2004－ＣＳＥＣ－26　(30)
2004／7／21

メール型コンピュータウィルス拡散モデル

佐藤　大輔[†]　内田　真人[†]　石橋　圭介[††]　小林　真[†††]

† NTT サービスインテグレーション基盤研究所
†† NTT 情報流通プラットフォーム研究所
〒 180-8585 東京都武蔵野市緑町 3-9-11
††† 日本ソフトウェアプロダクツ
〒 141-0022 東京都品川区東五反田 3-18-6
E-mail: †{satoh.daisuke,uchida.masato}@lab.ntt.co.jp, ††ishibashi.keisuke@lab.ntt.co.jp ,
†††m-kobayashi@mail.nsp.co.jp

あらまし　電子メールにより感染を拡大するコンピュータウィルスのモデルを提案する．まず，感染率が一定とし，電子メールによるウィルスの拡散過程を差分方程式でモデル化を行い，この差分方程式を連続化することによりリカッチ方程式を得る．リカッチ方程式において厳密解を持つ差分方程式を用いて，初期値が最終的な感染数に与える影響，ロジスティック曲線になる条件を示す．このモデルを用いて，現実のコンピュータウィルスの拡散過程のデータの分析を行い，モデルがデータをよく記述していることを示す．さらに，データの曲線から類推される他のモデルとの比較を赤池情報量基準 (AIC) により行い，本論提案モデル優位性を示す．また，新たに感染率が変動するモデルを提案し，このモデルが厳密解を持つ差分方程式で表されることを示す．

キーワード　コンピュータウィルス，ワームの感染拡大，電子メール，リカッチ方程式，差分方程式，厳密解

An E-mail Virus Diffusion Model

Daisuke SATOH[†], Masato UCHIDA[†], Keisuke ISHIBASHI[††], and Makoto KOBAYASHI[†††]

† NTT Service Integration Laboratories
†† NTT Information Sharing Platform Laboratories
3-9-11 Midori-cho, Musashino-shi, Tokyo, 180-8585 Japan
††† Nihon Software Products
3-18-6 Higashigotanda, Shinagawa-ku, Tokyo, 141-0022 Japan
E-mail: †{satoh.daisuke,uchida.masato}@lab.ntt.co.jp, ††ishibashi.keisuke@lab.ntt.co.jp ,
†††m-kobayashi@mail.nsp.co.jp

Abstract　We propose a model of computer viruses that infect other computers by using e-mail. We assume that infection-ratio is constant. We obtain an discrete equation by modeling the computer viruses that infect other computers by using e-mail. The discrete equation is reduced to the Riccati equation under a limit where the time interval equals 0. We analyze the effect of the initial number of infected computers to the total number of computers infected in the infinitely long duration by using a discrete equation that has an exact solution. We also analyze the condition that the infected curve becomes the logistic curve. We show that this model describes actual data of computer viruses. Akaike's information criteria (AIC) shows that the model is superior to a linear model. Finally, we propose another model whose infect ratio is varying. We propose the equation of the model is a discrete equation that has an exact solution.

Key words　Computer virus, Worm propagation, E-mail, Riccati equation, Discrete equation, Exact solution

1. はじめに

コンピュータネットワークが普及するに伴い，コンピュータウィルスの脅威は深刻なものになっている．それに伴い，コンピュータウィルスについて様々なアプローチからの研究がなされている．Kephart ら [4] ～ [6] は，コンピュータウィルスの振る舞いを理解するためにはミクロレベル，マクロレベル双方での理解が必要であるが，ミクロレベルの振る舞いの研究が数多くなされている一方，マクロレベルの研究はなかなか進んでいないと指摘している．昨今，コンピュータウィルスと実際のウィルスとの類似性から疫学を応用した研究 [8], [11] がなされている．Staniford ら [8] は，Code Red の感染過程をロジスティック方程式によりモデル化しており，実データと符合することを示している．Code Red は，ランダムに発生させられた IP アドレスに対して攻撃を仕掛けるもので，電子メールにより感染を拡大するウィルスとは感染の方法が異なる．そのため，電子メールにより感染を拡大するウィルスには，ロジスティック曲線モデルとは異なるモデルが適合すると思われる．電子メールにより感染を拡大するウィルスの研究には，Zou ら [12] の研究があり，ユーザの行動を考慮したモデルを構築し，トポロジーによる影響などを分析しているが，シミュレーションによる分析で数理モデルの構築には至っていない．

本論では，コンピュータウィルスの感染方法に注目し，電子メールを使った感染方法についてウィルスの拡散過程の数理モデルの構築を行う．また，実データによる比較検証を行う．

2. ネットワーク型とメール型モデル

鈴木 [9] によれば，コンピュータウィルスは，大きくファイルに感染するもの，ファイルに感染しないものの 2 種類に分類され，ファイルに感染しないものは，さらにトロイの木馬，コンピュータワームの 2 つに分類される．現在，特に問題視されているのは，コンピュータワームに分類されるものである．

コンピュータワームには，電子メールによって感染していく電子メール型と電子メールとは異なる通信プロトコルによって感染していくネットワーク型がある [9]．本論では，この電子メール型とネットワーク型の感染形態のモデル化について議論する．

ネットワーク型のコンピュータウィルスのひとつである CodeRed の拡散がロジスティック方程式

$$\frac{dM(t)}{dt} = \alpha M(t)(N - M(t)) \tag{1}$$

$$M(t) = \frac{N}{1 + m\exp(-\alpha N t)} \tag{2}$$

によってモデル化されることは既に報告されている [8]．このことは，感染した全てのノードが継続的に未感染ノードの感染に寄与していることを意味している．

一方，電子メール型では，感染したノードのアドレス帳を元に感染を拡大していくもので，感染したノードは，アドレス帳に載っている他のノードにウィルスを感染した後は，継続的に未感染ノードの感染に寄与しない．そこで，次のような差分方程式によるモデルを提案する．

$$M_{n+1} - M_n = \delta\alpha(N - M_n)(M_n - M_{n-1}) \tag{3}$$

ここで，M_n は n ステップ時における既感染ノード数であり，N はネットワーク内の全ノード数である．このモデルは，ロジスティック方程式を前進差分した式

$$M_{n+1} - M_n = \delta\alpha(N - M_n)M_n \tag{4}$$

と比較してわかるように，新規感染ノード数がロジスティック方程式の場合，既に感染したノード数と未感染ノード数との積に比例するのに対して，本提案モデルは，未感染ノード数と直近に感染したノード数の積に比例するモデルになっている．このことは，メール型ウィルスの感染形態をモデル化したものであると言える．

3. 微分方程式とその解

式 (3) において，

$$k = N - \frac{1}{\delta\alpha} \tag{5}$$

とおき，両辺から $M_n - M_{n-1}$ を引いて，両辺を δ^2 で割ると

$$\frac{M_{n+1} - 2M_n + M_{n-1}}{\delta^2} = \alpha(k - M_n)\left(\frac{M_n - M_{n-1}}{\delta}\right) \tag{6}$$

となる．ここで $\delta \to 0$ とすると次の微分方程式

$$\frac{d^2M(t)}{dt^2} = \alpha(k - M(t))\frac{dM(t)}{dt} \tag{7}$$

が得られる．ここで

$$K = \sqrt{k^2 + \frac{2C_1}{\alpha}} \tag{8}$$

とおくと，この微分方程式の解は

$$M(t) = k + K\tanh\left(\frac{\alpha K}{2}(t + C_2)\right) \tag{9}$$

となる．ここで，C_1, C_2 は積分定数である．解は，以下のような極限値をとる．

$$\lim_{t\to\infty} M(t) = k + K \tag{10}$$

$$\lim_{t\to-\infty} M(t) = k - K \tag{11}$$

また，式 (7) は，

$$\frac{d^2M(t)}{dt^2} = \frac{d}{dt}\left(-\frac{\alpha}{2}(M(t) - k)^2\right) \tag{12}$$

と書き直せるので，両辺積分して

$$\frac{dM(t)}{dt} = \frac{\alpha}{2}\left(K^2 - (M(t) - k)^2\right) \tag{13}$$

となる．式 (13) は定数係数の Riccati 方程式である．

4. 厳密解を持つ差分方程式

実データによる評価を行うためには，パラメータ推定をする必要がある．式 (13) を基にパラメータ推定を行う．Riccati 方程式の精度の高いパラメータ推定については，厳密解を持つ差

分方程式[1],[3]によるパラメータ推定法が提案されている[7]. 式(13)に対応する厳密解を持つ差分方程式は

$$M_{n+1} - M_n = \delta\alpha\left(K^2 - (M_{n+1} - k)(M_n - k)\right) \quad (14)$$

であり，厳密解は，

$$M_n = k + K\frac{m\left(\frac{1-\delta\alpha K}{1+\delta\alpha K}\right)^{-\frac{n}{2}} - \frac{1}{m}\left(\frac{1-\delta\alpha K}{1+\delta\alpha K}\right)^{\frac{n}{2}}}{m\left(\frac{1-\delta\alpha K}{1+\delta\alpha K}\right)^{-\frac{n}{2}} + \frac{1}{m}\left(\frac{1-\delta\alpha K}{1+\delta\alpha K}\right)^{\frac{n}{2}}} \quad (15)$$

となる．式(15)より

$$\lim_{n\to\infty} M_n = k + K \quad (16)$$

$$\lim_{n\to-\infty} M_n = k - K \quad (17)$$

が確認でき，微分方程式の解と対応がつく．

ちなみに，式(3)，式(7)に対応する厳密解を持つ差分方程式はそれぞれ以下のようになる．

$$M_{n+1} - M_n = \delta\alpha(k + \frac{1}{\delta\alpha} - M_n)\left(\frac{M_{n+1} - M_{n-1}}{2}\right) \quad (18)$$

$$\frac{1}{\delta^2}(M_{n+1} - 2M_n + M_{n-1}) = \alpha(k - M_n)\left(\frac{M_{n+1} - M_{n-1}}{\delta}\right) \quad (19)$$

初期値を M_0, M_1 として与えたときに飽和値 $k+K$ は以下の式のようになる．

$$k + K = k + \sqrt{(k - M_1)(k - M_0) + \frac{1}{\delta\alpha}(M_1 - M_0)} \quad (20)$$

M_0, M_1 は，式(20)の根号内が非負かつ $M_1 - M_0 > 0$ より以下の条件を満たす必要がある．

$$0 < M_1 - M_0 \leqq \frac{1}{\delta\alpha} \quad (21)$$

5. ロジスティック曲線になる条件

式(18)がロジスティック曲線になる条件は，

$$K = k \quad (22)$$

であるから，ロジスティック曲線になるとき，初期値に関して以下の関係式

$$M_1 = \frac{M_0(k + \frac{1}{\delta\alpha})}{k - \frac{1}{\delta\alpha} - M_0} \quad (23)$$

が成り立つ必要がある．

これと先ほどの初期値 M_0, M_1 が満たす条件を合わせ，さらに $M_0 > 0$ を合わせると

$$0 < M_0 \leqq \frac{-3 + \sqrt{5 + 4k\delta\alpha}}{2\delta\alpha} \quad (24)$$

となる．式(23)と式(24)の両方が成り立つときにロジスティック曲線モデルになる．

微分方程式でも同様の条件を出すことは可能であるが，微分方程式では，初期値 $M(t_0)$ と拡散の速度 $\frac{dM(t)}{dt}|_{t=t_0}$ との条件になる．通常，データは離散的に与えられるため，拡散の速度は近似的に与えるしかなくなる．それに対して，厳密解を持つ差分方程式では，式(23)のように，条件は，2つの時点のデータ値によって与えられるため，正確に求めることができる．

6. パラメータ推定法

提案モデルが，実際のコンピュータウィルスの拡散に対して良いモデルになっているかどうかを調べる．まず始めに，実データに対して，モデルのパラメータを推定する必要がある．Riccati方程式のパラメータ推定について，厳密解を持つ差分方程式を利用したパラメータ推定法[7],[10]が提案されており，既存の方法よりも精度が高いことが示されている．

式(14)から次のような回帰式を導く．

$$Y_n = A + B(M_n + M_{n+1}) + CM_nM_{n+1} \quad (25)$$

ここで，

$$Y_n = M_{n+1} - M_n \quad (26)$$

である．回帰分析により得られた A, B, C の各推定値 $\hat{A}, \hat{B}, \hat{C}$ からパラメータは次のように求まる．

$$K = -\frac{\sqrt{\hat{B}^2 - \hat{A}\hat{C}}}{\hat{C}} \quad (27)$$

$$k = -\frac{\hat{B}}{\hat{C}} \quad (28)$$

$$\delta\alpha = -\hat{C} \quad (29)$$

$$m = \left(\sum_{i=0}^{n}\frac{K - k + M_i}{K + k - M_i}\right) \Big/ \left(\sum_{i=0}^{n}\left(\frac{1-\delta\alpha K}{1+\delta\alpha K}\right)^{-\frac{i}{2}}\right) \quad (30)$$

7. コンピュータウィルス：Aliz

Alizは，既知のセキュリティホールを利用して，メールを大量送信するワームであり，具体的には，OutlookExpressのアドレス帳に登録されているすべてのアドレス宛にウィルスを添付したメールを送信することによって感染を拡大させていく．つまり，感染方法が，本論で扱う感染方法と一致している．

7.1 実データによる評価

実データは，メール型ウィルスであるAlizのデータ(データ提供：トレンドマイクロ社)を使用した．式(25)の回帰式からパラメータを求めると

$$K = 1355.4 \quad (31)$$

$$k = -455.76 \quad (32)$$

$$\delta\alpha = 1.2402E-4 \quad (33)$$

$$m = 1.2170 \quad (34)$$

となる．これらのパラメータの値を式(15)の厳密解に代入して得られた累積感染数の推定値と実データとの比較を図1に示す．図1からわかるように，実データは，明らかにロジスティック曲線モデルとは異なる成長曲線である．提案モデルによる推定値は，実データと良く適合していることがわかる．

7.2 赤池情報量基準によるモデル評価

図1の形状から次のような線形微分方程式モデルが考えられる．

$$\frac{dM}{dt} = r(K + k - M(t)) \quad (35)$$

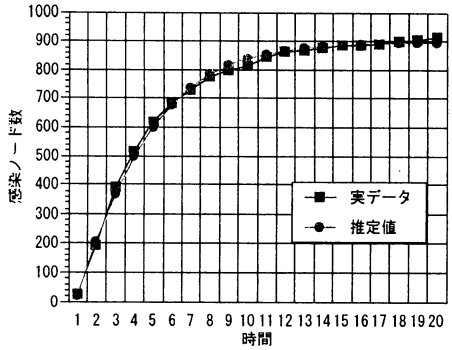

図 1　実データとモデルによる推定値との比較

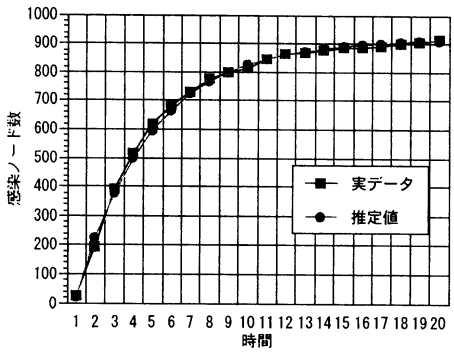

図 2　実データとモデルによる推定値との比較

厳密解は,

$$M(t) = (K+k)\,(1-\mu\exp(-rt)) \tag{36}$$

である．式 (35) を差分化して

$$M_{n+1} - M_n = 2\delta r\left(K + k - \frac{1}{2}(M_n + M_{n+1})\right) \tag{37}$$

を得る．式 (35) は線形微分方程式なので，その差分方程式の厳密解は容易に導けて

$$M_n = K + k - \mu\left(\frac{1-\delta r}{1+\delta r}\right) \tag{38}$$

となる．差分方程式 (37) から以下のような回帰式を得る．

$$Y_n = \alpha + \beta(M_n + M_{n+1}) \tag{39}$$

ここで,

$$Y_n = M_{n+1} - M_n \tag{40}$$

である．回帰分析により得られた α, β の各推定値 $\hat{\alpha}, \hat{\beta}$ からパラメータは次のように求まる．

$$K + k = -\frac{\hat{\alpha}}{2\hat{\beta}} \tag{41}$$

$$r = -2\hat{\beta} \tag{42}$$

$$\mu = \frac{\sum_{i=0}^{n}(N - M_i)}{\sum_{i=0}^{n}\left(\frac{1-\delta r}{1+\delta r}\right)^i} \tag{43}$$

式 (39) を使って Aliz のデータを回帰分析すると

$$K + k = 917.45 \tag{44}$$

$$\delta r = 0.12664 \tag{45}$$

$$\mu = 1154.9 \tag{46}$$

となり，実データとモデルによる推定値との比較結果は，図 2 のようになる．

表 1　AIC の比較

モデル	AIC
提案モデル	155.90
線型モデル	158.22

Aliz のデータに対して，式 (25) の提案モデルと式 (39) の線形微分方程式モデルの AIC(赤池情報量基準) を比較し結果を表 1 に示す.

図 1 の形状から，式 (35) の微分方程式を基にした式 (39) のモデルが予想されるが，表 1 から式 (13) の Riccati 方程式を基にした式 (25) のモデルの方が実データを記述するのに良いモデルであることがわかる．さらに，式 (25) のモデルは，ウィルスの感染拡大の仕方を表現している．

8.　コンピュータウィルス：Sircum

Sircum は，添付ファイルにウィルス以外のファイルと結合し，偽装するという特徴を持つワームである．感染方法としては，アドレス帳ばかりでなく Internet Explorer のキャッシュファイルからもメールアドレスを抽出し，上記ファイルを添付しメールを大量送信する．受信側では，この添付ファイルを開くことにより感染し，感染したコンピュータから同様にしてメールを大量送信して感染を拡大するものである．ただし，Aliz とは異なり，メールのみによって感染を拡大するわけではなく，LAN 上のあるコンピュータが感染すると，LAN 上の他のコンピュータにも感染を拡大させる特徴を持つ.

8.1　実データによる評価

実データは，メール型ウィルスである Sircum のデータ (データ提供：トレンドマイクロ社) を使用した．式 (25) の回帰式からパラメータを求めると

$$K = 515.16 \tag{47}$$

$$k = 252.04 \tag{48}$$

$$\delta\alpha = 2.3538E-4 \tag{49}$$

$$m = 1.9410 \tag{50}$$

となる．これらのパラメータの値を式 (15) の厳密解に代入して得られた累積感染数の推定値と実データとの比較を図 3 に示す．図 3 より Aliz 同様，Sircum でも推定値は実データに良く適合していることがわかる．実データは，S 字曲線を示してはいるが，変曲点に対して対象ではなく，ロジスティック曲線モデルとは異なる曲線を示していることが分かる.

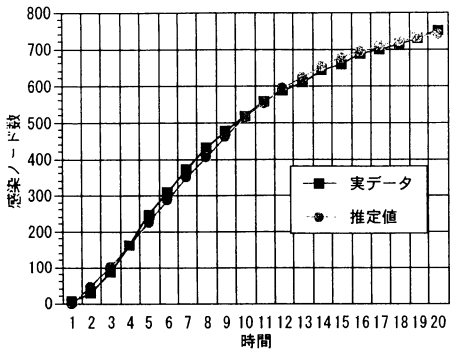

図 3　実データとモデルによる推定値との比較

8.2 AIC によるモデル評価

Aliz 同様に，図 3 の形状から式 (35) の線形微分方程式モデルが考えられる．Aliz のときと同様に差分方程式 (37) およびそれを用いた回帰式 (39) をあてはめてみる．回帰分析により求められたパラメータは以下の通りである．

$$K + k = 1049.1 \tag{51}$$

$$\delta r = 3.3791E - 2 \tag{52}$$

$$\mu = 1104.82 \tag{53}$$

これらパラメータの値と式 (38) から推定値を求め，実データと比較したものを図 4 に示す．

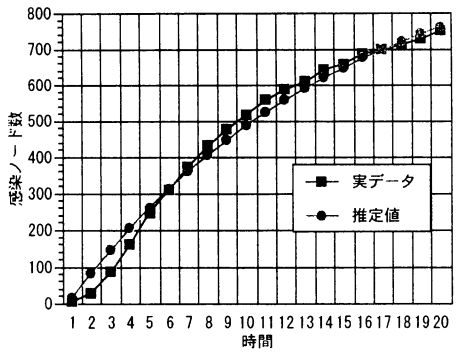

図 4　実データとモデルによる推定値との比較

Sircum のデータに対して，式 (25) の提案モデルと式 (39) の線形微分方程式モデルの AIC を比較し結果を表 2 に示す．Sircum についても，表 1 から式 (35) の微分方程式を基にした

表 2　AIC の比較

モデル	AIC
提案モデル	154.08
線型モデル	163.58

式 (39) のモデルよりも式 (13) の Riccati 方程式を基にした式 (25) のモデルの方が実データを記述するのに良いモデルであるといえる．

9. 感染率変動モデル

コンピュータウィルス用ワクチンソフトの使用により，感染率は時間経過と共に低下することが考えられる．今までのモデルは，感染率が一定としていたため，これを表現できない．そこで，感染率の変化を表現可能なモデルの提案を行う．A_n を n に応じて変化する値として差分方程式は

$$M_{n+1} - M_n = \delta A_n (k + \frac{1}{\delta A_n} - M_n) \left(\frac{M_{n+1} - M_{n-1}}{2} \right) \tag{54}$$

$$M_{n+1} - M_n = \delta A_n \left(K^2 - (M_{n+1} - k)(M_n - k) \right) \tag{55}$$

となる．式 (55) の厳密解は，

$$M_n = k + K \frac{mP - \frac{1}{mP}}{mP + \frac{1}{mP}} \tag{56}$$

$$P = m \prod_{i=0}^{n} \left(\frac{1 - \delta A_i K}{1 + \delta A_i K} \right)^{-\frac{1}{2}} \tag{57}$$

となる．式 (56) により，ワクチンソフトなどにより感染率が時間と共に低下したときに感染ノード数がどのように変化するかを分析することができる．

10. ま と め

本論では，メール型ウィルスの感染方法をモデル化し，その感染方法に基づき差分方程式モデルを提案し，その差分方程式の時間を連続化することにより Riccati 方程式が導き出されることを示した．厳密解を持つ差分方程式を用いて，感染過程がロジスティック曲線で表されるときの条件を明らかにした．Riccati 方程式が実データをうまく記述できているかを調べるために，厳密解を持つ差分方程式を用いて，パラメータ推定を行い，Riccati 方程式が実データをうまく記述できていることを確認した．さらに，実データのグラフ形状から類推される線形微分方程式モデルと提案モデルとを AIC で比較し，提案モデルの方が優れたモデルであること示した．最後に，感染率が変動する場合のモデルを提案し，差分方程式およびその厳密解を示した．これらのモデル化により最終的な感染ノード数の予測や感染率の推定，感染率の変化が感染ノード数に与える影響などの分析が可能となる．

文　献

[1] R. Hirota: Nonlinear partial difference equations. V. Nonlinear equations reducible to linear equations. *Journal of the Physical Society of Japan*, **46** (1979) 312–319.

[2] R. Hirota, Conserved Quantities of "Rondom-Time Toda Equation", J. Phys. Soc. Jpn, **66** (1997) 283–284.

[3] 広田: 差分方程式講義 -連続より離散へ-, サイエンス社, 2000.

[4] J. O. Kephart and S. R. White: Directed-graph Epidemiological Models of Computer Viruses, *Proceedings of the IEEE Symposium on Security and Privacy*, (IEEE, Oakland, May 20–22, 1991) 343–359.

[5] J. O. Kephart, S. R. White and D. M. Chess: Computers and epidemiology, *IEEE Spectrum*, **30-5** (1993) 20–26.

[6] J. O. Kephart and S. R. White: Measuring and Modeling Computer Virus Prevalence, *Proceedings of the IEEE Symposium on Security and Privacy*, , (IEEE, Oakland, May

24–26, 1993) 2–15

[7] D. Satoh: A Discrete Bass Model and its Parameter Estimation, *Journal of the Operations Research Society of Japan*, **44-1** (2001) 1–18.

[8] S. Staniford, V. Paxson, N. Weaver: How to Own the Internet in Your Spare Time, *Proceedings of the 11th USENIX Security Symposium*, (San Francisco , 2002) 149–167.

[9] 鈴木光勇: なぜコンピュータウィルスは悪さが出来るのか？, 毎日コミュニケーションズ (2003).

[10] 山田, 井上, 佐藤: ソフトウェア信頼性評価のための差分方程式に基づく統計的データ解析, 日本応用数理学会論文誌, **12-2** (2002) 155–168.

[11] C. C. Zou, W. Gong, and D. Towsley: Worm Propagation Modeling and Analysis under Dynamic Quarantine Defense, *Proceedings of the 2003 ACM Workshop on Rapid Malcode (WORM)*, (ACM, Washington DC, Oct 27, 2003) 51–60.

[12] C. C. Zou, D. Towsley, and W. Gong: Email Virus Propagation Modeling and Analysis, *Umass ECE Technical Report TR-03-CSE-04*, May, 2003.