SEPTEMBER 19, 2018

Web malware detection with Markov random fields and belief propagation

Matteo Dora • matteo.dora@ens.fr

As modern society has become increasingly dependent on digital techologies and communication networks, malicious software emerges as a major threat. Nefariously famous computer worms such as Code Red or Storm have caused economic losses estimated in the billions of USD [10] and sophisticated state-sponsored attacks have been used to leak business secrets [24] or damage military targets [14]. Due to their increasing popularity, websites have become one of the most frequent attack vectors. Extensive research has been conducted to detect malicious web domains based on their behaviour [25]. Here instead I follow a different approach. First, I analyse the topology of the interaction network between users and web domains, then I review and extend a simple Markov random field (MRF) model for malware epidemics. Finally, I use belief propagation to detect malicious domains given partial observation of user hosts, evaluating the quality of predictions via numerical simulations.

INTRODUCTION

Malicious software has been around since the days of yore, but it is only in the last two decades that it became a highly profitable business. The massive amount of sensible information that is nowadays available on digital devices make for an attractive loot to cybercriminals, and the internet provides a very effective medium for propagation. With the profitability growth, more and more resources have been dedicated to malware development, leading to increasingly advanced tools. Modern malware can be roughly divided into two main categories. On one hand, criminal organisations and intelligence agencies sponsored the creation of highly sophisticated attack tools. Examples of this kind of malware are Stuxnet [14], a worm targeting Iran's nuclear infrastructure, and Zeus, a criminal botnet dismantled in 2010 with more than 100 arrests worldwide [17]. On the other hand, older versions of advanced malware and lesser evoluted tools began circulating in a thriving black market, allowing small criminal groups and individuals to approach the business. These kind of malware is usually directed at novice users and diffuses through email, websites and social networks. In the following, I will focus on this latter category of less sophisticated but widely spread malware, altough some of the methods described may be used in more advanced contexts.

In the recent years, the most popular threat is represented by based malware [15]. This breed of malware is injected into target devices by malicious web pages. It can spread by inducing users into following malicious hyperlinks, typically distributed via social networks or email. Moreover, legitimate websites compromised compromised with techniques such as cross site scripting (XSS) or SQL injection (sometimes performed by the infected devices themselves) can be exploited to diffuse malware. Malicious web domains have usually a relatively short life, as once identified they can immediately taken down by authorities or blocked by antivirus software. Nonetheless, cybercriminals are continuously registering new malicious domains (or compromising existing ones), in an infinite hide-and-seek game. How can web threats be prevented? Is it possible to identify malicious domains at early stage, before most of the damage is done?

In the following, I will present a simple model for malware spreading based on Markov random fields. Then I will analyse the network topology on which the model can be applied in the case of web domains. Finally I will use an inference method, previously explored by Manadhata et al. [16], and check with numerical simulations how well it performs in identifying malicious web domains.

MODELLING MALICIOUS WEB DOMAINS

Many approaches modelling malware epidemics have been proposed based on classical compartmental models such as SIS and SIR [15, 8, 10]. Although these deterministic models

are valid and have proven successful (for example in the case of the Code Red worm [26]), they can only describe the global characteristics of the infection diffusion. Stochastic epidemic models would overcome most of the limitations of deterministic ones, but analytical analysis (e.g. using queuing theory) is often intractable. Moreover, both categories are extremely sensible to variations in the topology of the system, which can affect the system ergodicity, making these models an inconvenient choice when dealing with complex networks.

Following the framework proposed by Karyotis [10, 11, 9], here I will focus on a simpler model constituted by a Markov random field (MRF). A MRF is network of random variables interacting only with their neighbours. Despite their simplicity, MRFs are good candidates to describe a malware propagation. The Markov spatial property is a sensible assumption, as malware can only spread by direct contact. Also, they are extremely robust to topological variations, as each variable, given its neighbours, is conditionally independent of any other node. This latter characteristic make them very useful to model processes on different kind of networks.

To describe malware diffusion we consider the SIS paradigm and build a MRF where nodes can assume two states: susceptible ($x = +1$) or infected ($x = -1$). Nodes are organised on a network where links represent data exchange between the nodes. Intuitively, susceptible nodes communicating with infected neighbours will turn infected with higher probability. This kind of interaction can be described by an Ising Hamiltonian:

$$p(X) = \frac{1}{Z} e^{-\beta H(X)} \tag{1}$$

$$H(X) = -J \sum_{\langle i,j \rangle} x_i x_j - \sum_i h_i x_i \tag{2}$$

where X denotes a sequence of node states $(x_0, x_1, x_2, \dots)$ (the system configuration), J is the interaction potential and h_i is the local field of node i. It is important to notice that this joint probability function factorises:

$$p(X) = \frac{1}{Z} \left(\prod_i e^{\beta h_i x_i} \right) \left(\prod_{\langle i,j \rangle} e^{\beta J x_i x_j} \right) \tag{3}$$

$$\doteq \frac{1}{Z} \left(\prod_i \phi_i(x_i) \right) \left(\prod_{\langle i,j \rangle} \phi_i j(x_i, x_j) \right) \tag{4}$$

In this setting, all the nodes are in principle free to switch between the two states. Yet, we want to represent attackers as infected nodes that never change state. We can do that by setting their local field to infinite (in this case, $h_{attacker} = -\infty$). The same technique can be used to fix the state of immune nodes. Finally, we add to this model a human factor. In fact, system administrators are usually checking their domains for malware and keep them healthy. Although history teaches us that no domains are immune to malware [13], regardless of their popularity, administrators of popular domains will likely detect attacks and respond more quickly. To consider this in the model, we add to domain nodes a small positive field $h_i \sim log(d_i)$, where d_i is the degree of node i.

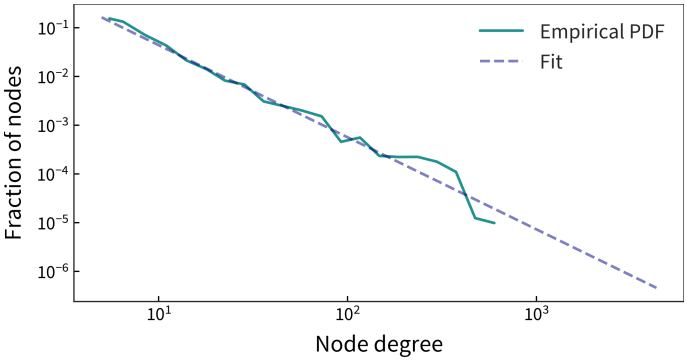


Figure 1 Degree distribution obtained by the analysis of the CTU-13 dataset (log-log plot). The result is typical of a scale-free network, showing the characteristic power law degree distribution. A power law fit (dashed line) in the form $p(x) \sim x^{-\alpha}$ results in exponent $\alpha = 1.89$.

Summarising, our model follows a *guilt by association* approach, where nodes tend to assume the same state of their neighbours. Whether this represents a good description of real malware processes is debatable, but previous works have shown that this kind of model can perform well for malware [23, 16, 3] and similar cases including fraud detection [19] and botnet detection [4].

Network topology

Let us suppose that, given a local network, we can intercept every request coming from local hosts and directed to domains in the outer web (e.g. through a HTTP proxy, as described in [16]). Using this knowledge, we can build a bipartite graph linking local hosts to the remote domains they interact with. In a real network, these links will not be uniformly distributed, but will likely depend on the popularity of the domains. Which are the characteristics of the interaction network between users (local hosts) and web domains? It has been shown that when popularity drives the node connectivity (the so called *preferential attachment* mechanism), a scale-free topology may emerge [2, 1]. Scale-free graphs are characterised by a power law node

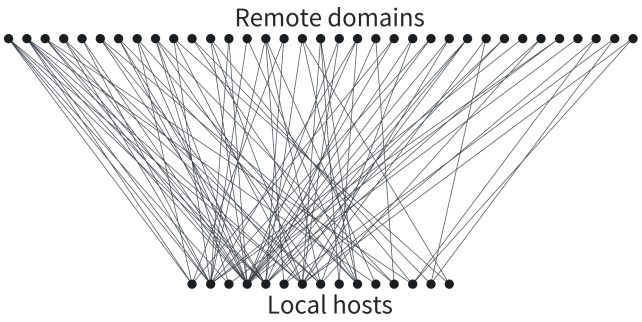


Figure 2 Example of bipartite scale-free graph for hosts and domains. The graph was generated using the Barábasi–Albert algorithm [1], conveniently adapted for the case of bipartite networks.

degree distribution. Examples are social networks, or the network of web pages connected by hyperlinks [12, 6]. To check whether the host-domain communicaton network belongs to this topological class, I analysed the CTU-13 dataset provided by the Czech Technical University [7]. The CTU-13 data contains the traffic flows between a local network and the outer internet. For privacy concerns, only a truncated version of IP packets is provided, so that information about the payload is lost. Nevertheless, it is possible to recover the traffic flow between IP addresses. I preprocessed the data by discarding the internal traffic flows (i.e. those between hosts of the local network) and considering only exchanges greater than 500 B (a sensible minimum for a web page request through HTTP). A network is built by linking hosts and domain based on the flows in the dataset after preprocessing. I found that the empirical degree distribution of this network is well described by a power law (fig. 1), confirming the initial hypothesis. The host-domain interaction graph is thus a bipartite scale-free network.

Epidemic behaviour

To analyse how a malware epidemic spread, I generated configurations for the MRF model previously described using Markov chain Monte Carlo (MCMC) on random bipartite scale-free networks. The networks were generated using the Barábasi–Albert algorithm [1], conveniently adapted for the case of bipartite networks (for details, see the supporting code [5]). A variable number of domains, randomly selected, were set as attackers (with fixed infected state). To improve the efficacy of the MCMC algorithm, I used simulated annealing (SA), decreasing temperature as $\beta \sim \log t$, where t is the simulation time. Then, I calculated the average number of infected hosts with respect to the number of malicious domains. The result (fig. 3), shows a clear transition to a completely infected state. The critical threshold is reached at around 25% infected domains. The shape of this transition is similar to the one observed in [26]. The macroscopic description of the epidemic is thus compatible with the one provided by classical SI compartmental models.

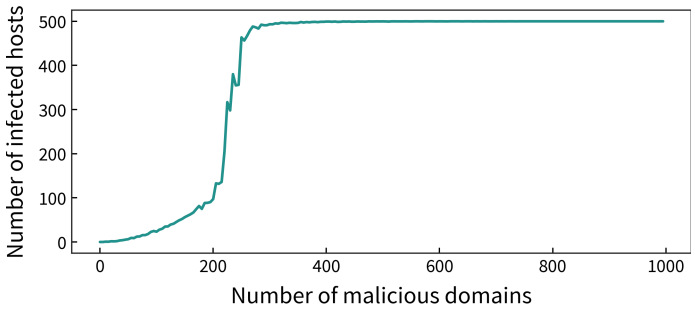
DETECTING MALICIOUS ACTORS

Let us assume that we can inspect the state of some of the hosts in the local network. Is it possible, given this partial knowledge, to identify malicious web domains? A known method to perform this kind of inference is belief propagation. This algorithm, initially proposed by Pearl for trees [21] and polytrees [20], can be used to estimate the marginal probability of hidden variables (given the known variables) in an arbitrary Markov random field. Belief propagation is a message-passing algorithm, where each node send to neighbours its estimate of their states (belief) in form of a *message*, and subsequently collects messages from neighbours to update its own belief. After a sufficient number of iterations the beliefs usually converge to an approximate estimate of marginals, although this cannot be proved in general.

I implemented a specific version of belief propagation known as max-product message passing, where the message from node i to its neighbour j is defined as:

$$m_{i\to j}(x_i) = \max_{x_i=\pm 1} \left\{ \phi_{ij}(x_i, x_j)\phi_i(x_i) \prod_{k\in\mathcal{N}(i)\setminus j} m_{k\to i}(x_i) \right\} \quad (5)$$

Figure 3 Critical behavious of the system. Increasing the number of actively malicious domains over a critical threshold, we obtain a complete infection.



and the belief assigned to the state x_i at node i is

$$b(x_i) = \phi_i(x_i) \prod_{k \in \mathcal{N}(i)} m_{k \to i}(x_i) \tag{6}$$

It must be noted that in this case the belief does not correspond to the marginal distribution, but can be considered a score function. The prediction for node i will be the state that maximises the score. Detailed implementation of the belief propagation algorithm can be found in the supporting code [5].

Prediction quality

To evaluate the quality of predictions I ran numerical simulations. First, I generated configurations with MCMC as described in the previous section. A fraction of the total domains (25%) were set as attackers (with fixed infected state). Then, I generated a prediction on the domain states given a variable number of known hosts using max-product belief propagation. Following the intuition that popular domains represent a more tough target for malware, a prior was set on the top 100 most popular domains mapping their popularity to an increasing probability of being in the non-infected state with a logistic function (a similar approach was proposed in [16]).

Evaluating the quality of prediction requires a bit of care, since mallicious domains are only a small fraction of the total number. In this situation, a biased estimator (which favours non infected state) would obtain good accuracy. It is thus required to check both the *true positive rate* (TPR) and the *false positive rate* (FPR). In particular for our application we want to avoid false positives (i.e. marking a legit website as malicious), while maximising the TPR. To help maintaining low the FPR, a slightly positive prior was set on all unknown nodes. Results of the simulations are shown in fig. 4.

RESULTS

Based on the CTU-13 dataset, I have found that host-domain interaction networks may exhibit a bipartite scale-free topology. In the literature, a similar structure was observed in collaboration networks [22, 18]. This organisation is likely to emerge in many other context where interaction by popularity is mediated by common factors. Nonetheless, the topology does not seem to affect the global behaviour of the model, as I found a critical infection threshold characterised by a transition compatible with a classical compartmental description.

Belief propagation proved to be quite effective in inferring states even for a very small number of observed nodes. The results of fig. 4 show that, with only a partial knowledge of the states of hosts and no information about the domains, it is still possible to detect a fraction of malicious actors while keeping FPR to an extremely low level. Moreover, the belief propagation algorithm has been found to perform extremely well on the bipartite scale-free graph, requiring few iterations to converge.

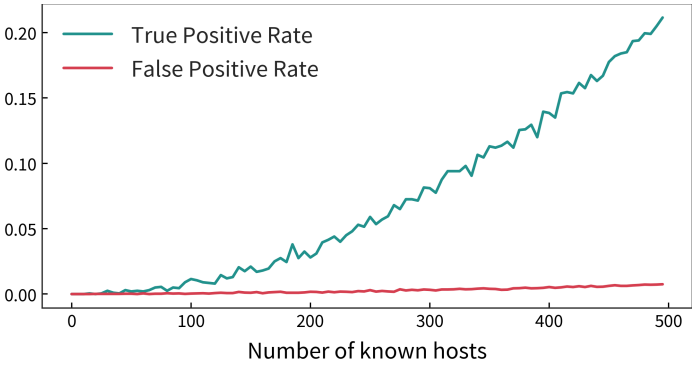


Figure 4 Quality of detection. TPR and FPR for a variable number of known hosts. Simulations were performed on a network of 1000 domains and 500 nodes, with 200 actively malicious actors. No information on the domains was given except a popularity based prior.

These results encourage to deepen the analysis with real data to thoroughly test the validity of the proposed model.

References

[1] Réka Albert and Albert-László Barabási. "Statistical mechanics of complex networks". In: *Reviews of modern physics* 74.1 (2002), p. 47.

[2] Albert-László Barabási and Réka Albert. "Emergence of scaling in random networks". In: *science* 286.5439 (1999), pp. 509–512.

[3] Duen Horng Chau and Adam Wright. "(54) INFERRING FILE AND WEBSITE". In: (). 00000, p. 17.

[4] Baris Coskun, Sven Dietrich, and Nasir Memon. "Friends of an enemy: identifying local members of peer-to-peer botnets using mutual contacts". In: *Proceedings of the 26th Annual Computer Security Applications Conference on - ACSAC '10*. the 26th Annual Computer Security Applications Conference. 00088. Austin, Texas: ACM Press, 2010, p. 131. ISBN: 978-1-4503-0133-6. DOI: 10.1145/1920261.1920283. URL: http://portal.acm.org/citation.cfm?doid=1920261.1920283 (visited on 08/31/2018).

[5] Matteo Dora. *Web malware detection with Markov random fields and belief propagation*. Supporting code. 2018. URL: https://github.com/mattbit/malware.

[6] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos. "On power-law relationships of the internet topology". In: *ACM SIGCOMM computer communication review*. Vol. 29. 06460. ACM, 1999, pp. 251–262.

[7] Sebastian Garcia et al. "An empirical comparison of botnet detection methods". In: *computers & security* 45 (2014), pp. 100–123.

[8] Santiago Gil, Alexander Kott, and Albert-László Barabási. "A genetic epidemiology approach to cyber-security". In: *Scientific reports* 4 (2014), p. 5659.

[9] Vasileios Karyotis. "A Markov Random Field Framework for Modeling Malware Propagation in Complex Communications Networks". In: *IEEE Transactions on Dependable and Secure Computing* (2017).

[10] Vasileios Karyotis and M. H. R. Khouzani. *Malware diffusion models for modern complex networks: theory and applications*. 00012. Amsterdam Boston Heidelberg und 9 weitere]: Elsevier, MK Morgan Kaufman, Morgan Kaufman Publishers is an imprint of Elsevier, 2016. 301 pp. ISBN: 978-0-12-802714-1.

[11] Vasileios Karyotis et al. "A novel framework for mobile attack strategy modelling and vulnerability analysis in wireless ad hoc networks". In: *International Journal of Security and Networks* 1.3 (Jan. 1, 2006). 00024, pp. 255–265. ISSN: 1747-8405. DOI: 10.1504/IJSN.2006.011785. URL: https://www.inderscienceonline.com/doi/abs/10.1504/IJSN.2006.011785 (visited on 09/10/2018).

[12] Serge A. Krashakov, Anton B. Teslyuk, and Lev N. Shchur. "On the universality of rank distributions of website popularity". In: *arXiv:cs/0404010* (Apr. 5, 2004). arXiv: cs/0404010. URL: http://arxiv.org/abs/cs/0404010 (visited on 09/10/2018).

[13] Brian Krebs. "Hacked ad seen on myspace served spyware to a million". In: *WashingtonPost* (2006).

[14] David Kushner. "The real story of stuxnet". In: *ieee Spectrum* 3.50 (2013), pp. 48–53.

[15] Wanping Liu and Shouming Zhong. "Web malware spread modelling and optimal control strategies". In: *Scientific Reports* 7 (Feb. 10, 2017). 00015, p. 42308. ISSN: 2045-2322. DOI: 10.1038/srep42308. URL: https://www.nature.com/articles/srep42308 (visited on 09/06/2018).

[16] Pratyusa K Manadhata et al. "Detecting malicious domains via graph inference". In: *European Symposium on Research in Computer Security*. Springer. 2014, pp. 1–18.

[17] *More than 100 arrests, as FBI uncovers cyber crime ring*. BBC. URL: https://www.bbc.com/news/world-us-canada-11457611.

[18] Jun Ohkubo, Kazuyuki Tanaka, and Tsuyoshi Horiguchi. "Generation of complex bipartite graphs by using a preferential rewiring process". In: *Physical Review E* 72.3 (Sept. 21, 2005). 00055. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.72.036120. URL: https://link.aps.org/doi/10.1103/PhysRevE.72.036120 (visited on 09/13/2018).

[19] Shashank Pandit et al. "Netprobe: a fast and scalable system for fraud detection in online auction networks". In: *Proceedings of the 16th international conference on World Wide Web - WWW '07*. the 16th international conference. 00355. Banff, Alberta, Canada: ACM Press, 2007, p. 201. ISBN: 978-1-59593-654-7. DOI: 10.1145/1242572.1242600. URL: http://portal.acm.org/citation.cfm?doid=1242572.1242600 (visited on 08/31/2018).

[20] Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Elsevier, 2014.

[21] Judea Pearl. *Reverend Bayes on inference engines: A distributed hierarchical approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science, University of California, Los Angeles, 1982.

[22] José J. Ramasco, S. N. Dorogovtsev, and Romualdo Pastor-Satorras. "Self-organization of collaboration networks". In: *Physical Review E* 70.3 (Sept. 14, 2004). 00318. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.70.036106. URL: https://link.aps.org/doi/10.1103/PhysRevE.70.036106 (visited on 09/13/2018).

[23] Acar Tamersoy, Kevin Roundy, and Duen Horng Chau. "Guilt by association: large scale malware detection by mining file-relation graphs". In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*. the 20th ACM SIGKDD international conference. 00064. New York, New York, USA: ACM Press, 2014, pp. 1524–1533. ISBN: 978-1-4503-2956-9. DOI: 10.1145/2623330.2623342. URL: http://dl.acm.org/citation.cfm?doid=2623330.2623342 (visited on 08/31/2018).

[24] *The Sony Pictures hack, explained*. Washington Post. 00021. URL: https://www.washingtonpost.com/news/the-switch/wp/2014/12/18/the-sony-pictures-hack-explained/ (visited on 09/17/2018).

[25] Yury Zhauniarovich et al. "A Survey on Malicious Domains Detection through DNS Data Analysis". In: *arXiv:1805.08426 [cs]* (May 22, 2018). 00000. arXiv: 1805.08426. URL: http://arxiv.org/abs/1805.08426 (visited on 09/13/2018).

[26] Cliff Changchun Zou, Weibo Gong, and Don Towsley. "Code Red Worm Propagation Modeling and Analysis". In: (), p. 10.