Modeling Malware Propagation With Peer-to-Peer Networks

Dr.Gaurav Kumar Jain,
Dr. Akash Saxena,
Mr. Ajay Sharma,
Mr. Indra Kishor

- **INTRODUCTION**

Peer to peer networks provide a paradigm shift from the traditional client server model of most networking applications by allowing all users to act as both clients and servers. The primary use of such networks so far, has been to swap media les within a local network or over the Internet These networks have grown in their popularity in the recent past and the fraction of network trap c originating from these networks has consistently increased .The growing popularity and high penetration of P2P clients such as KaZaa, Gnutella and BitTorrent have provided virus writers with a potent means of compromising hosts on a large scale.

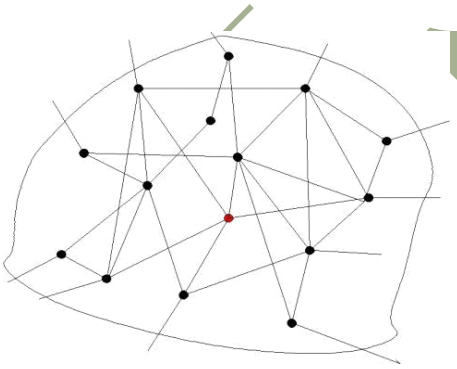

Fig. 1. Infection radius when T T L = 2.

The use of P2P networks as a vehicle to spread malware enjoys some important advantages over worms that spread by scanning for vulnerable hosts. This is primarily due to the methodology employed by the peers to search for content. For instance, in decentralized P2P architectures such as Gnutella, where search is done by coding the network, a peer forwards the query to its immediate neighbors and the process is repeated until a specified threshold T T L is reached. Here T T L is the threshold representing the number of overlay links that a search query travels. A sample scenario of the above description is depicted in Fig. (1), wherein the malicious host, labeled I, can potentially infect all peers that are within a distance of T T L = 2 hops from it. A relevant example here is the M andragore worm [that affected Gnutella users. Having infected a host in the network, the worm cloaks itself for other Gnutella users, leading them also to believe that it is actually an MP3 music le or an image le. Every time a Gnutella user searches for media les in the infected computer, the virus always appears as an answer to the request. The design of the search technique has the following implications: rst, the worms can spread much faster, since they do not have to probe for susceptible hosts and second, the rate of failed connections is less. Thus, rapid proliferation of malware can pose a serious security threat to the functioning of P2P networks.

Understanding the factors affecting the malware spread can help facilitate network designs that are resilient to such attacks, thereby ensuring proper protection of the networking infrastructure. In this

paper, we address this issue and develop an analytic framework for modeling the spread of malware in a peer-to-peer environment while accounting for the architectural, topological and user related factors.

Having motivated our work, we proceed to explore various facets of the problem. The rest of the paper is organized as follows: Section II differentiates the work presented in this paper from existing literature; in Section III, we lay the grounds for our modeling work and present the analytic framework in Section IV. We analyze the model in detail and Section V and present the numerical results validating our theory in Section VI. Finally, Section VII presents the concluding remarks.

- **RELATIONSHIP TO PRIOR WORK**

In this section, we provide a brief overview of modeling literature in P2P networks, not necessarily in the realm of malware spread, and differentiate the current work from ex-isting ones. Though the initial thrust in P2P research was measurement oriented, recent works, have proposed analytical models for the temporal evolution of information in the network. a branching process approximation characterizing the le transfer was presented, a stochastic uid model for BitTorrent-like networks is formulated and the steady state properies of the system are analyzed. A limitation the above works is that they are specialized to Bit-torrent like networks and the framework cannot be extended to analyze P2P networks such as Gnutella or KaZaa. Although, the authors do not model the of ine/online transition, their framework is more representative of a Bit-torrent network than existing ones. Again, the model's applicability is limited and cannot be extended to a Gnutella like network.

The issue of worms in peer-to-peer networks is addressed in wherein the authors perform a simulation study of the dangers posed by P2P worms and proceed to outline possible mitigation mechanisms. Modeling studies addressing malware spread in P2P networks appear wherein the authors formulate a deterministic model having it's basis in the eld of epidemiology. In formulating the equations for the various classes of peers, the authors assume that a vulnerable peer can be infected by any of the infected ones in the network. This assumption is certainly not true since the likely candidates for infected peers are limited to those present T T L hops away from it and not the entire P2P network. Incorporating this detail in the model is imperative since it gurus in the expression for the *basic reproduction number*, a metric that determines the presence/absence of an epidemic. Another important omission is the incorporation of user behavior in the analytic framework. Typically, users in a P2P network, alternate between two states: the on state, where they are connected to other peers and partake in network activities such as query forwarding/response, query initiation etc. and the off state wherein they are disconnected from the network. Peers going of ine act as an inherent deterrent to the density of infected hosts and can help check the rapid proliferation of the virus. This is so because a susceptible peer going of inner implies one lesser candidate choice for infection and an infected peer going of inner naturally lessens the intensity of malware spread.

In the current work, we formulate a comprehensive model for malware spread in Gnutella type P2P networks that ad-dresses the above shortcomings. We develop the model in two stages: rst, we quantify the average number of peers within T T L hops from any give peer and in the second stage incorporate the neighborhood information into the nal model for malware spread. While determining the average number of peers that are within k hops away is not feasible for arbitrary networks, the fact that the degree distribution of peers in Gnutella follows a power law distribution [7], makes the task realizable for such networks. In the next section, we report our simulation result that questions the validity of the bound on the spectral radius of the P2P adjacency matrix that is widely accepted to hold true in the presence/absence of a large scale infection. This nding, further substantiates the need to incorporate the limited view of a peer in a P2P network into the analytic model.

- **VIRUS PROPAGATION IN P2P GRAPHS**

Hypercube have often been chosen as a graph model for P2P networks and the authors derive a limiting condition on the spectral radius of the adjacency

graph, for a virus/worm to be prevalent in the network For instance, in the authors, when deriving the threshold for P2P like graphs, do not consider the fact that once a peer is infected, any susceptible peer within a T T L hop radius becomes a likely candidate for a virus attack.

In order to arrive at the threshold estimate for the virus spread, one need to look at the spectral radius of the *modified adjacency matrix*, M. Speciacally, this is a graph constructed from the original adjacency matrix, wherein an edge exists between two peers as long as there are within T T L hops from each other.

- **P2P MODEL**

In this section, we present our analytic framework for modeling the spread of information, in our case in the form of malware, in peer-to-peer networks. While the framework we develop is robust and is applicable across varied architectures such as Bit-torrent networks, we con ne ourselves to the analysis of Gnutella like networks. We rest describe the search process and the likelihood of le transfer and then present the model for the spread of les based on a compartmental model.

A. Search Mechanism

The transfer of information in a P2P network is initiated with a search request for it. There exist several search mechanisms, popular among which are coding and the random walk. In this section, we derive an expression for the search neighborhood, z_{av} , under the assumption that the search mechanism employed is f looding, as is the case in Gnutella networks. In this scenario, a peer searching for a le forwards a query to all its neighbors. A peer receiving such a request rest responds affirmatively if in possession of the le and then checks the hop count of the query. If this value is greater than zero, it forwards the query outwards to its neighbors, else, the query is discarded.

B. Compartmental Model

We formulate our model for the P2P network as a compartmental model, with the peers divided into compartments, each our formulation is based on the principle of mass action, wherein the behavior of each class is approximated by the mean number in the class at that instant of time. By employing the maenad approach to characterize each compartment, we make the following assumptions about the system:

- **MODEL ANALYSIS**

In this section, we analyze the model presented in the previous section, in totality and specie c illustrative cases, and obtain the necessary conditions for the global stability of the malware free equilibrium.

Malware Free Equilibrium

We now proceed with the derivation of the *basic reproduction number*, R_0, a metric that governs the global stability of the malware free equilibrium (henceforth termed VFE). Here, R_0 quantizes the number of vulnerable peers whose security is compromised by an infected host during its lifetime. It is an established result in epidemiology, that $R_0 < 1$ ensures that the epidemic dies out fast and does not attain an endemic state. Stability information of the VFE is important since this guarantees that the system continues to be malware free even if newly infected peers are introduced.

We follow the methodology presented in, where "next generation matrices" have been proposed to derive the basic reproduction number. In this method, the town of individuals (cell phones in our case) between the states are written in the form of two vectors F and V which describe the in own of new infected individuals and all other own in the system, respectively. These vectors are then differentiated with respect to the state variables, evaluated at the disease (malware) free equilibrium, and only the part corresponding to the infected classes are then kept to form the matrices F and V.

- **RESULTS**

In this section, we demonstrate the essence of our analysis presented thus far through numerical simulations. We rest present our numerical results for the simple SIR epidemic described in the latter

part of the previous section. The reason behind this is that the qualitative behavior of the model in is similar to that presented in Further, since the sampled model has a closed form expression for R_0, it is easy to see its dependence on various model parameters and this relationship can be extended to the more detailed model. The experiments were carried out using parameters emulating a 20000 node. The initial number of infective was set at 50. We see that R_0 is directly proportional to ¸on. The essence of this equation is that, nodes staying on-line for long periods as compared to their off-line durations result in a higher intensity of malware presence in the network. Numerical simulations concurred with the above observation. A similar trend was observed for the detailed model .The curve at the bottom corresponds to ¸t_{on} = 0:1 and the intensity of the epidemic increases monotonically with an increase in

- **CONCLUSION**

In the current work, we motivated the need to understand the dynamics of malware spread, especially in the context of inter-acting heterogeneous environments such as peer-to-peer net-works. The need for an analytic framework incorporating user characteristics (e.g. off-line to on-line transitional behavior) and communication patterns (e.g. the average neighborhood size) was put forth by quantifying there influence on the basic reproduction ratio. It was proved analytically that a model that does not incorporate the above features runs the risk of grossly overestimating R_0 and thereby falsely reporting the presence of an epidemic. Further, our simulations show that the bound on the spectral radius for the spread of malware needs to take into account, the underlying communication pattern, especially in a P2P kind of setting so as arrive at an accurate estimate. The model was also extended to characterize the dynamics of malware spread in networks of smart cell phones.

- **REFERENCES**

[1] O. Diekmann, J. A. P. Heeterbeek, "Mathematical Epidemiology of Infectious Diseases: Model Builind, Analysis and 1999 Interpretation," Wiley,

[2] Celery Search Engine, http://site.n.ml.org/info/ celery/

[3] "Napster Protocol Specification," March 12 2001, http://opennap.sourceforge.net/napster.txt

[4] Characterization of Internet traf c loads, segregated by application, byapplicat http://www.caida.org/analysis/workload/ ion/ .

[5] Kazaa. http://www.kazaa.com

[6] Clip2, "The Gnutella Protocol Speci cation v0.4," March 2001, http://www.clip2.com/GnutellaProtocol04.pdf. [7]Clip2 Company, Gnutella. http://www.clip2.com/gnutella.html

[8]B. Cohen, "Incentives Build Trust in BitTorrent," May 2003, http: //bitconjurer.org/BitTorrent/bittorrentecon.pdf

[9] P. van den Driessche and J. Watmough,"Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission," *Mathematical Biosciences*, vol. 180, pp. 29-48, 2002.

[10] J. Arnio, J. Davis, D. Hartley, R. Jordan, J. Miller and P. van den Driessche, "A multi-species epidemic model with spatial dynamics," *Mathematical Medicine and Biology*, March 2005.

[11] L. Zhou, L. Zhang, F. McSherry, N. Immorlica, M.Costa and S. Chien, "A First Look at Peer-to-Peer 2005.

[12] A. J. Ganesh, L. Massoulie and D. Towsley, "The Effect of Network Topology on the Spread of Epidemics," *Proceedings of IEEE INFOCOM*, Miami, USA, March 2005

[13]Y. Wang, D. Chakrabarti, C. Wang and C. Faloutsos, "Epidemic spread-ing in real networks: An eigenvalue viewpoint," *SRDS 2003*, (pages 25- 34), Florence, Italy

[14]M.E.J Newman, S.H. Strogatz, and D.J. Watts, "Random graphs with arbitrary degree distribution and their applications," *Physical Review E*, vol. 64, no. 026118, 2001

[15] D. Qiu and R. Srikant, "Modeling and performance analysis of BitTorrent-like peer-to-peer networks," *Proceedings of ACM SIGCOMM*, Portland, OR, August 2004.

[16] X. Yang and G. de Veciana, "Service capacity in peer-to-peer networks," *Proceedings of IEEE INFOCOM*, pp. 1-11, Hong Kong, China, March 2004.

[17] M. Costa, J. Crowcroft, M. Castro and A. Rowstron, "Can we contain Internet worms ?," *HotNets-III: Third Workshop on Hot Topics in Net-works*, San Diego, USA, 2003

[18]J. Mundinger and R. R. Weber, "Ef cient File Dissemination using Peer-to-Peer Technology," *Technical Report, Statistical Laboratory Research Reports 2004-01*, 2004.

[19]http://www.infoworld.com/articles/hn/xml/01/02/27/010227hnp2pvirus.html?0227alert

[20]R.W. Thommes and M.J. Coates, "Epidemiological Models of Peer-to-Peer Viruses and Pollution," *Technical Report, Department of Electrical and Computer Engineering, McGill University*, June, 2005.

[21] R.W. Thommes and M.J. Coates, "Modeling Virus Propagation in Peer-to-Peer Networks," *Technical Report, Department of Electrical and Computer Engineering, McGill University*, June, 2005.