一种无尺度网络上垃圾邮件蠕虫的传播模型*)

王长广[1,2] 王方伟[3] 张运凯[3] 马建峰[1]
(西安电子科技大学计算机学院 西安 710071)[1]
(河北师范大学物理科学与信息工程学院 石家庄 050016)[2]
(河北师范大学网络信息中心 石家庄 050016)[3]

摘 要 用有向图描述了电子邮件网络的结构,并分析了电子邮件网络的无尺度特性。在此基础上,通过用户检查邮件的频率和打开邮件附件的概率建立了一种电子邮件蠕虫的传播模型。分别仿真了电子邮件蠕虫在无尺度网络和随机网络中的传播,结果表明,邮件蠕虫在无尺度网络中的传播速度比在随机网络中更快,与理论分析相一致。

关键词 无尺度网络,幂律,随机网络,垃圾邮件蠕虫,传播模型

A Spam Worm Propagation Model in Scale-free Networks

WANG Chang-Guang[1,2] WANG Fang-Wei[3] ZHANG Yun-Kai[3] MA Jian-Feng[1]
(School of Computer, Xidian University, Xi'an 710071)[1]
(College of Physics Science & Information Engineering, Hebei Normal University, Shijiazhuang 050016)[2]
(Network Center, Hebei Normal University, Shijiazhuang 050016)[3]

Abstract A propagation model of spam worms in scale-free networks is presented and the directed graph is used to describe the e-mail networks. This model can explain the behaviors of email users by the frequency of their checking emails and the probability of their opening email attachments. The simulation results show that spam worms spread more quickly in a scale-free network than in a random network, which is much more important for the study of the defense strategies against spam worms.

Keywords Scale-free netwrok, Power law, Random networks, Spam worms, Propagation model

1 引言

随着 Internet 在人们生活中越来越普及,人们对网络的依赖程度也越来越高。在众多的通信工具中,电子邮件以其方便、快捷、价格低廉等特点已成了广大网络用户的首选。而近年来出现的垃圾邮件蠕虫对我们的攻击强度越来越大,所采用的技术也越来越先进。如今泛滥的垃圾邮件在很大程度上也是由电子邮件蠕虫引起的。近来爆发的比较著名的电子邮件蠕虫有:1999 年的"Melissa",2000 年的"Lover Letter",2001 年的"W32/Sircam",2003 年的"SoBig"系列,2004 年的"Bagle","Netsky","Plexus",2005 年的"MyDoom","Worm_Wurmark. J(蒙面者)"及"Sober"系列等[1]。

"电子邮件蠕虫"是通过电子邮件传播的恶意计算机代码。当邮件用户用鼠标单击蠕虫邮件附件中的蠕虫程序时,蠕虫就会破坏用户的计算机,然后找到存储在用户计算机中的所有邮件地址,并向这些地址分别发送蠕虫邮件[2]。

蠕虫在计算机网络上的蔓延可以看作是服从某种规律的网络传播行为。一个精确的蠕虫传播模型可以使人们对蠕虫有更清楚的认识,能确定其在传播过程中的弱点,而且能更精确的预测蠕虫所造成的损失。目前已有很多学者对蠕虫进行了深入研究,提出了一些模型[3,4]。但这些模型大都是针对基于扫描的蠕虫传播,它们的传播不用考虑网络拓扑结构问题。而电子邮件蠕虫是通过邮件附件传播的,也就是电子邮件蠕虫的传播不需要扫描即可实现,但需要考虑网络的实际拓扑结构问题。因此,基于扫描的蠕虫传播模型不适合描述电子邮件蠕虫的传播。好的邮件蠕虫模型有助于我们更深刻地理解电子邮件蠕虫的传播行为、评估邮件蠕虫防御机制的效率以及对蠕虫潜在的破坏提出早期预警。

2 电子邮件网络的无尺度特性

Holger 等人研究了电子邮件网络的拓扑结构,结果表明电子邮件网络表现出了小世界特性和无尺度行为[4]。Newman 等人收集了一所大学的电子邮件地址簿数据,并证明了校园级电子邮件拓扑具有一个指数形式,并且其入度和出度分别具有延伸的指数分布。但它们都没有考虑电子邮件列表,这个列表能极大地增加电子邮件网络的边数。而且,大部分电子邮件蠕虫都会往计算机中所储存的邮件地址发送蠕虫邮件,而不只是发往邮件地址簿中的邮件地址。

我们采用有向图描述电子邮件网络的拓扑结构:$G=\langle V, E\rangle$,$\forall v\in V$,v 表示一个电子邮件用户或一个电子邮件地址;$\forall e=(u,v)\in E, u,v\in V$,表示用户 u 给用户 v 发送了一封邮件。一个结点的"入度"λ 表示发给该结点邮件的用户有 λ 个;一个结点的"出度"d 表示该结点发出邮件的目的结点数。$|V|$表示电子邮件用户的总数。

Holger 给出了电子邮件网络"结点度"的表达式为,其中结点的"入度"表达式为:

*)基金项目:国家自然科学基金重大计划(No. 90204012);国家 863 高技术研究发展计划(2002AA143021);河北省自然科学基金(F2006000117)。王长广 副教授,博士生,主研方向:网络信息安全;王方伟 讲师,硕士;张运凯 教授,博士;马建峰 教授,博导,CCF 会员。

$$n(\lambda)\propto\lambda^{-1.49} \quad (1)$$

这是一个幂律表达式。虽然结点的“出度”不像“入度”那样容易确定，但它也给出了结点的“出度”的近似表达式，即：

$$n(d)\propto d^{-2.03} \quad (2)$$

由此可见，电子邮件网络表现出了无尺度网络的幂律特性，因此，电子邮件网络属于无尺度网络。尽管实际的电子邮件网络的结点度并不一定是严格幂律分布的，但采用幂律拓扑结构生成器生成电子邮件网络却是一个最佳选择。

目前，“电子邮件组”或“电子邮件列表”已经变得非常流行。如果用户的计算机或邮件地址簿中存有某个电子邮件组的地址，那么从邮件蠕虫的角度来看，这个用户实际上就拥有了该电子邮件组内所有的邮件地址。因此，即使一个用户的计算机可能只包含十几个电子邮件地址，但若其中的一个邮件地址是非常流行的电子邮件组，则用户在上述电子邮件网络中的结点度将会达到几千甚至上万。

3 无尺度网络上电子邮件蠕虫的传播模型

电子邮件蠕虫依靠邮件用户之间的交互来传播。我们认为，有两种主要的用户行为影响电子邮件蠕虫的传播：一个是用户检查电子邮件的频率，表示为 $F_i, i=1,2,\cdots,|V|$，它是用户 i 每天检查电子邮件的次数；另一个是邮件的打开概率，表示为 P_i，即用户 i 打开一个蠕虫附件的概率。

某些电子邮件蠕虫利用了邮件客户端的漏洞，这样不需要用户执行任何附件它们就能破坏用户的计算机。对于那些易感染的计算机来说，可以将这样的邮件蠕虫模型化为 $P_i=1$。

一个用户的电子邮件检查频率是一个由用户习惯决定的随机变量。为了简单起见，我们可以用电子邮件的检查时间间隔 $T_i, i=1,2,\cdots,|V|$ 来表示电子邮件的检查频率 $F_i, i=1,2,\cdots,|V|$，而二者的关系是确定的。我们用 $AVER[T_i]$ 表示用户 i 的检查时间间隔 $T_i(i=1,2,\cdots,|V|)$ 的平均值。T_i 可能遵循不同的分布。例如，当用户每天上午检查一次邮件或以固定的时间间隔用邮件客户程序收取并检查邮件时，T_i 就是一个常量；如果用户随机地检查邮件，那么 T_i 就是一个指数分布，即是说，检查行为是一个泊松过程。

由于因特网上的电子邮件用户数目非常庞大，而用户的行为也是独立的，因此我们假定用户 i 的邮件检查时间间隔 T_i 由高斯分布的随机变量 T 生成，即 $T\sim N(\mu_T,\sigma_T^2)$（当 $T<0$ 时 $AVER[T_i]=0$）。同时假定当某个用户检查他的邮件时，他将检查邮箱中所有的新邮件。

用户 i 打开附件的概率取决于用户的安全意识和邮件蠕虫所采用的技术（例如，由于“MyDoom”采用了先进的技术，它感染的用户数多于以前的任何邮件蠕虫[1]）。因此，对于某个电子邮件蠕虫的传播，我们假定用户 i 的打开概率 P_i 是不变的。类似于 $AVER[T_i]$，我们假定用户 i 的 P_i 由一个高斯分布的随机变量 P 产生，即 $P\sim N(\mu_T,\sigma_T^2)$（当 $P<0$ 时 $P_i=0$，而当 $P>0$ 时 $P_i=1$）。

同时，我们规定，当电子邮件用户打开了邮件蠕虫附件时，用户就被感染；当一个被感染的用户打开一个蠕虫附件时，他就立即往他所有的邻居发送蠕虫邮件。令 N_0 表示初始感染的用户数，令随机变量 N_t 表示蠕虫邮件传播期间 t 时刻所感染的用户数，则对于 $\forall t>0$，有 $N_0\leqslant N_t\leqslant|V|$。感染用户所发出的蠕虫邮件要经过一段时间才能传输到收件人，但邮件的传输时间相对于用户的邮件检查时间间隔来说通常是可以忽略不计的，因此，在我们的模型中忽略了邮件的传输时间。

4 仿真试验及分析

我们用 $AVER[N_t]$ 表示在任意时刻 t 感染用户的平均数。在仿真试验中，我们用不同的种子产生的随机数进行了100次仿真，并对得到的结果 N_t 取平均以求得 $AVER[N_t]$。该幂律网络具有100,000个结点，平均结点度为8，幂律系数 $\alpha=1.7$。除电子邮件检查时间间隔的分布试验外，在其它的试验中我们假定用户 i 的电子邮件检查时间间隔 T_i 遵循比率为 $1/AVER[T_i], i=1,2,\cdots,|V|$ 的泊松过程，其中 $AVER[T_i]$ 遵循 $T\sim N(40,20^2)$。其它的仿真参数为：$P\sim N(0.5,0.3^2)$ 且 $N_0=2$。在每次仿真试验中，初始感染的结点都是随机选择的。

4.1 邮件蠕虫感染情况分析

首先，我们将被感染的用户计算机发送蠕虫邮件的情况划分为以下两种：重复感染与不重复感染。重复感染意味着每当被感染的用户打开蠕虫附件时，都会往其所有的邻居发送一次蠕虫副本；而不重复感染意味着被感染用户只往其所有邻居发送一次蠕虫副本，之后即使用户重复地打开蠕虫附件，它也不再发送蠕虫邮件。图1说明了在幂律拓扑的电子邮件网络上以时间 t 为参数的 $AVER[N_t]$ 曲线。

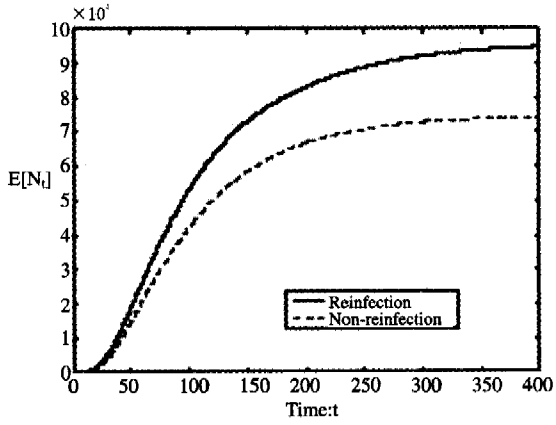

图1 重复感染与不重复感染

在我们的模型中，当用户 i 检查蠕虫邮件时，他打开蠕虫附件的概率为 P_i。因此当用户 i 收到 n 个蠕虫邮件时，他被感染的概率为 $1-(1-P_i)^n$，这就是图1所示的曲线中重复感染的用户比不重复感染的多得原因。

我们用 N_∞^h 表示蠕虫传播结束时未被感染的用户数。在不重复感染的情况下，具有 n 条边（邻居）的用户 i 收到蠕虫邮件副本的数目最多为 n_i，即用户 i 未被感染的概率至少为 $(1-P_i)^{n_i}$。令 $G(x)$ 表示电子邮件网络结点度的概率产生函数，则：

$$G(x)=\sum_{k=1}^{\infty}P(d=k)x^k \quad (3)$$

其中 $P(d=k)$ 表示一个结点具有度 k 的概率。当所有用户打开蠕虫附件的可能性相同时，即 $P_i=p, \forall i\in\{1,2,\cdots,|V|\}$，我们可以导出 $AVER[N_\infty^h]$ 的下界如下：

$$AVER[N_\infty^h]\geqslant|V|\sum_{k=1}P(k=d)(1-p)^k=|V|G(1-p) \quad (4)$$

由于重复感染邮件蠕虫的传播速度更快，因此，我们只考虑重复感染邮件蠕虫的传播情况。

4.2 拓扑结构对邮件蠕虫传播的影响

电子邮件网络的拓扑在确定邮件蠕虫传播行为中起着重要的作用。我们分别在幂律拓扑结构和和随机网络拓扑结构上执行了邮件蠕虫的仿真。这两种网络具有相同的平均结点度 8 和 100,000 个节点。图 2 表明了这两种拓扑以时间 t 为参数的 $AVER[N_t]$ 曲线。由此可见，电子邮件蠕虫在幂律上的传播速度更快。

为了研究蠕虫在幂律网络中传播速度更快的原因，我们令 D_t 表示在 t 时刻之前未被感染而在 t 时刻被感染的那些结点的平均度。D_t 告诉我们在每一时刻 t，$t=1,2,3\cdots$ 有哪些结点正在被感染。我们在图 3 中画出了这两种网络拓扑的 D_t，它以时间 t 为参数。

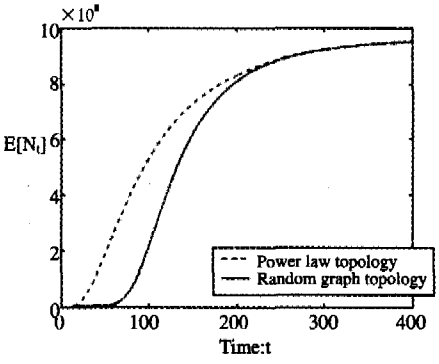

图 2 拓扑结构对邮件蠕虫传播的影响

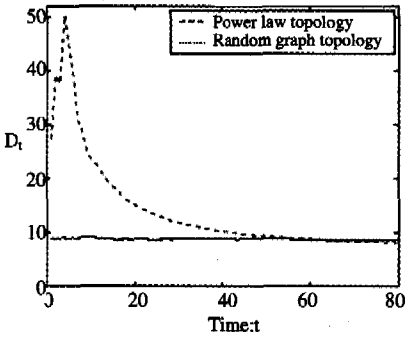

图 3 在每一时刻正被感染的结点的平均度

图 2 清楚地表明，在幂律网络上，电子邮件蠕虫倾向于首先感染连通性最强的结点，这些结点比其它被感染的结点发出的蠕虫邮件副本要多得多。因此，开始时感染速度将会被这些结点放大。由于随机图网络拓扑的所有结点都具有相似的结点度，因此它们都不会表现出这种的放大效果。此外，这两种拓扑具有不同的“特征路径长度”，而幂律拓扑具有更小的特征路径长度[5]。对电子邮件网络来说，更小的特征路径长度意味着被感染的用户到达其它用户所需的跳数更少，即电子邮件蠕虫在幂律拓扑中传播的速度更快。

4.3 用户检查频率对邮件蠕虫传播的影响

下面我们来研究 T_i 的不同分布如何影响电子邮件蠕虫的传播。我们研究了 T_i 的三种分布：指数分布；3 阶尔朗分布；固定的用户电子邮件检查时间间隔。为了便于比较，假定在这三种情况下，$AVER[T_i]$ 遵循相同的高斯分布 $T\sim N(40,20^2)$。

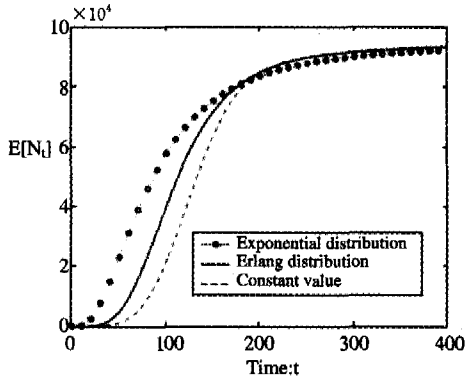

图 4 用户检查频率对电子邮件蠕虫传播的影响

图 4 给出了在幂律电子邮件网络上被感染用户的平均数 $AVER[N_t]$。给定同样的平均值，指数分布比 $k(k>1)$ 阶尔朗分布的变化更随机一些，它们的变化都比固定时间间隔更随机一些。图 4 表明，电子邮件检查时间间隔变化越大，邮件蠕虫的传播就越快。其原因是，在 T_i 变化不大的系统中，当蠕虫副本感染某些新用户时，T_i 变化较大的系统中的蠕虫副本早已对某些用户感染过多次了。

总结 我们给出了一个电子邮件蠕虫的传播模型，并比较了电子邮件蠕虫在无尺度网络和随机网络上的传播，结果表明，电子邮件蠕虫在幂律拓扑结构上比在随机网络拓扑上传播得更快。下一步将研究在校园网环境中如何对电子邮件蠕虫进行静态和动态免疫防御。

参 考 文 献

1 CERT. CERT/CC advisories. http://www.cert.org/advisories/.

2 Newman M, Forrest S, Balthrop J. Email networks and the spread of computer viruses. http://www.cs.unm.edu/~judd/papers/email.pdf. 2005-11-06

3 文伟平，卿斯汉，蒋建春，等. 网络蠕虫研究与进展[J]. 软件学报，2004，15(4)：1208~1218

4 Ebel H, Mielsch L, Bornholdt S. Scale-free topology of e-mail networks. http://www.theo-physik.uni-kiel.de/~bornhol/pre035103.pdf, 2005-12-20

5 Bu T, Towsley D. On distingishing between internet power law topology generators . http://ieeexplore.ieee.org/iel5/7943/21922/01019309.pdf? arnumber=1019309, 2005-12-03