Estimating Systematic Risk in Real-World Networks

Aron Laszka[1,2], Benjamin Johnson[3], Jens Grossklags[1], and Mark Felegyhazi[2]

[1] College of Information Sciences and Technology,
Pennsylvania State University, USA
[2] Department of Networked Systems and Services,
Budapest University of Technology and Economics, Hungary
[3] School of Information, University of California, Berkeley, USA

Abstract. Social, technical and business connections can all give rise to security risks. These risks can be substantial when individual compromises occur in combinations, and difficult to predict when some connections are not easily observed. A significant and relevant challenge is to predict these risks using only locally-derivable information.
We illustrate by example that this challenge can be met if some general topological features of the connection network are known. By simulating an attack propagation on two large real-world networks, we identify structural regularities in the resulting loss distributions, from which we can relate various measures of a network's risks to its topology. While deriving these formulae requires knowing or approximating the connective structure of the network, applying them requires only locally-derivable information.
On the theoretical side, we show that our risk-estimating methodology gives good approximations on randomly-generated scale-free networks with parameters approximating those in our study. Since many real-world networks are formed through preferential attachment mechanisms that yield similar scale-free topologies, we expect this methodology to have a wider range of applications to risk management whenever a large number of connections is involved.

Keywords: networks, security, topology, Internet, cyber-insurance

1 Introduction

Networks arise from many different type of real world connections. 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, their members may also increase their exposure to risks through participation. For example, phishing attacks against individuals on Facebook leverage the fact that you are more likely to click on a link that originates from a friend. Such attacks leverage the existing trust relation represented by the connections in the social-networking platform. Online social networks are especially vulnerable to

this type of attack, because the information accessible to our connections can be collected and used in subsequent attacks.

Businesses and organizations may increase their risk exposure from networks too. For example, Autonomous Systems (ASs) controlled by Internet Service Providers (ISPs) routinely form peering relationships in which they agree to provide transit service to their peers' customers. These connections enable each ISP to provide better service to its customers, but the connections also entail added risk in case one of their peers' customers is subject to a Denial of Service (DoS) attack. This was exactly the case when Spamhaus, a major player in the network security business, received an enormous DoS attack that affected the upstream ISPs providing Internet access to the company [1]. Fortunately, Spamhaus was able to combat this attack with the help of the ISPs.

Due to the ubiquity and magnitude of risks related to participation in networks, especially computer networks, businesses have become increasingly interested in the availability of insurance policies to mitigate against such risks. Unfortunately, the emergence of a market for cyber-insurance over the last decade has been painfully slow, motivating calls for a better understanding of risk propagation in networks [2].

To understand the nature of these types of risks, we need to understand both the risk propagation mechanism that affects two connected entities, and the topological structure of the connective network. For many networks, this latter problem is quite challenging. To give a sense of the complexity from an insurer's perspective, suppose that an insurer wants to provide insurance coverage to a subset of the nodes within a network, covering all risks that arise within this network. She may obtain data from all the nodes in the subset including connections between these nodes. However, because the risk exposure includes connections outside this subnetwork, in order to calculate the insurance premiums, an insurer would have to know the topology of a much larger part of the network [3]. This is obviously a very challenging task in practice, as the insurer would have to collect risk assessment data regarding entities to which she has no business connection at all.

Our goal in this paper is to find general rules for calculating the risk exposure of sets of nodes within a connected system, that can apply to a wide-range of networks that emerge in practice. To accomplish this goal, we analyze the topological structure of two independent real-world networks – one based on the business relationships between the Internet's autonomous systems, and the other based on a subnetwork of the Facebook friendship network. We also generate random scale-free networks with evolutionary parameters set to approximate these real-world networks. Finally, we simulate propagation attacks on each network and analyze the resulting loss distributions. We find structural regularities that apply to all four networks and that can be used to predict the risk very well. Moreover, we find ways to generate the parameters for these regularities by only using data collected from small samples of the network. This implies that these results can be applied in contexts with little information, as long as the

network in question has similar scale-free properties to the networks examined in our study.

The rest of the paper is organized as follows. In Section 2, we review related work. In Section 3, we describe the network risk propagation model, the two real-world networks, and the methodology used in our analysis. Section 4 contains numerical illustrations and results. We discuss these results in Section 5. Finally, we conclude in Section 6.

2 Related work

We review related work in the areas of interdependent security, scale-free networks, and cyber-insurance. Interdependent security literature addresses ways in which risks propagate within a network; and our risk propagation model is taken from this literature. We use randomly-generated scale-free networks – in addition to real networks – to validate our structural formulae. Finally, cyber-insurance serves as a key motivation for our goal of understanding the risk portfolio of networks in general.

Interdependent Security The prevalence of risk correlation in network systems 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 [4]. Varian, for example, studied security compromises that result from the failure of independently-owned systems to contribute to an overall prevention objective (i.e., a public good) [5]. In this model, security compromises are often the result of misaligned incentives. 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 [6, 7, 8]. 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 [9]. 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 the partitioning of a network in order to contain a propagation. 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 included in those models would limit the generality of the results [10, 11]. Those limitations include the assumption that every infected node deterministically infects all unprotected neighbors.

A third class of propagation models is the class of epidemic models, which describe how a virus spreads or extinguishes in a network. The results of Kephart and White [12] are the closest to our analysis. They 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 formulae. 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. 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 [13]. 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 [14]. 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 [15].

Finally, a popular approach to model interdependent risk is taken by Kunreuther and Heal, and forms the basis for our analysis [16, 17, 18]. The basic premise of this work is to separately consider the impact of direct attacks and propagated attacks. We explain the propagation details of this model in Section 3.1. The model has been generalized to consider distributions of attack probabilities [19] and strategic attackers [20]. Similarly, Ogut et al. proposed a related model that allows for continuous (rather than binary) security investments [21]. Our analysis draws from these extensions by implicitly considering a continuum of risk parameters to study the distribution of outcomes.

Scale-Free Networks Many real-world networks are believed to be scale-free, including social, financial, and biological networks, and the Internet at the AS level [22]. 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 [23], 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 4. Li et al. introduce a new, mathematically more precise, and structural definition of "scale-free" graphs [24]. Their approach promises to offer rigorous and quantitative alternatives to many sensational qualitative claims found in the literature. The networks discussed in our paper satisfy this definition as well.

One 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 [25]; thus, subnet data cannot be naïvely extrapolated to every property of the entire network.

Cyber-Insurance A key objective of our work is to allow for a better assessment of the insurability of a networked resource. 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 [26, 27]. However, the market for cyber-insurance is developing at a frustratingly slow pace due to several key challenges [2].

First, a group of defenders might appear as a particularly appealing target to an attacker because of a high correlation in their risk profiles. For example, even though systems may be independently owned and administrated, they may exhibit similar software configurations leading to so-called monoculture risks [28, 29]. Böhme and Kataria study the impact of correlation which is readily observable for an insurer and found that the resulting insurance premiums to make the risks insurable would likely endanger a market for cyber-insurance [30]. Chen et al. study correlated risks by endogenizing node failure distribution and node correlation distribution [31]. In their work, they allow for different risk mitigation measures, but do not consider the impact on the insurability of risks, different cases of interdependence, or whether an insurer would be able to collect the necessary data to infer a distribution of failures (i.e., sampling).

Related work on insurance pricing models also informs our analysis of network insurability. Basic pricing literature points to some simple premium calculation principles [32, 33]. The simplest premium calculation principle is the *net premium principle* (or *pure risk premium*), which gives the risk premium as exactly the expected loss. This principle is commonly used in the literature [32], because actuaries assume that there is no risk if enough independent and identically-distributed policies are sold. Obviously, the pure risk premium without any (direct or indirect) loading is impractical, as it leads to unacceptably high probabilities of ruin. The expected value premium principle, the variance principle, and the standard deviation principle all build on the net premium principle by adding a constant fraction of the relevant metric (expected value, variance, or standard deviation, respectively) to the premium. The quantile premium for a risk threshold ϵ is the premium required to ensure that the probability of ruin is at most ϵ. More modern treatments of insurance often employ the *capital asset pricing model*, in which additional time-relative considerations such as reinvestment of premiums in a risk-free market are considered [34]. As our network model is not time-sensitive, we do not use capital asset pricing, but rather rely primarily on the more intuitive quantile premium principle.

3 Network Risk Model and Methodology

In this section, we describe our model and methodology. We begin by introducing the network risk model grounding our analysis. Then, we introduce two large real-world networks and two additional generated networks. We proceed to discuss two methods for selecting subsets of nodes from these networks; and finally, we address computational aspects of the node loss distributions.

3.1 Network Risk Model

Our risk propagation model builds on the framework for interdependent security games introduced by Kunreuther and Heal [16, 17]. This model gives loss probabilities for each node in a network based on a simple risk transfer process.

Risk Propagation Consider a network of N nodes. Each node is subject to some direct risk of compromise from outside the network. Node i is directly compromised with probability p_i. If node i becomes directly compromised, this failure can propagate at most one hop within the network to i's direct neighbors. If node i is compromised, this failure propagates indirectly to node j with probability q_{ij}. A node that is not directly compromised, but only indirectly compromised, cannot propagate failure to its neighboring nodes.

Loss Outcomes A loss outcome is an event in which some nodes are compromised and others are not. This loss outcome can be specified by listing the compromised nodes; and a complete distribution over loss outcomes is a probability distribution over the subsets of nodes. To make the analysis tractable, we focus on the projection of this distribution onto the number of compromised nodes.

To make things more formal, let N be the number of nodes, and suppose that the model is in a fixed configuration with given probabilities p_i and q_{ij} for $i, j = 1, \ldots N$. Let TL be the random variable which counts the number of compromised nodes in an outcome of the model. Then, a *loss distribution* (over the number of compromised nodes) is a set of $N + 1$ probabilities giving $\Pr[TL = k]$ for $k = 0, \ldots, N$.

3.2 Real-World Networks

Network of Autonomous Systems In the context of the Internet, an autonomous system (AS) is a collection of IP routing prefixes having a clearly-defined routing policy. By analyzing these routing policies, it is possible to construct a network in which each autonomous system is a node, and edges of various types correspond to traffic-sharing relationships between ASes.

One focus of our study is the network whose nodes consists of autonomous systems, and whose edges consist of business relationships between them. The graph is obtained from the Cooperative Association for Internet Data Analysis (CAIDA) [35]. This network consists of 41 thousand nodes and 121 thousand links, which results in an average degree of 5.9.

It can be useful to associate autonomous systems with Internet Service Providers (ISPs), although the comparison is not perfect, as some autonomous systems are controlled by more than one entity, and some ISPs control multiple autonomous systems. Nevertheless, the AS network structure has been studied by many researchers, largely because it serves as a good approximation of the connective architecture of the Internet at the organizational level.

Network of Facebook Friends Facebook is a social-networking platform that was founded in 2004, and it is the largest of its kind today. A second focus of our study is the network whose nodes consist of an anonymized collection of 1.2 million Facebook users, where the edges of the network represent friend relationships between these users [36, 37]. The sample was constructed in a way to ensure that it is an approximately uniform sample of the entire network. There are a total of 29.8 million edges between the 1.2 million users, which results in an average degree of 50.

Random Scale-Free Networks To frame our network analysis in the greatest possible generality, we also study randomly-generated scale-free networks whose parameters are chosen to approximate the two real-world networks described above. Prior work has established that many real-world networks have scale-free properties, meaning roughly that their degree distributions satisfy a power law. The two real-world networks under our consideration can be easily shown to have this property. Our generated networks behave in some ways similar to the real-world networks, although they differ in their construction and in a few key measures. We use these generated networks as additional validation tools for testing the feasibility of our risk prediction formulae.

To generate random networks, we use the Barabási-Albert (BA) model, which is based on two concepts: *network growth* and *preferential attachment* [23]. Network growth means that the number of nodes increases over time, while preferential attachment means that when a node is added to the network, it is more likely to connect to nodes that already have a lot of connections. More formally, given parameters N, m_0, and m, the BA model generates random scale-free networks as follows. First, an initial clique is created by connecting the first m_0 nodes to each other. Then, the remaining $N - m_0$ nodes are added to the network one by one. Each new node is connected to m existing nodes, each of which is chosen with a probability proportional to its degree.

3.3 Subsets of Nodes

We study the risk of node subsets in two contexts. First, we assume that, in practice, we are able to measure the risk of a small number of nodes. For example, we can use incident reports to this end, which originate from only these nodes. Second, based on the measured risks of small subsets of nodes, we aim to reliably predict the risk of larger subsets of nodes, including the whole network.[4] We consider two types of node subsets: random samples and geographical subsets.

We focus primarily on uniform random samples of nodes. These types of samples can model voluntary incident reports originating from a few nodes, or they can model the selected clients of an insurance provider. In both cases, we

[4] Note that we intentionally do not refer to these subsets of nodes as subnetworks. The reason for this distinction is that the term subnetwork would suggest that the links inside the subset inherently play a more important role than links connecting to the outside, or that these subsets are isolated from the rest of the network.

assume that the underlying network structure does not affect the node selection. Consequently, we choose a random sample of n nodes in a very straightforward way: we draw n nodes without replacement from the set of all nodes, in such a way that each node has the same probability of being drawn.

Unfortunately, random sampling does not model every scenario. For example, companies that are located in the same country, or persons with some common attribute, are more likely to choose the same insurer. In the autonomous systems network, there is a country identifier for each node. We use these identifiers to select country subsets, which consist of all the nodes from a single country. In the Facebook network, there are no such attributes, as the dataset has been thoroughly anonymized. Thus, we restrict our analysis to random samples in the Facebook network.

3.4 Computing the Loss Distributions

We determine the probability that a given number of nodes is lost by counting the number of losses in an outcome of the propagation model many times, and continuing until the probability for each such number approaches a fixed limit.

More formally, an empirical loss distribution $\hat{F}_{TL}$ can be efficiently computed as follows:

- Generate n independent loss outcomes $TL_1, \ldots, TL_n$, each using the following simulation:
  - For each node i, decide randomly whether node i is directly compromised (or not) according to p_i.
  - For each directly compromised node i, iterate over all of its non-compromised neighbors. For each non-compromised neighbor j, decide randomly whether there is a propagation from node i to node j according to q_{ij}.
  - The loss outcome is the number of compromised nodes.
- Compute the empirical loss distribution as $\hat{F}_{TL}(k) = \frac{\text{the number of outcomes in which at most } k \text{ nodes are compromised}}{n}$.

While prior work has established that directly computing the true distribution $F_{TL}(k)$ for an arbitrary network is NP-hard [3], in the examples we have studied, these estimators converge efficiently for both the real-world networks and their theoretical approximations. Once we know that our simulations converge, the strong law of large numbers then tells us that our results arbitrarily approximate the true distribution.

We also compute the loss distributions of subsets of nodes. In this case, we simulate the propagation model for the entire network, but only count the compromised nodes in the subset. Note that this differs from computing the loss distribution of the subnetwork induced by the subset, which would incorrectly assume that the given subset is isolated from the rest of the network.

4 Analysis and Results

In this section, we analyze each of the two real-world networks (denoted *CAIDA* and *Facebook*, respectively) and the random scale-free networks described in the previous section (denoted *BA CAIDA* and *BA Facebook*, respectively). We simulate the loss distribution for each network using the Kunreuther-Heal model with $p_i = 0.005$ for each i, and $q_{ij} = 0.1$ for each i and j. In [38] it is shown that the loss distributions retain similar structural properties when varying homogeneous parameters, with the differences being quantitative rather than qualitative.

We provide a variety of graphs for numerical illustration to facilitate maximum understanding of risks, but we concentrate our attention in the discussion on features most relevant to insurance. We focus on the right hand side of the distribution which indicates the probability of realizing large catastrophic network losses, and for the values of parameters in the charts we concentrate on the *safety loading* parameter which shows how much additional capital must be set aside by the insurer to cover a maximum number of compromised nodes up to a certain tolerable amount of risk.

We use the binomial distribution as a baseline compared to the loss distributions, for the purpose of measuring the risk of networks. The binomial distribution serves as a good baseline because this distribution has no correlation between loss events, and consequently no *non-diversifiable risk*. Non-diversifiable risks are caused by correlated events, where the probability of some nodes being compromised depends on whether another set of nodes has been compromised. The binomial distribution appears as the loss distribution of a network in which there are no connections, because in such a system, loss events are independent. For a fair comparison, we compare each network's or subset's loss distribution to the binomial distribution that has the same size and the same expected number of compromised nodes.

4.1 Overall Network Loss Distributions

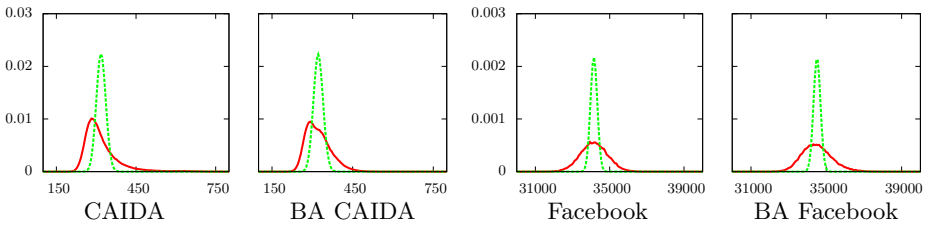

Fig. 1. Loss distribution of the whole network (solid red) compared to the binomial distribution that has the same expected number of compromised nodes (dotted green).

We begin with studying the loss distributions for the complete networks. These distributions can be seen in Figure 1. We find that, for every network, the loss distribution differs substantially from the binomial distribution with the

same mean. Recall that a binomial distribution would arise if the propagation probabilities were all zero, so that risks to individual nodes were independent.

Table 1. Statistics of the Loss Distributions Compared to Binomial Distributions

	CAIDA		BA CAIDA		Facebook		BA Facebook	
	actual	binom.	actual	binom.	actual	binom.	actual	binom.
Mean	319	319	322	322	34149	34149	34506	34506
Standard deviation	**67.3**	**17.8**	**45.9**	**17.9**	**723.1**	**182.1**	**794.5**	**183.0**
Quantile $Q(0.999)$	740	375	508	379	36414	34712	37487	35071
Safety loading for 0.999	**421**	**56**	**186**	**57**	**2265**	**563**	**2981**	**565**
Variance-to-mean ratio	14.23	0.993	6.53	0.992	15.31	0.971	18.29	0.971

Table 1 compares the networks' loss distributions to the binomial distributions having the same expected values. For every network, we see a substantial risk that a large number of nodes is compromised, compared to the binomial distributions. This indicates that the individual node compromise events are highly non-independent, resulting in correlations that are not non-negligible even for large networks. It is also interesting to note that the randomly-generated scale-free network's statistics are surprisingly close to the two real-world networks, especially for the Facebook network.

To illustrate the effect of this additional risk, consider an insurance premium for the Facebook network based on the naïve assumption of independent events.[5] Suppose that the insurance provider would like to keep her probability of ruin (i.e. the probability that the number of compromised nodes exceeds its expected value by more than her safety loading) below 0.1%. Thus, she wants to compute the insurance premium based on the quantile $Q(0.999)$, which means that her safety loading should be 2265. However, if she uses the binomial distribution instead, her safety loading is only 563. This has very severe consequences, as her probability of ruin with this safety loading is two orders of magnitude higher at 30.6%.

4.2 Loss Distributions of Subsets of Nodes

In the following, we study characteristic properties of our distributions based on subsets of varying size. Recall that we are not computing loss distributions on induced subnets, but are rather considering how risk propagation from the entire network affects a subset of nodes.

Table 2. Measured Constants for the Networks

		CAIDA	BA CAIDA	Facebook	BA Facebook
Average risk constant	$C =$	0.0077	0.0078	0.0287	0.0290
Dispersion constant	$A =$	0.000322	0.000134	0.000012	0.000015

[5] As we will later show, this assumption could be wrongly justified by the loss distribution measured on small sample.

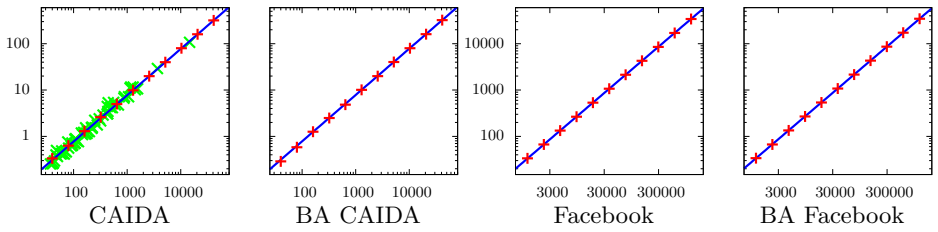

Fig. 2. Expected number of compromised nodes as a function of the number of nodes for random samples (red +) and countries (green `x`), and trendlines based on the formulae (solid blue line).

Number of Compromised Nodes Versus Number of Nodes We begin our analysis with the first moment of the loss distribution, the expected value of the number of compromised nodes. For the binomial distribution with parameter C, the expected number of compromised nodes is a linear function of the number of nodes, with linear slope C. In Figure 2, we analyze the relationship between the expected number of compromised nodes and the number of nodes in the subset. We find that if the nodes are chosen either as a random sample, or on a per country basis, then there is still a direct linear relationship similar to the relationship for the binomial distribution with the same mean. In particular, the ratio between the number of compromised nodes and the number of nodes is a constant, denoted by C, whose value for each network can be found in Table 2. We refer to C as the *average risk constant*. Formally,

$$\mu_{loss}(n) = Cn \ . \tag{1}$$

For random samples, there is very little deviation from this constant in every network. This shows that the average risk of random samples is an unbiased estimator of the average risk of the entire network. Recall that an unbiased estimator is an estimator whose expected value is equal to the parameter that it estimates.

For countries, however, there is some variation in average risk. This variation depends primarily on the average degree of the nodes in the country, and it is not correlated to the number of nodes in the country. This can be explained by the close relationship between a node's degree and risk due to indirect compromise.

Variance in Number of Compromised Nodes Versus Number of Nodes The variance of the binomial distribution with probability C and size n is $\sigma^2_{binomial} = C(1-C)n$. We analyze the relationship between the variance in the number of compromised nodes and the number of nodes in the subset using Figure 3. We find that for random samples, variance is a quadratic function of the sample size. The function is given by

$$\sigma^2_{loss}(n) = ACn^2 + C(1-C)n \ , \tag{2}$$

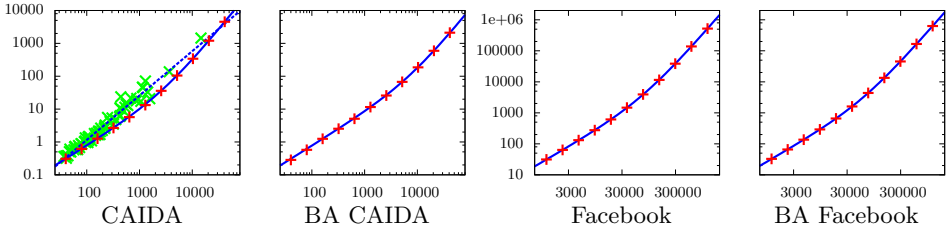

Fig. 3. Variance in the number of compromised nodes as a function of the number of nodes for random samples (red +), countries (green `x`), and trendlines for random samples (solid blue line) and for countries (dotted blue line) based on the formulae .

where C is the average risk constant defined above, and A is another constant, which we refer to subsequently as the *dispersion constant*, whose value for each network can be found in Table 2.

Notice that the right hand side of Equation (2) consists of two terms, and that the second term is equal to the variance of a binomial distribution with the same mean. This means that the variance of a random sample can be decomposed into two parts: a quadratic term and the variance of a binomial distribution. The second one is the inherent variance arising from having multiple nodes in the sample. This is a baseline variance, which we would see if the nodes were independent. Since, for risk-mitigation, this is the optimal case where all the risk is diversifiable, we will refer to this as the *diversifiable* part of the variance. The first part, on the other hand, is an extra quadratic term, which is a result of the risk correlations caused by the network structure. Hence, we will refer to this as the *non-diversifiable* part of the variance. Formally,

$$\sigma^2_{loss}(n) = \underbrace{ACn^2}_{\text{non-diversifiable risk}} + \underbrace{\sigma^2_{binomial}(n)}_{\text{diversifiable risk}} . \tag{3}$$

The relationship between variance in the number of compromised nodes and the number of nodes in country samples does not follow the same trend. These relationships are more noisy, and are better approximated by a power law of the form

$$Dn^E , \tag{4}$$

where

$$D \approx 0.0022971091 \text{ and } E \approx 1.3504067782.$$

Variance-to-Mean Ratio The *variance-to-mean ratio* (VMR) (also called the index of dispersion) is a normalized measure of the dispersion (i.e., variability or spread) of a probability distribution. Normalization means that the measure is independent of the expected value for many distributions (e.g., binomial or negative binomial distributions), and even independent of any parameters for some distributions (e.g., Poisson distribution). The variance-to-mean ratio of the binomial distribution with probability parameter C is $\text{VMR}_{binomial} = 1 - C$, regardless of the size of the distribution.

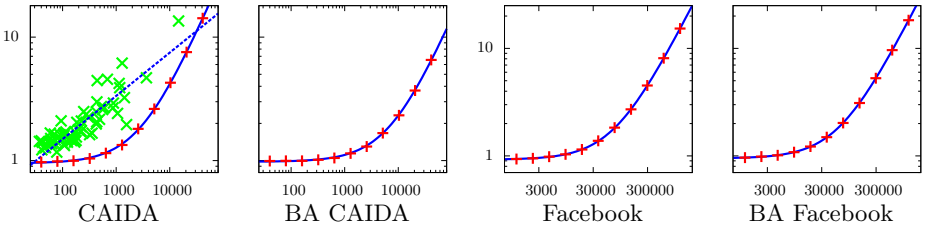

Fig. 4. Variance-to-mean ratio as a function of the number of nodes for random samples (red +), countries (green x), and trendlines for random samples (solid blue line) and for countries (dotted blue line) based on the formulae.

In Figure 4, we analyze the relationship between the variance-to-mean ratio and the number of nodes in the subset. We find that for random samples, the relationship is affine (but non-constant) with slope A and intercept $1 - C$. Formally, the variance-to-mean ratio for random samples of size n is given by

$$\mathrm{VMR}_{loss}(n) = An + 1 - C \tag{5}$$

$$= \underbrace{An}_{\text{non-diversifiable risk}} + \underbrace{\mathrm{VMR}_{binomial}}_{\text{diversifiable risk}} , \tag{6}$$

where C and A are the average risk constant and the dispersion constant, respectively.

For country samples, the relationship between VMR and the number of nodes in the country is again noisy and the relationship is again best approximated by a power function. Formally, the variance-to-mean ratio for countries of n nodes is approximated by

$$\frac{D}{C} n^{E-1} , \tag{7}$$

where C is the average risk constant, and D, E are the constants defined in Section 4.2 above.

4.3 Quantifying Insurability

Safety Loading Let μ be the expected number of compromised nodes, and let $Q(0.999)$ denote the number of compromised nodes such that with 99.9% probability, fewer or equal losses occur. Recall that the safety loading $Q(0.999) - \mu$ is the minimum amount of excess capital required to ensure that the probability of ruin is at most 0.001. Thus, safety loading is a good measure of how expensive a subset of nodes is to insure.

Figure 5 shows the value of safety loading as a function of the number of nodes in a subset of the network. Note that the safety loading increases, since the larger the subset, the more expensive it is to insure.

Safety Loading Versus Standard Deviation In Figure 6, we analyze the relationship between the number of nodes in the subset and the ratio of safety

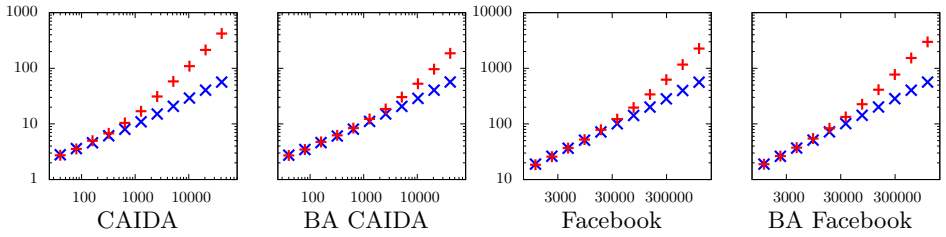

Fig. 5. Safety loading (for 0.999) as a function of size for random samples (red +) and for binomial distributions having the same average risk (blue x).

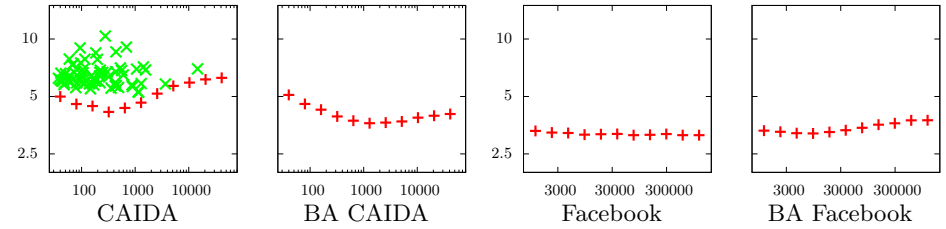

Fig. 6. Ratio of safety loading (for 0.999) to standard deviation for random samples (red +) and countries (green x).

loading to standard deviation. The results suggest that we can get a reasonable approximation of safety loading by considering only the standard deviation and multiplying it by a constant.

In the CAIDA network, the multiplicative constant is between 4 and 6.5 for all random samples and it is between 5 and 10 for countries. The average ratio is about 4.9 for random samples and about 6.5 for countries. In the Facebook network, the ratio is less noisy, the constant is between 3.1 and 3.3 for all samples sizes.

Since standard deviation is simply the square root of variance, its formula can be obtained from Equation (2) and is given by

$$\sigma_{loss}(n) = \sqrt{ACn^2 + C(1-C)n}\ , \tag{8}$$

and hence we can estimate safety loading by multiplying this value by an experimentally-determined constant K (that also depends on the maximum tolerable probability of ruin).

The safety loading can thus be estimated by the formula:

$$[Q-\mu]_{loss}(n) = K\sqrt{ACn^2 + C(1-C)n}\ . \tag{9}$$

Relative Safety Loading Relative safety loading is defined as the ratio of safety loading to the expected number of compromised nodes. Relative safety loading is a normalized measure of how expensive the subset is to insure. Figure 7 compares the relative safety loading for 0.999 of random samples and of binomial distributions.

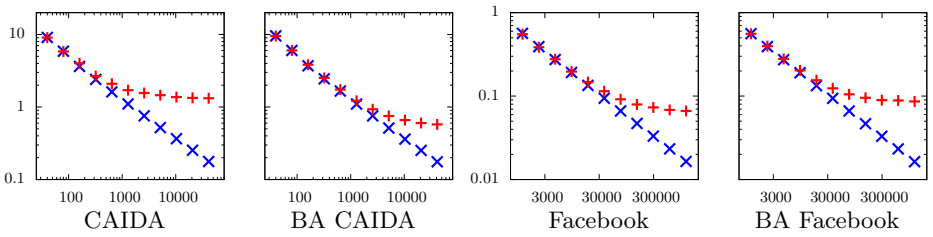

Fig. 7. Relative safety loading (for 0.999) as a function of size for random samples (red +) and for binomial distributions having the same average risk (blue `x`).

We can see that the relative safety loading for binomial distributions is steadily decreasing. For random samples, on the other hand, relative safety loading starts to decrease at the same rate as for the binomial distribution, but the curve flattens out after the sample size reaches about 2.5% of the complete network. The reason is that for smaller sample sizes, the dispersion of the loss distributions is dominated and determined by the diversifiable terms; however, as the sample size increases, the non-diversifiable terms – which have higher exponents – become relatively larger and cause substantial "extra" risk.

5 Discussion

The goal of our analysis is to show how to estimate risk in large networks, using information from small subsets of the network. We focused on cyber-insurance as the primary application, but our results can be applied to risk assessment and mitigation in general. While we analyzed only two real-world networks, the ubiquity of scale-free properties in many networks suggests our results yield additional applications.

We confirm the results of [3], which showed that the systematic risk estimated from even moderately-sized samples in scale-free networks is substantially lower than that of the complete network; so that naïve extrapolation underestimates the network's risk. In this paper, we study the problem in more detail for specific networks and find structural regularities that can aid in predicting the risk of a complete network (or larger subsets of it) from information that can be obtained from smaller samples.

Specifically, applying the formula for safety loading requires approximating the constants C, A, and K in Equation (9). The average risk constant C can be approximated from a small number of random samples, because the average risk in a random sample is an unbiased estimator of average risk in the whole network. The dispersion constant A can be determined experimentally from a small number of random samples using any two different sample sizes, since it is the slope of the trendline for the variance-to-mean ratio as a function of sample size. Finally, the constant K can also be estimated from a small number of random samples because the ratio of safety loading to standard deviation is roughly constant for all sample sizes. In summary, to estimate safety loading for

any desired number of nodes, first estimate C, A, and K using small random samples, and then substitute these values into Equation (9).

From a cyber-insurance provider's point of view, our findings can be summarized as follows. First, extreme care has to be taken when estimating the systematic risk of networks. Learning a complete network's topology is in practice impossible as this would require collecting data not only from the insured nodes, but also from their neighbors, with whom the insurer has no business relationship. Thus, one has to resort to predicting risk from small samples of historical data, such as incident reports. We show that this is very challenging, but nevertheless possible. Second, the insurer's portfolio should be chosen as close to a random sample as possible. For example, in the AS network example, this means that the insurer should aim for a geographically diverse portfolio.

6 Conclusions and Future Work

Our goal in this paper was to identify general rules practitioners can use to better estimate risks in networks. To achieve this goal, we used the connective structure of both real-world and randomly-generated scale-free networks to simulate attacks in which risk propagates subsequently through connections. The real-world networks – one involving social connections between users of Facebook, and the other involving business connections between the Internet's autonomous systems – had a known structure, but could otherwise be considered somewhat general representation of real-world networks. We identified structural regularities in these distributions, that allowed us to give predicting formulae for a variety of network risk measures; and we showed how to apply these formulae to estimate several risk measures for a large network even when one has only limited information about the network.

In this paper, our primary analysis of networks was limited to random samples. In future work, we intend to expand this study to other kinds of samples, for example, breadth-first search or other forms of grouping similar to our country samples. We would also like to expand our analysis to consider additional types of real-world networks whose structure differs from the scale-free variety used in our study. Finally, we intend to investigate the computability of additional risk metrics for networks.

Acknowledgements This research was partly supported by the Penn State Institute for CyberScience, CyLab at Carnegie Mellon under grant DAAD19-02-1-0389 from the Army Research Office, and the National Science Foundation under ITR award CCF-0424422 (TRUST). We also thank the reviewers for their comments on an earlier draft of the paper.

References

1. Markoff, J., Perlroth, N.: Firm is accused of sending spam, and fight jams Internet. The New York Times (March 26, 2013)

2. Böhme, R., Schwartz, G.: Modeling cyber-insurance: Towards a unifying framework. In: Workshop on the Economics of Information Security (WEIS). (2010)
3. Johnson, B., Laszka, A., Grossklags, J.: The complexity of estimating systematic risk in networks. In: Proceedings of the 27th IEEE Computer Security Foundations Symposium (CSF). (2014)
4. Laszka, A., Felegyhazi, M., Buttyán, L.: A survey of interdependent security games. Technical Report CRYSYS-TR-2012-11-15, CrySyS Lab, Budapest University of Technology and Economics (Nov 2012)
5. Varian, H.: System reliability and free riding. In Camp, L., Lewis, S., eds.: Economics of Information Security (Advances in Information Security, Volume 12). Kluwer Academic Publishers, Dordrecht, The Netherlands (2004) 1–15
6. Grossklags, J., Christin, N., Chuang, J.: Secure or insure?: A game-theoretic analysis of information security games. In: Proceedings of the 17th International World Wide Web Conference (WWW). (2008) 209–218
7. Fultz, N., Grossklags, J.: Blue versus red: Towards a model of distributed security attacks. In: Proceedings of the 13th International Conference on Financial Cryptography and Data Security (FC). (2009) 167–183
8. Grossklags, J., Johnson, B., Christin, N.: When information improves information security. Proceedings of the 14th International Conference on Financial Cryptography and Data Security (FC) (2010) 416–423
9. Johnson, B., Böhme, R., Grossklags, J.: Security games with market insurance. Decision and Game Theory for Security (2011) 117–130
10. Aspnes, J., Chang, K., Yampolskiy, A.: Inoculation strategies for victims of viruses and the sum-of-squares partition problem. Journal of Computer and System Sciences **72**(6) (September 2006) 1077–1093
11. Moscibroda, T., Schmid, S., Wattenhofer, R.: When selfish meets evil: Byzantine players in a virus inoculation game. In: Proceedings of the Twenty-Fifth Annual ACM Symposium on Principles of Distributed Computing. (2006) 35–44
12. Kephart, J., White, S.: Directed-graph epidemiological models of computer viruses. In: Proceedings of the IEEE Computer Society Symposium on Research in Security and Privacy. (1991) 343–359
13. Pastor-Satorras, R., Vespignani, A.: Epidemic spreading in scale-free networks. Physical Review Letters **86**(14) (2001) 3200–3203
14. Eguiluz, V., Klemm, K.: Epidemic threshold in structured scale-free networks. Physical Review Letters **89**(10) (2002) Article No. 108701
15. Pastor-Satorras, R., Vespignani, A.: Epidemic dynamics in finite size scale-free networks. Physical Review E **65**(3) (2002) Article No. 035108(R)
16. Kunreuther, H., Heal, G.: Interdependent security. Journal of Risk and Uncertainty **26**(2) (2003) 231–249
17. Heal, G., Kunreuther, H.: Interdependent security: A general model. Working Paper No. 10706, National Bureau of Economic Research (August 2004)
18. Kearns, M., Ortiz, L.: Algorithms for interdependent security games. In Thrun, S., Saul, L., Schölkopf, B., eds.: Advances in Neural Information Processing Systems 16. MIT Press, Cambridge, MA (2004) 561–568
19. Johnson, B., Grossklags, J., Christin, N., Chuang, J.: Uncertainty in interdependent security games. Decision and Game Theory for Security (2010) 234–244
20. Chan, H., Ceyko, M., Ortiz, L.: 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 (August 2012) 152–162

21. Ogut, H., Menon, N., Raghunathan, S.: Cyber insurance and IT security investment: Impact of interdependent risk. In: Workshop on the Economics of Information Security (WEIS). (2005)
22. Barabási, A.L.: Scale-free networks: A decade and beyond. Science **325**(5939) (July 2009) 412–413
23. Barabási, A.L., Albert, R.: Emergence of scaling in random networks. Science **286**(5439) (October 1999) 509–512
24. Li, L., Alderson, D., Doyle, J.C., Willinger, W.: Towards a theory of scale-free graphs: Definition, properties, and implications. Internet Mathematics **2**(4) (2005) 431–523
25. Stumpf, M., Wiuf, C., May, R.: 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 **102**(12) (2005) 4221–4224
26. Anderson, R.: Liability and computer security: Nine principles. In: Proceedings of the Third European Symposium on Research in Computer Security (ESORICS). (November 1994) 231–245
27. Böhme, R.: Towards insurable network architectures. it - Information Technology **52**(5) (2010) 290–293
28. Birman, K., Schneider, F.: The monoculture risk put into context. IEEE Security and Privacy **7**(1) (January 2009) 14–17
29. Geer, D., Pfleeger, C., Schneier, B., Quarterman, J., Metzger, P., Bace, R., Gutmann, P.: Cyberinsecurity: The cost of monopoly. How the dominance of Microsoft's products poses a risk to society (2003)
30. Böhme, R., Kataria, G.: Models and measures for correlation in cyber-insurance. In: Workshop on the Economics of Information Security (WEIS). (2006)
31. Chen, P.Y., Kataria, G., Krishnan, R.: Correlated failures, diversification, and information security risk management. MIS Quarterly **35**(2) (June 2011) 397–422
32. Čížek, P., Härdle, W., Weron, R.: Statistical tools for finance and insurance. Springer (2005)
33. Laeven, R., Goovaerts, M.: Premium calculation and insurance pricing. Encyclopedia of Quantitative Risk Analysis and Assessment (2008)
34. Sharpe, W.: Capital asset prices: A theory of market equilibrium under conditions of risk. Journal of Finance **19**(3) (1964) 425–442
35. The Cooperative Association for Internet Data Analysis (CAIDA): AS rank and AS relationship datasets `http://as-rank.caida.org/`, `http://www.caida.org/data/active/as-relationships/index.xml`
36. Gjoka, M., Kurant, M., Butts, C., Markopoulou, A.: Walking in Facebook: A case study of unbiased sampling of OSNs. In: Proceedings of the 29th IEEE Conference on Computer Communications (INFOCOM). (2010)
37. Gjoka, M., Kurant, M., Butts, C., Markopoulou, A.: Practical recommendations on crawling online social networks. IEEE Journal on Selected Areas in Communications **29**(9) (2011) 1872–1892
38. Johnson, B., Laszka, A., Grossklags, J.: How many down? Toward understanding systematic risk in networks. In: Proceedings of the 9th ACM Symposium on Information, Computer and Communications Security (ASIACCS). (2014)