软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2012,23(12):3161−3174 [doi: 10.3724/SP.J.1001.2012.04186]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

结构化对等网络中 P2P 僵尸网络传播模型*

钱 权[1,2+], 萧超杰[1,2], 张 瑞[1,2]

[1](上海大学 计算机工程与科学学院,上海 200072)

[2](中国科学院 软件研究所 信息安全国家重点实验室,北京 100190)

Propagation Modeling for P2P Botnet in Structured P2P Network

QIAN Quan[1,2+], XIAO Chao-Jie[1,2], ZHANG Rui[1,2]

[1](School of Computer Science and Engineering, Shanghai University, Shanghai 200072, China)

[2](State Key Laboratory of Information Security, Institute of Software, The Chinese Academy of Sciences, Beijing 100190, China)

+ Corresponding author: E-mail: qqian@shu.edu.cn

Qian Q, Xiao CJ, Zhang R. Propagation modeling for P2P botnet in structured P2P network. *Journal of Software*, 2012,23(12):3161−3174 (in Chinese). http://www.jos.org.cn/1000-9825/4186.htm

Abstract: Depending on the structured peer-to-peer networks, the P2P botnets are the main threats of the Internet in the future. In this paper, a formal mathematical P2P botnet propagation model is built based on a deep analysis of two typical structured P2P protocols, Chord and Kademlia. This model, it integrates different factors, such as structured P2P protocols, two-factor immunizations, and host online rates to describe the structured P2P botnet propagation mechanisms comprehensively. Meanwhile, in order to evaluate the model effectiveness, simulate the P2P botnets propagation, the difference between the theoretical model and the simulation is used to verify the model efficiency. The experiments prove the correctness of the theoretical model also verify the different influences of structured P2P protocols, immunization mechanisms, and the host online rates on P2P botnet propagation. Moreover, through simulating P2P network with millions of nodes, it can be shown that the propagation model is correct and valid in large scale network, which provides a theoretical basis for botnet detection and prevention.

Key words: network security; structured P2P network; P2P botnet; propagation model

摘 要: 依赖结构化对等网传播的 P2P 僵尸是未来互联网面临的重要威胁.详细分析了两种典型的结构化 P2P 协议 Chord 和 Kademlia 的工作原理,在此基础上,使用数学建模的方法建立了结构化 P2P 僵尸网络的传播模型.该模型将 Kademlia,Chord 协议与双因子免疫机制、主机在线率等因素相结合,较为全面地研究了两种典型的结构化 P2P 网络中僵尸的传播机理,并使用软件仿真的方法模拟了节点超过百万时,结构化 P2P 网络中僵尸的传播行为,通过软件仿真得出的数据与理论数据进行对比,验证了模型的正确性.从实验结果可以看出:对于 Kademlia 和 Chord 两种结构化 P2P 网络,僵尸传播无论是双因子免疫模型还是结合双因子与主机在线率的模型,理论模型与仿真结果都非常吻合,体现了模型的准确性,为僵尸的检测与防御提供了理论依据.

关键词: 网络安全;结构化对等网;P2P 僵尸网络;传播模型

* 基金项目: 国家自然科学基金(61003248); 国家教育部博士点基金(20093108120016); 上海教委重点学科(J50103); 上海教委创新基金(09YZ05)

收稿时间: 2011-03-07; 修改时间: 2011-09-22; 定稿时间: 2011-12-31

© 中国科学院软件研究所 http://www.jos.org.cn

中图法分类号: TP393 文献标识码: A

如今,随着互联网的日益普及,网络安全变得越来越重要.僵尸网络是一种从传统恶意代码进化而来的新型攻击方式,攻击者通过控制大量僵尸主机,实现信息窃取、分布式拒绝服务攻击和发送垃圾邮件等攻击.传统的僵尸网络采用 IRC 或 HTTP 协议的集中式结构,网络依赖中心节点,网络结构简单稳定,但健壮性不高,一旦中心节点失效,整个僵尸网络将无法工作.而基于 P2P 协议的僵尸网络由于没有中心节点,网络中的每个节点既是客户端又是服务端,极大地提高了网络的可扩展性、健壮性以及网络的负载均衡和资源利用率.根据 P2P 网络节点搜索邻居节点的规则不同,P2P 网络可以分为非结构化 P2P 网络与结构化 P2P 网络.非结构化 P2P 网络符合随机模型,僵尸的传播效率不高;而结构化 P2P 网络利用分布式哈希表技术进行网络路由,能够高效地进行信息查找与分发,是僵尸网络发展的主流.目前,使用 P2P 技术的僵尸网络有 Peacomm,Storm 及其更新版本 Waledac[1],这些新型的僵尸网络依赖 P2P 网络中的众多节点进行传输,不仅提高了节点的隐蔽性,还提高了僵尸的传播速度,为检测和防御带来巨大的挑战.因此,研究 P2P 僵尸的传播模型,对于充分理解 P2P 僵尸的传播原理,获得其发展趋势、传播范围,对于评价网络健壮性、不同因素对传播的作用,找出僵尸网络传播过程中的弱点,从而为检测与防御提供指导措施等都具有重要参考价值.

僵尸网络的传播模型源于蠕虫的传播模型,并在其基础上加入僵尸网路的传播特征.网络蠕虫的传播一般适合传染病模型,比较著名的网络蠕虫传播模型有 SI 模型[2]、SIR 模型[2]以及双因子模型[4]等.而对于 P2P 蠕虫的传播模型研究,目前还是针对非结构化 P2P 网络[5−10],主要是在经典传染病模型的基础上加入非结构化 P2P 网络因素(如节点度对传播的影响等).由于非结构化 P2P 网络蠕虫符合随机网络模型,而如今的 P2P 僵尸网络主要使用结构化的 P2P 协议,如 Chord,Kademlia 等进行传播与控制,非结构化 P2P 蠕虫传播模型不能反映 P2P 僵尸网络的传播特征.

结构化的 P2P 网络中,僵尸传播不仅要遵循特定的 P2P 协议,还与僵尸、主机等多因素相关,是个较为复杂的研究问题.这方面的研究成果相对较少,主要有:Grizard 等人对 P2P 僵尸网络的历史进行综述[11],并就基于 Overnet 协议(一种 P2P 协议)的 Peacomm 僵尸采用黑盒技术的逆向工程方法,从初始的 Bot 程序启动、通信机制以及如何利用二次注入来进行命令与控制等进行详细分析,其不足之处是并没有给出此类 P2P 僵尸网络的理论传播模型.

Dagon 等人结合计算机在不同时间内的易感染状态不同(例如夜间大多是关机下线),加上僵尸在感染过程中存在区域偏好的特性,提出基于时区的僵尸网络模型[12].研究了在单时区封闭网络情况下的传播模型,并进一步扩展到多时区的传播模型.研究结果表明,基于时区的传播模型较传统的 SIR 模型好,不足之处有:(1) 该模型仅对僵尸传播的时间特性建模,而对区域传播特性没有考虑;(2) 仅考虑了依赖传统的远程攻击漏洞进行传播的方式,而目前的 P2P 僵尸的传播方式大多依赖垃圾邮件、P2P 文件共享等方式;(3) 不同 P2P 协议的僵尸网络结构不同,不同结构的 P2P 网络对僵尸传播的影响也需要进一步加以研究.

Holz 等人通过二进制代码分析和网络流量跟踪方法对当下最为流行的 Peacomm 僵尸网络进行跟踪,在扩充 IRC 僵尸跟踪方法的基础上,给出了 P2P 僵尸网络的跟踪步骤,并提出使用隐匿和污染文件内容(eclipsing content and polluting)的方法来破坏僵尸间的通信,以清除僵尸网络[13].

Yu 等人对基于 P2P 系统的主动蠕虫攻击进行建模和分析[14],尽管主动蠕虫具有同僵尸网络的相似特征,然而论文是从攻击的角度建模,其不足在于:

(1) 所建立的模型不考虑检测器之间协作以及对蠕虫信息的共享;

(2) 所建立的模型是以疾病传播模型为基础的,与实际的 P2P 僵尸网络传播还是有一定差距.

北京航空航天大学的夏春和等人对 3 种结构化对等网(Chord,CAN 和 Pastry)中的 P2P 蠕虫传播模型进行研究[15],并在此基础上使用 NS-2 仿真建立了 P2P 蠕虫传播模型,通过比对仿真数据和理论数据分析模型的准确度.因此,从现有的公开文献看,尽管在蠕虫的传播研究方面已经取得了一定的研究成果[16−19],然而在僵尸传播模型,尤其是 P2P 僵尸网络的传播模型方面,国内的研究成果还很少.

© 中国科学院软件研究所 http://www.jos.org.cn

本文分析了结构化 P2P 协议 Chord 与 Kademlia 的工作原理,考虑双因子免疫、主机在线率等因素提出结构化 P2P 僵尸网络传播模型,并通过实验仿真对理论模型进行验证.通过实验分析 Chord 协议与 kademlia 协议在传播中的不同特征以及各种因素对传播的作用与影响.

本文第 1 节简单介绍 Kademlia 和 Chord 协议.第 2 节对基于 Kademlia 协议 P2P 传播模型进行详细理论分析.第 3 节研究基于 Kademlia 和 Chord 协议的僵尸传播模型.第 4 节是仿真实验结果及分析.第 5 节总结全文.

1 Kad 和 Chord 协议介绍

Kademlia 协议(简称 Kad)[20]是纽约大学的 Petar 和 David 在 2002 年为非集中式 P2P 网络而设计的,采用分散式杂凑表实现的算法.而 Chord 协议[21]是麻省理工学院提出的一种结构化 P2P 路由协议,具有完全分布式、负载均衡、可用性及可扩展性好、命名方式灵活等特点.Kad 和 Chord 协议十分相似.Kad 协议中每个节点都随机获得一个 m 位的节点号.Kad 协议将所有网络节点用一棵树表示,每个节点作为树的叶子节点.根据各个节点的最长前缀决定各个节点之间的距离.同时,每个节点内部维护一个路由表,包含 $\log_2 N$ 条路由信息(N 为网络中的节点数).此外,每个节点还保存了距离本节点距离为 $2^i\sim 2^{i+1}(0\leqslant i\leqslant m-1)$的节点信息,这些列表组成一个 k-桶.Kad 协议使用递归算法进行节点查询.当开始查询某个目标节点时,在查询节点信息列表中查找到距离最近的节点,然后向它们发出查询请求;若得到一些距离更近的节点信息,则再向这些距离更近的节点发送查询请求;如此不断重复查询,就可得到距离目标节点最近的那些节点的信息.Kad 由于每次都是从更接近目标节点的 k-桶中获取信息,因此,每次递归查询操作至少使距离减半.

Chord 协议构建的网络中,每个节点有一个 m 位的唯一节点标识符,该标识符通过对节点信息(如节点 IP)进行散列运算后得到.将所有节点根据节点标识符大小按顺时针方向排列形成一个环形标识符,称为标识符空间.Chord 中每个节点需要维护 3 项信息,即前向节点、后向节点列表以及一张由 m 个指向其他在线节点的指针组成的路由表.节点 n 的第 i 个入口为自节点$(n+2^{i-1})$起顺时针方向上具有最小标识符的节点.Chord 协议以一种去中心化的方式解决了分布式 P2P 网络中资源定位问题,在节点数为 N 的 Chord 网络中,其路由复杂度为 $O(\log N)$.

Kad 协议和 Chord 协议的相似处在于他们都使用了 2^i 作为下一跳的距离,但他们之间的不同之处主要在两个方面:

(1) Chord 协议必须访问距离自身 2^i 的节点,而 Kad 协议可以访问距离自身 $2^i\sim 2^{i+1}$ 区间内的 k 个节点(k 为可调的系统参数);

(2) Chord 协议访问是不对称的,每个节点只能直接访问节点号比自己大的节点.比如,节点 A 知道如何到达节点 B(说明节点 B 相对于节点 A 是顺时针的),但是节点 B 却不能直接知道节点 A 的信息(因为节点 A 相对于节点 B 是逆时针的).节点 B 需要通过其他节点路由到节点 A,这种非对称的算法造成了 Chord 协议的路由单向性.但是 Kad 协议却不是这样,Kad 协议使用异或运算,而对于异或运算,$A\oplus B=B\oplus A$,所以算法是对称的.节点 A 既可以访问顺时针方向上的节点,也可以访问逆时针方向上的节点.

2 基于 Kad 协议的 P2P 传播模型

设初始节点为 O,它在其路由表中选择 m 个节点,且这 m 个节点从其路由表的 m 个 k 桶中选出.

对 $0\leqslant k\leqslant m-1$,从距离节点自身$\{2^k,2^{k+1}\}$的节点中选出一个节点作为下一个目的节点.每个节点以此方式向外传播.

根据文献[15]的研究思想,在计算每个时刻新增节点个数时要避免重复计算.在 Chord 协议中的重复计算有两种可能:

- 一个节点两次访问了同一距离的节点造成重复.例如,两次访问了距离为 2 的节点与一次访问距离为 4 的节点是重复的;
- 两个节点以不同的次序访问了同一个节点.比如,一个节点访问距离为 7 的节点的访问次序为

© 中国科学院软件研究所 http://www.jos.org.cn

1→2→4→7,而另一个节点访问的顺序为 4→2→1,这两个节点访问的次序不同,但是最后访问的都是节点 7,所以造成了重复.

文献[15]十分巧妙地使用了组合排列的方式解决了节点重复的问题.

同样,对于 Kad 协议而言也存在重复计算的问题.由于 Kad 协议使用的是异或运算,同样存在两种情况会造成重复访问:

- 一个节点连续异或两个相同数,比如 1⊕4⊕4=1,这样会造成对节点 1 的重复访问;
- 两个节点以不同的顺序访问不同的距离,但是最后异或的结果相同.例如,1⊕3⊕6=4,而 1⊕2⊕7=4,节点 1 以不同顺序不同距离两次访问了同一个节点 4,从而造成重复.

根据 Kad 协议上述两个情况提出如下 5 个命题.

命题 1. 初始节点个数为 1,即源点 O.初始节点通过 Kad 协议向外传播.

设$\Delta\Phi_i$表示在第 i 个时刻网络中新增的被访问节点个数,则有:

$$\Delta\Phi_i=\left|\left\{x\left|x=\left(\bigoplus_j^i y_j\right)\bmod(2^m)\right.\right\}\right| \tag{1}$$

$$y_j\in[2^k,2^{k+1}),(k=0,1,2,...,m-1)且y_j\neq y_{j'},1\leqslant j,j'\leqslant i,j\neq j'$$

其中,x 表示集合中每个节点的节点号;$\bigoplus_j^i y_j$ 表示 $y_j⊕y_{j+1}⊕y_{j+2}⊕…⊕y_i$,$i\geqslant j$;$m$ 表示节点标识符长度.

证明:

对于一个节点,它每次从 m 个 k 桶中取出一个元素作为与目的地的距离进行异或,由于 Kad 协议选择节点的方式不同于Chord协议那样必须选择距离自身大于或等于2^i的最小节点,而是可以随机选择$[2^i,2^{i+1}]$区间中的任意元素,所以必须保证一个节点两次在同一个区间内取出的元素不相同.

所以命题 1 中有 $y_j\neq y_{j'},1\leqslant j,j'\leqslant i,j\neq j'$. □

命题 2. 设 $V(i)$表示第 i 时刻总的被访问过节点个数,$\Delta\Phi_i$ 表示在第 i 个时刻新增的被访问的节点个数,则在时刻 i 第 K 个区间被访问过的次数为

$$K(i)=V(i-1) \tag{2}$$

证明:

第 1 个时刻,$\Delta\Phi_1=1,V(1)=1,K(1)=0$;

第 2 个时刻,$\Delta\Phi_2=1\times m,V(2)=V(1)+\Delta\Phi_2,K(2)=K(1)+\Delta\Phi_1=V(1)$;

设第 n 个时刻,$\Delta\Phi_n=\Delta\Phi_{n-1}\times m',V(n)=V(n-1)+\Delta\Phi_{n\Delta t},K(n)=V(n-1)$,

则在第 n+1 时刻,$\Delta\Phi_{n+1}=\Delta\Phi_n\times m',V(n+1)=V(n)+\Delta\Phi_{(n+1)\Delta t},K(n+1)=K(n)+\Delta\Phi_{n\Delta t}=V(n-1)+\Delta\Phi_{n\Delta t}=V(n)$.

因此可得出,在 i 时刻,第 k 个区间被访问过的次数为 $K(i)=V(i-1)$. □

命题 3. 设 D_i 表示事件,表示在第 i 个时刻访问已被访问过的节点;$D_i|k$ 表示事件,表示在第 k 个区间访问了已被访问过的节点;设 $P(D_i|k)$表示 i 时刻在路由表的第 k 个区间访问已被访问节点的概率,则有

$$P(D_i\mid k)=\frac{V(i-1)}{2^{k-1}},若V(i-1)>2^{k-1},则\frac{V(i-1)}{2^{k-1}}=1 \tag{3}$$

证明:

根据 Kad 协议,在路由表的第 k 个区间的节点个数为 2^{k-1};根据命题 2,在 i 时刻路由表的第 k 个区间被访问过的次数为 $V(i-1)$.所以, $P(D_t\mid k)=\frac{V(i-1)}{2^{k-1}}$.由于重复概率不会大于 1,所以当$V(i-1)>2^{k-1},\frac{V(i-1)}{2^{k-1}}=1$. □

从广义角度,基于 Kad 协议的 P2P 网络节点的增长是一个随机过程.但是它不同于传统的随机过程如马尔可夫链,因为在基于 Kad 协议的 P2P 网络中,每个节点都有一个路由表,节点根据路由表访问其他节点.但是在不同节点的路由表中可能存在相同节点,这样,该节点会被访问多次,所以需要计算某个节点已经被访问的概率,从而去除重复访问同一个节点.而基于 Chord 协议的 P2P 网络,网络节点的访问增加数量符合一定的排列组合规律,因此不是随机过程.

© 中国科学院软件研究所 http://www.jos.org.cn

命题 4. 设 D_i 表示事件,表示在第 i 个时刻访问已被访问过的节点;$P(D_i)$表示在 i 时刻一个节点访问已被访问节点的概率,即重复访问概率;$p(k)$表示访问路由表第 k 个区间的概率,则有

$$P(D_i)=P(D_i\mid 1)p(1)+P(D_i\mid 2)p(2)+\cdots+P(D_i\mid k)p(k)+\cdots+P(D_i\mid m)p(m)=\sum_{j=1}^{m}P(D_i\mid j)p(j) \tag{4}$$

证明:

设 E 为"访问 Kad 路由表"实验,S 为实验 E 的样本空间.设 B_i={访问第 i 个区间},$1\leqslant i\leqslant m$,则 B_1~B_m 是对路由表的访问事件的一种划分.因为 B_i~B_m 满足:

$$B_iB_j=\phi, i\neq j, i,j=1,2,...,m,$$
$$B_1\cup B_2\cup...\cup B_m=S.$$

根据全概率公式可知:

$$\begin{aligned}P(D_i)&=P(D_i\mid B_1)p(B_1)+P(D_i\mid B_2)p(B_2)+\cdots+P(D_i\mid B_k)p(B_k)+\cdots+P(D_i\mid B_m)p(B_m)\\&=P(D_i)=P(D_i\mid 1)p(1)+P(D_i\mid 2)p(2)+\cdots+P(D_i\mid k)p(k)+\cdots+P(D_i\mid m)p(m)\\&=\sum_{j=1}^{m}P(D_i\mid j)p(j).\end{aligned}$$ □

命题 5. 设$|\Delta\Phi_{i\Delta t}|$表示第 i 个时刻可以传播的节点个数,则有

$$\begin{aligned}|\Delta\Phi_{(i+1)\Delta t}|&=|\Delta\Phi_{i\Delta t}|\times m\times(1-P(D_i))\\&=|\Delta\Phi_{i\Delta t}|\times m\times\left(1-\sum_{j=1}^{m}P(D_i\mid j)p(j)\right)\\&=|\Delta\Phi_{i\Delta t}|\times m\times\left(1-\left(\frac{1}{m}\frac{V(i-1)}{1}+\frac{1}{m}\frac{V(i-1)}{2}+...+\frac{1}{m}\frac{V(i-1)}{2^{m-1}}\right)\right)\\&=|\Delta\Phi_{i\Delta t}|\times m\times\left(1-\sum_{j=0}^{m-1}\left(\frac{1}{m}\frac{V(i-1)}{2^j}\right)\right)\end{aligned} \tag{5}$$

若$V(i-1)>2^j$,则$\dfrac{V(i-1)}{2^j}=1$.

证明:

根据命题 3 可知,在第 i 时刻,访问已被访问节点的概率为 $P(D_i)$,即重复访问的概率为 $P(D_i)$,所以不产生重复访问的概率是 $1-P(D_i)$.在第 i 个时刻,有$|\Delta\Phi_{i\Delta t}|$个节点可以进行传播,每个节点可以选择 m 个节点作为传播目标.所以在第 i+1 个时刻,可传播节点个数为$|\Delta\Phi_{(i+1)\Delta t}|=|\Delta\Phi_{i\Delta t}|\times m\times(1-P(D_i))$. □

3 基于 Kad 和 Chord 协议的僵尸传播

在描述网络蠕虫、病毒等恶意代码的传播方面,已有的模型包括 SEM,KM 和双因子模型等[2].其中,

- 简单传染病模型(SEM)中,节点只有两个状态:易感染状态和传染状态.而状态转移是易感染状态向传染状态转化;
- KM 模型在 SEM 模型的基础上考虑了受感染主机由于用户意识到了恶意代码的感染而进行打补丁、杀毒、重启系统等操作,使得主机不再具有传染性.因此,KM 模型中的主机包含 3 种不同的状态:易感染 S(susceptible)、可传染 I(infectious)和被移除 R(removed);
- 双因子模型是在 KM 模型的基础上提出了两个会影响恶意代码传播的因素:首先,人为的抵抗行为导致主机从易感染状态和可传染状态到被移除状态的转移;其次,大量恶意代码的传播会引起网络拥塞,降低其扫描率,从而影响其传播速率.由于双因子模型综合考虑了人为抵抗和自身传播导致网络拥塞等双重因素,类似于传播到一定程度节点产生了免疫效果,因此有其独特的优势.双因子模型的状态转换如图 1 所示.

© 中国科学院软件研究所 http://www.jos.org.cn

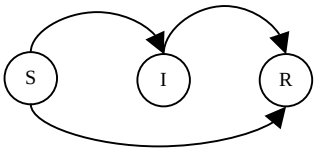

Fig.1 State transition diagram of two-factor-model

图 1 双因子模型状态转换图

如前所述,僵尸的传播表示出很强的时间特性[12],即由于不同时间主机的上线率不同,因而僵尸传播存在一定差异.一般而言,白天的主机上线率高,而在深夜由于大部分主机已下线,因此,一天中不同时刻由于主机上线率不同,僵尸的传播(易感染、已感染、被移除的免疫节点)自然受到很大的影响.本文将同时考虑双因子和主机上线率等不同因素来研究依赖 Kad 和 Chord 的 P2P 僵尸的传播模型.模型中所用到的数学符号和含义见表 1.

Table 1 Symbols and their meanings in model

表 1 模型使用的符号和含义

符号	含义
$S(t)$	在t时刻,处于易感染状态的节点数
$I(t)$	在t时刻,处于可传染状态的节点数
$R(t)$	在t时刻,新增的被移除的可传染节点数
$E(t)$	在t时刻,新增的可传染节点数
$Q(t)$	在t时刻,新增的被移除的易感染节点数
N	总节点数N,即$N=I(t)+R(t)+Q(t)+S(t)$
$J(t)$	在t时刻,被感染的节点总数,即$J(t)=I(t)+R(t)$
$C(t)$	在t时刻,处于被移除状态的节点总数,即$C(t)=R(t)+Q(t)$
γ	可传染节点的平均移除率
μ	易感染节点的平均移除率
$O(t)$	在t时刻,节点的平均在线率

3.1 基于Kad协议的僵尸网络传播

通过公式(5),可以计算出每个时间间隔内的新增节点个数,即

$$|\Delta\Phi_{(i+1)\Delta t}|=|\Delta\Phi_{i\Delta t}|\times m\times\left(1-\sum_{j=0}^{m-1}\left(\frac{1}{m}\frac{V(i-1)}{2^j}\right)\right),\text{若}V(i-1)>2^j,\text{则}\frac{V(i-1)}{2^j}=1$$

加入双因子免疫效果后,有

$$|\Delta\Phi_{(i+1)\Delta t}|=|\Delta\Phi_{i\Delta t}|\times m\times\left(1-\sum_{j=0}^{m-1}\left(\frac{1}{m}\frac{V(i-1)}{2^j}\right)\right)\times(1-\mu) \tag{6}$$

加入主机在线率因素后,有

$$|\Delta\Phi_{(i+1)\Delta t}|=|\Delta\Phi_{i\Delta t}|\times m\times\left(1-\sum_{j=0}^{m-1}\left(\frac{1}{m}\frac{V(i-1)}{2^j}\right)\right)\times(1-\mu\times O(i))\times O(i) \tag{7}$$

即在 t+1 时刻,新增的被感染节点个数为

$$E(t+1)=E(t)\times m\times\left(1-\sum_{j=0}^{m-1}\left(\frac{1}{m}\frac{V(i-1)}{2^j}\right)\right)\times(1-\mu\times O(t))\times O(t) \tag{8}$$

3.2 基于Chord协议的僵尸网络传播

文献[15]已说明了 Chord 协议在每个时刻增加的访问节点个数符合组合排列的规律,即 C_m^i.其中,m 表示节点标识符的长度.在此基础上加入免疫因素.设 0 时刻只有原点 O,网络中节点个数为 $N=2^m$,下面将计算从 t 时刻到 t+1 时刻新增的感染节点数.

根据组合排列可知,在不考虑双因子模型的情况下,每个时刻的新增节点数的递推方程为

© 中国科学院软件研究所 http://www.jos.org.cn

$$E(t+1)=\frac{E(t)(m-t)}{(t+1)} \tag{9}$$

加入免疫率以及在线率因素后,可得每个时刻新增节点数递推方程为

$$E(t+1)=\frac{E(t)(m-t\times(1-\mu\times O(t)/t))}{(t+1)}(1-\mu\times O(t)) \tag{10}$$

证明:在 t 时刻,根据 Chord 协议,每个节点已经从 m 个节点中选择了 t 个进行传播,所以剩下(m–t)个节点可以选择.但是由于双因子因素的存在,前面 t 个时刻有一定比例的节点没有被访问到,需要将这部分没有被访问到节点个数减去.

未被访问的节点主要是由于免疫的作用以及主机不在线造成的,因此该数量正比于μ以及 $O(t)$,而免疫的节点个数随着时间的变化不断减少,所以该数量反比于时间 t,即公式(10).

下面将给出该递推方程的极限平衡态形式.在公示(10)中:μ表示易感染节点平均移除率,是个常数;$O(t)$表示节点在线率,是关于 t 的一个函数,且当 t 较大时,可以认为 $O(t)$是一个关于 t 的常数.所以设$\mu\times O(t)=a$,即

$$E(t+1)=\frac{E(t)(m-t\times(1-a/t))}{(t+1)}(1-a)=\frac{E(t)(m-t+a)}{(t+1)}(1-a).$$

因为,

$$\frac{E(t+1)}{E(t)}=\frac{(m-t+a)}{(t+1)}(1-a),\frac{E(t)}{E(t-1)}=\frac{(m-t+1+a)}{(t)}(1-a),...,\frac{E(2)}{E(1)}=\frac{(m-1+a)}{(1)}(1-a).$$

所以有,

$$\frac{E(t+1)}{E(t)}\times\frac{E(t)}{E(t-1)}\times...\times\frac{E(2)}{E(1)}=\frac{E(t+1)}{E(1)}=(1-a)^t\frac{(m+a-1)!}{(t+1)!(m+a-t-1)!},$$

其中,$E(1)=1$,所以有,

$$E(t+1)=(1-a)^t\frac{(m+a-1)!}{(t+1)!(m+a-t-1)!}.$$

当 $t=m+a-1$ 时,$E(t)$不再增长达到平衡,可被感染节点个数为

$$Sum=\sum_{t=1}^{t=m+a-1}\left((1-a)^t\frac{(m+a-1)!}{(t+1)!(m+a-t-1)!}\right).$$

在上述两个协议的每个时刻新增感染节点个数递推方程基础上,结合双因子和在线率因素,可以得到 P2P 僵尸的传播其他信息,如下:

- 易感染节点个数:$S(t+1)=S(t)-E(t+1)-Q(t)$;
- 感染节点个数:$I(t+1)=I(t)+E(t+1)-R(t)$;
- 感染节点被移除个数:$R(t)=\gamma Q(t)I(t)$;
- 易感染节点被移除个数:$Q(t)=\mu O(t)S(t)J(t)$.

4 实验仿真与分析

4.1 实验环境与参数说明

本文使用 Peersim 软件进行结构化 P2P 僵尸网络传播模型仿真.Peersim 是一个用来进行 P2P 网络仿真的软件[22],该软件基于接口化设计,将 P2P 网络分为节点、协议、控制器和网络连接这 4 个接口,用户只需根据不同协议实现这 4 个接口,即可进行 P2P 网络仿真.

本实验使用的参数为 $N=2^m$,m 表示路由表的区间个数;免疫率 $u=0.05$;隔离率 $r=0.06$;模拟时间间隔为 $T=100$ 个时间单元,每个时间单元表示一个僵尸进行传播的时间.初始参数 $I(0)=1$,$S(0)=N$,$E(t)=1$.另外,实验中考虑 0 点~23 点每个时刻的主机在线率,当时间 $t>23$ 点后,使用 $t\%24$ 时刻的在线率.一天中不同时刻主机上线率曲线如图 2 所示,具体上线率数值见文献[12],为

© 中国科学院软件研究所 http://www.jos.org.cn

$O(t)$={0.40,0.58,0.77,0.91,0.99,0.995,0.985,0.974,0.957,0.965,0.97,0.978,
0.985,0.91,0.8,0.71,0.62,0.51,0.38,0.28,0.21,0.20,0.22,0.3}.

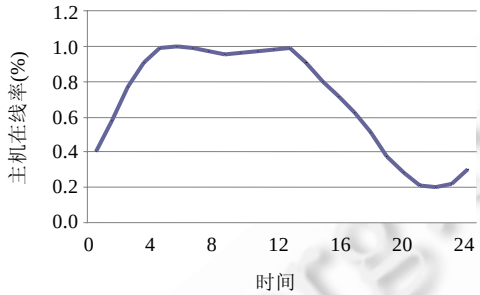

Fig.2 Host online rate at different time

图 2 不同时刻主机在线率

4.2 基于Kad协议的僵尸网络传播仿真

4.2.1 双因子模型下的 Kad 协议仿真

设 m=19,即网络节点个数为 $N=2^m$,实验结果如图 3 所示.

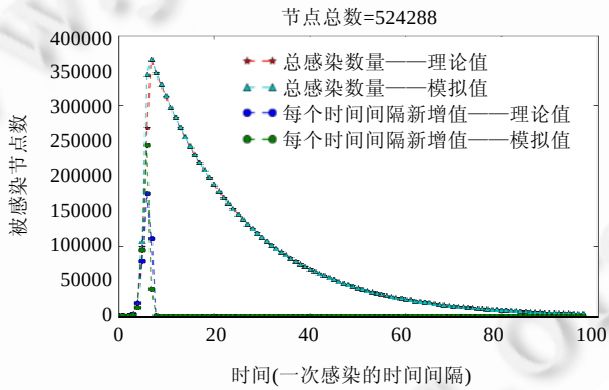

Fig.3 Comparison between two-factor model and simulation for Kad botnet (2^{19} nodes)

图 3 双因子 Kad 僵尸传播模型与仿真结果比较(节点个数 2^{19})

从图 3 可以看出,Kad 协议在双因子模型下的仿真效果十分理想,理论模型和实验仿真在数值上十分接近,并且从曲线走势上能够大致反映出 Kad 协议在双因子模型下的传播趋势.另外,从图 3 可以看出,曲线的走势符合我们在理论分析中提到的,基于 Kad 协议的 P2P 网络传播速度很快,从第 0 个时刻第 1 个节点开始传播,到传播到达最高点只用了 7 个时刻.一旦传播过了高峰期后,由于双因子免疫的作用,被感染节点数开始下降,最终被感染节点个数趋于 0.这说明了双因子免疫因素对传播能够起到抑制作用.

4.2.2 结合双因子和主机上线率的 Kad 协议仿真

加入每个时刻的主机上线率以及双因子模型相互结合后对 Kad 协议传播的影响.同样用 m=19 进行实验,实验结果如图 4 所示.

从图 4 可以看出,理论模型基本和模拟结果吻合,说明理论模型的准确性.

© 中国科学院软件研究所 http://www.jos.org.cn

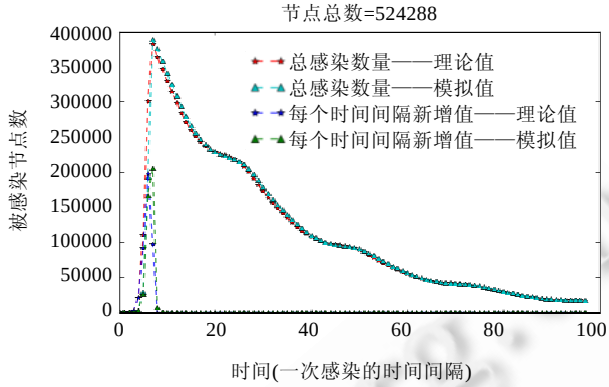

Fig.4 Comparison between two-factor with host online rate model and simulation for Kad botnet (2^{19} nodes)

图 4 结合主机上线率和双因子 Kad 僵尸传播模型与仿真结果比较(节点个数 2^{19})

4.2.3 Kad 协议中主机在线率对传播的影响

为了进一步理解主机在线率对传播的影响,我们将图 3 和图 4 进行对比,对比结果如图 5 和图 6 所示.

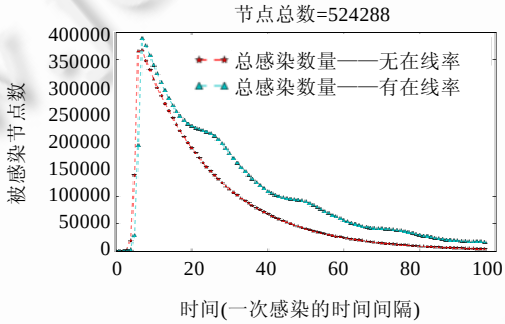

Fig.5 Influences of host online rate for total infected nodes in Kad

图 5 Kad 协议中在线率对总感染节点数量的影响

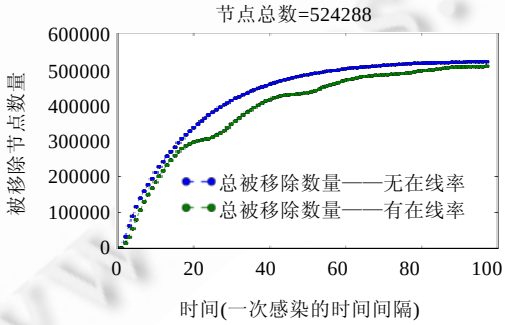

Fig.6 Influences of host online rate for recovered nodes in Kad

图 6 Kad 协议中在线率对被移除节点数量的影响

图 5 说明了主机在线率对总的感染节点个数的影响.可以看出,相对于无在线率的总感染节点个数,有在线率的总感染节点个数可以达到的最高点要大于前者,并且到达最高点的时间相比前者略慢.

速度较前者慢的原因在于不在线的主机无法感染,所以感染速度会降低.总的感染节点个数增加的原因在

© 中国科学院软件研究所 http://www.jos.org.cn

于免疫节点个数下降造成总的被感染节点个数上升.当考虑在线率因素后,双因子因素的免疫速度会由于主机在线率而有所降低,因为不在线的主机无法免疫.

图 6 进一步证实了这一结论.可以看出,每一个时刻考虑在线率因素下,总的被移除节点个数相比没有考虑在线率因素下总的被移除节点个数要少.一旦被免疫的节点个数减少了,被感染节点个数相对就会增加.

再次观察图 5 可以发现,当传播过了高峰期后,总感染节点个数由于双因子免疫作用开始下降.有在线率情况的曲线会产生波动,这是因为不同时刻的主机在线率不同造成免疫速度的不同,从而形成了波动.而在传播的初期,主机在线率并没有对传播造成太大波动,原因在于传播初期可感染节点数据占了绝大多数,大体上以上升趋势为主,主机在线率无法体现其作用.而下降过程中由于可感染节点数量不多,主机在线率开始体现其作用.

4.3 基于Chord协议的僵尸网络传播仿真

4.3.1 双因子模型的 Chord 协议仿真

加入双子模型后的 Chord 模型同样分两次进行实验,节点个数为 2^{21}=2097152.仿真结果如图 7 所示,双因子参数为 u=0.06,r=0.05.

从图 7 可以看出,加入了双因子因素后,感染节点的个数到达最高峰后开始下降.总体的模拟效果较为理想,理论模型的整体趋势和仿真模拟得到的基本一致,各个时间节点上的理论值与模拟值相差均小于 1%,总体上达到了准确描绘 Chord 协议在双因子模型下的传播趋势.

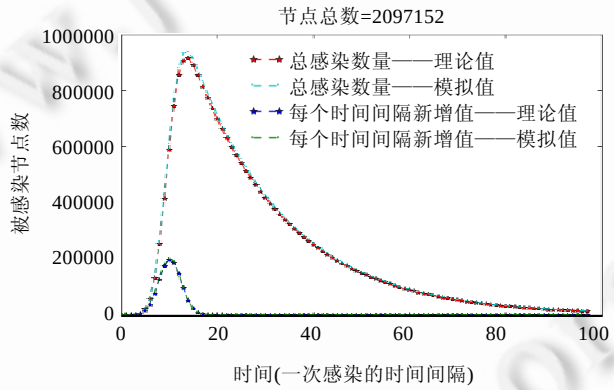

Fig.7 Comparison between two-factor model and simulation for Chord botnet (2^{21} nodes)

图 7 Chord 僵尸双因子传播模型与仿真结果比较(节点数 2^{21})

4.3.2 结合双因子和主机上线率的仿真实验

结合主机上线率以及免疫效果进行 Chord 协议的仿真实验,实验结果如图 8 所示.

同第 4.2.3 节中分析主机在线率对 Kad 协议的作用类似,对比图 7 和图 8 分析在线率因素对 Chord 协议的影响,结果如图 9 所示.

从图 9 可以看出,加入了在线率因素后,总的被感染节点数有所上升.并且从图中可以看出,在传播的初期,在线率并没有对传播造成太大影响;但是一旦传播过了高峰期,处于下降过程时,在线率对传播造成很大影响.从图 9 的曲线可以看出,在传播曲线的下降过程中出现间隔性的波动,这个波动正是由于不同时刻的主机上线率不同而造成的.

© 中国科学院软件研究所 http://www.jos.org.cn

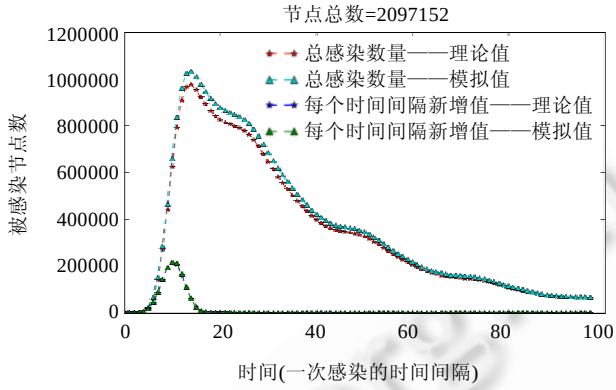

Fig.8 Comparison between two-factor with host online rate model and simulation for Chord botnet (2^{21} nodes)

图 8 结合主机上线率以及双因子的 Chord 僵尸传播模型与仿真结果比较(2^{21} 节点)

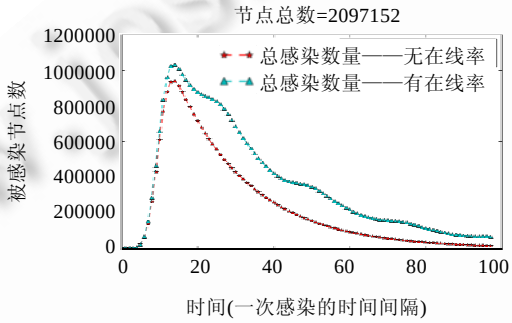

Fig.9 Influences of host online rate for botnet propagation in Chord

图 9 Chord 协议中在线率对传播的影响

4.4 Kad与Chord对比分析

在第 2 节已经提到,Kad 协议与 Chord 协议的不同之处在于:Chord 协议中,一个节点必须访问距离自身大于等于 2^i 的最近节点;而 Kad 协议中,一个节点则可以访问距离自身 2^i~2^{i+1} 区间内的 k 个节点(k 为系统参数并且可调).所以理论上,Kad 协议的传播速度和传播节点个数要大于 Chord 协议.下面通过实验进行验证.

设 m=19,即网络中节点个数为 N=2^{19}=524288.分别使用 Kad 和 Chord 协议进行实验,实验对比结果如图 10 所示.

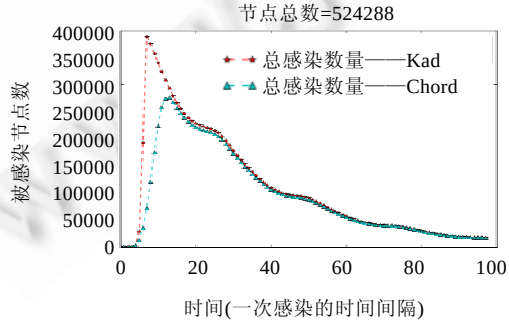

Fig.10 Comparison between Kad and Chord for total infected nodes

图 10 Kad 协议与 Chord 协议总感染量对比

© 中国科学院软件研究所 http://www.jos.org.cn

从图 10 可以看出:使用 Kad 协议的总感染节点个数要比基于 Chord 协议的大很多,Kad 协议最多可感染将近 40 万个节点,而 Chord 协议最多感染近 28 万个节点.Kad 协议的传播速度要比 Chord 协议快,Kad 协议需要 7 个时间间隔达到最高点,而 Chord 协议需要 12 个时间间隔.图 10 充分说明了 Kad 协议传播速度快、传播范围大的特点.

4.5 实验小结

通过上述实验,可以得出如下结论:

(1) 使用仿真软件 Peersim 进行结构化 P2P 网络仿真,实验结果表明,模拟结果与理论模型数值相吻合,一定程度上说明了本文提出的 P2P 结构化僵尸网络的传播模型的准确性;

(2) 通过将 Chord 与 Kad 协议与双因子模型相结合进行实验仿真,最终被感染节点个数趋于 0,说明双因子免疫机制能够有效地抑制结构化 P2P 僵尸网络的传播;

(3) 分析了主机在线率对传播的影响.在考虑主机在线率的情况下,结构化 P2P 僵尸网络的传播速度会有所减慢,但是总的被感染节点个数有所上升;

(4) 理论和实验都说明,Kad 协议相比 Chord 协议的传播速度更快、范围更大;

(5) 实验进行了大规模的模拟,Kad 协议的模拟节点个数超过了 50 万个,而 Chord 协议的模拟节点个数超过了 200 万个;并且,模拟结果与理论模型基本吻合,说明本文提出的理论模型在大规模网络中是有效的;

(6) 由实验结果可以看出,P2P 僵尸网络在传播的初期,其传播速度非常迅速,可以在短时间内感染大量节点;到了峰值后,其感染速度开始变缓.所以,如果能够在 P2P 僵尸网络传播的初期就遏制它的快速传播,对于防止 P2P 僵尸程序大范围传播有很大作用.

5 结束语

目前,P2P 网络在科研、教育、军事等领域都有大规模的应用,然而 P2P 网络因其每个节点具有多重角色(客户机、服务器和路由器)、网络自身的灵活性以及 P2P 网络自身的安全漏洞等,为僵尸的大范围传播提供了便利条件.本文针对两种典型的结构化对等网络 Chord 和 Kad 进行形式化分析,建立了 P2P 僵尸传播的理论模型.该模型将结构化 P2P 协议 Kad,Chord 与免疫率、在线率等因素相结合,分别建立了 Kad 和 Chord 协议在不同时刻的新增感染节点数、易感染节点数、已感染节点数、感染节点被移除数目、易感染节点被移除的数目等数学模型.从现有文献看,该模型是目前较为全面地描述 P2P 僵尸传播的理论模型.论文通过仿真实验对模型进行了验证,从实验结果可以看出:对于 Kad 和 Chord 协议,无论是双因子模型还是结合双因子与主机上线率的模型,理论模型与仿真结果都非常吻合,误差非常小,体现了模型的准确性,为相关的检测防御和结构化 P2P 安全研究提供了理论依据.当然,P2P 僵尸的传播除了与免疫率、移除率、主机上线率等因素外,与结构化 P2P 协议本身密切相关,本文仅讨论了 Kad 和 Chord 协议下僵尸传播的特征,对于其他结构化 P2P 协议下僵尸传播的规律性特征还需做进一步研究.另外,对于结构化 P2P 僵尸网络的其他传播特征,如网络延迟、丢包率以及动态网络对传播的影响等,也需要做进一步研究.

References:

[1] Wang P, Wu L, Aslam B, Zou CC. A systematic study on peer-to-peer botnets. In: Byrav R, ed. Proc. of the 18th Int'l Conf. on Computer Communications and Networks (ICCCN 2009). San Francisco: IEEE Press, 2009. 1–8. [doi: 10.1109/ICCCN.2009.5235360]

[2] Zou CC, Gong WB, Towsley D. Code red worm propagation modeling and analysis. In: Atluri V, ed. Proc. of the 9th ACM Conf. on Computer and Communications Security (CCS 2002). New York: ACM Press, 2002. 138–147. [doi: 10.1145/586110.586130]

© 中国科学院软件研究所 http://www.jos.org.cn

[3] Kim D, Radhakrishnan S, Dhall SK. Measurement and analysis of worm propagation on Internet network topology. In: Ronald PL, ed. Proc. of the Int'l Conf. on Computer Communications and Networks (ICCCN 2004). Chicago: IEEE Press, 2004. 495−500. [doi: 10.1109/ICCCN.2004.1401716]

[4] Zou CC, Gong WB, Towsley D. Worm propagation modeling and analysis under dynamic quarantine defense. In: Staniford S, ed. Proc. of the ACM CCS Workshop on Rapid Malcode (WORM 2003). New York: ACM Press, 2003. 51−60. [doi: 10.1145/948187.948197]

[5] Li H, Zheng Q, Pan XH, Zhang XS. Propagation model of non-scanning active worm in unstructured P2P network. In: Lina W, ed. Proc. of the 2009 Int'l Conf. on Multimedia Information Networking and Security (NINES 2009). IEEE Press, 2009. 378−381. [doi: 10.1109/MINES.2009.109]

[6] Yu W, Chellappan S, Wang X, Xuan D. Peer-to-Peer system-based active worm attacks: Modeling, analysis and defense. Computer Communications, 2008,31(17):4005−4017. [doi: 10.1016/j.comcom.2008.08.008]

[7] Luo XR, Yao Y, Gao FX. Research of a potential worm propagation model based on pure P2P principle. Journal of Communications, 2006,27(11A):53−58 (in Chinese with English abstract).

[8] Zhang XS, Chen T, Zheng J, Li H. Active worm propagation modeling in unstructured P2P network. In: Yu F, ed. Proc. of the 2nd Int'l Symp. on Computer Science and Computational Technology (ISCSCT 2009). Huangshan: Academy Publisher, 2009. 035−038.

[9] Feng CS, Qin ZG, Laurence C, Laurissa T. Reactive worms propagation modeling and analysis in peer-to-peer networks. Journal of Computer Research and Development, 2010,47(3):500−507 (in Chinese with English abstract).

[10] Zhang XS, Chen T, Zheng J, Li H. Proactive worm propagation modeling and analysis in unstructured peer-to-peer networks. Journal of Zhejiang University—Science C (Computer & Electronics), 2010,11(2):119−129.

[11] Grizzard JB, Sharma V, Nunnery C. Peer-to-Peer botnets: Overview and case study. In: Niels P, ed. Proc. of the 1st Workshop on Hot Topics in Understanding Botnets. Cambridge: USENIX Association Berkeley Publisher, 2007. 1−8.

[12] Dagon D, Zou C, Lee W. Modeling botnet propagation using time zone. In: Eric H, ed. Proc. of the 13th Annual Network and Distributed System Security Symp. (NDSS 2006). San Diego: The Internet Society Press, 2006. 1−15.

[13] Holz T, Steiner M, Dahl F, Biersack E, Freiling F. Measurement and mitigation of peer-to-peer-based botnets: A case study on storm worm. In: Fabian M, ed. Proc. of the 1st Usenix Workshop on Large-Scale Exploits and Emergent Threats (LEET 2008). San Francisco: USENIX Association Berkeley Publisher, 2008. 1−9.

[14] Yu W, Chellappan S, Wang X, Xuan D. Peer-to-Peer system-based active worm attacks: Modeling and analysis. Computer Communications, 2008,31(17):4005−4017. [doi: 10.1016/j.comcom.2008.08.008]

[15] Xia CH, Shi YP, Li XJ. Research on epidemic models of P2P worm in structured peer-to-peer networks. Chinese Journal of Computers, 2006,29(6):952−959 (in Chinese with English abstract).

[16] Wen WP, Qing SH, Jiang JC, Wang YJ. Research and development of Internet worms. Journal of Software, 2004,15(8):1208−1219 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/15/1208.htm

[17] 杨峰,段海新,李星.网络蠕虫扩散中蠕虫和良性蠕虫交互过程建模与分析.中国科学(E 辑),2004,34(8):841−856.

[18] Gao CX, Zhang FY, Xin Y, Niu XX, Yang YX. Research on worm's propagation and defense model in different P2P networks. Journal of Beijing University of Posts and Telecommunications, 2006,29(z2):49−53 (in Chinese with English abstract).

[19] Wu CJ, Zhou SJ, Xiao CJ, Wu Y. Simulation of epidemic of P2P worms in BitTorrent networks. Journal of University of Electronic Science and Technology of China, 2007,36(6):1206−1210 (in Chinese with English abstract).

[20] Maymounkov P, Mazieres D. Kademlia: A peer to peer information systems based on the XOR metric. In: Peter D, ed. Proc. of the 1st Int'l Workshop on Peer-to-Peer Systems (IPTPS 2002). Cambridge: Springer-Verlag, 2002. 53−65.

[21] Stoica I, Morris R, Karger D, Kaashoek MF, Balakrishnan H. Chord: A scalable peer-to-peer lookup service for internet applications. In: Rene C, ed. Proc. of the 2001 Conf.on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM 2001). New York: ACM Press, 2001. 149−160. [doi: 10.1145/383059.383071]

[22] Montresor A, Jelasity M. PeerSim: A scalable P2P simulator. In: Henning S, ed. Proc. of the 9th Int'l Conf. on Peer-to-Peer Computing (P2P 2009). Seattle: IEEE Computer Society Press, 2009. 99−100. http://peersim.sourceforge.net/ [doi: 10.1109/P2P.2009.5284506]

© 中国科学院软件研究所 http://www.jos.org.cn

附中文参考文献:

[7] 罗兴睿,姚羽,高福祥.基于纯 P2P 原理的蠕虫传播模型的研究.通信学报,2006,27(11A):53−58.

[9] 冯朝胜,秦志光,劳伦斯·库珀特,罗瑞莎·托卡库克.P2P 网络中沉默型蠕虫传播建模与分析.计算机研究与发展,2010,47(3):500−507.

[15] 夏春和,石昀平,李肖坚.结构化对等网中的 P2P 蠕虫传播模型研究.计算机学报,2006,29(6):952−959.

[16] 文伟平,卿斯汉,蒋建春,王业君.网络蠕虫研究与进展.软件学报,2004,15(8):1208−1219. http://www.jos.org.cn/1000-9825/15/1208.htm

[18] 高长喜,章甫源,辛阳,钮心忻,杨义先.P2P 网络中蠕虫传播与防治模型的研究.北京邮电大学学报,2006,29(z2):49−53.

[19] 吴春江,周世杰,肖春静,吴跃. BitTorrent 网络中的 P2P 蠕虫传播仿真分析.电子科技大学学报,2007,36(6):1206−1210.

钱权(1972－),男,安徽怀宁人,博士,副研究员,主要研究领域为计算机网络,网络安全.

张瑞(1981－),女,博士,讲师,主要研究领域为计算机网络,信息安全.

萧超杰(1987－),男,硕士生,主要研究领域为网络安全.

© 中国科学院软件研究所 http://www.jos.org.cn