一种恶意软件传播的离散概率模型

顾亦然,王锁萍
(南京邮电大学控制与智能技术研究中心,江苏南京 210046)

摘 要: 复杂网络理论为恶意软件传播的研究提供了新的思路和方法.本文针对恶意软件的实际传播机制,提出一种新的离散概率DP-SI模型,该模型可适用于任意网络拓扑.同时提出了一种节点信息网络模型方法,为大规模复杂网络及复杂网络上的传播动力学的仿真,以及离散传播动力学模型的建立,提供了有效的研究平台.仿真结果表明本模型比传统模型更接近现实,对恶意软件的控制具有一定指导意义.

关键词: 恶意软件传播; SI模型; 免疫; 节点信息网络模型

中图分类号: TP393.08 **文献标识码:** A **文章编号:** 0372-2112 (2010) 04-0894-05

A Discrete Probabilistic Model of Malware Propagation

GU Yi-ran, WANG Suo-ping
(*Center for Control & Intelligence Technology, Nanjing University of Posts and Telecommunications, Nanjing, Jiangsu* 210046, *China*)

Abstract: Complex network theory has provided the new train of thought and method to the research of malware's propagation. In this paper, against malicious software spread of the actual mechanism, we develop a new model, called the discrete probability susceptible-infectious (short for DP-SI), which could be applied to any network topology. At the same time, the node information network model method have been proposed for investigating complex network simulations, provided an efficiency research flat for simulating of propagation dynamics in the cosmically complex network as well as building discrete propagation dynamics model. The results show that this model is closer to reality than the traditional model, and it has some significance in the control of malicious software.

Key words: malware propagation; SI model; immunization; node information network model

1 引言

恶意软件(Malware)是对非用户期望运行的、怀有恶意目的或完成恶意功能的软件的统称,这类软件可以是计算机病毒、间谍软件、木马,也可以是网络蠕虫等.当前,网络中的恶意软件传播已成为网络安全的首要问题.卡巴斯基实验室发布的2007年调查报告显示,互联网已成为恶意软件作者首选的传播媒介[1].

迅速发展的复杂网络理论为生物和计算机病毒的传染机制研究提供了新的思路和方法.尤其是基于通信网络的,诸如Internet、WWW、P2P和E-mail等网络的恶意软件传播模型的建立及动力学分析问题已被一些研究者所关注[2],国内也有研究者从传感器网络节点重要性的角度来研究病毒传播的局域控制[3].实际上,网络拓扑对恶意软件的传播产生重要的影响,然而现有的传播模型大多基于网络同质性的假设,忽视了拓扑对传播行为的影响或只针对一种网络拓扑进行研究[4].

互联网是一个典型的复杂网络.不仅结构复杂,而且不断演化,具有多样性的连接,并具有复杂的动力学行为.目前为止,现有的研究主要从两种不同的角度来描绘互联网的结构[5,6].一种是基于路由器,把每个路由器看作节点,路由器之间的连接看作边.另一种基于自治系统(Autonomous System, AS指各个采用不同内部路由算法的子网络),每个AS代表一个节点,如果两个AS之间存在BGP(Border Gateway Protocol)对等连接,则表示这两个节点之间有一条边相连.统计表明,互联网是个无标度网络,表现出很强的幂律分布的特点,其幂指数 $r = 2.5$[7].所以它是一种非均匀网络并且节点的度有很大的波动性[5,8].

本文提出一种恶意软件传播的离散概率模型,简称DP-SI模型,该模型可适用于任意已知网络拓扑,同时,针对互联网模型,分析了DP-SI模型上的随机免疫策略和目标免疫策略的不同效果.为仿真提出的模型,提出一种新的节点信息网络模型方法,克服了邻接矩阵方式表达网络拓扑信息时计算机处理能力的局限性.

收稿日期:2008-12-24;修回日期:2009-06-25

2　恶意软件传播的离散概率 SI 模型(DP-SI)

根据流行病经典 SI(Susceptible- Infected)模型的定义,网络中的节点只能处于两个状态:S 状态即易感状态(也即健康状态),I 状态即感染状态.我们提出一种新的恶意软件传播的离散概率 SI 模型(DP-SI),该模型的建立由网络固有的性质及恶意软件传播的概率特性所决定.

定义 1　定义连通网络 $G=(V,E;N;Q)$,其中:V 表示网络的节点的集合,E 表示网络的边的集合;N 表示网络中节点的个数;Q 是网络中各个节点的状态的集合.

设 $\boldsymbol{A}$ 是一个与连通网络 G 对应的 $N\times N$ 的矩阵,反映网络拓扑信息的邻接关系,其中:$\boldsymbol{A}$ 的元素 $a_{ij}\in\{0,1\}$,$i,j\in\{1,\cdots,N\}$,a_{ij}为 0 表示节点 i 和节点 j 不相邻,为 1 表示相邻.

定义 2　网络 G 中任意节点 v 的度(或邻居数)k_v 定义为与此节点相连的边的个数.

定义 3　网络 G 中任意节点 v 的邻居定义为 $\Gamma_v=\{u:d(u,v)=1\}$,其中 $d(u,v)$为节点 u 和节点 v 之间的最短路径.

定义 4　用 $I_v\subseteq\Gamma v$ 表示节点 v 的邻居中处于感染状态的节点的集合.

定义 5　$x_{vn}\in Q$ 表示节 v 在离散时刻 n 的状态,$Q=\{0,1\}$,x_{vn}为"0"表示节点 v 在 n 时刻处于 I 状态,为"1"表示节点 v 处于 S 状态.

任何节点仅能被其邻居感染,$x_{v,n+1}$依赖于节点在上一时刻的状态 x_{vn}及其邻居的状态,易感个体 v 在单位时间内被某个染病邻居传染的概率为 p_{vi},由此建立如下模型:

定义 6　任意节点 v 在时刻 $n+1$ 依据概率 p_{vn}决定是否被传染.

$$p_{vn}=(1-p_{vi})^{|I_v|} \tag{1}$$

其中:$|I_v|$为 I_v 中元素的个数.

即若 $x_{vn}=0$ 则 $x_{v,n+1}=0$;若 $x_{vn}=1$ 则 $x_{v,n+1}$依据概率 p_{vn}决定是否维持原状态.由式(1)可以看出,邻居中受感染的节点越多,健康节点改变为感染状态的可能性越大.

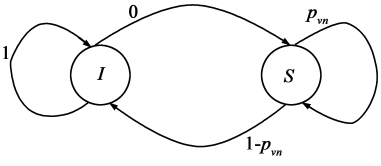


图1　DP-SI模型中节点的状态转移图

由此得到节点 v 对应的状态转移图如图 1 所示.

易知,$X_v=\{x_{vn},n=0,1,2,\cdots\}$是一个齐次马尔可夫链,其一步转移矩阵为:

$$P=\begin{pmatrix}1 & 0\\ 1-p_{vn} & p_{vn}\end{pmatrix} \tag{2}$$

定义 7　I_n 表示 n 时刻网络 G 中感染者的数目,I_0 表示初始时刻的感染者数目.显然:

$$I_n=N-\sum_{v=1}^{N}x_{vn} \tag{3}$$

归一化的感染节点比例 i_n 定义为:

$$i_n=I_n/N \tag{4}$$

在 DP-SI 传播模型中,节点存在两种状态:S 状态和 I 状态.初始时刻网络中有 Q_0 个节点处于 I 状态,其他节点处于 S 状态.当任一健康状态的节点 v 的邻居中有处于 I 状态的节点时,该节点依据概率 p_{vn}决定是否转为 I 状态.演化算法如下:

算法 1　离散概率 SI 模型(DP-SI)演化算法

① 初始化初始感染节点 Q_0,初始化节点变量 j 为 1,初始化所有节点的状态和相关信息.初始化计算步长 *STEP* 为需要计算的步长,设置步长初值 $n=0$;

② 提取节点 j 的状态与相关信息;

③ 根据定义 6 计算该节点的下一状态 $x_{v,n+1}$;

④ $j=j+1$;

⑤ 若 $j=N+1$,则更新所有节点的当前状态,即用 $x_{v,n+1}$替代 x_{vn},同时计算当前感染节点比率 i_n,$j=1$;否则,转到②;

⑥ $n=n+1$;

⑦ 若 $n=STEP$ 或者系统中没有处于健康状态的节点时结束;否则重复②~⑥循环.

3　仿真

3.1　节点信息网络模型

为了仿真提出的 DP-SI 模型,考虑到邻接矩阵方式表达网络拓扑信息时计算机处理能力的局限性(难以处理较大规模网络的邻接矩阵),我们提出了节点信息网络模型方法,用一个节点信息网络模型来表达一个网络.节点信息网络模型的具体描述如下:

采用两个表格存放对应的网络拓扑数据.表 1 为节点信息表 Net,存放每个节点对应的相关信息;表 2 为节点邻居表 Node,存放每个节点的邻居的信息,作为对 Net 表的补充.其中 Net 表中包含节点名 v(即序号),节点 v 的前一状态 $x_{v,n-1}$,节点 v 的当前状态 x_{vn},节点 v 的度 k_v,节点 v 的关联指针 R_v,传染率 p_i 等.Node 表是一个一维数组,从第一个节点的所有邻居(按照序号从小到大排列),第二个节点的所有邻居……依次存放所有节点的邻居名.利用表 Net 中的节点 v 的的关联指针 R_v 和节点 v 的度 k_v,可以从表 Node 中找出节点 v 的所有邻居.

下面举例说明.图 2 是一个包含 5 个节点和 7 条边

的连通网络 DEMONET 的示意图.

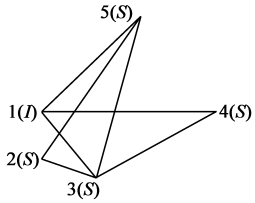


图2 含有五个节点的网络DEMONET的示意图

表 1 和表 2 是采用节点信息网络模型方法表示网络 DEMONET 的结果.

从 Net 表中第一行可以看出,节点 1 的前一状态为 1(代表 S),当前状态为 0(代表 I).相应的如果要做其他传播模型的仿真,可以增加状态集 Q 的元素,例如 SIR 模型的仿真可以增加一个状态 2(代表"R"状态).之所以在 Net 表中既有当前状态又有前一状态,是考虑到在进行仿真运算时,是一个节点一个节点逐次处理的,例如先处理节点 1,再处理节点 2,然后处理节点 3,4,5.如果处理节点 1 时直接修改了节点 1 的当前状态,则在处理节点 3 的时候,所利用到节点 3 的邻居节点 1 和节点 2 的状态就不是同步的状态了.当所有节点都处理完以后,用 Net 表中的所有节点的当前状态更新前一状态,为下次处理做准备.

表 1 DEMONET 的 Net 表

节点名 v	前一状态 $x_{v,n-1}$	当前状态 x_{vn}	节点度 k_v	关联指针 R_v	传染率 p_{vi}
1	1	0	3	0	0.2
2	1	1	2	3	0.4
3	1	1	4	5	0.3
4	1	1	2	9	0.1
5	1	1	3	11	0.1

表 2 DEMONET 的 Node 表

3	4	5	3	5	1	2	4	5	1	3	1	2	3

从 Net 表中可以知道,节点 1 的度为 3,关联指针为 0,从 Node 表中的第 1 + 0 = 1 列开始的三列内容 3、4、5 即为节点 1 的邻居;节点 2 的度为 2,关联指针为 3,从 Node 表中的第 1 + 3 = 4 列开始的两列内容 3、5 即为节点 2 的邻居.在仿真运算中,随时可以查到任何一个节点的所有邻居,以及这些节点和邻居的所有信息.

Net 表中的传染率可以根据需要设置,可以设成相同的值也可以设置成为不同的值.同时,也可以针对具体模型根据变化规律设置为可以动态改变的一个变量,另由公式表达.

Net 表中还可以根据需要为每个节点增加其他状态信息栏.

3.2 ER 模型、NW 模型与 BA 模型上的 DP-SI 模型仿真

复杂网络根据网络节点之间的连接属性主要分为均匀网络和非均匀网络两大类[9],这两大类复杂网络又分别称之为指数网络和幂率网络.

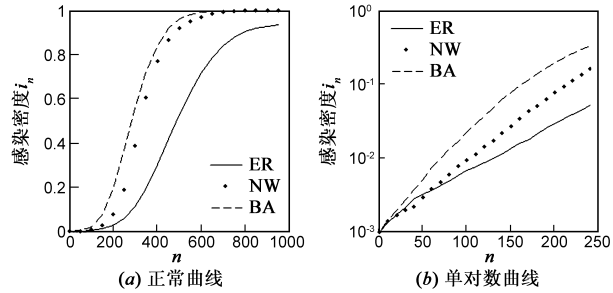
为了验证本文提出的 DP-SI 模型的普适性和正确性,本文选取均匀网络中的 Erdös-Rényi 随机网络(简称 ER 模型)、非均匀网络中的 Newman-Watts 小世界网络(简称 NW 模型)和 Barabási-Albert 无标度幂律网络(简称 BA 模型)三个典型的复杂网络模型作为仿真研究对象,依据 ER 模型算法[10]、NW 模型算法[11]和 BA 模型算法[12],生成这三种网络模型进行仿真验证.

定义 8 网络 G 中所有节点 v 的度 k_v 的平均值被称为网络的(节点)平均度 $< k >$.

对于仿真所用的 ER 网络模型、NW 网络模型和 BA 网络模型,我们取相同的仿真参数.仿真参数设置如下:网络规模(即节点总数)$N = 5000$,平均度 $< k > = 4$,初始时刻感染节点的比例 $i_0 = 0.001$.

同时在不影响仿真结果正确性的前提下,为了简化仿真的算法,本文做如下简化:所有易感个体在单位时间内被某个邻居感染节点传染的概率为 β,即对于所有的 $v \in \{1, \cdots, N\}$,$p_{vi} = \beta$.

图 3 给出了 $\beta = 0.005$ 时,DP-SI 模型在 ER 网络、NW 网络和 BA 网络上的仿真,每根曲线都是独立仿真 100 次的平均值.



(a) 正常曲线 (b) 单对数曲线

图3 无免疫策略时ER模型、NW模型与BA模型上DP-SI模型感染密度i_n与n的关系曲线

可以看出,在传播初期,恶意软件呈指数规律扩散,而对于幂律分布的 BA 无标度网络,传播速率更快.该结论与文献[13,14]是一致的.可见,由于模型建立的时候,没有任何网络拓扑的前提假设,该模型可适用于任意网络拓扑.

3.3 DP-SI 模型的免疫策略仿真

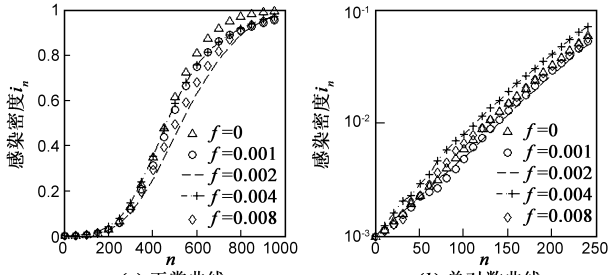
免疫是抑制计算机病毒的一种重要方法.例如运行 Unix 操作系统的主机对基于 Windows 操作系统的病毒就具有免疫作用.目前常用的主要有两种免疫策略:随机免疫(均匀免疫)和选择免疫(目标免疫)[15,16].

研究最多的免疫策略是随机免疫.随机免疫是指完全随机地选取网络中的一部分节点进行免疫,它对度大的节点(被感染的风险高)和度小的节点(相对安全)是平等对待的.已经有专家证明[17],对于无标度网络采取随机免疫策略,需要对网络中几乎所有的节点都实施免疫才能保证最终消灭病毒感染,也就是说随机免疫对无标度网络没有多大作用.

统计表明,互联网是个无标度网络,其幂指数 $r=2.5$[7],仿真时我们采用了基于统计数据构建的互联网模型.1998 年开始,贝尔实验室开展了互联网映射工程(Internet mapping project),由 Burch 和 Cheswick 等人完成了路由器级的互联网拓扑结构图描述.本文抽取了由 Cheswick 等人采用踪迹路线法收集的实际数据[18]进行仿真研究.由于数据量非常大,这里按照实际数据,构成了一个子网络进行模拟计算.该子网络含有 10132 个节点,采用节点信息网络模型方法表示,该无标度网络的幂指数 $r=2.46$.

定义 9　被选择预防接种的节点占节点总数的比例被定义为免疫比例 f.

我们在图 4 中给出了针对不同的免疫比例 f,感染密度 i_n 与 n 的关系曲线.所有的数据都是 100 次独立运行的平均结果.传播率 $p_{vi}=\beta=0.005$,初始感染节点比例为 0.001.从图中可以看出,如果只有很少的节点被选择预防接种,恶意软件的传播速率几乎没有改变.可见,对于互联网而言,采用随机免疫策略对于遏制恶意软件的传播速度没有明显的作用.



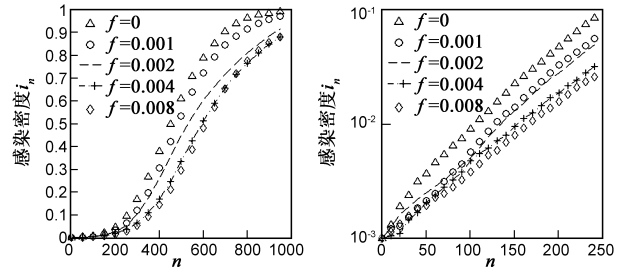
(*a*) 正常曲线　　(*b*) 单对数曲线

图4　应用随机免疫策略时互联网模型的DP-SI模型中感染密度i_n与n的关系曲线

根据无标度网络的不均匀性,可以进行有选择的目标免疫,即选取少量度最大的节点进行免疫.而一旦这些节点被免疫后,就意味着它们所连的边可以从网络中去除,使得恶意软件传播的可能的连接途径大大减少.图 5 给出了针对不同的免疫比例 f,采用目标免疫的效果.所有的数据都是 100 次独立运行的平均结果.传播率 $p_{vi}=\beta=0.005$,初始感染节点比例为 0.001.由图可见,即使在免疫比例很小的 $f=0.001$ 时,扩散的速度也显著下降了.从图 4(*b*)中还可以很明显地看出,即使只有很小的免疫比例($f=0.001$ 即只有 2 个节点免疫),在恶意软件传播的初期,传播的速度仍然得到了较好的控制.

文献[19]研究了在 BA 无标度网络上 SI 模型的免疫策略,文中利用平均场方程推导出 i_n 与 n 的关系方程,得出的结论与本文也是一致的.

互联网用户会安装和更新防病毒软件,但恶意软件的生存期还是很长的,原因就在于这些软件采用的文件扫描和防病毒更新的过程实际上是一种随机免疫过程.在互联网范围内,即使大量节点都被免疫,仍无法根除恶意软件的传播.而目标免疫则是一种更有效的方法,可以大大降低恶意软件传播的速度.



(*a*) 正常曲线　　(*b*) 单对数曲线

图5　应用目标免疫策略时DP-SI模型中感染密度i_n与n的关系曲线

4　结论与展望

复杂网络理论为恶意软件传播的研究提供了新的思路和方法.本文针对恶意软件的传播机制,提出了一种新的恶意软件传播的离散概率模型 DP-SI.仿真时将该模型应用于 ER 网络、NW 网络和 BA 网络,并在基于统计数据构建的互联网模型上,研究了 DP-SI 模型在互联网上的免疫行为.可以看出,由于 DP-SI 模型的构建过程不依赖于某种特定结构的网络,该模型可用于研究任意已知网络拓扑上恶意软件传播的动力学行为.当突发事件发生时,可以利用该模型快速模拟传播行为,以利于防范和预测.同时,该模型结构灵活,能够在演化过程中随时改变控制策略,这是传统微分方程模型所不能比拟的.

同时本文为仿真提出的节点信息网络模型方法,占用存储空间小,结构灵活,解决了邻接矩阵方式表达网络拓扑信息时计算机处理能力的局限性,为大规模复杂网络及复杂网络上的传播动力学的仿真,以及离散传播动力学模型的建立,提供了有效的研究平台.

以上模型特征和结果能够帮助理解恶意软件传播的特征以及拓扑对传播的影响,从而建立有效的对抗措施控制恶意软件的传播,这也将是我们下一步的工作.

参考文献:

[1] Kaspersky Security Bulletin 2007: Malware evolution in 2007 [OL]. http://www.viruslist.com/en/analysis?pubid=204791987, 2008-03-15.

[2] Eben Kenah, James M Robins. Second look at the spread of epidemics on networks [J]. Physical Review E, 2007, 76: 036113.1 – 036113.12.

[3] 张书奎,崔志明,龚声蓉,孙涌.传感器网络病毒感染传播局域控制研究[J].电子学报,2009,37(4):877 – 883.

ZHANG Shu-kui, CUI Zhi-ming, GONG Sheng-rong, SUN Yong. An Investigation on Local Area Control of Compromised Nodes Spreading in Wireless Sensor Networks[J]. Acta Electronica Sinica, 2009, 37(4): 877 – 883. (in Chinese)

[4] Zou C C, Towsley D, and Gong W B. Modeling and simulation study of the propagation and defense of internet e-mail worms [J]. IEEE Transactions on Dependable and Secure Computing, 2007, 4(2): 105 – 118.

[5] Vázquez A, Pastor-Satorras R, Vespignani A. Large scale topological and dynamical properties of the Internet[J]. Physical Review E, 2002, 65: 066130.

[6] Chen G, Fan Z, Li X. Complex Dynamics in Communication Networks[M]. London: Springer Publisher, 2005. 213 – 234.

[7] Newman M E J. The structure and function of complex networks[J]. SIAM Review, 2003, 45(2): 167 – 256.

[8] Faloutsos M, Faloutsos P, Faloutsos C. On power law relationships of the Internet topology[J]. Computer Communication Review, 1999, 29(4): 251 – 262.

[9] Pastor-Satorras R and Vespignani A. Epidemic dynamics and endemic states in complex networks[J]. Physical Review E, 2001, 63(6): 066117.

[10] Erd? s P, Rényi A. On the evolution of random graphs. Publications of Mathematical Institute of the Hungarian Academy of Science, 1960(5): 17 – 61.

[11] Newman M E J, Watts D J. Renormalization group analysis of the small-world network model. Physics Letters A, 1999, 263: 341 – 346.

[12] Barabási A L and Albert R. Emergence of Scaling in Random Networks[J]. Science, 1999, 286: 509 – 512.

[13] Boguñá M, Pastor-Satorras R, Vespignani A. Absence of epidemic threshold in scale-free networks with connectivity correlations[J]. Physical Review Letters, 2003, 90: 028701.

[14] Pastor-Satorras R, Vespignani A. Epidemic dynamics in finite size scale-free networks[J]. Physical Review E, 2002, 65: 035108.

[15] Wang C, Knight J C, Elder M C. On computer viral infection and the effect of immunization[C]. Proceedings of the 16th Annual Computer Security Applications Conference, Washington DC, USA, 2000. 246 – 256.

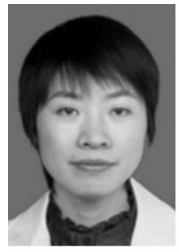
[16] Pastor-Satorras R, Vespignani A. Epidemics and immunization in scale free networks[Z]. Bornholdt S. Handbook of Graphs and Networks: From the Genome to the Internet[M]. Berlin: Wiley VCH, 2003. 111 – 130.

[17] Callaway D S, Newman M E J, Strogatz S H, Watts D J. Network Robustness and Fragility: Percolation on Random Graphs [J]. Physical Review Letters, 2000, 85(25): 5468.

[18] Burch H, Cheswick B. Internet mapping project[OL]. http://www. cheswick. com/ches/map/dbs/index. html, 2007-12-31.

[19] Bai W J, Zhou T, Wang B H. Immunization of susceptible-infected model on scale-free networks[J]. Physica A: Statistical Mechanics and its Applications, 2007, 384(2): 656 – 662.

作者简介:



顾亦然 女, 1972 年出生于江苏金坛, 副教授, 南京邮电大学控制与智能技术研究中心副主任. 主要研究方向为通信网络的性能分析与控制, 复杂网络理论在通信中的应用, 嵌入式系统开发等.

E-mail: guyr@njupt.edu.cn

王锁萍 男, 1946 年生于江苏丹阳, 教授, 博士生导师. 主要研究方向为通信网络的性能分析、流量控制、QoS 理论与技术、单播组播路由算法、任意通信及网络安全理论与技术, 以及信源编码、信道编码的理论与技术.

E-mail: wangsp@njupt.edu.cn