An Effective Scheme for Email Virus Detection and Containment

Jintao Xiong, PhD
Assistant Professor, Universidad del Turabo, Gurabo, PR, USA

Abstract
Email has evolved to be a convenient and important communication media. It greatly facilitates the communication among people from all over the world. The existence of all kinds of free email services makes email communication even more attractive. The convenience and popularity of emails has also made them ideal vehicles to spread computer worms and viruses. Email viruses and worms have appeared more frequently in recent months. How to detect and contain the spread of email viruses and worms while minimizing the side effects on normal email communications is a very important topic. This paper proposes an email virus detection and containment scheme to address this issue. Our scheme is inspired by the infectious disease control schemes used in real world. An infected computer is identified if it meets two conditions. Firstly, the computer has to demonstrate abnormal behavior by sending out large number of emails with attachments. Secondly, some of the computers which receive these emails have to demonstrate similar abnormal behavior. We propose to apply the "stone-in-the-pond" principle to reduce the overhead of the email virus detection system. The identified infected computers can then be isolated. Simulation experiments demonstrate that our proposed scheme is effective in controlling the spread of email viruses.

Keywords
Computer virus, network security, detection, containment

1. Introduction

Background
The rapid growth of computer networks have greatly changed the way people work and live. The Internet infrastructure and the huge information that can be obtained from it make the exchange of information much more convenient than a decade ago. However, unwanted things have also been created. Computer network viruses and worms are among those “un-welcomed”. These are self-replicating programs which can take advantage of the vulnerability of certain software and spread themselves to tens of thousands or even millions computers in a short time. In recent years, the threat of these worms and viruses have become more severe. Some have caused huge damage to the society. For example, the Code Red worm of 2001 caused over 2.6 billions of dollars of loss to the United States. Nimda of 2001 caused half a billion dollars in the first week in US. The blaster worm of August 2003 has infected over half a million users. The recent worms have demonstrated faster and faster propagation speed. For example, the Slammer worm infected 90% of the vulnerable computers within 10 minutes in January 2003. The SoBig.F virus has broken the email virus speed records.
There are many ways through which worms and viruses can spread themselves. Email is a major medium that worms and viruses use to spread themselves. To spread by email, a virus disguises itself as attachments of emails and send these emails with virus attachments to people. It can activate itself when a recipient opens the virus attachment. When an email virus infects a computer, it usually will search the computer for email addresses. It can find email addresses stored in a user’s address book. Some email viruses can also find email addresses stored in other directories, such as temporary Internet page folders. Email is a favorite of the hackers for several reasons. Firstly, since email is fast, convenient, it is widely used by millions of people in the world. Many people have more than one email addresses. The availability of all kinds of free email service providers certainly accelerated the adoption of email as a main communication media. Secondly, with governments and corporations beefing up their network security investment, emails seem to be a good way to penetrate those well guarded firewalls. Email viruses have appeared more frequently recently. Within the first several months of this year alone, there have been several well-known email virus incidents which include W32/Bagle and W32/Mydoom.

Related work

The severity and frequency of network worm and virus incidents have revived research in this area in the last several years. Models have been developed to study the propagation of network worms. (Staniford et al. 2001) used a simple epidemiological logistic equation to model the propagation of Code Red worm. (Zou et al. 2002) proposed a “two-factor” model to model the effect of human countermeasures and congestion caused by worm traffic. Ways to contain the propagation of worms have been proposed by some researchers. (Moore et al. 2003) studied the impacts of several design factors on the dynamics of a worm epidemic. (Zou et al. 2003a) proposed a dynamic quarantine method and analyzed its impact on worm propagation. (Wang et al. 2000) studied the effect of immunization on virus propagation. (Williamson 2003) proposed a behavior-based detection and control method to contain worm propagation by restricting the probing rate of infected hosts.

Some studied have been done on email virus modeling and control. (Zou et al. 2003b) studied email virus propagation on power law, small world and random graph topologies. (Swab 2001) proposed SMTP gateway virus filtering techniques.

2. Objectives

Objective

The objective of this paper is to develop an effective method to automatically detect the propagation of an email virus and contain or stop the propagation if such a virus propagation exists.

Original aspects

This paper has several original aspects. Firstly, unlike most other virus detection tools, this paper proposes a method that is effective toward both known and unknown email viruses. Most available anti-virus tools rely on virus signatures to detect the existence of viruses. Since it takes time to obtain the signatures of a virus, they are usually only effective toward known viruses. The time it takes to obtain the virus signatures of a virus is usually much longer than the time it takes for a virus to infect most vulnerable machines. Those approaches are usually not helpful in detecting and mitigating the spread of new worms and viruses. Secondly, the method proposed in this paper does not require human intervention. A system built upon the method proposed in this paper can run fully automatically. Both the detection method and the countermeasure can be implemented automatically. It does not need install patches or update software. Thus it can respond quickly to a worm incident. Thirdly, the method proposed
in this paper makes efforts to detect and contain the virus propagation as early and effective as possible but also disrupt normal operations of the system as little as possible. Although it is mostly desirable to detect and control the spread of a virus early, a anti-virus system won’t be welcomed if it constantly interferes with the normal operations of the systems. Our method takes steps to reduce the frequency of false warnings which is a major weakness of many anomaly based virus detection method.

3. Scope of Work

This paper proposes an effective method for email virus detection and containment. Although the method is developed within the context of email viruses, similar ideas can be extended to other network virus and worm detection and containment. The scope of work presented in this paper includes the following.

- This paper proposes a method for email virus detection.
- This paper proposes a method for email virus containment which can be integrated with the proposed email virus detection method.
- This paper sets up simulation experiments and use them to verify the effectiveness of the proposed methods.
- This paper proposes recommendations to reduce the load of the proposed email virus detection method.

4. Methodologies

Email virus detection

The target of our detection and containment are email viruses and worms. We assume that a local email network is the network that we want to protect. The key of a universal detection scheme is to identify the symptoms that are common to most, if not all, viruses. Currently, an email virus is identified by the subject line of the email, the file name of the attachment, the size of the attachment, etc. However, these are different for different viruses. Also, one email virus can use arbitrary subject line, attachment file name. By zip the virus program and other irrelevant files together, a virus can also freely change the size of the attachment. Obviously, none of these is credible detection criterion.

A more credible way is to identify a virus by its behavior. Most viruses exhibit two symptoms. Firstly, when they find the email addresses store on an infected computer and send copies of itself to those email addresses, they usually send emails with virus attachments to all the email addresses at once. The infected computer exhibits the symptoms of sending out a large number of emails with attachments in a short time. To detect this symptom, we can choose a threshold $K$. If a computer sends out more than $K$ emails in a given time period $T$, this computer could be an infected computer. However, this symptom alone is not very reliable. For example, a user sending emails with attachments to an email list will also be detected by this criterion, although this behavior is legitimate and often happens.

But there is a big difference between a virus spreading itself and a user sending emails to an email list. If emails were sent by a virus, some of the recipients of the virus emails will open it and their computers will be infected. After their computers are infected, the virus will spread themselves from these newly infected computers which means some of these computers will also send more than $K$ emails within time $T$. This phenomenon is rarely true if the emails were sent by a user to an email list. Combing these two observations, we have the virus detection algorithm as follows.

Email virus detection algorithm
1. Periodically monitor the number of emails with attachments sent by the computers (hosts) in the network.
2. If a host sent more than \( K \) emails within one period, put all the recipients of the emails in a special watch list.
3. If any host in the watch list sent more than \( K \) emails within one period, a virus propagation is said to be identified.
4. If none of the host in a special watch list exhibits abnormal behavior within time \( D \) (\( D \gg K \)), the watch list is canceled.

Since the detection algorithm need the ability to collect the receivers of emails, email servers are naturally the ideal place to implement the detection algorithm.

**Email virus containment**

After an email virus propagation is detected, the next step is to contain or stop the virus propagation as soon as possible. Classic epidemic theory has shown that in order to completely stop the propagation of an epidemic, the average number of new infective that an infected host can infect must be less than 1. Otherwise, the epidemic will finally infect the majority of the population. Since an email virus can contact many computers in a short time, it is a daunting task to stop the propagation of it. In order to achieve the goal, dramatic countermeasures have to be taken.

Firstly, we decide the hosts that the virus control system has to take action on. Obviously, these hosts include all the hosts that have received virus emails. This means that the virus containment algorithm must record all these hosts. After the hosts are identified, they will be isolated. By isolation we mean that all the emails with attachments sent from these hosts are blocked until the virus email is deleted. In order not to isolate hosts that have already deleted virus emails, we can use the following mechanisms if the users can be assumed to be honest and cooperative. After the virus control system block these identified host, the system sends an email to these hosts telling them the reason for the isolation. It also asks the user to delete virus emails. Once the user reads the email and deletes the virus email, he will send a reply to the virus control system and the system will remove his computer from the isolation list. If a user has already opened the virus emails, he can then seek system administrator for help.

**5. Results**

In this section, we use simulation to verify the effectiveness of the email virus detection and containment method.

**5.1 Experiment Setup**

We first simulate an local email network. We model a local email network as a graph \( G = (V, E) \) which consists of \( V \) nodes and \( E \) links. Each node in the network denotes a host and a link from node \( a \) to node \( b \) means that the email address of \( b \) is stored on node \( a \). It is not feasible to obtain the topology of real email networks. We use two topologies in our simulations. One topology is the random topology where each link connects randomly chosen two nodes. The other topology is the power law topology where the degrees of nodes have power law distributions. Each topology has 5000 nodes and 15000 links.

We describe a node \( i \)'s behavior using a quadruplet \( (c_i, t_i, p_i, a_i) \). \( a_i \) is total number of email addresses stored on node \( i \). \( p_i \) is the probability that the user opens a virus email attachment. It is modeled as a
Gaussian random variable with distribution $p \sim N(0.5,0.4)$. $t_i$ is the time interval that the user checks his emails. It has exponential distribution with mean $E[T_i]$. $E[T_i]$ has Gaussian distribution.

We simulate two types of virus behaviors. One is the nonreinfection virus. This type of virus only spreads itself once after it infects a computer. The other is the reinfection virus. This type of virus spreads itself repeatedly after it infects a computer. Naturally, reinfection viruses are more dangerous than nonreinfection viruses.

5.2 Experiment results

The experiment results are shown in Figure 1 and Figure 2. In both figures, the x-axis is the simulation time and the y-axis is the number of infected hosts in the network. The results are the averages of 100 simulation runs.

Figure 1: Virus Propagation in a Power Law Network

Figure 2: Virus Propagation in a Random Network
6. Conclusions

This paper proposed an email virus detection and containment method. The effects of the method was verified by simulation experiments. These experiments show that the proposed method can effectively stop the propagation of nonreinfection and reinfection viruses under different network topologies. By taking into consideration both the abnormal behavior of a suspicious host and the hosts that have email links with it, we can reduce the false alarm probability. By incorporating user responses into the virus control system, we can minimize the inconveniences caused by the necessity to isolate infected computers.

7. Recommendations

Since one infected machine may send emails to many machines, the virus detection algorithm has to monitor many possible infected machines simultaneously. This may not scale well for large networks. To reduce the load of the detection system, we recommend to apply the “stone-in-the-pond” principle. This principle says that we only need to monitor the most likely infected subjects first. If these subjects are infected, we can then expand the scope of monitoring to other suspected subjects. It is difficult to prove which computers are most likely infected. Our recommendation is to monitor the accounts with the smallest number of new emails. We recommend this for two reasons. Firstly, Fewer new emails usually mean that the user checks emails more often. Secondly, fewer new emails means that it is easier for the virus email to catch the attention of the user. We acknowledge these are empirical arguments.

By only monitoring a small portion of all the receivers of the possible virus emails, we can greatly alleviate the burden of the virus detection system.

8. References