Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 00 (2011) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

`The 9th International Conference on Mobile Web Information Systems (MobiWIS)

The Early (tweet-ing) Bird Spreads the Worm: An Assessment of Twitter for Malware Propagation

Ameya Sanzgiri [a*], Jacob Joyce, Shambhu Upadhyaya

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY-14260

Abstract

Social Networks have rapidly become one of the most used Internet based applications. The structure and ease of information dissemination provides an opportunity for adversaries to use it for their own malicious purpose. In this paper we investigate a popular social network – Twitter as a malware propagation medium. We present a basic model for Twitter-based malware propagation using epidemic theory. Our analysis shows that even with a low degree of connectivity and a low probability of clicking links, Twitter and its structure can be exploited to infect many nodes.

Keywords: Malware; Malware propagation; Model; Social networks; Twitter

1. Introduction

Malware has traditionally been the bane of the Internet and its end users. The traditional ways of spreading malware has been through spamming sites via bulletin boards and email. The aim of any virus/worm is to propagate quickly and infect as many nodes as it can before they are detected and systems are patched. One traditional way of propagation is to use an infected user's email contact list [1] and to mail either the software itself or links to the software to other users. This method of malware propagation is reaching a dead-end, thanks to better spam filters of today, forcing spammers to resort to novel techniques to propagate malware. One such technique is to use Wikis and blogs by posting malware links to unsuspecting users [1, 2]. Other techniques have been to hide malware programs disguised as expensive legitimate software names such as Adobe Creative Suite, and MatLab which users download illegally. The real goal in malware propagation is to minimize the time for propagation while maximizing the malware download. With traditional email the malware provider has the added burden of making the user "want" to click on links. The result was seen in the spread of the *Anna Kournikova* worm and the

* Corresponding author. Tel.: +1-716-548-8126
E-mail address:ams76@buffalo.edu

ILOVEYOU virus. Even, in the latest malware spreading methodology of posting links on Wikis and boards, the onus is to make the user want to click the links. Further, it is difficult to increase malware propagation unless more complex code is written that also accesses, say, a user's email contact list or similar programs. In view of these facts, we can conclude that the main points of interest for malware propagation are: 1) how one can post links such that a user is unaware of the link contents, 2) how to maximize the chance of or entice a user to click on the link, and 3) what methodology is best suited to increase the propagation effect while decreasing the time.

Today, social networks have become ubiquitous. Online social networks provide the ideal way to propagate malware since they are Internet based schemes and paths are already set up for the malware to propagate rapidly, if desired. Given this aspect of online social networks, malware has already targeted Facebook accounts. The famous "Click-jacking" worm and Koobface worm [7] are two *famous* worms that have used Facebook as a malware propagation medium [8]. In this paper, we investigate the use of Twitter as a malware propagation medium. Twitter has already been under various attacks, mostly by spammers who have targeted Twitter's severs. Recently, Twitter has also been targeted for malware, though the outbreak details are not known [8]. Our contributions are as follows:

- We systematically study the different aspects related to propagating malware in the Twitter domain.
- We assess the use of Twitter for malware propagation.
- We present and assess a basic conceptual model of malware propagation for Twitter or Twitter based communication.

The paper is organized as follows – Section 2 discusses related work and Section 3 outlines preliminaries of social networks. Section 4 presents our attack model and threats related to Twitter. Section 5 presents a basic model to analyze malware propagation in Twitter. Section 6 assesses the model and provides some quantitative results. In Section 7, we outline the mitigation techniques related to this work. Finally Section 8 provides a summary and some insights into future work.

2. Related Work

Malware propagation has been a long studied topic in network security. Malware propagation in scale free networks [3, 4] has been investigated. Similarly propagation of malware in unconventional networks such as Wireless Sensor networks [5, 6] and cellular networks (using MMS and Bluetooth) [7] has been studied. Malware propagation in old world "social networks" such as email and instant messaging networks has been studied in [8, 9]. The attacks by worms and spammers on social networks have recently led researchers to focus on online social networks. In [10], [11] the authors investigate malware propagation using simulated topologies and user activities. Similarly, the authors of [12] use correlation techniques based on user activities to suggest some mitigation schemes for worms in the online social networks context. A focus of researchers has been the understanding of how information flows in social networks such as Facebook and Twitter [13]. Authors of [14] and [15] have used this information to detect spammers in online social networks. Our work differs from these works in two aspects. First, we present a model that can be used to analyze information as well as malware propagation in Twitter-like models. Second, we capture specific aspects of Twitter such as trending topics and '#-tags' that can be maliciously exploited to accelerate the propagation rate of worms.

3. Background and Preliminaries

3.1. Background

Social Networks, Twitter and Tweets. The main purpose of social networks is to bring people together, organize them in "networks" and provide mechanisms to disseminate information easily and

quickly in these networks. The current popular social networks are Facebook, Google+, Twitter, LinkedIn, and MySpace. Twitter is focused on disseminating information through conversations as is symbolized by their slogan posted on fly.twitter.com – "A faster, simpler way to stay close to everything you care about." Tweets, which are basically a broadcast to the world, have a limit of 140 characters, and can contain URLs. By using the string "@username", a tweet is classified either as a direct message or a mention which results in bringing the specific message to the attention of the user. In cases of such messages, the tweet is treated as unicast. Tweets can also be appended by "#-tags" to broadcast information about specific topics. Twitter and some other services use and mine information based on these #-tags in data heuristics to find 'trending topics.' Since users have a strict character limit on their tweets, users can encode URLs to "short-URLs" using third party services such as bit.ly and tinyURL. Privacy on Twitter is very simplistic. It is possible to flag an account as protected (where follow requests are required to view tweets, and requests need approval), otherwise all accounts are public and can be followed by anyone. Tweets are an unprotected/unencrypted broadcast and have been used for plain communication, marketing and obtaining current news data sets. Tweets which are part of conversation, or published in many places, can be found via searches thereby gaining many possible views.

Short-URLs. Short-URLs are normal URLs that are encoded into URLS with fewer characters, and can thus be used in tweets. However, short-URLs have some inherent issues. First, some services encode the same input URL into different (unique) short-URLs for different users. Figure 1 shows screenshots of the same URL being encoded as different Short-URLs. Second, unlike traditional systems, a user cannot follow the target of the short-URL (by hovering their mouse over the URL). The short-URL providers such as bit.ly or tinyURL services are required to decode them. Thus, a user has a very limited knowledge of the target of the short-URLs.

Figure 1. Same URL shortens as two different short-URL

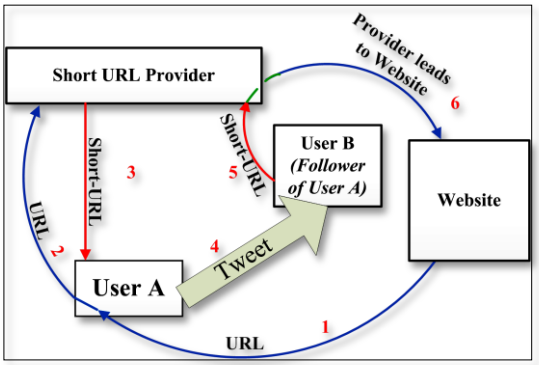

Figure 2. The process of tweeting links and following links (numbers show the sequence of operation)

The nature of short-URLs by itself presents a very potent threat to any user. The encoding of URLs is a method of obfuscating information, which can be exploited as a redirection attack to force *some* unwilling users to download/spread malicious software without their knowledge. However, when viewed in the context of Twitter this threat becomes much more potent and far reaching. Figure 2 shows the process of short-URLs being used in Twitter by a user and how a follower requires the provider to go to the website.

3.2. Problem Description

Twitter, like many other social networks, inherently possesses the risk of some malicious entities using the medium to share malicious ideas/executable/injects. However, the model of *user and followers* also presents a tree structure where information can spread rapidly as depicted in Figure 3. Further, in such a structure and with the large flow of information it is hard to detect new attacks and block them. These two

factors coupled together present an opportunity for any malicious attacker to masquerade an attack with relevant looking data, thus causing it to spread much faster than any detection techniques. In addition to these factors some additional factors that make Twitter a target attack medium are:

- Twitter is a new concept; most users aren't educated on risks and potential threats unlike other well-known technologies.
- Malicious users can leverage ongoing conversations and use them to gain personal information of a user.

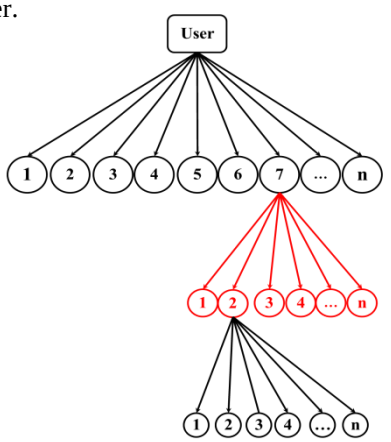

Figure 3. Tree structure depicting Twitter's User-Follower Model

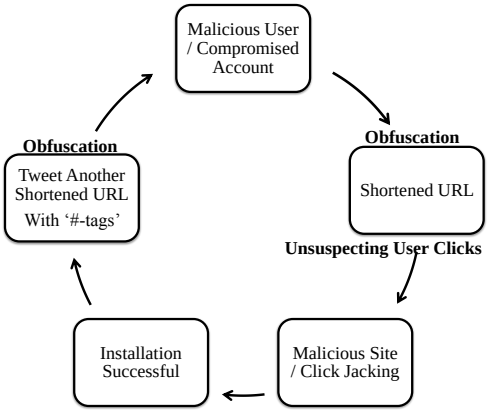

Figure 4. Cycle depicting our Twitter attack

4. Threat Model

4.1. Attack Scenario

In this paper, we consider a variant of the *Koobface* botnet attack [16]. A Twitter user account gets compromised by an adversary. The adversary, who has a malware site, creates a short-URL and tweets the short-URL. The user's followers see the tweet and click on it, which sets off a reaction as in click-jacking. The follower is taken to the malware site via the short-link provider (where the malware may be downloaded) and at the same time, a new short-URL is created and tweeted by the follower to his/her followers. This new tweet would be visible to his/her followers, and this way the attack propagates down the Twitter tree. In the context of our attack, all compromised tweets are also appended with a trending topics #-tag as depicted in Figure 4.

Of course, the success of the attack depends on the number of followers of a user and the probability of a follower clicking on a link. However, as detailed next the inherent Twitter structure presents an adversary with a very good medium for malware propagation.

4.2. Twitter Structure in the Context of our Attack

In our attack context, the entire Twitter structure can be thought of in terms of two distinct models –1) User → Follower, 2) #-Tag/Trending Topic. The first model provides the classical Twitter use where user(s) follows the tweets of another user and has his own followers who follow him. This model provides a better chance for a follower to click on a link simply because there is a high amount of trust between a user and his followers due to their interactions over time. Hence for penetrating a network, this provides the best chance for an adversary. However, the second model is where Twitter and similar structures distinguish themselves from other online social networks – here users are following trending topics, which are based on the #-tags. This model represents the use of Twitter in marketing, collecting opinions,

debating, etc. In the context of our attack, it allows an adversary to infect group(s) that is not connected to the compromised user or his followers. Further, this particular attack, allows a particular #-tag(s) to be in the top trending topics list, thus increasing the chances of a user clicking on the link.

5. A Basic Model to Analyze Twitter

Epidemic theory aims to measure how infections spread and assessing the risk it presents to a population. Epidemic theory builds on parameters such as the number of people who are already infected, people who are exposed, the spread rate of the infection, etc. Epidemic models [17] have been used to model malware spread in a network [5, 6].

The two popular models in epidemic theory are the Susceptible-Infected-Recovered (SIR) and Susceptible-Infection-Susceptible (SIS) models. In the SIR model, there is a chance for an infected person to recover from the infection, whereas in the SIS model, the infected person after what is known as the incubation period becomes susceptible to infection again. In our model we use the more generic SIR model, to show that once infected and recovered, a user can no longer become a part to infect more people. This would hold true to a large extent in most cases without loss of generality. Of course if an infected user were to get many more followers in this duration, this supposition would not hold true, but we believe the chances of a user quickly gaining followers during an attack are minimal. In the SIR model, $S(t)$, $I(t)$ and $R(t)$ denote the number of susceptible, infected and recovered nodes at time t, respectively. Every member of the population belongs to one of these groups. Thus, if $N(t)$ is the total population we consider then, $N(t) = S(t) + I(t) + R(t)$.

We also follow the standard convention of denoting the infection rate as β and the recovery rate as γ. The recovery rate denotes the removal rate of the infected users [5].

Since Twitter [18]'s method of disseminating information is a broadcast we assume that each of the susceptible users (the followers of an infected user) can get in contact with the infectious members and thus get infected. Given these criteria, we have the following equations for the rate of change of susceptible, infected, and recovered members, respectively:

$$\frac{dS(t)}{dt} = -\beta \cdot S(t) \cdot I(t), \quad \frac{dI(t)}{dt} = \beta \cdot S(t) \cdot I(t) - \gamma \cdot I(t) \quad \text{and} \quad \frac{dR(t)}{dt} = \gamma I(t)$$

The infection rate β denotes the probabilistic rate at which an infected or malicious user broadcasts his tweet containing a link. It is to be noted that based on our attack model, each tweet is also appended with the '#-tag' that is trending at that point. This considers those people who are not in the network of the followers of the infected user(s) and yet have a (probabilistic) chance of getting the infection by clicking on the link in the malicious tweets.

Thus, $No.\,of\ Infected\ users = Infection\ to\ Followers + Infected\ by\ "\#-tags"$

However, not all users are equally susceptible; the degree of susceptibility depends on the average degree of connectivity, the rate of trending topic, the probability of malware infection, and the probability of a link being clicked.

Similarly, if we consider N number of Twitter users, they can be divided into two categories – users who are in some way connected to an infected user (N_1) and users who are not connected to the infected users but are following the trend the malware uses via '#-tags' (N_2). Thus, $N = N_1 + N_2$ (where N_1 and N_2 are assumed to be large integers). In this analysis, we assume a no recovery model, i.e., once infected, the users are compromised and cannot be recovered. This assumption is justified since we do not want loops between infected → recovered → susceptible → infected and we clearly discount those compromised once, from getting compromised again. Now that we assume this particular model, we have a non-homogenous mixing, since only followers of a particular user (infected) have the potential to get infected themselves. Thus users that are outside the particular network, are inoperative and cannot get infected or spread the infection. We can visualize this as a circular region of infected nodes, centered

around the source user, which grows with time as the infection spreads based on the tweet broadcasts. Since we consider a "no recovery" model, we have $R(t) = 0$ and $\gamma = 0$ in all the equations above. Consequently, if left undetected and unchecked, the infected users will infect all the susceptible users. The number of infected users $I'(t)$ that lie in the network of an infected user is given by

$I'(t) = I(t) - \sigma\pi(r(t) - 1)^2$ where σ denotes the density of followers based on time and geographic locations [19] and $r(t)$ denotes the radius of the circle that contains the infected users. The parameter $r(t)$ is based on considerations such as time difference, geographic location, interest in a particular user's tweets, and that followers of an infected user (who are now infected) can be placed in a circular strip of thickness $r(t)$. This consideration is validated by the practicality of the assumptions we have made so far and the fact that while distance may not mean much in the Internet, time zones and time differences are a factor that determines how tweets are followed or get re-tweeted.

Based on this assumption, it is clear that $\sigma\pi r(t)^2 = I(t)$ and $I'(t) = \left(2\sqrt{\sigma\pi}\right)\sqrt{I(t)} - \sigma\pi$

$2\sqrt{\sigma\pi}$ is a proportionality constant that we are going to denote as c and thus we can state without loss of generality that $I'(t) = c\sqrt{I(t)}$.

It is important to note that, the average degree of followers of a user η plays an important role and that each user in $I'(t)$ is able to communicate only with η followers. The parameter η depends on the activity of a user, popularity and the interest generated by his/her tweets. Obviously, a user with a high degree of followers can spread the infection faster, than one with lesser number of followers. As mentioned above, the probability of a follower clicking on a link is the highest determining factor in the follower getting infected which we denote as τ.

The analysis so far only considers the user → follower model, i.e., users infecting followers. Since we are considering this to be N_1 users, we can write the relationships between the susceptible and infected users as:

$$N_1(t) = S_1(t) + I_1(t)$$

Thus, the equations for rate of change of infected and susceptible users become:

$$\frac{dI_1(t)}{dt} = \beta c\tau\sqrt{I_1}\,\frac{(N_1 - I_1)}{N_1}\,\eta \quad \text{and} \quad \frac{dS_1(t)}{dt} = \beta c\tau\sqrt{I_1}\,\frac{(N_1 - I_1)}{N_1}\,\eta$$

Solving these differential equations, the rate at which followers of users will be compromised, if they are all part of a network is:

$$I_1(t) = N_1\left(\frac{2}{1+\left(\frac{N_1-1}{N_1+1}\right)e^{\frac{\beta c\tau}{\sqrt{N_1}}t}} - 1\right)^2$$

Similarly, we now consider the rate of infection spread based on '*#-tags.'* The general form of the equation remains the same as the user→ follower model, except that the probability of infection is also determined on the rate of the trend and the probability of a (non-compromised) user clicking the link denoted by φ.

Thus combining the two factors the rate of infection is

$$I(t) = N_1\left(\frac{2}{1+\left(\frac{N_1-1}{N_1+1}\right)e^{\frac{\beta c\tau}{\sqrt{N_1}}t}} - 1\right)^2 + N_2\left(\frac{2}{1+\left(\frac{N_2-1}{N_2+1}\right)e^{\frac{\beta c\varphi}{\sqrt{N_2}}t}} - 1\right)^2 \quad \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

6. Model Assessment and Results

Figures 5 and 6 show the plots of the fractions of infected users as time progresses for different number of followers as well as different probabilities of clicking links. The plots are based on equation

(1) for different data trend rates namely, constant, linear, exponential and polynomial. Parameters p1 and p2 represents the probabilities of followers clicking links from users and trending topics respectively. For simplicity we assume all users to have the same number of followers. The x-axis represents the progression of time (in units) while the y-axis represents the fraction of infected users. In all the plots in figures 5 and 6, snapshots of the zoomed portion of the actual graphs are shown for clarity.

As can be seen from the plots the fraction of infected users increases directly based on the number of followers and the probability of them clicking the links. Further, the data trend rate also plays as important part, as can be observed from the graphs. Exponential data trend not only spreads the malware faster, but also has the highest initial amount of infection.

For the constant data trend, the spread of infection is very low as is the initial amount of infected users.

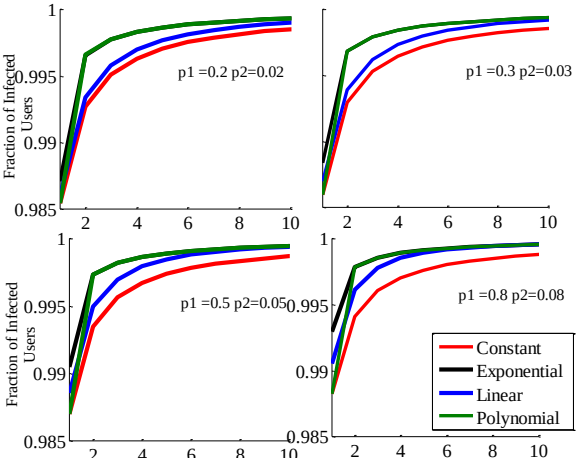

Figure 5. Factor of infected users, when followers =10, for different probabilities of clicking links

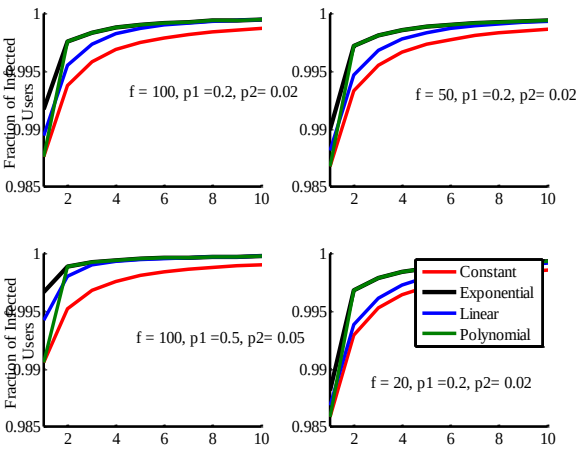

Figure 6. Infection Spread for different no. of followers (f) and different probabilities of clicking links

These plots validate our analysis of Twitter being a potent medium for malware propagation due to its inherent structure.

7. Mitigation Techniques

The attack we have presented in this paper is based on three important factors – 1) Obfuscation of information via short-URLs, 2) The broadcast nature of tweets, and 3) The probability of Twitter users clicking on the short-URLs. Any mitigation technique has to address these issues. Since changing broadcast nature of tweets would be detrimental to the popularity of Twitter, we are left with two options – either all followers/users do not click on links (which is detrimental to the online social network philosophy and cannot be enforced) or Twitter controls the way short-URLs are generated and used in tweets.

Twitter has recently created its own short-URL provider that to a certain degree does mitigate this issue. However, the authors have verified that Twitter still supports bit.ly and other short-URLs, which still maintains the threat of such an attack. It is important to note that, while Twitter users can term a short-URL as an abuse, the basic fact that most third-party short-URL providers generate unique short-URLs for different users, makes it difficult and time consuming for Twitter to appropriately rely on this mechanism. There could be other techniques such as "crowd sourcing" anti-virus software which could update itself on the final destination of the websites (verified via sandboxing techniques) and warns users, when they click on links. In such cases, the mitigation techniques might be simple; however they are outside the scope of this paper.

8. Conclusions

In this paper, we have presented an attack model and assessed Twitter as a malware propagation medium. The obfuscation of information by short-URLs, the broadcast nature of tweets and the user → follower model of Twitter presents an opportunity to exploit. Further, Twitter's trending topics present a unique way to infect people from different networks. We have also presented a conceptual model based on epidemic theory which shows how the malware infection could spread. We have analyzed and considered the different factors that an attacker must consider for his attack to be successful and have shown that even with a low degree of connectivity and a low probability of clicking links, the attack on Twitter can infect many people. We have also suggested some mitigation techniques to prevent such attacks. In the future we plan on running simulation, look at the challenges of performing real experiments to validate our model and investigate what would be the best methodology for an attacker to increase the impact of this attack in terms of the cost and time taken by malware to spread.

Acknowledgments

This research has been supported in part by National Science Foundation Grant DUE-0830814.

References

[1] S. Security. (2010). *New Attack Today against Wordpress*. Available: http://blog.sucuri.net/2010/05/new-attack-today-against-wordpress.html.
[2] S. Khare. (2008). *Wikipedia Hijacked to Spread Malware*. Available: http://tech2.in.com/news/telecom/wikipedia-hijacked-to-spread-malware/2667/0.
[3] L. Briesemeister, P. Lincoln and P. Porras, "Epidemic profiles and defense of scale-free networks," in *Proceedings of the 2003 ACM workshop on Rapid malcode*, New York, NY, USA, 2003, pp. 67-75.
[4] C. Griffin and R. Brooks, "A note on the spread of worms in scale-free networks," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 36, pp. 198-202, 2006.
[5] P. De, Y. Liu, and S. K. Das, "An Epidemic Theoretic Framework for Vulnerability Analysis of Broadcast Protocols in Wireless Sensor Networks," *Mobile Computing, IEEE Transactions on*, vol. 8, pp. 413-425, 2009.
[6] R. Di Pietro and N. V. Verde, "Introducing epidemic models for data survivability in Unattended Wireless Sensor Networks," in *WoWMoM, 2011 IEEE International Symposium on a*, 2011, pp. 1-6.
[7] S.M. Cheng , W. C. Ao, P.Y. Chen and K.C. Chen, "On Modeling Malware Propagation in Generalized Social Networks," *Communications Letters, IEEE*, vol. 15, pp. 25-27, 2011.
[8] M. Mannan and P. C. van Oorschot, "On instant messaging worms, analysis and countermeasures," in *Proceedings of the 2005 ACM workshop on Rapid malcode*, New York, NY, USA, 2005, pp. 2-11.
[9] C.C Zou , D. Towsley and W. Gong , "Email worm modeling and defense," in *Computer Communications and Networks, 2004. ICCCN 2004. Proceedings. 13th International Conference on*, 2004, pp. 409-414.
[10] M. R. Faghani and H. Saidi, "Malware propagation in Online Social Networks," in *MALWARE, 2009* pp. 8-14.
[11] G. Yan, G. Chen, S. Eidenbenz and N. Li, "Malware propagation in online social networks: nature, dynamics, and defense implications,"Proc. of the 6th ACM *Symp. on Information, Comp. and Comm. Security*, Hong Kong, China, 2011, pp. 196-206.
[12] W. Xu, F. Zhang and S. Zhu, "Toward worm detection in online social networks," in *Proceedings of the 26th Annual Computer Security Applications Conference*, New York, NY, USA, 2010, pp. 11-20.
[13] K. Lerman and R. Ghosh, "Information Contagion: an Empirical Study of the Spread of News on Digg and Twitter Social Networks," in *Proceedings of 4th International Conference on Weblogs and Social Media (ICWSM)*, 2010.
[14] F. Benevenuto, G. Magno, T.Rodrigues and V. Almeida, "Detecting Spammers on Twitter," in *Proceedings of the Seventh Annual Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)*, Washington, DC, USA, 2010.
[15] K. Beck, "Analyzing tweets to identify malicious messages," in *2011 IEEE International Conference on Electro/Information Technology* Piscataway, NJ, USA, 2011.
[16] K. Thomas and D. M. Nicol, "The Koobface Botnet and the Rise of Social Malware," in *MALWARE 2010*, Piscataway, NJ.
[17] D. Daley and J. M. Gani, *Epidemic modelling : an introduction*: Cambridge University Press, 1999.
[18] "Phishing for Twitter security," *Communications News*, vol. 46, pp. 6-6, 2009.
[19] T. Rao and S. Nagpal, "Real-time geo influence in social networks," in *Electronics Computer Technology (ICECT), 2011 3rd International Conference on*, 2011, pp. 246-250.