# Assessment of Security Threats via Network Topology Analysis: An Initial Investigation

Marcello Trovati[1(✉)], Win Thomas[2], Quanbin Sun[1], and Georgios Kontonatsios[1]

[1] Department of Computer Science, Edge Hill University, Ormskirk, UK
Marcello.Trovati@edgehill.ac.uk
[2] Department of Computer Science, Gloucestershire University, Cheltenham, UK

**Abstract.** Computer networks have increasingly been the focus of cyber attack, such as botnets, which have a variety of serious cybersecurity implications. As a consequence, understanding their behaviour is an important step towards the mitigation of such threat. In this paper, we propose a novel method based on network topology to assess the spreading and potential security impact of botnets. Our main motivation is to provide a toolbox to classify and analyse the security threats posed by botnets based on their dynamical and statistical behaviour. This would potentially lead to a better understanding and prediction of cybersecurity issues related to computer networks. Our initial validation shows the potential of our method providing relevant and accurate results.

**Keywords:** Botnets · Cybersecurity · Network theory

## 1 Introduction

Security threats have been steadily increasing due to the emergence of new technology and methodologies, which has led to an expanding research effort to detect and minimise such threats [1,2]. More specifically, *botnets* due to their unique structure based on distributed communication command patterns across networks, are widely regarded as a serious security issue. In fact, they can successfully carry out surveillance attacks, perform DDoS extortion, general spam, as well as phishing. Furthermore, some of them utilise structured overlay networks, whose lack of centralisation enhance the ability of a botnet to evade detection whilst retaining a good level of robustness with respect to a churn process, where single machines are frequently cleansed [6]. It is estimated that their use has led to malicious activity resulting in a loss of millions of dollars per year [5].

In this paper, we introduce a novel method to assess security threats, based on the dynamical properties associated with networks generated by computer communication. In fact, their topology can provide an insight into specific features exhibited by botnets across computer networks. How connections change,


M.H.A. Au et al. (Eds.): GPC 2017, LNCS 10232, pp. 416–425, 2017.
DOI: 10.1007/978-3-319-57186-7_31

**Table 1.** A selection of network connection flows

| Time | Source | Destination | Protocol | Length |
|---|---|---|---|---|
| 161.077519 | PcsCompu_b5:b7:19 | Broadcast | ARP | 60 |
| 162.079007 | PcsCompu_b5:b7:19 | Broadcast | ARP | 60 |
| 162.079013 | PcsCompu_b5:b7:19 | Broadcast | ARP | 60 |
| 162.765245 | 147.32.84.165 | 147.32.84.255 | NBNS | 110 |
| 162.765253 | 147.32.84.165 | 147.32.84.255 | NBNS | 110 |
| 166.206344 | 147.32.80.9 | 147.32.84.165 | DNS | 503 |
| 166.207297 | 147.32.84.165 | 74.125.232.195 | TCP | 62 |
| 166.207308 | 147.32.84.165 | 74.125.232.195 | TCP | 62 |
| 166.215343 | 74.125.232.195 | 147.32.84.165 | TCP | 62 |
| 166.21559 | 147.32.84.165 | 74.125.232.195 | TCP | 60 |

their types and length of communication can provide a deeper and more efficient approach to security threat detection and prediction.

To achieve this, we consider five main parameters: *time, source, destination, protocol*, and *length*. Table 1 depicts a small example of these parameters of the connection flows.

The main motivation is to provide a set of tools to assess the behaviour of host-to-host communication to allow an agile, real-time assessment. In contrast to the current state of the art approaches, which tend to focus on the different parameters based on whether or not they are present in the collected data, we are aiming to exploit the topology of the network and the probabilistic information related to botnets behaviour. In fact, a dynamical investigation of such networks, can lead to the assessment of the likelihood of the maliciousness of computer communications.

The paper is structured as follows. In Sects. 2 and 3 we provide a description of existing technology and theories, and in Sect. 4 we detail our approach. In Sect. 5 we discuss the validation process and finally, Sect. 6 concludes our work and prompts to future research directions.

## 2 Related Work

In [3], the authors propose a detection method for botnets from large datasets of Netflow data, based on a variety of cloud computing paradigms especially MapReduce for detecting densely interconnected hosts which are potential botnet members.

BotGrep [5] is a tool to identify peer-to-peer communication structures based on the information about communicating pairs of nodes. This type of P2P detection is defined as a (communication) network, which exploits the spatial relationships in communication traffic. Furthermore, the authors argue that subnetworks with different topological patterns can be partitioned by using random walks,

whilst comparing the relative mixing rates of the P2P subnetwork structure and the rest of the communication network. However, such approach is computationally expensive due to the typical size of such networks.

In [4], an approach based on a Markov chain model in introduced. In particular, botnet infection is modelled to identify behaviour that is likely to be associated with attacks, with a prediction rate over 98%. Another example of the utilisation of Markov chain for intrusion detection system is described in [8], where it is trained on a sequence of audit events. However, these types of approaches allow attack identification but they have limited intrusion prediction. In [9], the authors assess the set of bot lifecycle stages using Markov chains to identify the occurrence of infection. Similar to the previous approach, there is a focus on the identification of infection rather than on any predictive capability.

## 3 Network Theory

Networks have been extensively used to successfully model many complex systems, and their applications span across a variety of multidisciplinary research fields, ranging from mathematics and computer science, to biology, and the social sciences [12,13].

Networks are defined by a *node set* $V = \{v_i\}_{i=1}^{n}$, and the *edge set* $e_{v_i,v_j} \in E$, so that if $v_a$ and $v_b \in V$ are connected, then $e_{v_a,v_j} \in E$ [2]. Note that in this paper, we do not allow self-loops, or in other words, $e_{v_i,v_i} \notin E$.

Scale-free networks, in particular, appear in a numerous contexts, such as the World Wide Web links, biological and social networks [2]. The main property of scale-free networks is based on their node degree distribution, which follows a power law. More specifically, for large values of $k$, the fraction $p_k$ of nodes in the network having degree $k$, is modelled as

$$p_k \approx k^{-\gamma} \tag{1}$$

where $\gamma$ has been empirically shown to be typically in the range $2 < \gamma < 3$ [2].

From Eq. 1, it follows that a relatively small number of hubs occur, which define the topological properties of the corresponding networks, as well as the way information spreads across them [15].

An important property of such networks is related to the creation of new nodes over time, which are likely to be connected to existing nodes that are already well connected. Since the connectivity of nodes follows a distribution which is not purely random, the dynamical properties of such networks and their general topological properties can lead to predictive capabilities [13].

## 4 Description of the Method

In this section we introduce the model whose objective is to understand, assess and predict the type and severity of security threats. As discussed above, the dynamical properties of networks can provide a useful insight into the system they model. In this paper, we will focus on the following properties:

– The topology of the network, or in other words, the level of connectedness between nodes measured by joining paths, and
– Their dynamical properties.

Loosely speaking, we are interested in the properties exhibited by the single threats and how they change over a specific amount of time.

As defined in Sect. 3, let $G = G(V, E)$ be a directed network where $V$ is the node set and $E$ is the arc set. The former contains the nodes, and the latter contains the arcs, or directed edges corresponding to requests from the source node to the target node. Let $\deg^t_{in}(v_i)$ and $\deg^t_{out}(v_i)$ be the in and out degrees of the node $v_i$ at a given time $t$, that is the number of connection into and out of it, respectively. We then define the *maliciousness* of a node $v_i$ at the time $t$, as

$$P^t_M(v_i)_{in} = \frac{|\deg^{M,t}_{in}(v_i)|}{|\deg^t_{in}(v_i)|} \tag{2}$$

or

$$P^t_M(v_i)_{out} = \frac{|\deg^{M,t}_{out}(v_i)|}{|\deg^t_{in}(v_i)|}, \tag{3}$$

where $|\deg^{M,t}_{in}(v_i)|$ and $|\deg^{M,t}_{out}(v_i)|$ are the number of malicious connections into or out of $v_i$, at a given time $t$.

For a time $t$ and an arc $e_{v_i,v_j} \in E$, define its *weight* as

$$w_t(v_i, v_j) = f_t(r, p), \tag{4}$$

where $f_t(r, p)$ is a function of the *length of time* of a request $r$ and the *number of request protocols* $p$ from $v_i$ and $v_j$. In this paper, we define

$$f_t(r, p) = \frac{1}{2}(w^t_r + w^t_p), \tag{5}$$

where $w^t_r$ and $w^t_p$ are the length of the time and the number of protocols of different requests, respectively.

We then define the probability of a malicious request at the time $t$ from $v_j$ to $v_i$ as

$$P^t_M(v_i, v_j) = \frac{1}{3}(P^t_M(v_i)_{in} + P^t_M(v_j)_{out} + w_t(v_i, v_j)). \tag{6}$$

In order to consider the dynamics of this model, we assume that new requests arise according to time snapshots $t = 1, \ldots, T$. Let

$$\delta_T(v_i, v_j) = P^t_M(v_i, v_j) - P^{t-1}_M(v_i, v_j) \tag{7}$$

and define

$$\Delta_T(v_i, v_j) = \frac{1}{T-1}\sum_{t=2}^{T} \delta_T(v_i, v_j). \tag{8}$$

Finally, let the *probability of maliciousness* as

$$\tilde{P}_M^T(v_i, v_j) = \min\{\max\{\Delta_T(v_i, v_j), 0\}, 1\}. \tag{9}$$

Note the above equation can be extended to assess the (average) probability of malicious attacks from a set of nodes $\tilde{V}$ on a specific node as

$$\tilde{P}_M^T(v_i) = \frac{1}{|\tilde{V}|} \sum_{\tilde{v} \in \tilde{V}} \tilde{P}_M(v_i, \tilde{v}) \tag{10}$$

for $e_{v_i,\tilde{v}} \in E$ and $\tilde{v}$ is a node in $\tilde{V}$.

Algorithms 1 and 2 show the implementation of the above approach.

**Algorithm 1.** Evaluation of $\tilde{P}_M^T(v_i, v_j)$

```
Let t = 0
Determine P_M^{t=0}(v_i, v_j)
for t = 1, ... T do
    Find Δ_T(v_i, v_j) and P̃_M^T(v_i, v_j)
end for
return P̃_M^T(v_i, v_j)
```

**Algorithm 2.** Evaluation of malicious attacks on node $v_i$

```
Let t = 0
for ṽ ∈ V \ v_i do
    Determine P_M^{t=0}(v_i, ṽ)
    for t = 1, ... T do
        Find Δ_t(v_i, ṽ) and P̃_M^T(v_i, ṽ)
    end for
end for
return P̃_M^T(v_i, ṽ)
```

As discussed in Sect. 3, if the network $G$ follows a scale-free structure, new arcs are likely to be added to highly connected nodes. As a consequence, Eqs. 2 and 3 can be modified to incorporate this property. Recall that the fraction of nodes $p_k$ with degree $k$ is

$$p_k \approx k^{-\gamma}.$$

For a scale-free network $G$, we then assume that

$$P_M^t(v_i, v_j) = \frac{1}{2}(\deg(v_i)^{-\gamma} + w_t(v_i, v_j)). \tag{11}$$

Note that in this case, we are considering the overall degree of the destination node $v_i$, rather than distinguishing between the in and out degree values of the source and destination nodes. Although we are providing fewer parameters in

the model above, compared to Eq. 6, the initial validation appears to support the claim that (11) indeed provides good modelling capabilities.

The dynamics described by Eq. 9 can be used to provide some level of prediction of the number of malicious attacks. In this paper, we assume that the trend of $\tilde{P}_M^T(v_i, v_j)$ can give an insight into a "near future" behaviour of the communications from $v_j$ to $v_i$. In particular, we shall assume that $\tilde{P}_M^T(v_i, v_j) \approx \tilde{P}_M^{T+1}(v_i, v_j)$, or in other words, they exhibit a similar trend. We acknowledge this is a simplistic approach as it does not consider potential variations that could occur. However, our initial validation seems again to support the above. In future research, we are aiming to fully investigate and extend the predictive properties of our approach by fully analysing the topology of a large set of communication networks.

## 5 Results

In this section, we will discuss the validation process, which was based on the publicly available datasets offered by the Malware Capture Facility Project [11]. More specifically, we used the CTU-MALWARE-CAPTURE-BOTNET-42 dataset, which contains relevant data generated by a Neris botnet. It used an HTTP based C&C channel, and all the actions performed by the botnet were communicated via C&C channels containing specific "click-fraud" spam based on advertisement services.
This was subsequently preprocessed via WireShark [10] to capture all the parameters relevant to our approach.

A directed network $G = G(V, E)$ was defined, where the node-set $V$ contains the source and destination IPs mutually linked by a request. I particular, we had

- Number of nodes: 4247
- Number of arcs: 6588
- Average in and out degree: 1.5512

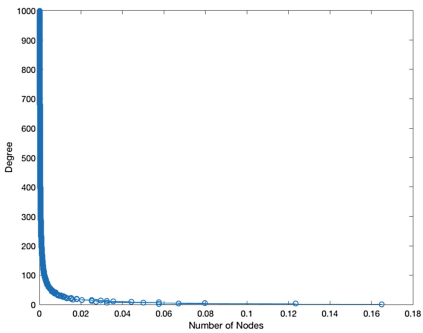



**Fig. 1.** The degree distribution of the network $G$.

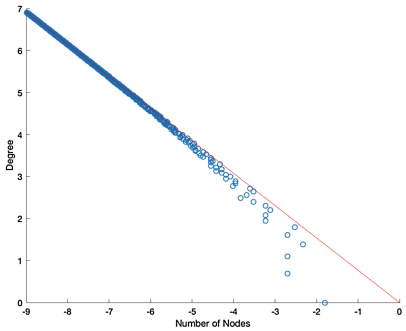



**Fig. 2.** The log node degree distribution of the network $G$ depicted in Fig. 1, which is compared with a (theoretical) scale-free network with $\gamma = 1.9$. As it can be clearly seen, this is a good approximation of the node degree distribution of $G$.

Figures 1 and 2 show the degree distribution of the network $G$, which indicates the existence of few highly connected hubs. Note that this behaviour is similar to scale-free networks, as described by Eq. 1. In [14], a method to topologically reduce complex networks is discussed. When such method is applied to the network $G$, a value of $\gamma = 1.9$ is determined. As discussed in Sect. 3, for many complex systems $\gamma$ is usually within the range $2 < \gamma < 3$, suggesting that the dataset used for the validation exhibits properties similar to many other systems from across various contexts. As discussed above, we are aiming to widen our investigation to a large set of malware botnet datasets to fully assess whether such behaviour can be indeed generalised.

In order to evaluate our approach, we trained the parameters of Eq. 5 on approximately 2000 malicious requests. First of all, we noticed that over 95% of the malicious requests had a TCP protocol, and among them we detected two main clusters for time length values in the interval $[0, 70]$ and $[950, 1400]$, as depicted in Fig. 3.

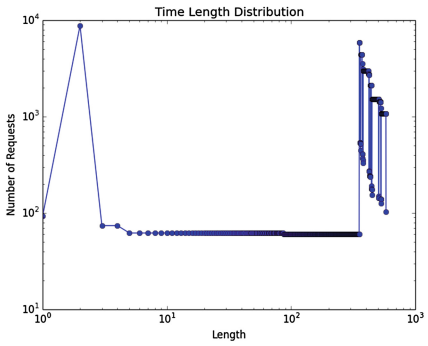



**Fig. 3.** The distribution of the time length requests.

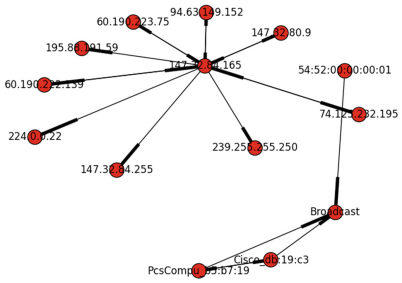
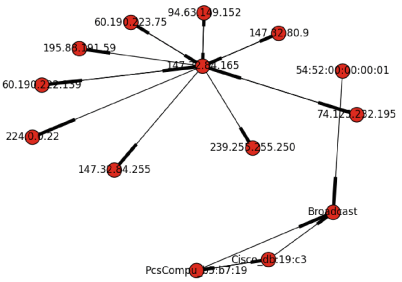
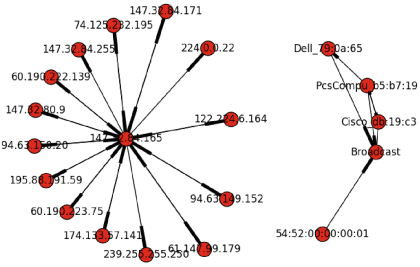



**Fig. 4.** A sub-network generated by the dataset described in Sect. 5.



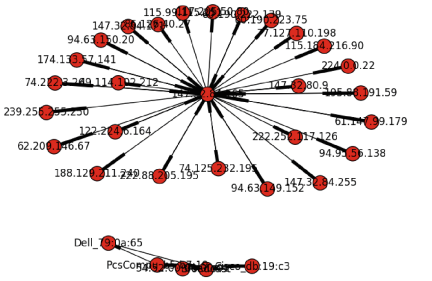




**Fig. 5.** A selection of the sub-networks generated by some time iterations on the dataset described in Sect. 5.

Therefore, we assumed that for malicious requests, $w_t = w_p = 0.6$ if the time length is within those intervals and the protocol is TCP, and $w_t = w_p = 0.2$ otherwise.

We subsequently considered the dynamics of the system, by analysing batches of approximately 300 requests per time iterations, and we assumed that a malicious request from $v_j$ to $v_i$ is associated with $\tilde{P}_M^T(v_i, v_j) > 0.7$. The analysis of the data produced that 71% of the malicious requests had indeed a $\tilde{P}_M(v_i) > 0.7$. Figures 4, 5 and 6 depict a small proportion of the network created in the first three iterations on the process. Furthermore, Fig. 6 also shows the malicious requests, which are depicted in red.

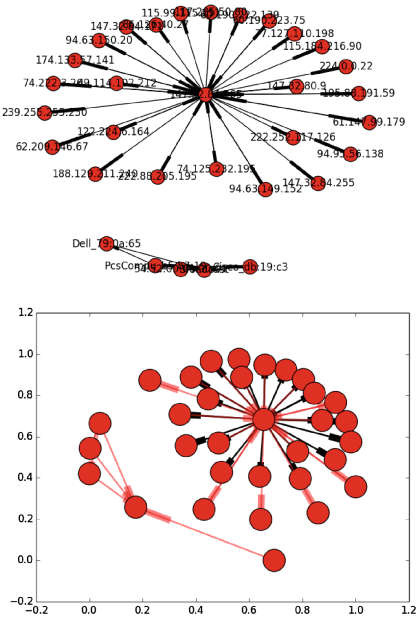



**Fig. 6.** The sub-networks generated by the fourth iteration, where the second figure highlights the malicious connections (Color figure online).

We subsequently evaluated the model defined by Eq. 11. In this case, $\tilde{P}_M(v_i, v_j) > 0.7$ for approximately 61% of the malicious requests. This decrease in accuracy was indeed expected due to the more general scope of the model, as discussed above.

Finally, we evaluated the level of prediction associated with our model, and we considered approximately 200 pairs of nodes exchanging request. Approximately 59% of the malicious requests exhibited the same trend $\tilde{P}_M^T(v_i, v_j) \approx \tilde{P}_M^{T+1}(v_i, v_j)$, and we noted this was particularly the case for larger values of $T$, as expected.

## 6 Conclusion

In this paper, we have discussed a method to assess and predict the malicious connection requests in terms of bonets. As indicated by the validation shows, this approach shows potential in providing a robust method to detect and predict malicious request activity. However, this is still at its infancy and in future research we are aiming to extend our investigation to consider more parameters and create a more comprehensive model. In particular, a full investigation of networks generated by such requests will require a deeper understanding of the topological properties of such networks to ensure a more comprehensive and accurate analysis, which will provide a robust, accurate and computationally effective approach.

## References


1. Wang, W., Daniels, T.E.: A graph based approach toward network forensics analysis. ACM Trans. Inf. Syst. Secur. **12**(1), 1–33 (2008)
2. Liao, N., Tian, S., Wang, T.: Network forensics based on fuzzy logic and expert system. Comput. Commun. **32**(17), 1881–1892 (2009)
3. Francois, J., Wang, S., Bronzi, W., State, R., Engel, T.: BotCloud: detecting botnets using mapreduce. In: IEEE International Workshop on Information Forensics and Security, WIFS, Foz do Iguacu, Brazil, November 2011
4. Abaid, Z., Sarkar, D., Ali Kaafar, M., Jha, S.: The early bird gets the Botnet: a markov chain based early warning system for Botnet attacks. In: 41st Conference on Local Computer Networks (LCN). IEEE (2016)
5. Nagaraja, S., Mittal, P., Hong, C., Caesar, M., Borisov, N.: BotGrep: finding P2P bots with structured graph analysis. In: Proceedings of the 19th USENIX Conference on Security (2010)
6. Stover, S., Dittrich, D., Hernandez, J., Dietrich, S.: Analysis of the storm, nugache trojans: P2P is here. Login **32**(6), 1–8 (2007)
7. Loguinov, D., Kumar, A., Rai, V., Ganesh, S.: Graph-theoretic analysis of structured peer-to-peer systems: routing distances and fault resilience. In: Proceedings of ACM SIGCOMM, August 2003
8. Ye, N., et al.: A markov chain model of temporal behaviour for anomaly detection. In: Proceedings of the 2000 IEEE Systems, Man, and Cybernetics Information Assurance and Security Workshop, West Point, NY, vol. 166, p. 169 (2000)
9. Kidmose, E.: Botnet detection using hidden markov models. Master's thesis. Aalborg University, Denmark (2014)
10. https://www.wireshark.org, (Accessed 10 Feb 2017)
11. Malware Capture Facility Project, http://mcfp.weebly.com/, (Accessed 10 Feb 2017)
12. Palmieri, F.: Percolation-based routing in the internet. J. Syst. Softw. **85**(11), 2559–2573 (2012)
13. Trovati, M., Bessis, N.: An influence assessment method based on co-occurrence for topologically reduced big data sets. Soft Comput. **20**(5), 2021–2030 (2015)
14. Trovati, M.: Reduced topologically real-world networks: a big-data approach. Int. J. Distrib. Syst. Technol. (IJDST) **6**(2), 45–62 (2015)
15. Ebel, H., Mielsch, L.I., Bornholdt, S.: Scale-free topology of e-mail networks. Phys. Rev. E **66**, 035103 (2002)