コンピュータウィルスの動的拡散過程の観測とそのモデル化

Obeservation and Modeling Method of Dynamics of Computer Virus Spread

豊泉　洋 [1]　　海和　建太 [2]
Hiroshi Toyoizumi　　Kaiwa Tatehiro

早稲田大学 [1]　　会津大学 [2]
Waseda University　　University of Aizu

1　まえがき

メールに添付される形で増殖するコンピュータウィルスやワームは、アンチウィルスソフトの導入や一般ユーザーへの啓蒙活動などの防御活動にもかかわらず増加の傾向にある。

本論文では、発信源の PC に保存されているアドレスを検索し、次の感染先を探し出すといったメール型のコンピュータウィルスの特徴を踏まえ、コンピュータウィルスの動的拡散過程をモデル化する。特に、メールによるコミュニケーションのネットワークを scale-free network[1, 2] としてモデル化し、同時に、実際のコンピュータウィルスの拡散状況をモニターして得られた観測結果をこのモデルに基づいて考察する。

コンピュータウィルス等の malicious mobile code のインターネット上での拡散を分析、防御する研究としては [3, 4, 5, 6, 7] などがすでにある、また、メールのネットワークを scale-free network としてモデル化しているものもあるが [8, 10]、社内ネットワークのアドレスブックを調査するなど限定的であり、ウィルスがそのようなネットワーク上、実際にどのような形で拡散しているのかを実証的に調査し、モデル化している研究は少ない。

本論文では、メールネットワークを scale-free network としてモデル化して得られた知見と会津大学で実際に得られたコンピュータウィルスの活動状況を比較し、scale-free network でのメール型のコンピュータウィルスの特性を明らかにする。

2　Scale-Free Network と Outbreak

Scale-free network とは、各ノードのリンク数が特定の分布（例　指数分布、ベキ法則分布）に従うようなネットワークであり、近年インターネット、感染症の伝染、遺伝子のネットワークなどが、scale-free network に従うことが実証され、盛んに研究がなされている [1, 2, 9]。ここでは [10] のモデル化に従い、メールネットワークをモデル化する。

個人のメールアドレスをノードとして考え、同一のマシン上に保存されているメールアドレスにはリンクが張られると考え、メールアドレスの空間をネットワーク化し、そのネットワーク上をコンピュータウィルスが拡散していくと考える。各ノードのリンク数は、それぞれ独立で、同一な確率分布に従うと仮定する。ネットワーク上でランダムに選ばれたノードのリンク数を K とする。さらに、ランダムに選ばれたリンクの先につながっているノードのリンク数を K_e とする。リンクの多いノードは、そのリンク数に比例して選ばれやすいため、K_e の分布は K の分布を用い、次式で与えられる。

$$P\{K_e = k\} = \frac{kP\{K = k\}}{E[K]}. \tag{1}$$

特にその期待値は、

$$E[K_e] = \frac{E[K^2]}{E[K]} \tag{2}$$

となる。ベキ分布のような裾の長い分布では、$E[K_e] \geq E[K]$ となり、裾が長くなるに従い、$E[K_e]$ は大きくなることが知られている。

ノードがコンピュータウィルスに汚染された場合に、リンクごとに確率 p でリンク先にあるノードが感染すると仮定し、ランダムに選んだノードから感染が始まった場合の outbreak のサイズ（最終的な感染ノード数）を S とし、その大きさを評価する。この問題は、物理学では percolation 問題として知られている [2, 10]。

ここで、ランダムにリンクを選んだ場合の outbreak のサイズを S_e とする。ネットワークのサイズが十分大きく、感染のループができないと仮定すると、S と S_e の関係は次のように表すことができる。

$$S = 1 + \sum_{m=0}^{M} S_{e,m} \tag{3}$$

但し、M は最初に選んだノードの汚染されたリンク数で、$S_{e,m}$ は S_e と同一で独立な分布である。一方、最初に選んだリンクに接続されたノードの感染リンク数が M_e の場合には、次のような再帰式が得られる。

$$S_e = 1 + \sum_{m=0}^{M_e-1} S_{e,m}. \tag{4}$$

(3) と (4) の両辺の期待値をとって、Wald の方程式 [11] を使い、適当に整理すると、outbreak のサイズの期待値が次のように与えられる。

$$\begin{aligned} E[S] &= 1 + E[M]E[S_e] \\ &= 1 + \frac{E[M]}{2 - E[M_e]} \\ &= 1 + \frac{pE[K]}{2 - pE[K_e]}. \end{aligned} \tag{5}$$

ここで、$E[K_e]$ は (2) で与えられる。(5) より $E[M_e] = pE[K_e] = pE[K^2]/E[K] = 2$ のときに、outbreak サイズの期待値 $E[S]$ は発散することがわかる。これは、ちょうど感染拡大が起こる条件と一致する。また、裾の長いベキ分布のようなリンク数分布を持つ場合には、感染率 p が $p < 2/E[K]$ であっても、$p \geq 2/E[K_e]$ となり、感染爆発が起こりうることを示している。

次に、感染が広がるスピードについて考える。時刻 t における感染ノード数を $S(t)$ とすると、平均的には次のような指数的な感染拡大が予想される。

$$E[S(t)] = \{E[M_e] - 1\}^{\alpha t}. \qquad (6)$$

ここで、α はある正定数である。この場合も、$E[M_e] = pE[K_e] = pE[K^2]/E[K]$ によって感染拡大の様子が劇的に変化することがわかる。

3　感染拡大の観測

最近のウィルスは from フィールドを偽装し、感染源を特定させないようにしているため、どのアドレスが感染源であるかは容易に特定できない。しかし、感染しているメールの到着間隔を観測することで、メールネットワークの中で、どの程度感染拡大が起こっているかを知ることができる。

図 1 は、典型的な 3 種類のウィルスに感染したメールの到着間隔を会津大学のゲートウェーで観測したデータである。観測ウィルスと観測日については表 1 参照していただきたい。到着間隔が短くなるということは、それだけ多くのメールアドレスが感染していること、観測ネットワークとリンクがつながっているノードが感染しはじめていることを示している。

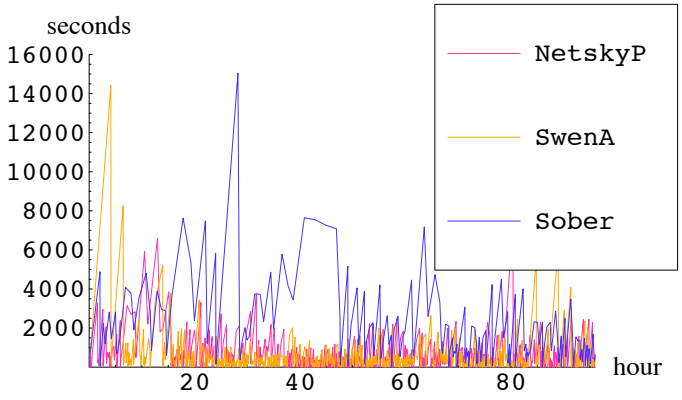

図 1　コンピュータウィルスの到着間隔の時間経過。会津大学のゲートウェイで観測されたウィルス付きのメールの到着間隔の記録より解析した。

表 1　観測ウィルス

ウィルス名	発見日 (米国時間)	会津大初到着時間
Netsky.P	2004/03/21	2004/03/23 01:37:20
SwenA	2003/09/18	2003/09/18 23:07:31
Sober.I	2004/11/19	2004/11/19 21:11:56

3 つのウィルスは、まったく別の時期に発生しているが、類似した感染拡大の経過をたどっている。このことから、同じメールネットワークを経由してきていることが予想できる。どのウィルスも、最終的には、ほぼ同じ程度のウィルスの到着間隔になっていることから、同程度のウィルス感染率を持つことも予想される。また、(6) から予想されるように、感染数は指数関数的に増え、それに伴い到着間隔は指数関数的に減少することが予想される。実際、図 1 より Sober.I（このウィルスは感染数のサンプルが少ない）以外については、指数関数的な減少が見られる。

4　今後の研究の方向性

今後は、観測された感染拡大の様子と scale-free network での感染拡大を比較することにより、実メールネットワークの構造、特に感染を支配するパラメータ $E[M_e] = pE[K_e] = pE[K^2]/E[K]$ を推定し、ユーザーへの影響を最小限に押さえるようなウィルスの効率的な撃退戦略の研究に役立てる予定である。

参考文献

[1] Albert-Laszlo Barabasi and Reka Albert. *Science*, Vol. 286, No. 5439, pp. 509–512, 1999.

[2] Albert-Laszlo Barabasi. Plume, 2003.

[3] Carey Nachenberg. *Commun. ACM*, Vol. 40, No. 1, pp. 46–51, 1997.

[4] Prabhat K. Singh and Arun Lakhotia. *SIGPLAN Not.*, Vol. 37, No. 2, pp. 29–35, 2002.

[5] Harold Thimbleby, Stuart Anderson, and Paul Cairns. *The Computer Journal*, Vol. 41, No. 7, pp. 445–458, 1998.

[6] K. G. Anagnostakis, M. B. Greenwald, S. Ioannidis, A. D. Keromytis, and D. Li. In *Proceedings of the 11th IEEE International Conference on Networks (ICON'03)*. IEEE, October 2003.

[7] Hiroshi Toyoizumi and Atsuhi Kara. In *Proceedings of the 2002 workshop on New security paradigms*, pp. 11–17. ACM Press, 2002.

[8] Justin Balthrop, Stephanie Forrest, M. E. J. Newman, and Matthew M. Williamson. *Science*, Vol. 304, No. 5670, pp. 527–529, 2004.

[9] Stephanie Forrest M. E. J. Newman and Justin Balthrop. *Phys. Rev.*, Vol. E 66, p. 035101, 2002.

[10] M E J Newman. *Physical Review E*, Vol. 66, p. 016128, 2002.

[11] Sheldon M. Ross. *Applied Probability Models With Optimization Applications*. Dover Pubns, 1992.