Monitoring Social Centrality for Peer-To-Peer Network Protection

Miray Kas, L. Richard Carley, and Kathleen M. Carley, Carnegie Mellon University

ABSTRACT

Resilience of networking infrastructures is crucial for ensuring protection and readiness in the case of an emergency. Enabling Peer-to-Peer (P2P) communication is one way of alleviating potential outcomes of depending on a centralized system with a single point of failure, which would be hard to recover in the case of cyber-attack. In this article, we consider P2P-enabled networks and discuss how it is possible to benefit from social centrality metrics to prevent catastrophic spreading of malicious content. Using classical Susceptible-Infected- Recovered (SIR) simulation models designed for disease propagation in a social community on two different real-life P2P network topologies, we demonstrate that the nodes that are ranked highly by social centrality metrics should be better protected to prevent malicious content propagation. Our results suggest that nodes that are central are significantly more effective than random nodes in spreading malicious content across the network and among different centrality metrics, Eigenvector centrality is more useful for identifying the nodes that should be better protected in P2P networks.

INTRODUCTION

A Peer-to-Peer (P2P) network refers to a computer network where each node can act as a client or server for other computers in the network, providing a decentralized, self-organizing, and self-healing networking infrastructure [1]. To date, P2P technology has found itself several applications in business including

- Collaboration among geographically distributed teams
- Edge services by network caching (e.g. moving the data closer to the point where it is actually consumed)
- Distributed computing where P2P is used to share idle CPU power and storage space, enabling large-scale computation [2]

Recently, with the advances in networking technology and the need for sharing of resources to tackle large-scale computing problems, networks are becoming increasingly interconnected while facing an increasing number of cyber-attacks. Due to the larger scale of today's networks, a cyber-attack targeting the central coordinator for a network now has an even bigger impact if it is successful. To increase readiness and resilience against such scenarios, enabling P2P overlay structure is of significant benefit.

Enabling P2P communication is beneficial for a number of reasons. First, P2P networks provide important advantages as they can be implemented using a diverse collection of hardware and software, making their deployment relatively inexpensive. Second, P2P networks allow communication with peripheral nodes that would otherwise not be possible in the case of central server disfunction. Third, P2P communication provides an agile structure that is quite good at coping with dynamic, heterogeneous topologies, which in turn makes them stand as a good option for ensuring resilience, responsiveness, and readiness of a network in emergency cases, making the network highly robust [3].

On the flip side of the coin, enabling P2P networking makes networks more prone to various cyber-attacks including generic network attacks (e.g. identity tracking/theft for Sybil attacks, spamming, denial of services) and more specific attacks related to file transmissions such as

- Poisoning (disseminating files whose contents do not match their description)
- Polluting (inserting bad packets into files)
- Defection (free-riding; using service without contributing to the system)
- Malware spread (originally attached to the P2P software or files)
- insertion of viruses (attached to other files) [4]

This potential tradeoff raises the need for careful assessment of P2P topologies to ensure maximum benefit and protection. Network topologies have been assessed in various disparate research fields including graph theory, computer networking, and social network analysis. Social network analysis, which is the main methodology we use in this article, is a methodology used in many sciences to model populations and organizations as networks of actors. In a social network, nodes represent social actors (e.g., humans, organizations, computers, or other agents) while edges represent the relationships among these social actors.

Within the field of social network analysis, a number of well-established metrics have been developed to characterize network topologies, assess prominence/importance of nodes in social networks, identify groups or communities, and

0163-6804/13/$25.00 © 2013 IEEE

forecast information flow or performance. Centrality metrics are node level metrics that identify those nodes that stand out because they are more central or connected in some way to other nodes. Although there exist hundreds of metrics, we focus on four main centrality metrics used in many publications in multiple fields: degree centrality, Eigenvector centrality, betweenness centrality, and closeness centrality.

In this article, we discuss how social centrality metrics can be used to increase the security of P2P enabled networks by identifying the nodes could efficiently spreading malicious content across the network. In other words, we propose using social centrality metrics for vulnerability and risk assessment in P2P networks. The resemblance between agents in social networks and P2P networks has long been discussed, especially in the context of making P2P networks more resilient and selforganizing [5].

To illustrate how we can benefit from social centrality metrics in monitoring P2P network security, we use the SIR (Susceptible-Infected-Recovered) class of simulation models, which are primarily used by sociologists and epidemiologists for modeling disease spread among agents in a social community. More precisely, we draw an analogy between diseases in the public health community and infection in the cyber world. We perform SIR simulations to observe the spread of malicious content around the network when the socially central nodes are infected by malicious content they do not have any prior knowledge and resistance.

The rest of the article is organized as follows. We briefly review the four major social centrality metrics. We describe SIR disease propagation simulation models and mentions example studies that have adapted them to the cyber security field. We describe the real life P2P datasets we use, and report our simulation results. We discuss potential future research, and finally conclude the article highlighting our key findings.

Nodes that have high betweenness centrality are usually the nodes that are gateways or that connect different clusters in a clustered network topology, or ones whose removal may partition the network.

Background

In the social network analysis literature, centrality metrics can be loosely classified into two groups: metrics that are based on the number of connections nodes have (e.g. degree) and metrics that are based on the shortest paths in the network. This section briefly reviews four commonly used centrality metrics. Among these four metrics, degree centrality and Eigenvector centrality are degree based while closeness centrality and betweenness centrality are shortest path based.

Degree Centrality

Degree centrality measures the number of immediate connections a node has (adjacent nodes), and it has a number of variations. In-degree centrality refers to the number of edges directed towards a node and it is usually associated with the prestige of the node in a social network. Out-degree centrality measures the number of links emanating from a node, which is usually interpreted as involvedness in social activity. Total degree centrality considers all immediate connections regardless of edge direction.

Eigenvector Centrality

Eigenvector centrality has a degree-based, recursive definition where a node is highly ranked if it is connected to highly ranked nodes. Eigenvector centrality is simply the dominant Eigen vector of the matrix. Philip Bonacich proposed Eigenvector centrality in 1987 [6] and Google's PageRank algorithm is a variant of it. It measures the influence of a node in the network and is useful for identifying cluster heads, and social agents that can mobilize others.

Closeness Centrality

In a given network, the distance between any two nodes is usually characterized in terms of their shortest-path distance. The mathematical definition for closeness centrality of node x, $C_c(x)$, is shown in Eq. 1 where $d(x, y)$ denotes the shortest distance from node x to y.

$$C_c(x) = \frac{1}{\sum_{x \neq y} d(x,y)} \tag{1}$$

A node has closeness values if its total distance to all other nodes in the network is low. Hence, closeness can be regarded as a metric to measure how long it will take for a node to spread a piece of information to spread in the network. H owever, closeness can be slightly misleading if not every node is reachable from every node. For instance, if a node y is not reachable from node x (i.e., $d(x, y)$ is infinity) then it is not taken into account in Eq. 1.

Betweenness Centrality

Betweenness centrality measures the fraction of the shortest paths a node is on when the shortest paths across all node pairs are considered [7]. Nodes that have high betweenness centrality are usually the nodes that are gateways or that connect different clusters in a clustered network topology, or ones whose removal may partition the network.

$$C_B(z) = \sum_{x \neq y \neq z} \frac{\sigma_{xy}(z)}{\sigma_{xy}} \tag{2}$$

Equation 2 presents the formulation of betweenness centrality where σ_{xy} denotes the number of shortest paths from x to y and $\sigma_{xy}(z)$ denotes the number of shortest paths from x to y passing through z.

Interpretations

The centrality metrics can be interpreted in different ways to answer P2P communication specific questions. Depending on the research question at hand, one of them can be more important than the others. Example uses include the following:

- *Degree Centrality*: How many different sources did this user download files from? How many unique users did this user serve files to?
- *Eigenvector Centrality*: Which server is surrounded by the other, most popular servers?
- *Closeness Centrality*: How fast a new piece of software will spread from this user to the other users in the network?

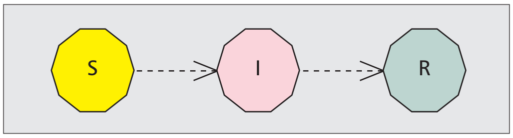

Figure 1. *State flow transition for S (Susceptible) — I (Infected) — R (Recovered) simulation model.*

- *Betweenness Centrality*: If you can sniff packets, through which user confidential information is most likely to flow?

The topic of this article, attempting to identify and protect the nodes that would be the most central in the case of a malicious spread, is not the first use of social centrality metrics in P2P networks. However, we believe that the work presented in this article is unique in the manner in which it employs social centrality metrics and in the scale and realworld nature of the P2P datasets to which we apply them.

Simulating for Malicious Content Spread

Susceptible-Infected-Recovered Simulation Models

One approach to simulation of disease spread is through using S (Susceptible) — I (Infected) — R (Recovered) models. Numerous variations of SIR model have been applied to study the spread of epidemics and immunization strategies for a population.

In SIR models, each social agent in the modelled population is in one of the three states: susceptible (S), infected (I) and recovered (R), typically progressing from susceptible to infectious to recovered as in Fig. 1.

To make simulations of SIR models tractable, the population is generally assumed to consist of a homogeneous mix of individuals that have predefined transition probabilities of moving from one state to another. SIR models were originally designed for public health communities as a way of displaying historic data. However, recently they have been adapted to the fields of cyber security and online social network data mining where they have been used to model the propagation of email worms, botnets, mutating P2P malware, increasing immunization by security patches, and how anti-virus companies respond to worm outbreaks within hours or propagation of information in online social networks such as Flickr [8].

Our Simulation Model

In this article, we use a simulation model that is, at its root, an SIR-based disease propagation model, implemented in ORA (http://www.casos.cs.cmu.edu/projects/ora/). An initial network of social agents is provided, on which the dissemination of a single "disease" (or malicious content) will be simulated. In the first round of the simulation a number of source (seed) nodes are selected and deliberately "infected".

In our case, the nodes represent computers and so the network represents the P2P network. The "disease" is the bad information, malware, or virus which we describe as "malicious content" to be generic. The nodes transition from "susceptible" into "infected" state when they receive the malicious content. Our evaluations are designed to examine how the infection of socially central nodes impacts the dissemination of the malicious content.

In our simulation model, we define the transmission events as the key events. At every transmission phase, a node checks all of its outgoing links to see if a transmission can happen down any one of its links based on the link weight. If the input network is a weighted network, all the link weights are normalized to [0..1] range. The normalized weight of a link is then used as the probability of transmission down that link. If the check on the link weight is positive, then another check is done to see if the transmission resistance of the receiving node can be beaten. Transmission resistance of nodes is another simulation parameter in the [0..1] range, set by the user. When transmission resistance is zero, the receiving node shows no resistance; accepting whatever is transmitted unless it has already been infected and reached the recovery state before.

During the simulations, each node can be in one of three states — S, I, R. A node is susceptible (S) if it has not yet received the malicious content. A node is infected (I) if it has received the malicious content. And a node is recovered (R) if it has been patched or has removed the malicious content. We assume that if the node is recovered (R) then it cannot re-get the malicious content nor can it give it out. The node can give out the malicious content to multiple other nodes while it is in the infected state. We assume that nodes recover, i.e., the infected computer is patched or removed after one time period, which we use to model one-hour time frame in our main simulation results. The number of time periods the node is infected for can be varied.

Peer-to-Peer Benchmarks

In this article, we use topologies from two real life P2P networks. We prefer using real life network topologies over stylized networks because stylized networks have been argued to model unrealistic assumptions about P2P topologies, potentially resulting in misleading conclusions [9].

Proximity Can-O-Sleep Dataset

The Privacy, Internetworking, Security, and Mobile Systems Laboratory at the University of Massachusetts Amherst collected the Proximity can-o-sleep dataset on a server running OpenNap file sharing system with additional preparation by the Knowledge Discovery Laboratory, University of Massachusetts Amherst [10].

The dataset covers ownership and transfer details of mp3 files across more than 6,000 uniquely identified users active between February 28, 2003 and May 21, 2003. Using file transfer information as the links connecting these users, we have constructed their social network graph. This social network contains 6,464 users, with 221,152 file transfers among these users,

SIR models were originally designed for public health communities as a way of displaying historic data. However, recently they have been adapted to the fields of cyber security and online social network data mining

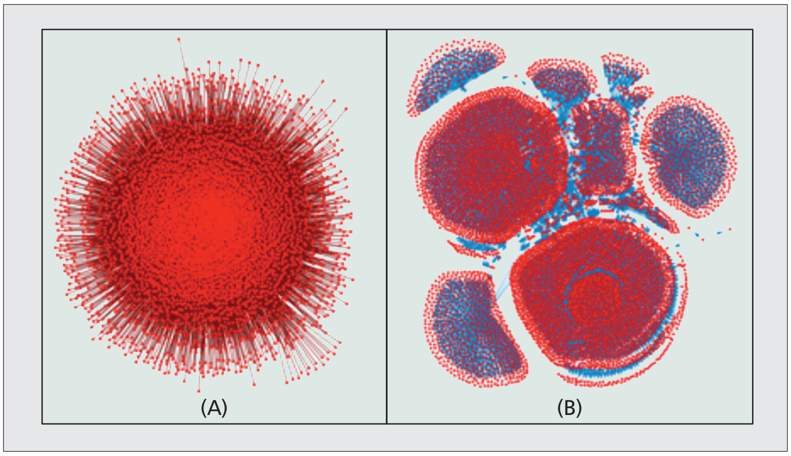

Figure 2. *Network topologies for proximity and GT datasets: a) proximity dataset; b) GT dataset.*

resulting in 98,680 links. The majority of the users join the system to receive the file/information they need while not contributing for the rest of the users. Out of 6,464 users, only 769 unique users act as sources for file transfers.

The overall topology of this dataset contains two disconnected clusters of users. When the smaller cluster is removed from the network, there are 4445 nodes, 89,725 links remaining. Out of 4445 nodes, only 504 of them act as sources for other users. In the rest of the article, we use the topology for only the large component, as transmissions cannot occur between components. Since the larger component covers 92.4% of the transmissions, most of the actual data is retained. The second column on Table 1 outlines its network-level statistics.

GT Dataset

The GT dataset was collected on two campus networks: the University of Brescia (UNIBS) and the Politecnico di Torino (POLITO). The traces contain the Internet usage data from informed participants who agreed to use the Internet as they usually do during the experiment [11]. Out of this dataset, we focus only on the P2P traffic, generated by several different P2P applications including amule, bittorrent, edonkey, instant messaging, and Skype. We derive P2P transmission data and model it as a network where nodes represent the users and links represent the communication/transmission between those two users. The third column on Table 1 outlines its network-level statistics.

Topological Evaluation

Next, in Table 1, we evaluate the topologies of both datasets using a number of metrics. Most real life P2P networks are known to suffer from the 'free-riding' phenomenon where a lot of users join the network to download files from others and provide nothing in return. We observe that both of the examined networks' topologies are governed by this phenomenon.

In both datasets, the maximum total degree values are around 2000. For instance, the user with the highest number of connections in the GT dataset serves 2185 out of 6843 users. Hence, there are very few nodes that serve most other users, while the majority of the users make a couple of file downloads which drags the average number of connections for a node to 2.21 despite very large number of connections (e.g. 2185) for some users. Similar reasoning holds for the Proximity dataset as well.

Network density is defined as the ratio of the number of edges in a network (m) over the total number of possible edges across all nodes in a network (n (n – 1)): m /(n (n – 1)). Both networks are very sparse, while the Proximity dataset is slightly denser. In general, density is inversely correlated with the number of nodes in a network.

Reciprocity refers to the symmetry of relationships. If there is an edge from node x to y, and an edge from node x to y also exists, then the relationship between x and y is called reciprocal (symmetric). This is because the majority of the users join the network to download files and do not reciprocate or contribute to the system. Hence, in both topologies, the overall reciprocity is very small and most relationships are one way as indicated by the reciprocity values.

The characteristic path length represents the average length of the shortest paths while the diameter represents the maximum of the shortest paths between any two nodes. The values in Table 1 represent the average and maximum distances in terms of number of hops. Despite the number of nodes in both networks, the numbers of hops the shortest paths contain are very low. This is especially true for GT dataset, which has an effect on the behavior of betweenness value as explained later.

Virtual Experiments and Simulation Results

Virtual Experiment Design

This section describes the setup for our virtual experiments and presents controlled, independent, and dependent parameters as outlined in Table 2. Our virtual experiments attempt to model a new infection, against which the nodes do not have pre-existing resistance. Since we are concerned about protection of networking infrastructures, even a single day is very important. Hence, we simulate for 24 time steps and assume that reaction to an infection will take place very quickly, within an hour of detection (e.g. a single time step).

In our SIR simulations, the number of interactions that took place over down a link in the P2P benchmarks is converted to the probability of transmission for that link. The assumption here is that the history of previous file transmissions indicates likelihood of future transmissions. In addition, since malicious content (disease) is transmitted probabilistically, we replicate each simulation 20 times to identify the average response. Many nodes in the P2P networks only download files from other nodes and do not contribute to the system. If such a node is selected as a source node then no matter how many times we simulate, there will be no transmissions/spreading as the node has no outgoing links. To avoid this condition, we use multiple sources, and for each replication, a new set of source nodes is chosen randomly. This removes pathologies due to odd starting conditions. In the experiments run with the other selection criteria, the corresponding centrality values are computed for each node in the network and sorted in descending order. Top-1, top-25, top-50, top-75, and top-100 nodes are selected, respectively.

Simulation Results on Proximity Can-O-Sleep Dataset

First, we report our simulation results on Proximity dataset. Figure 3 presents our simulation results obtained using different criteria for selecting the source nodes (i.e., the nodes that are initially infected).

According to the results in Fig. 3, there is a significant difference in the number of nodes socially central nodes can affect when compared to the randomly selected nodes. Overall, Eigenvector centrality is the most effective centrality metric. The randomly selected source nodes show high variation because some of the randomly selected nodes happened to be one of the central nodes while most other nodes have small impact areas.

Simulation Results on GT Dataset

Next, we report our simulation results obtained on the GT dataset (Fig. 4). The simulation results in Fig. 4 again suggest that socially central nodes are significantly more effective in spreading malicious content than randomly selected nodes.

In Fig. 4, while all other centrality metrics act similar to one another, betweenness centrality has a slightly different behavior, especially at lower number of source nodes. Considering the average number of hops in the shortest paths in the network is 1.24, not many nodes can lie on the shortest path between other nodes. In addition, in such networks, especially in clustered topologies similar to that of GT dataset as shown in Fig. 2b, betweenness centrality might identify funneling nodes that connect two components and may not necessarily have too many immediate connections. Such nodes turn out to be on a lot of the shortest paths to/from relatively isolated the components they are bridging to the rest of the network although the number of nodes that they can spread content to is relatively limited. However, similar to Fig. 3, closeness centrality, out degree centrality and Eigenvector centrality are again quite effective indicators of infection spread. This is observed due to two reasons. First, the nodes with high number of immediate connections cover a substantial portion of the network. Second, the range of link weights stem from a smaller range of values, resulting in stronger transmission probabilities on average. Comparing information presented in Table 1, it can be observed that GT network is substantially sparser than Proximity dataset. Hence, the total number of affected nodes is lower in the GT network when the absolute values on the y-axes of Fig. 3 and Fig. 4 are considered.

Common Results

We note that which centrality metric is most effective is to an extent a function of the topology of the P2P network. On the two networks examined, random nodes are far less effective at propagating malicious content than highly central nodes. In the simulated topologies, and in most real life P2P networks, a lot of users join the network to download files from other users while not providing any files in return, which is known as the 'free-riding' phenomenon. Since the number of users acting as servers for others is considerably lower, nodes selected at random are more likely to be receiving only than are nodes that are high on a centrality metric. Thus highly central nodes, even when there are fewer of them, are more capable of propagating malicious content. From a sensor-prevention perspective, fewer sensors on highly central nodes will be more effective than more sensors on random nodes.

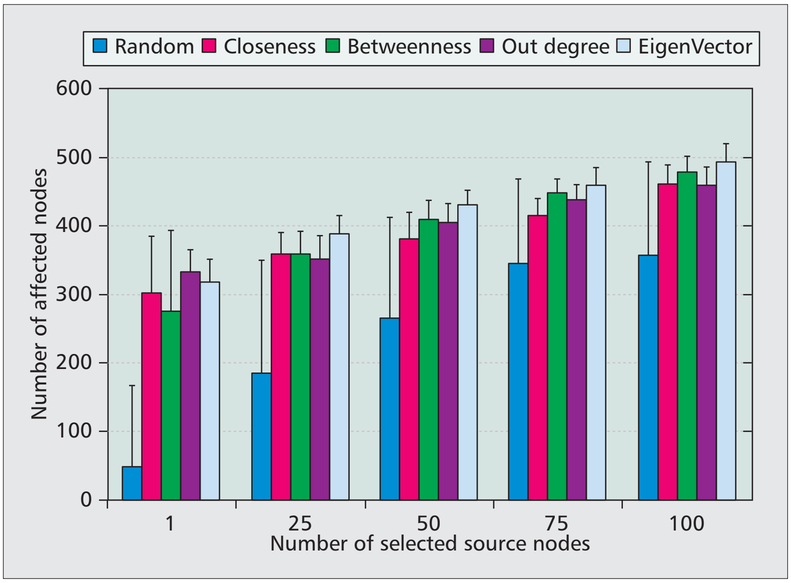

Figure 3. *Simulation results showing the number of affected nodes versus the number of initially selected source nodes, using different selection criteria (Proximity Can-O-Sleep Dataset).*

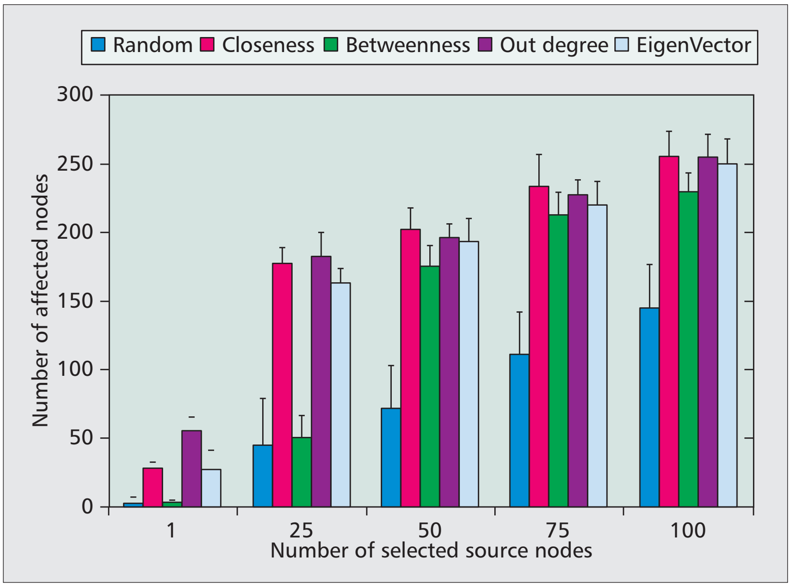

Figure 4. *Simulation results showing the number of affected nodes versus the number of initially selected source nodes, using different selection criteria (GT Dataset).*

Parameter	Proximity Dataset	GT Dataset
Number of nodes	4,445	6,843
Aggregated link count	89,725	7,572
Number of file transmissions	204,344	26,325
Link weights (Number of transmissions over a link)	Min: 1 Max: 413 Avg: 2.28 Stddev: 4.39	Min: 1 Max: 176 Avg: 3.46 Stddev: 5.89
Total Degree	Min: 1 Max: 1733 Avg: 40.16 Stddev: 100.66	Min: 1 Max: 2185 Avg: 2.21 Stddev: 38.33
Network density	0.0045194	0.0001617
Reciprocity	0.003	0.001
Characteristic Path Length	2.74	1.2481
Diameter	8	3

Table 1. *Topological properties of the main component in Proximity dataset and GT Dataset.*

On average, Eigenvector centrality is a better indicator of effectiveness. Eigenvector centrality is more effective at spreading malicious content than the other centrality metrics due to the echo chamber effect; i.e., nodes that are high in Eigenvector centrality are linked to other nodes that are also highly central allowing for the malicious content to move rapidly within the group and beyond. In a P2P network, there are a few nodes that act as popular hosts and have hundreds of out-going connections. Hence, in P2P topologies, the nodes ranked high by Eigenvector centrality are often popular source nodes surrounded by other popular source nodes, resulting in a larger spread in the network.

Closeness centrality, which is primarily known for measuring a node's effectiveness at information spread, tends to be less effective than Eigenvector centrality in P2P network topologies. Closeness is effective due to the reachability effect and nodes high in closeness have long communication tendrils that support information from them. Closeness centrality might outperform Eigenvector centrality when there are longer communication paths in the network.

As one final note, Eigenvector centrality, closeness centrality, and out-degree centrality produce results that are very close to one another. This is because both dataset are under the heavy influence of the "free-riding" phenomenon. Hence, there are very few nodes that serve almost everybody else in the network. Such nodes have high out-degrees because they provide immediate service/connection to a lot of users. Similarly, such nodes are the powerful nodes that have connections to other powerful nodes because users with large archives are usually tempted to broaden their archives. Hence they have high Eigenvector centralities as well. Since they are connected to most of the other nodes directly, they have relatively short paths to all other nodes in the network, which gives them relatively high closeness centralities as well. In cases where there is less of free-riding and more of Skype or instant messaging like communication in the network, Eigenvector centrality would remain more powerful as it is good at identifying powerful users that have connections to other powerful users in the network.

Future Search Directions

As presented in this article, social centrality metrics can be extended to or used in several future studies to provide better understanding of P2P network topologies and protection.

Learning Heuristics

One potential research direction is to design fast learning heuristics based on centrality metrics used in social network analysis using different goal functions such as preventing attacks, altering (e.g. reducing) cascade probabilities, increasing resilience, and maintaining network throughput during a cyber-attack.

Heterogeneous Topologies

This article focuses on P2Penabled networking infrastructures. However, there might be various combinations of inter-linked networks used jointly. One promising future research direction would be simulating for the malicious content propagation across heterogeneous networks where each node in the high-level network is a hyper node of some smaller network, connected to other networks through a backbone link.

Herein, our goal is to demonstrate the importance of socially central nodes for the protection of P2P-enabled networks. However, the SIR simulation model has other cyber applications. For example, transmission resistance might be used to model the effectiveness of anti-virus software or firewalls while the nodes that are immune initially might be used to model the nodes that have required patches to protect themselves. Hence, one interesting direction is to evaluate the implications and dynamics between the transmission resistance and the infection outcomes. Similarly, the time that nodes remain contagious can be used to represent the responsiveness of the systems, denoting how long it takes to release an update for a new threat and to make the required patches. More complicated virtual experiments can be designed to handle the simulation of benign and malicious traffic together or to incorporate the simulation of user behavior modeling with a distribution of users who fix or do not fix problems quickly when prompted by their computer. The proposed experimental framework is thus a valuable testbed allowing numerous other scenarios to be considered.

Conclusions

Enabling P2P networking is a way of creating decentralized communication, which reduces dependence on functionality of a centralized server and eliminates reliance on single point of failure while increasing resilience, responsiveness, and readiness of the network in emergency cases. In this article, our goal is to draw attention to potential security issues with P2P enabled networks and their protection by utilizing social centrality metrics. We propose close monitoring/guarding of highly central nodes in the network to ensure better protection. To assess centrality of nodes, we use metrics developed in the field of social network analysis. Our simulation results obtained on real life P2P network topologies suggest that nodes that are identified by social centrality metrics are more effective at spreading malicious content throughout the network, while Eigenvector centrality has higher spreading abilities.

Control Variables	#(Test Cases)	Values Used
Transmission resistance	1	0.0
Number of periods node remain contagious	1	1
Number of periods to run	1	24
Percentage of nodes that are immune at start	1	0.0
Independent Variables	**#(Test Cases)**	**Values Used**
Number of selected sources	5	1, 25, 50, 75, 100
Source nodes selection criteria	5	Out Degree Centrality, Eigenvector Centrality, Closeness Centrality, Betweenness Centrality, Random
Dependent Variables	**#(Test Cases)**	
Number of nodes that received disease	—	
This is a 5 × 5 design		
Simulations are run 20 times for each test case, resulting in 500 simulations per P2P benchmark, and 1000 simulations in total.		

Table 2. *Virtual experiment design and SIR simulation parameters.*

References

[1] A. Oram, *Peer-to-peer: Harnessing the Power of Disruptive Technologies*, Sebastopol, CA, USA: O'Reilly & Associates, Inc., 2001.

[2] D., Garcia, A., Kramer, and B. DeFigueiredo, "Analysis of Peer-to-Peer Network Security Using Gnutella," Berkeley, CA, 2002.

[3] A. T. Stephanos and S. Diomidis, "A Survey of Peer-topeer Content Distribution Technologies," *ACM Computing Surveys*, vol. 36, no. 4, 2004, pp. 335–71.

[4] I. Livenson, "Security Aspects of P2P Networks," University of Tartu, Estonia, TechReport 2006, http://courses.cs.ut.ee/2006/crypto-seminarfall/files/livenson1.pdf.

[5] R. A. Ghanea-Hercock, F. Wang, and Y. Sun, "Self-Organizing and Adaptive Peer-to-Peer Network," *IEEE Trans. SMCB: Cybernetics*, vol. 36, no. 6, 2006, pp. 1230–36.

[6] P. Bonacich, "Power and Centrality: A Family of Measures," *American J. Sociology*, vol. 92, no. 5, 1987, pp. 1170–82.

[7] L. C. Freeman, "A Set of Measures of Centrality based on Betweenness," *Sociometry*, 1977, pp. 35–41.

[8] M. Cha *et al.*, "Delayed Information Cascades in Flickr: Measurement, Analysis, and Modeling," *Computer Networks*, vol. 56, no. 3, 2012, pp. 1066–76.

[9] B. Zhang *et al.*, "The Peer-to-Peer Trace Archive: Design and Comparative Trace Analysis," *ACM CoNEXT Student Wksp*., Philadelphia, 2010.

[10] UMass Amherst. (2003, May) Dataset: Can-O-Sleep, http://kdl.cs.umass.edu/proximity/index.html.

[11] F. Gringoli et al., "GT: Picking Up the Truth from the Ground for Internet Traffic," *Comp. Commun. Rev.*, vol. 39, no. 5, 2009, pp. 13–18.

Biographies

MIRAY KAS (mkas@ece.cmu.edu) received her Ph.D. degree from Carnegie Mellon University (CMU) in electrical and computer engineering in 2013. Prior to joining CMU, she obtained her B.Sc. and M.Sc degrees from Bilkent University, Ankara, Turkey, both in computer engineering. She is currently with Google. Her current research interests are in the areas of algorithm design and trend analysis for largescale, dynamic social networks and wireless networks. Her previous publications focus on channel access scheduling for wireless mesh networks and on-chip networks.

KATHLEEN M. CARLEY (kathleen.carley@cs.cmu.edu) is a tenured full professor in the Institute for Software Research Department in the School of Computer Science of Carnegie Mellon University (CMU), Pittsburgh, Pennsylvania. She currently leads substantial research efforts in the areas of dynamic network analysis and information diffusion, develops new algorithms and technologies for this area ranging from text mining, to network and visual analytics, to agent-based models. She has published widely with over 350 articles in the areas of network science, organizations, simulation, and social change. Her tools, in particular AutoMap and ORA are used in a variety of settings to extract and analyze networks (see www.casos.cs.cmu.edu/tools). In 2008, she founded with L. Richard Carley, the CMU startup known as Netanomics (www.netanomics.com) which specializes in providing technologies to support complex socio-technical systems from a combined social and technical perspective. Her current research interests include information dynamics, social media, dynamic networks, extracting networks from massive data, complex systems, re-usable and interoperable simulations, organizational design, WMD deterrence, remote detection of CBRNE capability, and security. She currently serves, or has served, as a consultant for several companies, government agencies, and on multiple national research council panels.

L. RICHARD CARLEY (carley@ece.cmu.edu) received an S.B. in 1976, an M.S. in 1978, and a Ph.D. in 1984, all from the Massachusetts Institute of Technology. He joined Carnegie Mellon University in 1984, and in March 2001, he became the STMicroelectronics Professor of Engineering at CMU. His research interests include analog and RF integrated circuit design in deeply scaled CMOS technologies and novel nano-electro-mechanical device design and fabrication. He has been granted 15 patents, authored or co-authored over 120 technical papers, and authored or co-authored over 20 books and/or book chapters. He has won numerous awards including Best Technical Paper Awards at both the 1987 and the 2002 Design Automation Conference (DAC). In 1997, Dr. Carley co-founded the analog electronic design automation startup, Neolinear, which became part of Cadence in 2004. In 2001, he co-founded a MEMS sensor IC startup which morphed into a MEMS RF IC startup in 2005, and in 2007 he co-founded a Network Sciences Company — Netanomics.