Malware Propagation Modelling in Peer-to-Peer Networks: A Review

Ahmad Sanda Musa
School of Electrical Engineering and Computer Science
University of Bradford
Bradford, United Kingdom
asmusa1@brad.ac.uk

Hamad Al-Mohannadi
School of Electrical Engineering and Computer Science
University of Bradford
Bradford, United Kingdom
h.i.m.al-mohannadi@brad.ac.uk

Jassim Alhamar
Ministry of Interior
Doha, State of Qatar
j.alhamar@hotmail.com

***Abstract*— Peer-to-Peer (P2P) network is increasingly becoming the most important means of trading content throughout the last years due to the constant evolvement of the cyber world. This popularity made the P2P network susceptible to the spread of malware. The detection of the cause of malware propagation is now critical to the survival of P2P networks. This paper offers a review of the current relevant mathematical propagation models that have been proposed to date to predict the propagation behavior of a malware in a P2P network. We analyzed the models proposed by researchers and experts in the field by evaluating their limitations and a possible alternative for improving the analysis of the expected behavior of a malware spread.**

***Keywords*—** Malware, Peer-to Peer, SIR Model, Propagation

I. INTRODUCTION

For every pro, the con follows. The digital world is reliant on computer networks in sharing and processing data. A malware is a major threat to any network system. Malwares (such as viruses and worms) are a short form for malicious software which if gained access to a system, infects it or commands it to take malicious actions. They work for the purpose of gaining access to a system or infecting it with a program of malicious intent. Malwares infects computers in different type of ways, ranging from downloading unverifiable software, outdated browsers, or simply through a weak operating system with vulnerabilities, etc. The increasing acceptance and use of the World Wide Web resulted in the increase of malware threat. The malware threat gave birth to anti-virus companies that are trying to battle the threat.

Ravichandran and Xavier gives the following definition of peer-to-peer; "Peer-to-peer computing is the sharing of computer resources and services by direct exchange between systems" [1]. Unlike the traditional client and server communication networks, peer-to-peer (P2P) networks entails a set of users, or nodes, that can simultaneously request and provide information to other nodes [2]. In P2P networks, nodes form a loose group of virtual network to communicate and share processing cycles, network capacity, and storage for files.

The P2P model is convenient because of its importance and the ability of nodes to join and leave without affecting the functionality of the system. This ability made the P2P network vulnerable to attacks, as a malicious node can infect an entire network and spread malware posing security threat. As the use of computers and its applications are becoming an integral part of our daily lives, new online services for communication, file and information sharing came to everyday use. These services are of more help to computer users with limited resources. Understanding the P2P architecture and how it works is the first step to getting the better solution of limiting or stopping malware spread.

The broad progress of the computing paradigm leads to many new innovations as well as challenges that needs solving at the earliest. Another vulnerability to make note of is the issue of content verification. The reason why nodes are getting infected with malwares or downloading them is because of blind trust in the P2P network and its participants. Since the network is made up of nodes with equal rights i.e. there will be no way of verifying legitimate requests. Currently, most malware developers have migrated from destructive malwares to malwares that can gain and store information of users, which is more dangerous [3]. In order to find an effective countermeasure against the propagation in P2P, there is a need to review the existing work done on malware propagation modelling on P2P networks. The aim of this paper is to develop an in-depth understanding of P2P malware spread modelling. This review paper is organized as follows: Section II gives an overview of the disease spread and how it is used to study malware spread. While section III discusses the propagation modelling, section IV discusses the models with regards to prior research. Finally, the conclusion in section V states some suggestions and limitations of this work.

II. PROPAGATION MODELLING

This section gives an overview of the current types of malware propagation modelling on P2P networks. Our overview is based on the study of the process of disease transmission that has recently been used in modelling the P2P malware spread.

A. DISEASE SPREAD OVERVIEW

According to the epidemiological model [4] in which diseases find vulnerable hosts to infect, many propagation models for P2P worms have been studied. The two main broad categories are infectious diseases that require direct transferral of body fluids and diseases that infects the environment itself leaving trails of disease which can infect any susceptible person going through the infected site. Using this same model to help understand the behaviours of the P2P malwares, we can categorize the infection of systems into: malwares that get downloaded and executed along legitimate files and malwares that assume the identities of legitimate files. The latter is the most dangerous kind of malware. This can be visualized by imagining an executable file been allowed access into a system in the name of a good file.

III. THE SIR MODEL

Epidemiology is the study and analysis of disease spread in a population. Kermack-McKendrick [4] and his seminal gets the credit for the creation of the mathematical epidemiology. The Epidemiological model researchers' use in the study of the computer malware is the Susceptible-Infected (denoted by SI) and Susceptible-Infected-Removed model (denoted by SIR) [5]. The former model assumes that all peers in a P2P network stay as either susceptible or infected [6] while the latter assumes that a host can recover or be disconnected from the network. The difference between the SI and SIR models depends on whether infected peers can become susceptible again after recovery or be disconnected from the network. If peers can recover or be disconnected from the network, then the SIR model can be used. In a typical P2P network, infected peers can be susceptible again if the host delete the malware by the use of an effective antivirus or disconnect from the P2P network entirely. However, this model does assume that once a host is removed, it will stay in 'removed' state forever. The removed hosts cannot be infected anymore and they do not try to infect others. This model helps by classifying peers as either susceptible, infected or removed/recovered.

Susceptible – Class of peers that are clean but can be infected with malware, usually denoted by S.

Infected – Class of infected peers, now also known as malware. The size of peers are usually denoted by I.

Removed – Class of infected peers that are disconnected from the network are called removed/recovered peers, usually denoted by R.

The number of peers in each of these classes changes with time, that is, S(t), I(t), and R(t) are functions of time t. The size of P2P network N is the sum of the sizes of the three classes in the SIR model [6]:

$$N = S(t) + I(t) + R(t) \quad (1)$$

A. DETERMINISTIC PROPAGATION

Recent studies has proved that the epidemiological models are effective and applicable for the propagation models in vulnerable P2P systems [7]. Deterministic propagation modelling is usually classified into continuous time and discrete time models [8] [9]. Continuous time models are expressed by a set of differential equation while the discrete time models are expressed by a set of different equations [10]. A malware can be a target and directly infect a random susceptible node in the continuous time model. Except the internally generated target lists, all other target discovery techniques employed by active worms satisfy this condition [10] [16]. Deterministic propagation models are usually simpler in analysis for large networks [10]. Some notable examples of deterministic malware propagation models are:

- In the classical simple epidemic model [11] [12] [13] [7], a peer can only be in one of the only two states at any time: susceptible or infected, also known as the SI model. The model assumes that once a peer is infected by a malware, it will stay infected forever. Thus formal transition of any host can only be: susceptible to infected.

- The classical general Kermack-McKendrick epidemic model [8] [4] improves the classical simple epidemic model by considering removal of infected peers due to patching (installing software designed to fix security vulnerabilities).

- Zou et al. [13] presented a propagation model for the Code-Red worm, which is basically similar to the classical simple epidemic model.

- The two-factor worm model [14] extends the classical general epidemic model by considering the removal of infected peers and the susceptible peers getting infected.

- The discrete-time Analytical Active Worm Propagation [14] (AAWP), models the malware spread that employs random scanning and uniform scanning approach. It also measures the time an infected peer takes to infect other peers.

Pros:

- Suitable for the studies of computer malware at the early stage for large network.
- Can be classified as continuous and discrete time in nature.
- More accurate in estimating propagation behaviour than the stochastic model.

Cons:

- Not effective for malware in small networks

B. STOCHASTIC PROPAGATION

The theory of stochastic processes are used to study the stochastic malware propagation modelling. The stochastic modelling is also in discrete-time nature [10]. Two typical examples used in characterizing the stochastic malware modelling are the Density-Dependent Markov jump process modelling [10] [8] [15] and the Gatson-Watson branching process [16].

- The Density-Dependent Markov jump process model for malware propagation uses the uniform scanning approach derived from the famous epidemic modelling [4].

- The Gatson-Watson Markov model proposed by Sellke et al. [16] is a stochastic branching process model for characterizing malware propagation by employing the uniform scanning approach, which particularly focuses on the number of compromised peers against the number of malware scans, then offered a closed form expression for the relationship.

Pros:

- Can be used to characterize the early phase of malware propagation for small network.

- Containment models strategy can be used to avoid propagation [11].

- Uses Markov jump process propagation modelling.

Cons:

- Mostly only discrete time nature.

- Not effective for malwares in large network.

C. LOGISTIC PROPAGATION

Scenario**:** To get an idea of how a malware propagate, we will try to recreate the SIR model from a simple mathematical model that can capture the essential dynamics of some malwares as introduced by current work [17] [2] [7] [12]. Under the assumption of every peer downloads files from the P2P network, any host can contain a malware. Now imagine an infected node getting through into a clean P2P network. It doesn't matter how the one node got infected but keeping in mind that some malwares are designed to propagate and infect any node they come in contact with, the malware can now infect and disrupt the system. The clean nodes are susceptible of getting infected and thereby the network gets infected too as a whole with time. The type of malware plays a role on the rate at which the system gets infected. Malwares that propagates themselves are in the form of malicious executable files shared by P2P peers, they have enticing names of original files or might be infected files. The intent of the Logical model below is to predict the propagation complexity of a malware spread in a P2P network.

The classical simple epidemic model for a finite population can be represented by the differential equation below. Traditionally mathematical epidemic model [4] [12] describes the SIR model using a differential equation to measure the infected peers; let m be the fraction of a malware and $1 - m$ is the fraction of clean files which are susceptible, therefore

$$\frac{\Delta m}{\Delta t} = \alpha m(1 - m) - \beta m, \qquad (2)$$

Where α is the rate at which a peer infects the network and β is the rate at which the other peers in the network flag the malwares and gets them expose. For different types of P2P networks, α and β will vary. The Classical Simple Epidemic Model is the simplest and most popular differential equation model [1].

Let's suppose n is the number of files in a P2P network, and I_n is the number of infected peers in the nth system of the P2P network. Then one can assume

$$I_{n+1} = I_n + I_{fresh}, \qquad (3)$$

Where I_{fresh} is the number of peers that can be infected in the P2P network n. By assuming that any peer can be infected with the malware at random, if each malware starts propagating embedding i peers, and the P2P network contains K peers that are either infected or recovered, then an approximation for the total number of peers which are infected by the malware in the network n is

$$I_n \times {}^{i}/_{K}. \qquad (4)$$

The number of peers that are still susceptible of being infected in the network n, S_n, is the total (T) number of infected peers which will be

$$S_n = T - I_n. \qquad (5)$$

And to get the approximated number of the peers that can be infected in the network

$$I_{fresh} = S_n \times I_n \times {}^{i}/_{K}. \qquad (6)$$

Finally, computing this models together gives us the discrete logistic model for the propagation of malware in P2P networks:

$$I_{n+1} = I_n + I_n \times {}^{i}/_{K} \times (T - I_n). \qquad (7)$$

IV. DISCUSSIONS

Peer-to-Peer malware modelling is vital for understanding the dynamic impact of malware attacks. The modelling provides a comprehensive approach to help researchers study the fundamental propagation patterns that characterize malware propagation. On this basis, P2P networks can predict their potential damages and develop effective countermeasures.

Recently, authors are using the mathematical epidemiology models to study the spread of malware and its propagation in p2p networks [11] [5] [18]. Due to the complexity of the malware spread in the p2p network, researchers are still struggling to develop an airtight solution to its propagation. Rossow et al. [18] models the propagation of virus in p2p networks by developing mathematical equations that examines the expected behaviour of the p2p network.

In [5], different types of p2p malwares has been discussed. About seven different types of malwares were introduced along with their propagation methods and forms. The study was based on the file propagation model of susceptible-infected-recovered (SIR) model which was developed into the susceptible-infected-exposed-removed (SIER) model. The research analysis was based on assumptions and simulations. Such analysis require numerical calculations to be proved.

The modelling consists of understanding the different types of propagation modelling and their characteristics. Deterministic and Stochastic propagation models have emerged to characterize P2P malware spread in continuous and discrete-time nature. The mathematical models [10] [4] has helped researchers to derive models that can eradicate topology-based malwares as well as limit and control the malware spread, previous work [11] [19] [7] [16] presented certain strategies to prevent topology-based malwares from spreading. For example, Sellke et al [16] relied on the branching process model to characterize the propagation of random scanning malwares rather than on mathematical analysis. Their paper demonstrates a fairly comprehensive analysis on the condition that determines whether the malwares die out completely or the total amount of hosts that the malware will infect can be determined.

However, this model describes the containment process of uniform scan malwares instead of modelling the dynamic propagation modelling between each pair of nodes. Therefore, it poorly estimates the propagation of malwares. In addition, the study [17] [9] rely on simulations to model the propagation. They investigate deeply on simulating the best network environment to thwart the propagation of

topology-based malwares. The result of this paper supports the logical propagation modelling to facilitate determining the maximum number of peers in a P2P network that can be infected in a given period. The novel logic reviewed in this paper models the propagation process of a P2P malware by a differential equation of logic. The logical propagation model can be applied to the topology-based malwares as well as a discrete-time deterministic model.

Table I. gives a summary of the different types of the propagation modelling reviewed in this paper to compare the nature and accuracy of the models. The model type SI and SIR in the table represents Susceptible-Infected model and the Susceptible-Infected-Removed model. In this table, the deterministic models, stochastic propagation models and the logistic propagation models proposed to date are shown to differentiate between the models. The deterministic, stochastic and logistic propagation models functionality in this table were derived from prior research [17] [10] [13] [15] [9]. The use of Logistic malware propagation modelling is still developing in the computing paradigm. The approach's ease of use gives it an edge over other propagation models. It is a mathematical model that has not gotten enough attention from researchers to give rise to a better model that would help information security managers make better informed decisions.

TABLE I. Comparison of Propagation Model

Modelling Technique	Complexity	Accuracy	Nature	Model Type
Classical Simple Epidemic	Low	Poor	Continuous-Time	SI
General Epidemic	Low	Poor	Continuous-Time	SIR
Code Red Worm	Low	Poor	Continuous-Time	
Two Factor Worm	Low	Poor	Continuous-time	SIR
Analytical Active Worm Propagation	Medium	Good	Discrete-Time	SIR
Density-Dependent Markov Jump Process	Medium	Good	Discrete-Time	SIR
Gatson-Watson Markov	Medium	Good	Discrete-Time	SIR
Logistic Propagation	High	Good	Discrete-Time	SIR

V. CONCLUSION AND FUTURE WORK

Malwares are proving to be one of the most serious challenges in network security research. The behavior based study is the most current effective technique of malware spread, even though the propagation mechanisms used by malwares have evolved with addition of instant messages and other communication technologies. In order for P2P networks to devise effective defense strategies, the modelling of malware propagation must be used to understand malware spread. This review discusses variety of mathematical models that have been proposed to date for modelling the P2P propagation mechanism. Stochastic and Deterministic research [10] [15] [18] [12] on modelling the propagation of worms can be said to be existent and more progressive than the recent Logical approach. Our reason for trusting the Logical approach is the ease of derivation of the propagation model in general and this review paper. We used the discrete logistic model that shows how infected files can spread in a P2P network by deriving an expression from the SIR model [4].

Since P2P malware are similar to the viruses in the study of epidemiology [4], we used the epidemic model to predict the behavior of malware propagation. However, 'internet based' epidemic models have their limitations. They can only be used on P2P networks that allows peers to share unencrypted files. Encrypted file-share is an approach by the P2P network developers to make the system more secure by allowing peers to share encrypted messages using the Diffie-Helman key exchange between peers as some sort of verification between them [20]. Therefore, our review offers a predictive propagation model study that offers the following suggestive solutions. Firstly, P2P networks security should be revisited and given more attention by the Governments and Research bodies. Secondly, encryption should be introduced to make propagation more difficult or impossible in P2P networks. Lastly, P2P networks should develop a more trusted mechanism for verification of legitimate nodes in a network system.

This paper shows our initial review of malware propagation in P2P networks with debatable assumptions. Future work will be analyzing the logical matrix model to serve as a source model to develop more advanced models from. We plan to explore further the effects of malware propagation on the network performance and the possibility of securing the network from malware invasion and work towards developing a better model that will explain the propagation as well as its vulnerability. Such a model would help predict malware spread in a given protocol and help develop ways to increase P2P network security.

ACKNOWLEDGMENT

This work was supported by the Petroleum Technology Development Fund (PTDF) Nigeria PTDF/ED/PHD/MAS/179/17.

REFERENCES

[1] R. C.G and X. J. Lourdu, "A survey of data sharing and security issues in P2P networks," *Advances in Natural and Applied Sciences*, vol. 11, no. 7, pp. 329-335, 2017.

[2] W. Logan, "A survey of P2P Network Security," *arXiv preprint*, p. arXiv: 1504.01358, 6 April 2015.

[3] K. Hole, "Diversity Reduces the impact of Malware," *IEEE Security & Privacy*, vol. 13, no. 3, pp. 48-54, 2015.

[4] A. McKendrick, "Applications of Mathematics to Medical Problems," *Proc. Edinb. Math. Soc.*, vol. 44, pp. 98-130, 1926.

[5] M. Ebrahim, S. Khan and a. U. B. Khalid, "Security Risk Analysis in Peer 2 Peer System; An Approach towards Surmounting Security Challenges," *ASIAN JOURNAL OF ENGINEERING, SCIENCES & TECHNOLOGY,*, vol. 2, no. 2, pp. 94-101, 2012.

[6] S. Yu and e. al, "Malware Propagation in Large-Scale Networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 27, no. 1, pp. 170-179, 2015.

[7] R. M. M. R. M. Anderson, Infectious Diseases of Humans: Dynamics and Control, Oxford: Oxford University Press, 1991.

[8] X. Fan and Y. Xiang, "Propagation Modeling of Peer-to-Peer Worms," in *Advanced Information Networking and Applications (AINA), 2010 24th IEEE International Conference on*, Perth, WA, 2010.

[9] H. Andersson and T. Britton, Stochastic Epidemic Models and their Statistical Analysis, New York: Springer, 2000.

[10] "Stochastic Behavior of Random Constant Scanning Worms," in *14th International Conference on ICCCN 2005 Proceedings. . IEEE*, 2005.

[11] S. Sellke, H. Ness and B. S. a. S. Bagchi, "Modeling and Automated Containment of Worms," *IEEE Transactions on Dependable and Secure Computing*, vol. 5, no. 2, pp. 71-86, 2008.

[12] S. Eshgni, M. Khouzani, S. Sarkar and S. Venkatesh, "Optimal Patching in Clustered Malware Epidemics," *IEEE/ACM Transactions on Networking* , vol. 24, no. 1, pp. 238-298, 2016.

[13] Y. Wang, S. Wen and Y. X. a. W. Zhou, "Modeling the Propagation of Worms in Networks A Survey," *IEEE COMMUNICATIONS SURVEYS & TUTORIALS*, vol. 16, no. 2, pp. 942-960, 2014.

[14] X. Fan and Y. Xiang, "Propagation modeling of peer-to-peer worms," in *Advanced Information Networking and Applications (AINA)*, Australia, 2010.

[15] K. K. Ramachandran and B. Sikdar, "Dynamics of Malware Spread in Decentralized Peer-to-Peer Networks," *IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING*, vol. 8, no. X, pp. 1-7, 2011.

[16] R.W.Thommes and M. Coates, "Modeling Virus Propagation in Peer-to-Peer," in *In Information, Communications and Signal Processing*, Bangkok, 2005.

[17] T. Zink and M. Waldvogel, "BitTorrent Traffic Obfuscation: A chase towards semantic traffic identification," in *IEEE*, Tarragona, 2012.

[18] C. Rossow and e. al, "SoK: P2PWNED — Modeling and Evaluating the Resilience of Peer-to-Peer Botnets," *IEEE Symposium on Security and Privacy*, pp. 97-111, 2013.

[19] A. A. a. M. G. Andrew Kalafut, "A Study of Malware in Peer-to-Peer Networks," in *In Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, Rio de Janneriro, 2006.

[20] Y. Xiang, X. Fan and W. T. Zhu, "Propagation of Active Worms: A Survey," *International journal of computer systems science & engineering*, vol. 3, no. 24, pp. 1-30, 2009.