

[35] M. Stumpf, C. Wiuf, and R. May, "Subnets of scale-free networks are not scale-free: Sampling properties of networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 12, pp. 4221–4224, 2005.

[36] S. Dhall, S. Lakshmivarahan, and P. Verma, "On the number and the distribution of the Nash equilibria in supermodular games and their impact on the tipping set," in *Proceedings of the International Conference on Game Theory for Networks (GameNets)*, Istanbul, Turkey, May 2009, pp. 691–696.

[37] M. Kearns and L. Ortiz, "Algorithms for interdependent security games," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, Eds. MIT Press, 2004, pp. 561–568.

[38] P. Erdős and A. Rényi, "On the evolution of random graphs," *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, vol. 5, pp. 17–61, 1960.

## APPENDIX

*Proof:* Suppose that for each $i$ and $j$, $p_i = p$ and $q_{ij} = q$. Fix $C, D$ with $D \subseteq C \subseteq \{1, \ldots, N\}$, $|C| = k$ and $|D| = d$. Then

$$\prod_{i \in D} p_i = p^d, \tag{3}$$

$$\prod_{i \in C \setminus D} \Big((1-p_i)\Big(1-\prod_{j \in D}(1-q_{ji})\Big)\Big) = \big((1-p)(1-(1-q)^d)\big)^{k-d}, \tag{4}$$

and

$$\prod_{i \notin C} \Big((1-p_i)\prod_{j \in D}(1-q_{ji})\Big) = \big((1-p)(1-q)^d\big)^{N-k} \ . \tag{5}$$

From Theorem 1, $\Pr[NL = k]$ has the form

$$\sum_{d=0}^{k} \sum_{\substack{C,D: \\ C \subseteq \{1,\ldots,N\},\ D \subseteq C,\ |D|=d,\ |C|=k}} \Pr[(C,D)] \ , \tag{6}$$

where $\Pr[(C,D)]$ is the product of Equations (3), (4) and (5).

The number of pairs $(C, D)$ with $D \subseteq C \subseteq \{1, \ldots, N\}$, $|C| = k$, and $|D| = d$ is exactly $\binom{N}{k} \cdot \binom{k}{d}$; and $\Pr[(C,D)]$ is uniform over all pairs $(C, D)$ satisfying these properties. So we have

$$\begin{aligned}
&\Pr[NL = k] \\
=&\sum_{d=0}^{k} \binom{N}{k}\binom{k}{d} \Pr[(C,D)] \\
=&\binom{N}{k} \sum_{d=0}^{k} \Bigg[ \binom{k}{d} p^d (1-p)^{(N-d)} \\
&\qquad \cdot \big(1-(1-q)^d\big)^{k-d} \cdot \big((1-q)^d\big)^{(N-k)} \Bigg] \ .
\end{aligned}$$

■

*Proof:* We divide the set of outcomes into three possibilities. Either

1) the center node is directly compromised, or
2) the center node is not directly compromised, but is indirectly compromised, or
3) the center node is neither directly nor indirectly compromised.

We address each case separately, and then add their probabilities.

1) In the first case, we further subdivide the space according to the number $d$ of directly-compromised exterior nodes. Fix $k$ and $d$. In this sub-case we know that $k - d - 1$ exterior nodes were not directly but indirectly compromised, and $N - k$ nodes were not compromised at all. The total probability of this case happening is the product of the probabilities that
   a) the center node is directly compromised
   b) $d$ exterior nodes are directly compromised
   c) $k - d - 1$ exterior nodes are not directly compromised but are indirectly compromised
   d) $N - k$ exterior nodes are neither directly nor indirectly compromised

   which gives

$$\begin{aligned}
&p_0 \cdot p_1^d \cdot ((1-p_1)q_{out})^{k-d-1} \cdot ((1-p_1)(1-q_{out}))^{N-k} \\
=&p_0 p_1^d (1-p_1)^{N-1-d} \cdot q_{out}^{(k-1)-d} \cdot (1-q_{out})^{N-k} \ .
\end{aligned}$$

   The number of ways to choose $d$ and $k$ in this case is $\binom{N-1}{k-1} \cdot \binom{k-1}{d}$; and the total probability of this case is obtained by summing the probabilities over all possible values for $d$, i.e.,

$$\begin{aligned}
&\binom{N-1}{k-1} \sum_{d=0}^{k-1} \Bigg[ \binom{k-1}{d} \cdot p_0 p_1^d (1-p_1)^{N-1-d} \\
&\qquad \cdot q_{out}^{(k-1)-d} \cdot (1-q_{out})^{N-k} \Bigg] \ .
\end{aligned} \tag{7}$$

2) In the second case, each of the $k - 1$ external compromised nodes must be directly compromised, because the center node is not directly compromised, and only the center node can indirectly compromise external nodes. For a fixed choice of these $k - 1$ compromised external nodes, the probability of this configuration is the product of the probabilities that

a) the center node is not directly compromised, but is indirectly compromised
b) $k-1$ exterior nodes are directly compromised
c) $N-k$ exterior nodes are not directly compromised

which gives

$$\begin{aligned}&(1-p_0)\cdot(1-(1-q_{in})^{k-1})\cdot p_1^{k-1}\cdot(1-p_1)^{N-k}\\ =&p_1^{k-1}(1-p_0)(1-p_1)^{N-k}\cdot(1-(1-q_{in})^{k-1})\ .\end{aligned}$$

There are $\binom{N-1}{k-1}$ ways to choose the external compromised nodes, so the probability of this case is

$$\binom{N-1}{k-1}p_1^{k-1}(1-p_0)(1-p_1)^{N-k}\cdot(1-(1-q_{in})^{k-1})\ . \tag{8}$$

3) In the third case, there are $k$ external compromised nodes, each of which must be directly compromised; and for a fixed choice of these $k$ compromised external nodes, the probability of this configuration is the product of the probabilities that
   a) the center node is neither directly nor indirectly compromised
   b) $k$ exterior nodes are directly compromised
   c) $N-1-k$ exterior nodes are not directly compromised

   which gives

$$\begin{aligned}&(1-p_0)\cdot(1-q_{in})^k\cdot p_1^k\cdot(1-p_1)^{N-1-k}\\ &=p_1^k(1-p_0)(1-p_1)^{N-1-k}\cdot(1-q_{in})^k\ .\end{aligned}$$

   There are $\binom{N-1}{k}$ ways to choose the external compromised nodes, so the probability of this case is

$$\binom{N-1}{k}p_1^k(1-p_0)(1-p_1)^{N-1-k}\cdot(1-q_{in})^k\ . \tag{9}$$

Finally, the total probability of $k$ losses is the sum of Equations (7), (8) and (9).

$$\begin{aligned}&\binom{N-1}{k-1}\sum_{d=0}^{k-1}\Bigg[\binom{k-1}{d}\cdot p_0p_1^d(1-p_1)^{N-1-d}\\ &\qquad\qquad\cdot q_{out}^{(k-1)-d}\cdot(1-q_{out})^{N-k}\Bigg]\\ &+\binom{N-1}{k-1}p_1^{k-1}(1-p_0)(1-p_1)^{N-k}\cdot(1-(1-q_{in})^{k-1})\\ &+\binom{N-1}{k}p_1^k(1-p_0)(1-p_1)^{N-1-k}\cdot(1-q_{in})^k\ .\end{aligned}$$

■