Association for Information Systems
AIS Electronic Library (AISeL)

PACIS 2009 Proceedings

Pacific Asia Conference on Information Systems (PACIS)

July 2009

Internet Privacy Information Propagation Model

Han-Wei Hsiao
National University of Kaohsiung, Taiwan, hanwei@nuk.edu.tw

Kun-Yu Chen
National University of Kaohsiung, Taiwan, xspiritualx@gmail.com

Cathy S. Lin
National University of Kaohsiung, Taiwan, cathy@nuk.edu.tw

INTERNET PRIVACY INFORMATION PROPAGATION MODEL

Han-Wei Hsiao
Associate Professor of Information Management
National University of Kaohsiung
700, Kaohsiung University Rd., Nanzih District, 811. Kaohsiung, Taiwan, R.O.C.
hanwei@nuk.edu.tw

Kun-Yu Chen
Graduate Student of Information Management
National University of Kaohsiung
700, Kaohsiung University Rd., Nanzih District, 811. Kaohsiung, Taiwan, R.O.C.
m0973309@mail.nuk.edu.tw

Cathy S. Lin
Associate Professor of Information Management
National University of Kaohsiung
700, Kaohsiung University Rd., Nanzih District, 811. Kaohsiung, Taiwan, R.O.C.
Cathy@nuk.edu.tw

ABSTRACT

With the rapid growth of information and communication technology (ICT), the violation of information privacy has increased in recent years. The privacy concerns now re-emerge right because people perceives a threat from new ICT that are equipped with enhanced capabilities for surveillance, storage, retrieval, and diffusion of personal information. With the trend in the prevalence and the easy use of ICT, it is of necessary to pay much attention to the issue how the ICT can threaten the privacy of individuals on the Internet. While the Email and P2P tools are the most popular ICT, this paper aims at understanding their respectively dissemination patterns in spreading of personal private information. To this purpose, this paper using dynamic model technique to simulate the pattern of sensitive or personal private information propagating situation. In this study, an Email propagation model and a Susceptible-Infected-Removed (SIR) model are proposed to simulate the propagation patterns of Email and P2P network respectively. Knowing their dissemination patterns would be helpful for system designers, ICT manager, corporate IT personnel, educators, policy makers, and legislators to incorporate consciousness of social and ethical information issues into the protection of information privacy.

Keywords: Information Privacy, Propagation Model, Email, Peer-to-Peer (P2P)

INTRODUCTION

With the rapid growth of information and communication technology (ICT), the violation of information privacy has increased in recent years. While there is no affirmative fact to predicate if the ICT has the more mischief or beneficial to our human life, it is for sure that ICT can both benefit humans and come up the threat to the privacy of individuals.

Just as the saying "curiosity killed the cat"; the concept of privacy is a fragile perception, especially "other's privacy". For example, recent years in Chinese society we have several recognized cases concerning the spreading of private pictures on the Internet, such as 'Chu Mei-feng' sex scandal event in 2001; the sex scandal photos of 'Edison Chan' in 2008; and 'Zhang Ziyi' sexy beach photo scandal in 2009. Actually these events are quite personal private affairs, but through the ICT, these private pictures/films all have spread like "wildfire" on the Internet. Indeed, the ICT bringing much convenience in data processing, query efficiency, yet the privacy concerns now re-emerge right because people perceives a threat from new ICT that are equipped with enhanced capabilities for surveillance, storage, retrieval, and diffusion of personal information (Clarke, 1988; Gentile & Sviokla, 1990; Mason, 1986; Miller, 1971; Westin, 1967).

With the trend in the prevalence and the easy use of ICT, it is of necessary to pay much attention to the issue how the ICT can threaten the privacy of individuals on the Internet. To this purpose, this paper aims at investigating two kinds of ICT, the email and P2P tool, to see their respectively dissemination patterns in spreading of personal private information. The email has the personalized characteristic and the P2P is quite an influential and powerful ICT in sharing information, both of them are extremely pervasiveness Internet applications and almost everyone uses them. Therefore, this study proposes an email propagation model and a Susceptible-Infected-Removed (SIR) model to simulate the diffusion patterns of email and P2P network respectively. Knowing their dissemination patterns would be helpful for system designers, ICT manager, corporate IT personnel, educators, policy makers, and legislators to incorporate consciousness of social and ethical information issues into the protection of information privacy.

LITERATURE REVIEW

3.1. E-mail

According to the 'Pew Internet Project Data Memo' on January 28, 2009[1], email remains the most popular online activity for Internet users. Compared to the physical post mail service, Email has the advantages of no time limit, free space constraints and zero costs sending and receiving messages among people. The content of email now covers a variety forms can be delivered, including texts, all kinds of files such as music, video, audio, and even computer viruses. Specific to the viruses, many previous studies pointed that email has become a most serious dissemination channel and causes lots of torment to Internet users (Zou, Towsley et al. 2003; Jin, Liu et al. 2008). For the reasons mentioned above, since email is right the common ICT channel for most Internet users delivering their private message and files, there is of a need exploring the diffusion pattern of computer virus via email. Thus, a new email propagation model is proposed in this study to examine this point.

3.2. P2P Network

As the Network technique development, Peer-to-peer (P2P) network has now replaced traditional file sharing tools and become a major platform spreading all kinds of files, including the unauthorized and private files sharing. The P2P software, such as Kazaa, E-Donkey, E-Mule, Morpheus, and Napster, are the hotbed of private files spreading. The privacy risks on P2P since which is quite an easy way transmitting larger media files, which might include private videos and images being molested. The private and even extremely sensitive files are shared unintentionally by Internet users using P2P networks. A HP and the University of Minnesota survey[2] has confirmed this point that only few P2P users aware the files sharing, that means there are more files transmission without users discovered. Obviously, many privacy invasions occur without P2P users' knowledge. Previous study of Parameswaran, Susarla et al. (2001) exhibited a mathematical model to simulate the diffusion of private files in P2P network. Following the privacy concerns mentioned above, this study would propose a Susceptible-Infected-Removed (SIR) model to simulate the propagation patterns of P2P network.

3.3. P2P Network classification

Traditionally, the structure of client-server is one of the main ideas of P2P network. This innovative idea makes every peer both as a client and a server all at once. This has the advantages of making not only a lower cost of maintaining websites but also very efficient in spreading files by

[1] http://www.pewinternet.org/pdfs/PIP_Generations_2009.pdf

[2] http://www.hpl.hp.com/research/idl/papers/kazaa/index.html

separating the computing to each unit in the network. Thus, the ‘file index’ and ‘file distribution’ are considered as two dimensions in the P2P network. Overall, three distinct types of network can be implemented to achieve the network file-sharing: (1) Centralized index & Distributed data; (2) Centralized index & Centralized data; and (3) Distributed index & Distributed data. The first and the third types have been applied to the P2P network. Recent years, a mixed type of those two types has proposed, for example, the E-Mule.

3.4. P2P Network Features

According to Thommes and Coates (2005), P2P network has some unique features. That is, every peer in the P2P network has one file-sharing folder containing files which are publicly available for download by others on the network. When users try to download files, it began with sending out a searching query. Eventually, users will receive a list which contains the searching result that matches the criteria. The list was generated varies among the various P2P network. Then, when a file is decided to download, users would click the file on the list. The connections will be set up among peers. After that, users start downloading the file from others. The features of P2P network makes it available to download different parts of files from different peers at the same time. Finally, when parts of the file were downloaded, they would immediately be placed at the share folder in order to share with others immediately.

3.5. Epidemic models

Ages ago, diffusion of epidemic disease has been an important issue since 1927, the first propagation model, SIR (Susceptible, Infected, Removed) was proposed by Kermack & McKendrick. This model allowed the simulation and prediction of the diffusion of epidemic disease. With the rapid growth of information and communication networks, computers become prevalent in our daily life. In the P2P network, files are transmitted, large amount of computer viruses are also brought huge damages to Internet users. And the dissemination actions of computer viruses on the Internet is quite similar to the epidemic diseases, therefore, this study adopts epidemiological models for disease propagation to simulate P2P network. While the P2P network has the feature of time-dynamic (Leibnitz, Hoßfeld et al. 2006), this study adopts the SIR dynamic model to simulate the propagation patterns of P2P network.

PROPAGATION MODELS

4.1. E-mail Model

4.1.1 *Model Description*

To examine the diffusion pattern of private files through Email, a new email propagation model is proposed in this study. In the model, the variable θ represents the average amount of emails recipient sends per time. Also, this study considers the volume that recipient sends each time would decrease as the time increase. Hence, in this model, $D(i)$ is the degree formula so as to match the natural phenomenon of human behavior. And a probability mechanism is made to deduct the new recipient which can be received the privacy target more than twice. The above manipulation would make this model more realistic to the real world.

4.1.2 *Model Equation*

This model aims at investigating the volume changed of new email recipient in a given period. A general form of our model equation is as follows.

$$f(n) = [\sum_{i=1}^{n-1} D(i) f(n-i)] * [\frac{N-F(n-1)}{N}] \qquad (1)$$

To make a better understating of the model, the equation would separate into two distinct parts.

$$[\sum_{i=1}^{n-1} D(i) f(n-i)] \qquad (2)$$

$D(i)$ represents the volume of email each recipient sends whereas its decrease follows a linear or non-linear recession as the time increase. Due to the natural humility, we can be inferred that as the time increase the volume of sending out emails will decrease. It can be any formula that has this identity. For example, exponential distribution follows the non-linear recession. But here we simply just make it as θ divided by the times (i) the recipient sends;

$f(n-i)$ represents new recipients increased during the period from 1 to $(n-i)$. $[\sum_{i=1}^{n-1} D(i) f(n-i)]$ which means the summation of new recipients created. In other word, it's the total volume of new recipients created during the period from 1 to n.

$$D(i) = \frac{\theta}{i} \qquad (3)$$

To simplify the model, we assume that the average number of sending out a popular private file is two, despite of the fact that it would probably be sent in many times.

$$[\sum_{i=1}^{n-1} \frac{\theta}{i} f(n-i)] = [\frac{\theta}{1} f(n-1) + \frac{\theta}{2} f(n-2) + \frac{\theta}{3} f(n-3)... + \frac{\theta}{n-1} f(1)] \qquad \mathbf{(4)}$$

$$[\frac{\theta}{1} f(n-1) + \frac{\theta}{2} f(n-2)] \qquad (5)$$

$f(n-1)$ represents new recipients increased during the period from $(n-3)$ to $(n-1)$, while $f(n-2)$ is from $(n-4)$ to $(n-2)$.

After simplifying, our model would be like equation (6).

$$f(n) = [\frac{\theta}{1} f(n-1) + \frac{\theta}{2} f(n-2)] * [\frac{N-F(n-1)}{N}] \qquad (6)$$

In our simplified model, the variables are defined as follows.

- $f(n)$: New recipients increased during the period from $(n-2)$ to n.
- $D(i)$: A degree formula which has the identity of linear recession.
- θ : The average amount of emails recipient sent per time.
- N : Total number of users in the network.
- $f(n-i)$: represents new recipients increased during the period from $(n-i-2)$ to $(n-i)$.
- $F(n-1)$: Cumulative quantity of email recipients accumulate to time $n-1$.

The second part.

$$[\frac{N-F(n-1)}{N}] \quad (7)$$

One may notice that the new recipients would probably be counted repeatedly, so we set this part to deduct the number which was counted more than twice.

4.2. Peer-to-peer Model

4.2.1 *Epidemic Model of File Diffusion*

Among epidemic models, SIR (Susceptible, Infected, Removed) model is the commonly used. This model categorized the population into groups depending on their states, which are susceptible, infected, and removed. The transform rate α and β were used to transform one state to another. The model initially intended to discuss the propagation of epidemic diseases. Yet in this paper, SIR model were used to represent the network of P2P. It's obvious that the simplicity of the model can't take fully control of the complexity of P2P network. Still, due to the techniques of building dynamic model in SIR would demonstrate the time-dynamic features in P2P network (Leibnitz, Hoßfeld et al. 2006).

4.2.2 *Model Description*

The purpose of this model is to simulate the diffusion of private files in the network of P2P. Three distinctive states were recognized in the model. Each state will be corresponded respectively to the ones in the P2P network. We must draw attention to the definition of users and peers. Users were defined as the people who use P2P software to download files, and the peers were defined as the client software which was used by the users. The followings will explain a little further to the states corresponding to the three in P2P network.

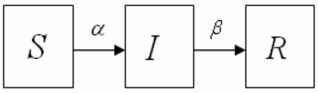

Figure 1. SIR model

- Susceptible

  The category represents the one who hasn't been infected. Corresponding to P2P network, it will be regarded as the peer that hasn't the private files downloaded. Volume of the category would change by the increase of time. The number of peers in this category at time t is denoted by $S(t)$.

- Infected

  The category represents the one who has already been infected. Corresponding to P2P network, it will be regarded as the peer that already has the file downloaded. Volume of the category would change by the increase of time. The number of peers in this category at time t is denoted by $I(t)$.

- Removed

  The category represents the one who has already died due to the illness. Corresponding to P2P network, then it will be regarded as the peer that already has the file removed from the share folder. Due to the mass volume of uploading the file might sacrifice the computation time of the computer. Volume Of the category would change by the increase of time. The number of peers in this category at time t is denoted by $R(t)$.

Assuming that the total number of the peers is N in any given time, and no other states except those three above. So the equation below can be inferred.

$$N = S(t) + I(t) + R(t)$$

From the equation above, we can know that the total volume of the states is zero at any given time.

$$\frac{dS}{dt} + \frac{dI}{dt} + \frac{dR}{dt} = 0$$

In the SIR model, states transformed from one to another by the transform rate α and β, which explains a little further in the following paragraphs.

- α : Transform rate that transforms the susceptible to infected. Corresponding to P2P network, it represents the rate that transforms the state from not having the file to file downloaded.
- β : Transform rate that transforms the infected to removed. Corresponding to P2P network, it represents the rate that transforms the state from having the file downloaded to file removed from the share folder.

4.2.3 *Model Equation*

- Differential equation of infected peers

  When transformed from 'infected' into 'removed', the number of the transformed (βI) will be deducted from the volume of infected peers. Moreover, when transformed from 'susceptible' into 'infected', the volume of the infected peers (αSI) will increase, and it will be deducted from the susceptible peers. The equation is shown below.

  $$\frac{dI}{dt} = \alpha SI - \beta I \qquad (1)$$

  Corresponding to P2P network, this equation represents the volume of the peers that have the private file downloaded.

- Differential equation of removed peers

  When transform from 'infected' into 'removed', the volume of removed peers (βI) will increase, and relatively be deducted from the infected peers. The equation is shown below.

  $$\frac{dR}{dt} = \beta I \qquad (2)$$

  Corresponding to P2P network, this equation represents the volume of the peers that have the private file removed from the share folder.

- Differential equation of susceptible peers

  We can recognize from the equations above, once the susceptible peers were infected, the volume (αSI) will be deducted from the susceptible. The equation is shown below.

  $$\frac{dS}{dt} = -\alpha SI \qquad (3)$$

  Corresponding to P2P network, this equation represents the volume of the peers that do not have the private file downloaded.

SIMULATION

5.1. Email

5.1.1 *Notion θ*

In the email model, the notion θ was given two different values which are respectively 2 and 4 to observe the diffusion around the network. Assume that only one person has the private files at the very beginning, and the total number of recipient in the network is N = 300,000. Considering one day as a unit 30 days in total. The results of the simulation are shown below.

5.1.2 *Result and Discussion*

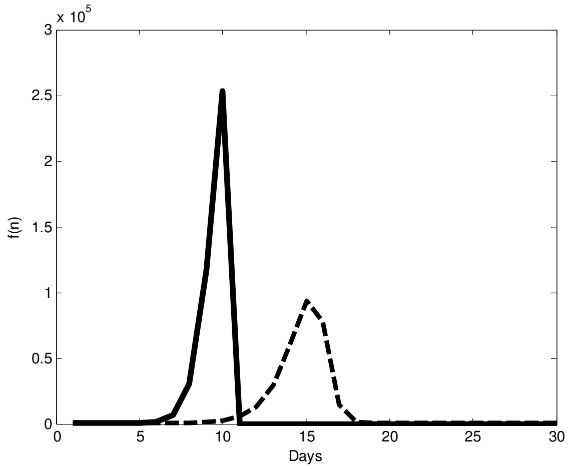

Figure 2. Dotted line: θ =2, Solid line θ =4

Notice that once the $f(n)$ reached their maximum; it'll start to go down. That's because the latter the diffusion, the popularity contains more recipients, so it must be deducted from $f(n)$.

As seen in figure 2, when the value of θ gets higher (which represents the average amount of email recipient sent per time is larger). The number of new recipient $f(n)$ per day is significant different. When θ =4, the maximum of new recipient is 250,000 and it takes 10 days to complete the files diffusion throughout the network. Yet when θ =2, the maximum of $f(n)$ is 90,000, which is smaller than the previous one and it takes 15 days to complete the file dissemination.

5.2. Peer-to-peer Network

5.2.1 *Notion α and β*

In this part, we'll discuss two notions respectively while the other remains unchanged. Our concern in this paper is the diffusion of private files, so the volume of the state "Infected" will be focused. As mentioned in the previous parts of this paper, the state "Infected" represents the peers already have the private files downloaded. Here we set the total number of peers in the P2P network is N = 300,000. We consider one day as a unit, and there are 30 days in total. And only one peer has the private files at the very beginning.

- Exploration in α while β remains unchanged.

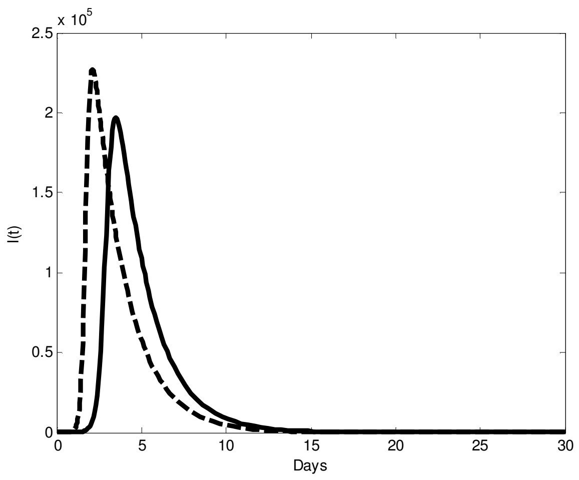

Figure 3. Solid line: α ***=.000016 ,Dotted line:*** α ***=.000026***

Under such scenario, we find out that when α gets larger, the volume of the peers will grow faster and the differences between downward slopes are small. The speed of growth was affected by α, but the influence is relatively small to the speed of decline.

- Exploration in β while α remains unchanged.

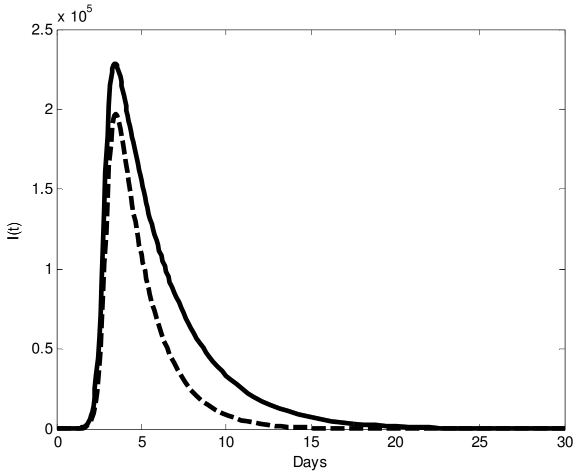

Figure 4. Solid line: β ***=.3058, Dotted line:*** β ***=.5058***

In figure 4, the results show that when β gets larger, the volume of the peers will decline faster and the differences between upward slopes are small. The speed of decline was affected by β, yet the influence is relatively small to the speed of growth.

5.3. Email & P2P

5.3.1 *Notions*

Due to the convenience of email and the innovative techniques on file-sharing of P2P network, this study considers two kinds of ICTs to observe their respectively dissemination patterns. Mathematical models were developed to simulate the diffusion of private files. The basic assumptions for these two models are as follows: (1) Units (people or peers) in each network is N= 15,000 in total. 2. Only one peer or person has the private files at the very beginning. 3. We consider one day as a unit, and there are 30 days in total; in the email model, θ the average amount of email recipient sent per time is five. And the notions α and β in the P2P model were based on(Thommes and Coates, 2005), which α =5.7870e-005 (corresponds to five download per day) and β=1.1574e-005 (the average time for a peer to remove a private file from the share folder is 24 hours). The figure below shows the result of the simulation.

5.3.2 *Result and Discussion*

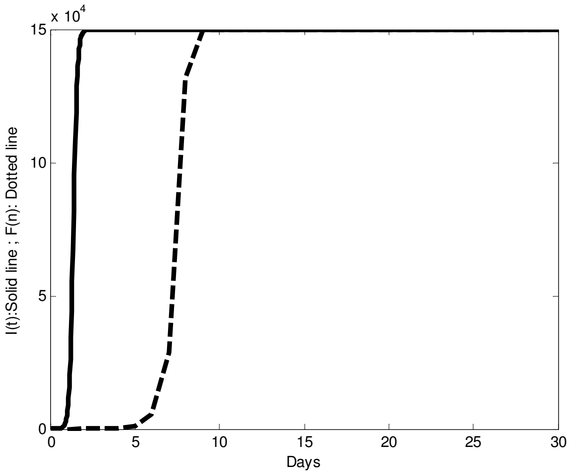

Figure 5. Solid line: P2P model; Dotted line: e-mail model

It is worthy notice that the two lines have slightly different (see Figure 5). The solid line represents the volume of the peers that own the file while the dotted line represents the cumulative quantity of Email recipients. The simulation results obviously exhibit that the P2P network has a dominant propagation power (in two-day) in dissemination files on the Internet than Email (in nine-day). After that, all Email users and P2P users in their respectively network have already received and/or downloaded the files.

CONCLUSION

While the ICT bringing much convenience in the information age, yet the privacy concerns now re-emerge right because people perceives a threat from new ICT that are equipped with enhanced capabilities for diffusion of personal information. But, how soon and how different the ICTs' dissemination patterns? This remains a doubt. The findings of this study demonstrate that the propagation speeds by using Email or P2P are rapidly. Specifically, the speed of spreading privacy information is far faster in P2P network than in Email. And most important, as the increasing of θ or α, the diffusion spreading speed will be enhanced, this finding illustrates that the spread speeds is dependent on the interesting of privacy targets. That is, the dissemination of privacy targets via ICTs, the different interesting level will result in dissimilar spread speeds.

Further, to the different kinds of ICTs, Email and P2P network, just as the simulation results in this study, P2P network has a dominant propagation power (in two-day) in dissemination files on the Internet than Email (in nine-day). This somehow demonstrates that P2P software has a much more enormous impact when dispersed privacy information.

The current study has responded the doubt concerning the ICTs actual impact to information privacy. In the future, a collection of real dataset is needed to empirical validate the two distinct model of Email and P2P network, to apply the models and see their respectively prediction power. With the empirical validation, the notion theta will be adjusted and justified with a more precise value. Also, we can analytically conduct sensitivity analysis because both models are well-formed. Overall, the proposed models in this study have their academic meanings to make a further understanding in these two model and parameters considered in these two models. Also, the practical implications would be that by knowing the propagation patterns of email and P2P network would provide the practicians such as policy makers, legislator, system designers, corporate IT personnel and educators to incorporate consciousness of social and ethical information issues into the protection of information privacy.

References

Clarke, R.A. (1988) Information Technology and Dataveillance, *Communications of the ACM,* 31, 5, 498-512

Gentile, M. and Sviokla, J.J. (1990) Information Technology in Organizations: Emerging Issues in Ethics & Policy, note, *Harvard Business School*, Boston, MA.

Jin, C., J. Liu, et al. (2008). A Novel Email Virus Propagation Model. *Proceedings of the 2008 Workshop on Power Electronics and Intelligent Transportation System* - Volume 00, IEEE Computer Society.

Kermack, W. O. and A. G. McKendrick (1927). "A Contribution to the Mathematical Theory of Epidemics." *Proceedings of the Royal Society of London.* Series A 115(772): 700-721.

Kermack, W. O. and A. G. McKendrick (1932). "Contributions to the Mathematical Theory of Epidemics. II. The Problem of Endemicity." *Proceedings of the Royal Society of London.* Series A 138(834): 55-83.

Kermack, W. O. and A. G. McKendrick (1933). "Contributions to the Mathematical Theory of Epidemics. III. Further Studies of the Problem of Endemicity." *Proceedings of the Royal Society of London.* Series A 141(843): 94-122.

Leibnitz, K., T. Hoßfeld, et al. (2006). Modeling of Epidemic Diffusion in Peer-to-Peer File-Sharing Networks. *Biologically Inspired Approaches to Advanced Information Technology*: 322-329.

Mason, R. O. (1986) Four Ethical Issues of the Information Age, *MIS Quarterly*, 10, 1, March, 5-12.

Miller, A. (1971) The Assault on Privacy: Computers, Data Banks and Dossiers, University of Michigan Press, Ann Arbor, MI.

Parameswaran, M., A. Susarla, et al. (2001). "P2P Networking: An Information-Sharing Alternative." *Computer 34(7)*: 31-38.

Thommes, R. W. and M. J. Coates (2005). Modeling Virus Propagation in Peer-to-Peer Networks. *Information, Communications and Signal Processing*, 2005 Fifth International Conference Bangkok.

Westin, A.F. (1967) Privacy and Freedom, Atheneum Publishers, New York.

Zou, C. C., D. Towsley, et al. (2003). Email Virus Propagation Modeling and Analysis. Technical Report: TR-CSE-03-04. Amherst, University of Massachusetts.