第35卷 第21期 计 算 机 工 程 2009年11月
Vol.35 No.21 Computer Engineering November 2009
·安全技术· 文章编号：1000—3428(2009)21—0131—03 文献标识码：A 中图分类号：TP393

无标度网络环境下 E-mail 病毒的传播模型

刘 俊，金 聪，邓清华
(华中师范大学计算机科学系，武汉 430079)

摘 要：提出无标度网络环境下 E-mail 病毒的传播模型。通过对模型的求解，得到 E-mail 用户感染密度随传播率、恢复率和网络平均度变化的计算表达式。实验表明，在反病毒技术未出现前，用户感染密度最终将达到一个稳定状态，并通过实验证明了传播率与网络平均度是影响 E-mail 病毒蔓延的关键性因素。

关键词：无标度网络；E-mail 病毒；用户感染密度；传播率；网络平均度

E-mail Virus Spreading Model in Scale-free Network

LIU Jun, JIN Cong, DENG Qing-hua
(Department of Computer Science, Central China Normal University, Wuhan 430079)

【Abstract】 An E-mail virus spreading model based on the scale-free network is proposed. By mathematical calculations, the density of infected users is obtained, and it is affected by many parameters, e.g., spreading rate, recovery rate and average degree of network, etc. Experiments show that the density of infected users finally becomes a stable state before anti-virus technique appearing. Moreover, the spreading rate and average of degree of network take the key important role in this phase.

【Key words】 scale-free network; E-mail virus; density of infected users; spreading rate; average degree of network

1 概述

目前，电子邮件病毒已成为网络信息安全的重要威胁之一。邮件病毒通过电子邮件的信息交换进行传播，用户一旦感染，强烈的自我复制能力使邮件病毒在 Internet 上迅速蔓延开来，给网络系统造成巨大的破坏。因此，研究邮件病毒的传播特性、建立邮件病毒的传播模型是很有意义的课题。

现有的计算机病毒传播模型是基于流行病毒学理论提出的[1]。在该模型中，用户被分为易感染态 S(Susceptible)、潜伏态 E(Exposed)、感染态 I(Infected)和免疫状态 R(Removed)。SI, SIS, SIR 和 SEIR 等是经典的计算机病毒传播模型[2-3]。2001 年，文献[4]将网络拓扑结构引入这些经典病毒传播模型中，给出了不同网络拓扑环境下的病毒传播特性和反病毒策略的差异性，为人们研究 E-mail 病毒传播模型指出了新的方法和思路。2003 年，文献[5]提出了相对完整的 E-mail 病毒传播模型，研究了 E-mail 病毒在小世界网络、随机图和无标度网络中的传播模型，并通过实验指出，E-mail 病毒传染机制和病毒控制策略均不同于其他计算机病毒。

基于不同网络拓扑结构的 E-mail 病毒传播模型的研究已经取得了很多重要进展[6]，但仍存在一些明显不足，主要有：(1)网络拓扑的均匀性。已有的 E-mail 病毒传播模型通常限制在一个局部范围内，由于用户之间联系相对紧密，因此假设网络具有均匀性，例如小世界网络等。事实上 Internet 上 E-mail 用户构成的逻辑网络是一个无标度网络，因此，更应从无标度网络的角度去研究邮件病毒传播模型。(2)缺少理论依据。现有的研究大多从实验的角度仿真网络拓扑结构，结论多为实验性的，未给出较为具体的数学模型。

针对上述不足，本文提出了无标度网络环境下 E-mail 病毒的传播模型，并给出了用户病毒感染密度的计算表达式。通过进一步分析传播率、恢复率和网络平均度等参数对用户感染密度的影响，给出了 E-mail 病毒在无标度网络中的传播新特性。

2 E-mail 病毒传播模型

2.1 模型的提出

在 Internet 上，当邮件用户感染某个 E-mail 病毒后，该病毒会以附件的形式将其发送给用户地址薄中的所有用户，若这些邻居用户接收到 E-mail 病毒附件并激活它，这些用户将迅速成为中毒状态，如此下去，E-mail 病毒就会在整个网络中蔓延开来。通常在恶性 E-mail 病毒爆发的早期，与之相应的杀毒技术不会很快出现，即用户没有杀毒技术支持，不能让计算机具有免疫功能。可用图 1 所示的状态图来描述这一阶段用户状态的变迁。其中，$S(t)$ 表示 t 时刻网络中健康用户密度；$I(t)$ 表示 t 时刻中毒用户密度。病毒的传播率 α 表示健康用户以速率 α 转变为中毒用户。另外中毒用户或许可以通过某种手段删除病毒，比如用现有的杀毒技术，即中毒用户会以一个较小速率 β 返回到健康状态，称 β 为恢复率。但删除病毒后用户并不具有免疫性。

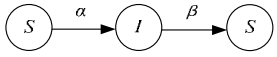


图 1 反病毒技术出现之前用户状态的变迁

上述分析指出了病毒的传播率和恢复率会直接影响 E-mail 病毒的蔓延情况和用户的感染程度。下面从网络的拓

基金项目：湖北省自然科学基金资助项目(2007ABA119)
作者简介：刘 俊(1981－)，男，硕士，主研方向：网络安全，多媒体技术；金 聪，教授、博士；邓清华，硕士
收稿日期：2009-04-08 **E-mail：**jincong26@yahoo.com.cn

扑结构给出进一步的考虑。通常网络中所有邮件用户都会有自己的 E-mail 地址簿，只是每个用户的地址簿大小不同。以 yahoo 邮箱为例，一般用户的地址簿都较小，网络中大部分是这些小地址簿用户；而对于企业用户，如 yahoo 的系统管理员，地址簿能达到几万甚至几十万，若这类用户感染了病毒，病毒副本会以巨大的数目传播出去，与普通用户相比，这类用户对网络的破坏性将是致命的。这样的用户在网络中的比例通常会很小。为反映这种分布不均匀特性对病毒传播的影响，不妨定义 E-mail 用户的地址簿大小为该用户(节点)的度，从而 E-mail 用户之间通过用户地址簿会形成一个逻辑网络。近年来的研究表明，该逻辑网络服从幂率分布，为无标度网络。

在图 2 中度为 k 的用户是健康状态，度为 k_1 和 k_2 的用户是中毒状态($k_1 < k_2$)。假设 k_1 和 k_2 用户与 k 用户均连接，即度为 k 的 E-mail 地址均包括在度为 k_1 和 k_2 的用户中，则 k_1 和 k_2 用户都存在将病毒感染给 k 的可能。而 t 时刻中毒用户 k_1 将病毒感染给用户 k 的概率可以表示为

$$\frac{(k_1-1)p(k_1)i_{k_1}(t)}{\langle k\rangle}$$

其中， $p(k')$ 为无标度网络的度分布函数；$\langle k\rangle$ 为用户的平均度； $i_{k_1}(t)$ 表示 t 时刻网络中所有度 k_1 的用户的感染密度。同理， t 时刻中毒用户 k_2 感染用户 k 的概率为

$$\frac{(k_2-1)p(k_2)i_{k_2}(t)}{\langle k\rangle}$$

然而度为 k 的用户除了有度为 k_1 和 k_2 的邻居用户外，可能还有其他的邻居用户，因此所有这些用户对度为 k 的用户感染的综合效应为

$$\frac{1}{\langle k\rangle}\sum_{k'}(k'-1)p(k')i_{k'}(t)$$

不妨用 $\theta_k(t)$ 表示，$\theta_k(t)$ 可以理解为度为 k 的用户在无标度网络中被感染的概率。

按照上面分析并综合图 1 和图 2，可以获得在反病毒技术出现之前，网络中度为 k 的用户在 t 时刻关于病毒感染密度 $i_k(t)$ 的微分方程式(1)和式(2)，其中， $(1-i_k(t))$ 为未被感染病毒的度为 k 的用户密度。

$$\frac{\mathrm{d}i_k(t)}{\mathrm{d}t}=\alpha k(1-i_k(t))\theta_k(t)-\beta i_k(t) \tag{1}$$

$$\theta_k(t)=\frac{1}{\langle k\rangle}\sum_{k'}(k'-1)p(k')i_{k'}(t) \tag{2}$$

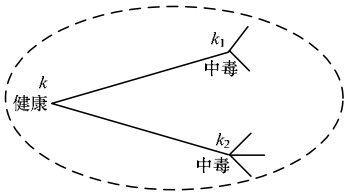


图 2　度分布不均匀的网络示意图

2.2　模型求解

式(1)、式(2)构成关于 $i_k(t)$ 的变系数微分方程组。本文先对式(2)求出关于 t 的导数：

$$\frac{\mathrm{d}\theta_k(t)}{\mathrm{d}t}=\frac{1}{\langle k\rangle}\sum_{k'}(k'-1)p(k')\frac{\mathrm{d}i_{k'}(t)}{\mathrm{d}t} \tag{3}$$

对式(1)，忽略掉 $O(i^2)$ 阶项，得

$$\frac{\mathrm{d}i_k(t)}{\mathrm{d}t}\approx(\alpha k-\beta)\theta_k(t) \tag{4}$$

将式(4)代入式(3)可得

$$\frac{\mathrm{d}\theta_k(t)}{\mathrm{d}t}=\frac{1}{\langle k\rangle}\sum_{k'}(k'-1)p(k')(\alpha k'-\beta)\theta_k(t) \tag{5}$$

即

$$\frac{\mathrm{d}\theta_{k'}(t)}{\mathrm{d}t}=\frac{1}{\langle k\rangle}\sum_{k'}[\alpha k'^2-k'(\beta+\alpha)+\beta]p(k') \tag{6}$$

而 $\sum_k p(k)=1$， $\sum_k kp(k)=\langle k\rangle$ 为网络节点的平均度，同时令 $\sum_k k^2p(k)=\langle k^2\rangle$，则式(5)可表示为

$$\frac{\mathrm{d}\theta_k(t)}{\mathrm{d}t}=\frac{\alpha\langle k^2\rangle-(\beta+\alpha)\langle k\rangle+\beta}{\langle k\rangle}\theta_k(t) \tag{7}$$

由于 α,β 可以视为常量，$\langle k^2\rangle$ 与 $\langle k\rangle$ 在具体网络拓扑中是定值，因此式(7)为 $\theta_k(t)$ 的一阶常系数齐次微分方程，可解得 $\theta_k(t)$ ：

$$\theta_k(t)=c'\mathrm{e}^{[\frac{\alpha\langle k^2\rangle-(\beta+\alpha)\langle k\rangle+\beta}{\langle k\rangle}]t} \tag{8}$$

其中， c' 为常数。

将式(8)代入式(4)可求得 $i_k(t)$ ：

$$i_k(t)=c(\alpha k-\beta)\frac{1}{\varphi}\mathrm{e}^{\varphi t} \tag{9}$$

其中，$\varphi=\dfrac{\alpha\langle k^2\rangle-(\beta+\alpha)\langle k\rangle+\beta}{\langle k\rangle}$ ；c 是常数；$i(t)=\sum_k p(k)i_k(t)$ ；从而

$$i(t)=c(\alpha\langle k\rangle-\beta)\frac{1}{\varphi}\mathrm{e}^{\varphi t} \tag{10}$$

式(9)、式(10)给出了无标度网络中度为 k 的用户的感染密度 $i_k(t)$ 以及所有用户的感染密度 $i(t)$ 的表达式。可以看出，两者均正比于 $\mathrm{e}^{\varphi t}$ 。为了仿真各参数对 $i_k(t)$ 与 $i(t)$ 的影响，网络拓扑采用无标度网络模型，而网络的度分布函数为

$$p(k)=2m^2k^{-3}$$

其中， m 是节点的最小度。有

$$\langle k\rangle=\sum_k kp(k)=\int_m^M kp(k)\mathrm{d}k=2m^2(\frac{1}{m}-\frac{1}{M})$$

$$\langle k^2\rangle=\sum_k k^2p(k)=\int_m^M k^2p(k)\mathrm{d}k=2m^2(\ln M-\ln m)$$

3　仿真实验

3.1　传播率 α 和恢复率 β 的仿真

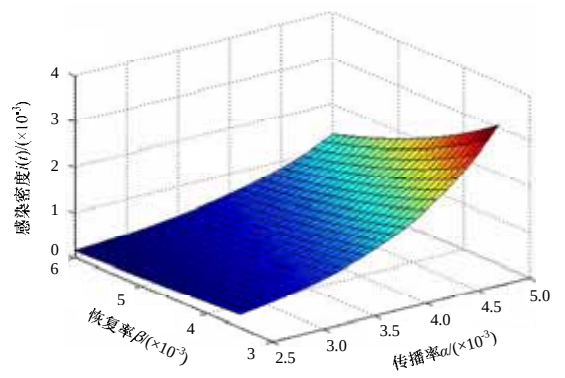
由微分方程式(1)可知，传播率 α 和恢复率 β 两因素共同影响 E-mail 病毒在网络中的传播，图 3 给出了感染用户密度 $i(t)$ 在参数 α 和 β 共同影响下的曲面图。

通过观察可知：

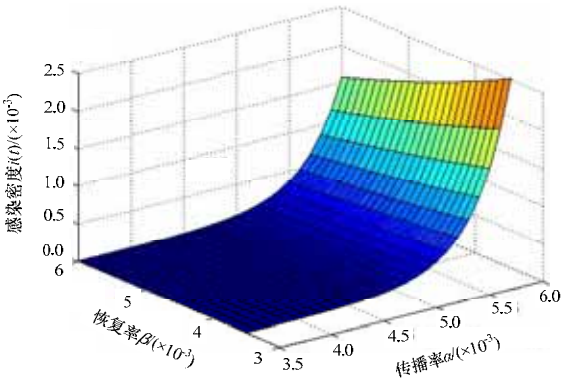
(1)如果 α 越大、 β 越小，则 $i(t)$ 越大，反之 $i(t)$ 越小。该结论表明，传播率越大且恢复率越小将加速病毒蔓延，相反将抑制病毒传播。

(2)在 α 和 β 的相互作用下， α 的增大会使 $i(t)$ 迅速变大，而 β 的增大将使 $i(t)$ 减小的速度要慢得多，即 α 对 $i(t)$ 的影响要比 β 对 $i(t)$ 的影响显著。这表明在反病毒技术还未出现以前，病毒扩散主要受传播率的影响，恢复率对病毒蔓延不构成决定性的影响。

通过上述分析可知，在杀毒软件出现前，仅仅依靠提高恢复率(如迅速删除 E-mail 病毒附件，以及用现有技术杀毒)来控制病毒传播是远远不够的，这进一步表明，当某种恶性 E-mail 病毒爆发后，加速研究对应的反病毒技术是有效控制病毒传播的重要手段。

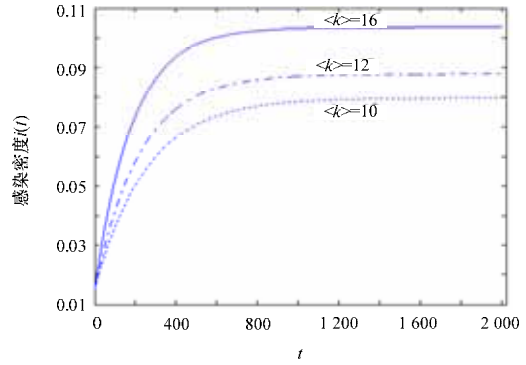


(a)t=50

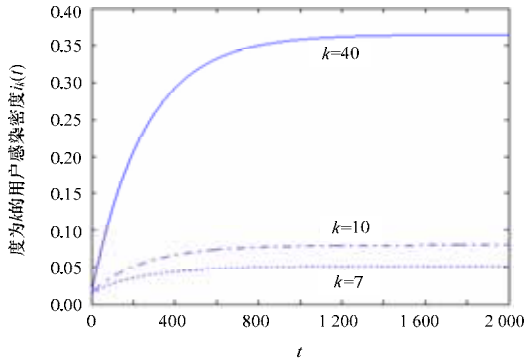


(b)t=200

图 3　参数 α 和β对 $i(t)$的影响情况



(a)不同网络平均度影响下的 $i(t)$曲线



(b)不同度的网络环境下病毒的感染情况

图 4　参数<k>和 k 对用户感染密度的影响

3.2　平均度<k>的仿真

通过观察式(10)可知，网络中所有用户的感染密度 $i(t)$ 除了受传播率 α 和恢复率 β 影响外，还受网络平均度 $\langle k\rangle$ 的影响。为了仿真参数 $\langle k\rangle$ 对 $i(t)$ 的影响，令 $\alpha=0.025,\ \beta=0.015$，图 4(a)给出了不同网络平均度影响下的 $i(t)$ 曲线。可以看出，$i(t)$ 的增长速度经历迅速增大 $\rightarrow$ 逐渐变缓 $\rightarrow$ 保持稳定 3 个阶段。这说明在无标度网络拓扑结构下，病毒传播早期(反病毒技术尚未出现，用户未采取免疫措施)病毒蔓延速度很快，随着病毒继续传播，网络中感染用户数目最终将达到一个稳定水平，而并非感染用户数目一直递增下去，也就是说存在某个时刻 t_o $(t_o>0)$ 使 $\dfrac{\mathrm{d}i(t)}{\mathrm{d}t}=0$。这个重要结论体现了 E-mail 病毒在无标度网络环境下的新特性。

通过对比图 4(a)中的 3 条曲线可知，$\langle k\rangle$ 越大，$i(t)$ 的增长速度越快，同时 $i(t)$ 对应的值也越大。这表明如果用户的 E-mail 地址簿都很大(即各用户之间的联系较紧密)，那么一旦 E-mail 病毒在网络中爆发，病毒就会在网络中迅速传播，这意味着将给整个网络用户造成极大的破坏。因此，E-mail 用户尽量减少邮箱地址簿地址数目对抑制病毒传播有一定的作用。

图 4(b)给出了在具有不同度的网络环境下病毒的感染情况。通过观察可知，网络中所有度为 7 的用户感染密度要远远小于度为 40 的用户感染密度，感染密度达到稳定水平的时刻也相应提前。这表明病毒传播主要集中在度较大的邮件用户(即用户地址簿较大)上，一旦这些用户中了毒，病毒将迅速在网络中传播；另一方面，这些度较大的用户往往也是黑客优先攻击的目标。因此，当某种 E-mail 病毒爆发后，那些地址簿较大的邮件用户应实时更新最新杀毒软件，及时实施杀毒，这能有效地控制病毒的大范围传播。

4　结束语

目前，有关 E-mail 病毒的研究主要将病毒传播所依赖的网络考虑成均匀网络，然而事实上邮件用户所形成的逻辑网络是具备无标度特性的非均匀网络。本文提出了反病毒技术出现前基于无标度网络的 E-mail 病毒传播模型，给出了 $i(t)$ 与 $i_k(t)$ 随时间 t 变化的表达式，并从实验的角度给出了无标度网络环境下 E-mail 病毒传播的如下结论：

(1)如果让其他参数均相同，考察传播率和恢复率，E-mail 病毒的扩散速度主要受传播率影响。也就是说，在反病毒技术出现之前，用户仅仅依靠提高恢复率，如主动删除 E-mail 病毒附件或者通过现有技术删除病毒等来控制病毒传播是不够的，而是传播率起着决定性的作用。因此，E-mail 用户应该从减小传播率的角度去抑制病毒蔓延。

(2)E-mail 用户地址簿的平均大小直接影响着病毒传播的速度和规模。如果所有用户的 E-mail 地址簿都很大，那么病毒将迅速蔓延，这意味着此时病毒传播率较大，地址簿大的用户感染密度要远高于地址簿小的用户。从这个意义上说，病毒爆发后，对于那些地址簿很大的用户应尽可能减少地址簿中用户数目，减少群发 E-mail 信息的次数，及时更行最新杀毒软件，这都将有益于减小传播率，从而有效控制病毒传播。

(3)在反病毒技术出现前，病毒经过一段时间的传播，用户的感染数目最终会达到一个稳定的状态，该结论不同于均匀网络中病毒一直扩散的结论。

本文给出了在反病毒技术出现前的病毒模型，将来的工作是继续提出反病毒技术出现后带免疫策略的 E-mail 病毒传播模型。

（下转第 137 页）