Markov Random Fields for Malware Propagation: The Case of Chain Networks

Vasileios Karyotis, *Member, IEEE*

***Abstract*—Epidemic and stochastic models have been employed for describing the dynamic behavior of malware outbreaks. However, most of them lack a holistic treatment of the problem. In this work, we model malware propagation as a Markov Random Field and employ Gibbs sampling for the analysis of the system. We demonstrate the proposed framework for the case of a chain network, a model often emerging in both wired and wireless multi-hop networks.**

***Index Terms*—Markov Random Fields; Gibbs sampling; malware propagation; chain networks.**

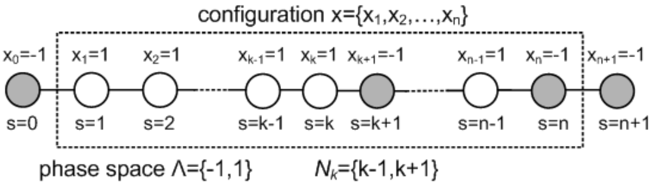

Fig. 1. SIS malware-propagative chain network and MRF notation.

I. Introduction

MALICIOUS software (malware) has become critical for network infrastructures, administrators and users. Traditional efforts in modeling malware propagation employ epidemics [1], where propagation is cast as a set of differential equations with properly defined node infection/recovery rates. However, epidemics are threat-specific and require prior knowledge of the involved rates.

Advanced epidemics (e.g. combined with Kalman estimation [2]) and probabilistic models (e.g based on Interactive Markov Chains [3]) have emerged as alternatives to the deterministic treatment of traditional epidemiological models. Advanced models intend to capture the probabilistic nature of malware propagation on arbitrary topologies. Furthermore, the model developed in [2] enables the design of a detection system, although, only for a specific type of malware threats, i.e. Internet worms. In [4] the impact of topology on the dynamics of the propagation was identified, through a Markovian approach and relevant approximations.

Motivated by the last observation, we propose an alternative stochastic framework that explicitly contains the impact of topology. Specifically, we model an attacked network as a Markov Random Field (MRF) [5], in which nodes oscillate between non-infection and infection, and then employ Gibbs sampling to analyze the propagation dynamics. Utilizing the proposed analytic framework, we focus on the aggregated behavior of the system. We demonstrate the effectiveness of the proposed framework for the case of a chain network (Fig. 1), a structure arising often in wired and wireless networks and can be also used as a building block for more complex topologies.

Our approach, aims at characterizing malware in the long-term, irrespectively of the specific type(s) of threat(s) propagating. However, most of the proposed models (especially epidemics) are devoted to the development of malware-specific models [1], [2]. Our work is closer to [3], [4] sharing the same objectives and essentially analyzing the 'macroscopic' behavior of the attacked system. Our framework differs, in that it is able to treat more general topologies and application scenarios (i.e. types of attack), as opposed to [3], [4]. Compared to [3], [4], we focus more on system transitions, the aggregated system state and its steady state, while topological information is embedded in the MRF framework by design.

II. Malware Propagative Markov Random Fields

We consider a malware propagative chain network with n nodes (Fig. 1), where infected legitimate nodes can further spread malware. Two malware sources at both ends infect the network. Users follow the Susceptible-Infected-Susceptible (SIS) infection paradigm [6], in which they oscillate between the non-infected and infected state. The SIS model is appropriate for the long-term behavior of a network, where new attacks emerge periodically without notification.

Assume a finite set S of cardinality n, with elements $s \in S$ referred to as sites (i.e. nodes). Let Λ be the set of possible states of each $s \in S$, called the phase space. A collection $X = \{X_s, s \in S\}$ of random variables with values in Λ is called a Random Field (RF) on S with phases in Λ and corresponds to a state of the system at each time instant. A configuration $x = \{x_s, s \in S\}$, where $x_s \in \Lambda$, corresponds to one of all possible states of the system. The product space Λ^n is called the configuration space.

A neighborhood system on S is defined as a family $\mathcal{N} = \{\mathcal{N}_s\}_{s \in S}$ of subsets $\mathcal{N}_s \subset S$, such that for every $s \in S$, $s \neq \mathcal{N}_s$ and $r \in \mathcal{N}_s$ if and only if $s \in \mathcal{N}_r$. $\mathcal{N}_s$ is called the neighborhood of site (node) s. The random field X is called a Markov Random Field (MRF) with respect to the neighborhood system $\mathcal{N}$, if for every site $s \in S$,

$$\mathbb{P}(X_s = x_s \mid X_r = x_r, r \neq s) = \\ = \mathbb{P}(X_s = x_s \mid X_r = x_r, r \in \mathcal{N}_s) \quad (1)$$

The aforementioned MRF framework abstracts connectivity and node interactions through the neighborhood system $\mathcal{N}$. The spatial Markov property expressed by a MRF describes

Manuscript received May 22, 2010. The associate editor coordinating the review of this letter and approving it for publication was F.-N. Pavlidou.

V. Karyotis is with the Department of Electrical and Computer Engineering, National Technical University of Athens (NTUA), Athens, 15780, Greece (e-mail: vassilis@netmode.ntua.gr).

Digital Object Identifier 10.1109/LCOMM.2010.072910.100866

the basic fact that malware infections propagate in a localized fashion that depends on 1-hop communications [7]. This explicitly excludes the cases of email contamination, which allows for a form of 'multi-hop' user infection. However, even in such cases, one can adapt the neighborhood definition and implicitly define a proper neighborhood system between the malware source and infected host, suppressing the intermediate non-infected nodes.

In this work, in order to demonstrate the effectiveness of the proposed framework, we detail the MRF for a chain network (Fig. 1). Chains (i.e. paths) arise often in wired networks, especially at the transport layer under the TCP protocol, and similarly in various types of multi-hop networks, such as sensor and ad hoc. Furthermore, the structure may be further extended to cover rings and lattices of various connectivity degrees. Every node of the chain corresponds to a site of a MRF, and for each site k, its neighborhood includes $\mathcal{N}_k = \{k-1, k+1\}$. For malware propagation, $\Lambda = \{-1, 1\}$, where -1 corresponds to the infected state and 1 corresponds to the non-infected state.

We employ Gibbs sampling for the analysis of the system. A random field X is called a Gibbs Random Field (GRF) if it satisfies:

$$\mathbb{P}\left(X = x\right) = \frac{1}{Z} e^{-\frac{U(x)}{T}} \tag{2}$$

where $Z := \sum_{x \in \Lambda^n} e^{-\frac{U(x)}{T}}$ is called the partition function of the system and $T = T(n)$ is called the temperature of the system. $U(x)$ is called the potential function and represents an 'energy' metric of configuration x. The potential function is not unique, however, a very useful class of potential functions is one in which $U(x)$ is decomposed into a sum of clique potentials:

$$U(x) = \sum_{c \in \mathcal{C}_s} \Phi_c(x) \tag{3}$$

where each clique potential depends only on the states of the cliques formed in the underlying graph of the system. $\mathcal{C}_s$ is the set of cliques formed. Clique-based representation of the potential function is very useful in malware propagation, when infections are transmitted through 1-hop neighbor interactions.

The Hammersley-Clifford theorem ensures that a GRF with distribution (2) and potential function expressed in terms of clique potentials leads to a MRF with conditional probabilities (1) and vice-versa [5]. In this work, we consider the potential function $U(x) = \sum_k \Phi_k(x)$, $x = \{x_k,\ 1 \le k \le n\}$, where $\Phi_k(x)$ depends only on x_k and the state random variables in $\mathcal{N}_k$, i.e. x_{k-1}, x_{k+1}. Thus, for the chain network under analysis:

$$\Phi_k(x) = \hat{\Phi}_k\left(x_k, \{x_{k'} : k' \in \mathcal{N}_k\}\right) = \hat{\Phi}_k\left(x_k, \{x_{k-1}, x_{k+1}\}\right) \tag{4}$$

We are interested in the interactions between neighboring nodes at different states, i.e. S(usceptible)-I(nfected) and I-S. Such node pairs essentially drive the evolution of malware propagation. By selecting $\Phi_k(x) = \Phi_k(x_k) = \sigma_k\sigma_{k-1} + \sigma_k\sigma_{k+1}$ for each clique potential, the overall system potential function may be obtained in the more general form: $U(x) = -D\sum_{(i,j)} \sigma_i\sigma_j$ for all neighboring pairs (i,j) in the chain network and $\sigma_i(x) = \sigma_i(x_i)$ is a function of the current state of site i. For simplicity and without loss of generality we assume $\sigma_i(x_i) = x_i$, but also any bijective function of x_i would work similarly under proper scaling. Thus, the form of the potential function of the chain network will be:

$$U(x) = -J\sum_{k=0}^{n} x_k x_{k+1} \tag{5}$$

where J is a proper scaling factor and so is D.

Simulated annealing can be combined with the Gibbs sampler in order to analyze the evolution of the system state [8]. Nodes are visited once in a random fashion (such a visit is called sweep) and the state of each one is updated. The process is repeated for a number of sweeps, until the system converges to its steady state. Prior to 'sweeping' the system, a temperature $T(\cdot)$ and the total number of sweeps I is selected. For each sweep a randomized visiting scheme is determined. For each node in the random sequence obtained, a decision is made on whether its state should remain the same for the next sweep or change. For demonstration purposes, we focus on binary decisions, but more states could have been employed reflecting different infection or non-infection states. For instance, if one is interested on how many infections a node currently suffers, $L+1$ states are required, state 0 corresponding to no infection and the rest $1, 2, \cdots, L$, denoting the number of different infections a node suffers (L being the maximum number of infections propagating in the network). For $\Lambda = \{-1, 1\}$ (two states) and $\ell \in \Lambda$ we have $\Phi_k(x_k^{\ell}) \doteq \hat{\Phi}_k\left(x_k = \ell, \{x_{k-1}, x_{k+1}\}\right)$. Then, from Gibbs distribution (2), the probability that the state of node k, will be $x_k = \ell$, may be obtained as:

$$\mathbb{P}\left(x_k = \ell\right) = \frac{e^{-\frac{\Phi_k(x_k^{\ell})}{T(n)}}}{\sum_{\ell' \in L_k} e^{-\frac{\Phi_k(x_k^{\ell'})}{T(n)}}} \tag{6}$$

where L_k refers to the possible states of the nodes $k' \in \mathcal{N}_k$. Furthermore, given the potential $\Phi_k(x_k)$, we have:

$$\begin{aligned} \Phi_k\left(x_k^{-1}\right) &= \hat{\Phi}_k(x_k = -1, x_{k-1}, x_{k+1}) = \\ &= \sigma_k(x)\sigma_{k-1}(x) + \sigma_k(x)\sigma_{k+1}(x) = \\ &= x_k x_{k-1} + x_k x_{k+1} = -\left(x_{k-1} + x_{k+1}\right) \end{aligned} \tag{7}$$

and

$$\Phi_k\left(x_k^{1}\right) = \hat{\Phi}_k(x_k = 1, x_{k-1}, x_{k+1}) = x_{k-1} + x_{k+1} \tag{8}$$

Consequently, the probability that the next state of node k is $x_k = 1$ is obtained as:

$$\mathbb{P}(x_k = 1) = \frac{1}{1 + e^{-\frac{\left(\Phi_k(x_k^{-1}) - \Phi_k(x_k^{1})\right)}{T(n)}}} = \frac{1}{1 + e^{\frac{2(x_{k-1} + x_{k+1})}{T(n)}}} \tag{9}$$

Similarly, the probability that the next state of node k is $x_k = -1$ is obtained as:

$$\mathbb{P}(x_k = -1) = \frac{1}{1 + e^{-\frac{\left(\Phi_k(x_k^{1}) - \Phi_k(x_k^{-1})\right)}{T(n)}}} = \frac{1}{1 + e^{-\frac{2(x_{k-1} + x_{k+1})}{T(n)}}} \tag{10}$$

where it is straightforward to verify that $\mathbb{P}(x_k = 1) = 1 - \mathbb{P}(x_k = -1)$. Consequently, in each sweep of the chain, the state of each site is updated according to $\mathbb{P}(x_k = 1)$ or $\mathbb{P}(x_k = -1)$. If one employed the mapping:

$$\sigma_k(x) = \sigma(x_k) = \begin{cases} 1, & \text{if } x_k = 1 \\ 0, & \text{if } x_k = -1 \end{cases} \tag{11}$$

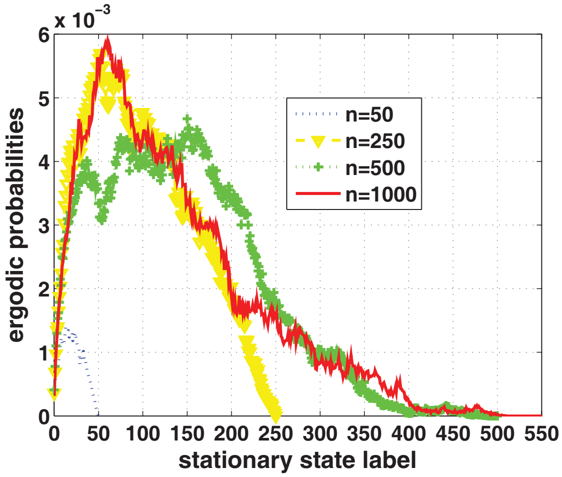

Fig. 2. Steady-state system distributions ($T/J = -0.2$).

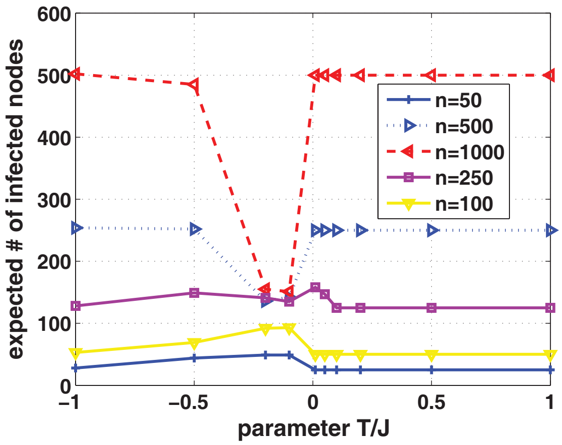

Fig. 3. Expected number of infected nodes.

the expressions obtained for $\mathbb{P}(x_k = 1)$, $\mathbb{P}(x_k = -1)$ would be similar to those of (9), (10) respectively, without the factor 2 in the exponential term of the denominator.

The effect of simulated annealing is more evident when the whole system is considered. Defining configuration $\bar{x}$ for a site k according to configuration x, where in $\bar{x}$ the state of site k has been switched, then

$$\frac{e^{-bU(\bar{x})}}{e^{-bU(\bar{x})} + e^{-bU(x)}} = \frac{1}{1 + e^{b(U(\bar{x})-U(x))}} \qquad (12)$$

may be used to determine the next state of k as well. Parameter $b = J/T$ is a proper constant.

The case of a ring network is identical to the chain, where now only one malware source is needed. Thus if nodes $s = 0,\ n+1$ are considered as one in Fig. 1, the above analysis holds and the results obtained apply.

III. Results and Discussion

Without loss of generality we simulate with random sweeps the chain in Fig. 1, i.e. nodes are visited once, randomly in each sweep. Totally $S = 10000$ sweeps are performed in each scenario and results are averaged over 50 different scenarios.

Figure 2 presents the distribution of the ergodic probabilities of the system, i.e. probabilities $\pi(i)$ that the system is at state i. The state of the system is defined as the number of infected nodes. The main observation is that as the size of the chain increases, it becomes tougher to increase the number of infected nodes (non-vanishing $\pi(i)$'s remain concentrated at lower states). This is because in a lengthy chain, the state of a node in a specific location, has little effect on the state of another node far away. Thus, it is tough for sources at the two ends to eventually propagate malware to nodes located in the middle of the chain. In relatively small chains, such as with $n = 50$ nodes, propagation becomes easier with non-vanishing values for all states.

Figure 3 shows the average number of infected nodes for various chain sizes and different temperatures T/J (infection-recovery rate combinations). The greater n is, the greater the expected number of infected nodes. However, this does not mean necessarily a greater percentage of infected nodes. In fact, there exist critical n, T/J values, for which system behavior changes drastically in the range $-0.5 < T/J < 0$ and depends significantly on T/J. Above the critical n threshold the percentage of infected nodes drops and increases sharply, whereas below the threshold, only a small drop is observed as $T/J \to 0^-$. Since T/J may be considered as the analog of the infection over recovery rate, indicating different propagation dynamics/capabilities, it is shown that for the range $-0.5 < T/J < 0$ attack potentials are practically limited, and almost independent of network size. For $T/J > 0.1$ propagation depends mainly on network size. These outcomes may be taken into account in designing efficient countermeasures.

IV. Conclusion and Future Work

We presented a framework for malware spreading in SIS propagative networks when infections propagate through direct node interactions. A MRF based approach was introduced for arbitrary networks and solved specifically for a chain network. MRFs are suitable for analyzing the spatial and contextual dependencies of malware propagation. Our future work, will be focused on two dimensional settings, which are expected to be more complicated, but also more fruitful.

References

[1] C. C. Zou, W. Gong, and D. Towsley, “Code red worm propagation modeling and analysis,” in *Proc. 9th ACM Conf. on Computer and Communications Security (CCS)*, pp. 138-147, Nov. 2002.

[2] C. C. Zou, W. Gong, D. Towsley, and L. Gao, “The monitoring and early detection of Internet worms,” *IEEE/ACM Trans. Networking*, vol. 13, no. 5, pp. 961-974, Oct. 2005.

[3] M. Garetto, W. Gong, and D. Towsley, “Modeling malware spreading dynamics,” in *Proc. 22nd Annual Joint Conf. of IEEE Comp. and Comm. Societies (INFOCOM)*, vol. 3, pp. 1869-1879, Mar. 2003.

[4] A. Ganesh, L. Massoulié, and D. Towsley, “The effect of network topology on the spread of epidemics,” in *Proc. 24th Annual Joint Conf. of IEEE Comp. and Comm. Societies (INFOCOM)*, vol. 2, pp. 1455-1466, Mar. 2005.

[5] R. Kindermann and J. L. Snell, *Markov Random Fields and Their Applications*. Providence, RI: American Mathematical Society, 1980.

[6] R. Pastor-Satorras and A.Vespignani, “Epidemic spreading in scale-free networks,” *Phys. Rev. Lett.*, vol. 86, pp. 3200-3203, Apr. 2001.

[7] M. H. R. Khouzani and S. Sarkar, “Dynamic malware attack in energy-constrained mobile wireless networks,” in *Proc. 5th Symposium on Information Theory and Applications*, UCSD, Feb. 2009.

[8] S. Geman and D. Geman, “Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 6, no. 6, pp. 721-741, Nov. 1984.