

Kad 网络的联合污染模型

孔 劼,蔡皖东

(西北工业大学 计算机学院,西安 710129)

(jackeykong@mail.nwpu.edu.cn)

摘 要:将 Kad 网络中的关键词污染和文件源污染结合起来,使用状态转移分析的方法构造了一种联合污染模型。模型中综合考虑了污染程度、退出率、等待率等因素。对模型的仿真实验数据显示,受到联合污染时,Kad 网络中查询失败的用户数远大于查询成功的用户数,并随着时间的增加而趋于稳定。在影响联合污染效果的若干因素中,污染程度对联合污染的效果有决定性的影响,退出率的影响次之,等待率的影响最小。

关键词:Kad 网络;对等网;关键词污染;文件源污染;联合污染模型;状态转移

中图分类号:TP393.08 **文献标志码:**A

Joint pollution model in Kad network

KONG Jie, CAI Wan-dong

(School of Computer Science and Technology, Northwestern Polytechnical University, Xi'an Shanxi 710129, China)

Abstract: In this paper, a joint pollution model, which combined the pollution of keyword and the pollution of location, was proposed. The degree of pollution, the rate of exit and the rate of waiting were taken into account in the model. The simulation results show that the quantity of user of querying failed is much larger than the quantity of user of querying successfully by the impact of joint pollution and become stable with time increasing. The degree of pollution is the key factor which influence the effect of the joint pollution, the effect of exit rate is smaller than the degree of pollution and the effect of waiting rate is the smallest.

Key words: Kad network; Peer-to-Peer (P2P); keyword pollution; file source pollution; joint pollution model; state transfer

0 引言

近年来,为了提高系统的鲁棒性,基于分布式散列表(Distributed Hash Table, DHT)技术的 Kademlia 算法^[1]被引入到 Peer-to-Peer (P2P)文件共享系统中,使用户不查询中心服务器即可找到资源的提供者。采用 Kademlia 算法构建的网络,其拓扑结构属于全分布式结构化拓扑,在 eMule 系统中被称为 Kad 网络,在 BitTorrent 系统中被称为 DHT 网络。

P2P 污染是一种用于控制特定文件在 P2P 文件共享系统中传播的文件下载控制技术,最早用于控制盗版音像制品通过 P2P 文件共享系统非法传播。2005 年,Liang 等人^[2]对当时已发现的 P2P 污染给出了详细的描述。在接下来几年中,又派生出多种 P2P 污染技术,主要可分为内容污染与索引污染。内容污染通过在 P2P 网络中传播虚假的文件内容实现对文件传播的控制。Dhungel 等人^[3]介绍了被称为 Fake-Block Attack 和 Uncooperative-Peer Attack 的内容污染方法,并对这两种污染方法在 BitTorrent 系统中的效果进行了评估。索引污染则通过干扰 P2P 网络的索引系统,使其无法提供资源索引服务到控制特定文件传播的目的。Liang 等人^[4]介绍了非结构化的 FastTrack 系统与结构化的 Overnet 系统中的索引污染,并对污染效果进行了测量分析。目前学术界对 P2P 污染的研究一方面是对 P2P 污染效果进行测量分析,文献[1-4]都属于此类研究;另一方面是对 P2P 污染进行建模分析。胡辛遥^[5]提出了污染统一模型,将内容污染与索引污染结合起来进行建模分析。左敏等人^[6]用状态转移模型描述 P2P 文件污染现象以及污染散播过程的时间特性。方群等人^[7]基于 Markov 生灭过程建立了污染传播模型,另外方

群^[8]基于古典概率模型创建了文件污染模型。云燕^[9]对 P2P 文件共享中污染版本的传播情况进行建模。Mao 等人^[10]通过建模分析了资源流行度和传播时间对污染效果的影响。文献[5-10]的建模分析主要针对非结构化的 P2P 系统,没有考虑结构化 P2P 系统的污染建模问题。本文对结构化的 Kad 网络中的 P2P 污染进行建模分析,得到污染效果的时间特性,并对影响 Kad 网络污染效果的因素进行分析。

1 面向 Kad 网络的 P2P 污染

1.1 Kad 网络的资源发布与查询机制

在 eMule 系统中,Kad 主要充当文件信息检索协议的角色。Kad 网络中有 2 种重要的散列值:关键词散列值和文件散列值。对 eMule 客户端的共享资源的描述信息进行分词,得到共享资源的关键词,再使用 SHA1 算法对关键词进行计算得到的散列值被称为关键词散列值。对 eMule 客户端的共享资源使用 SHA1 算法计算出的散列值被称为文件散列值。

eMule 客户端在 Kad 网络中发布资源信息时,首先将关键词信息发布到 Kad 网络中节点 ID 等于或最接近关键词散列值的若干个节点(具体查询过程可参阅文献[1]),将这些节点称为“关键词信息中间节点”。关键词信息是一个(key, value)的属性对,其中 key 的值等于关键词散列值,value 为一个列表,该列表给出了关键词对应的文件信息,格式为:(文件名,文件长度,文件散列值)。随后 eMule 客户端还需要发布文件源信息。客户端首先文件源信息发布到 Kad 网络中节点 ID 等于或最接近文件散列值的节点上,这些节点被称为“文件源信息中间节点”。文件源信息也是一个(key, value)的属性对,其中 key 的值等于所发布资源的文件散列值,value

收稿日期:2011-02-18;修回日期:2011-04-01。 基金项目:国家 863 计划项目(2009AA01Z424)。

作者简介:孔劼(1981-),男,陕西西安人,博士研究生,主要研究方向:网络信息安全、P2P 网络污染; 蔡皖东(1955-),男,陕西西安人,教授,博士生导师,主要研究方向:网络安全、信息对抗。

值为拥有该文件的节点(即发布者)的网址信息,格式为:(拥有者 IP,端口号,拥有者节点 ID)。

资源请求者通过输入关键词查询资源时,搜索 Kad 网络中节点 ID 等于或最接近关键词散列值的节点,可从关键词信息中间节点获得与关键词对应的文件散列值。随后根据获得的文件散列值,搜索 Kad 网络中节点 ID 与之相等或最接近的节点,即可找到文件源信息中间节点,得到资源提供者的地址信息,进而请求与资源提供者建立连接,开始数据传输。

1.2 Kad 网络的关键词污染和文件源污染

污染者进行关键词污染时,首先提取欲污染文件的关键词信息,并计算关键词散列值,根据关键词散列值查询到关键词信息中间节点。随后,污染者向这些节点发布大量虚假的关键词信息,即(文件名,文件长度,文件散列值)三元组中文件散列值是无效的。由于 Kad 网络的资源查询过程中,资源请求者从关键词信息中间节点获取正确的文件散列值是进行文件源查询的前提条件,因此关键词污染将导致资源请求节点无法找到正确的资源提供节点。

污染者进行文件源污染时,首先查询关键词或解析文件 eD2k 链接获得文件的文件散列值,并根据文件散列值查询到文件源信息中间节点。随后,污染者向这些节点发布大量虚假的文件源信息,即将虚假的(拥有者 IP,端口号,拥有者节点 ID)三元组发布到文件源信息中间节点。文件源信息中间节点受文件源污染后,资源请求者无法由文件散列值得到正确的资源提供节点的地址信息。

2 Kad 网络的联合污染模型

根据前述 Kad 网络污染的基本原理,本文提出了一种 Kad 网络的联合污染模型,模型同时考虑了 Kad 网络中关键词污染与文件源污染对查询结果带来的影响。模型从用户的角度进行建模,通过模拟用户在 Kad 查询过程中的状态转移来反映污染在 Kad 网络中的效果。

2.1 Kad 网络受到污染时的状态转移分析

受到联合污染时,用户通过 Kad 网络查询资源的状态转移情况如图 1 所示。在不同状态之间转移所需的时间被标注在连接状态的线条上,没有标注数字的线条代表两个状态之间的转换时间可以忽略不计。

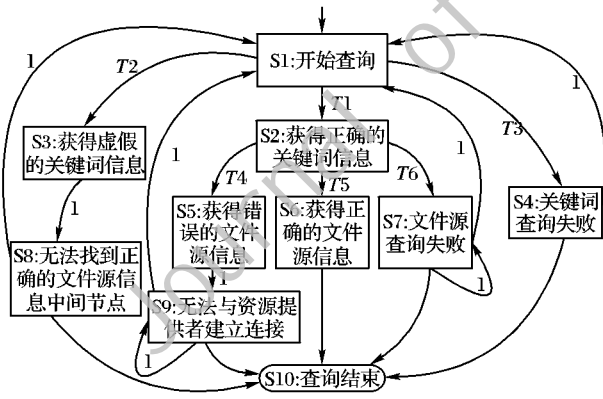


图1 受联合污染的 Kad 网络的户状态转移图

在状态 S1,设用户到达 Kad 网络的到达率服从参数为 λ 的泊松分布。用户输入欲搜索文件的关键词,开始关键词信息查询。设查询成功的概率为 P_c ,查询到虚假关键词信息的概率为 P_f ,以概率 P_c 在 T_1 个时间单位后转入状态 S2;以概率 P_f 在 T_2 个时间单位后转入状态 S3;以概率 $1 - P_c - P_f$ 在 T_3 个时间单位后转入状态 S4。

在状态 S2,用户的客户端获得关键词信息,并根据关键词信息查询文件源信息。设查询成功的概率为 P_{ic} ,查询到虚假文件源信息的概率为 P_{if} ;以概率 P_{if} 在 T_4 个时间单位后转入状态 S5;以概率 P_{ic} 在 T_5 个时间单位后转入状态 S6;以概率

$1 - P_{ic} - P_{if}$ 在 T_6 个时间单位后转入状态 S7。

在状态 S3,由于受到关键词污染的影响,用户的客户端获得虚假的关键词信息,并根据虚假的关键词信息查询对应的文件源信息,经过 1 个时间单位后转向状态 S8。

在状态 S4,由于关键词查询失败,用户的客户端将不显示任何与关键词相关的资源信息。设用户选择结束查询的概率为 θ_4 ,则以概率 θ_4 转向状态 S10;以概率 $1 - \theta_4$ 在 1 个时间单位后转向状态 S1。

在状态 S5,由于受到文件源污染的影响,用户的客户端获得虚假的文件源信息,并根据虚假的文件源信息查询资源提供者的地址信息,经过 1 个时间单位后转向状态 S9。

在状态 S6,文件源信息查询成功,用户的客户端获得资源提供者的地址信息,查询结束,转向状态 S10。

在状态 S7,文件源信息查询失败,用户的客户端会显示出资源的名称,但可用源数为 0。由于 Kad 网络的查询是一个逐渐逼近目标节点的过程,可用源信息的显示可能会存在延迟,因此当可用源数为 0 时用户可能选择等待,也可能选择结束查询或重新查询。设用户选择等待的概率为 ω_7 ,选择退出查询的概率为 θ_7 ,则用户选择结束查询,转向状态 S10 的概率为 θ_7 ;选择重新查询,以概率 $1 - \omega_7 - \theta_7$ 在 1 个时间单位后转向状态 S1;选择等待,1 个时间单位后用户停留在状态 S7 的概率为 ω_7 。

在状态 S8,由于根据虚假的关键词信息无法查询到正确的文件源信息中间节点,用户的客户端将不显示任何资源信息,设用户选择结束查询的概率为 θ_8 ,则以概率 θ_8 转向状态 S10,以概率 $1 - \theta_8$ 在 1 个时间单位后转向状态 S1。

在状态 S9,由于获得的资源提供者地址信息是虚假的,用户所在的客户端无法与正确的资源提供者建立连接。设用户选择继续等待建立起连接的概率为 ω_9 ,选择退出查询的概率为 θ_9 ,则用户选择结束查询,转向状态 S10 的概率为 θ_9 ;选择重新查询,以概率 $1 - \omega_9 - \theta_9$ 在 1 个时间单位后转向状态 S1;选择等待,1 个时间单位后用户停留在状态 S9 的概率为 ω_9 。

2.2 模型的数学表示

设 $X_n(t)$ 为 t 时刻状态 S_n 的用户数, $Y_s(t)$ 与 $Y_b(t)$ 分别为 t 时刻通过 Kad 网络查询资源成功与失败的用户数,根据马尔可夫链的知识,得到联合污染模型的数学表达式为:

$$X_1(t) = (1 - \theta_7 - \omega_7) \cdot X_7(t-1) + (1 - \theta_4) \cdot X_4(t-1) + (1 - \theta_8) \cdot X_8(t-1) + (1 - \theta_9 - \omega_9) \cdot X_9(t-1) + \lambda \quad (1)$$

$$X_2(t) = P_c \cdot X_1(t-T_1) \quad (2)$$

$$X_3(t) = P_f \cdot X_1(t-T_2) \quad (3)$$

$$X_4(t) = (1 - P_c - P_f) \cdot X_1(t-T_3) \quad (4)$$

$$X_5(t) = P_{if} \cdot X_2(t-T_4) \quad (5)$$

$$X_6(t) = P_{ic} \cdot X_2(t-T_5) \quad (6)$$

$$X_7(t) = (1 - P_{ic} - P_{if}) \cdot X_2(t-T_6) + \omega_7 \cdot X_7(t-1) \quad (7)$$

$$X_8(t) = X_3(t-1) \quad (8)$$

$$X_9(t) = X_5(t-1) + \omega_9 \cdot X_9(t-1) \quad (9)$$

$$Y_s(t) = X_6(t) \quad (10)$$

$$Y_f(t) = X_5(t) + X_3(t) \quad (11)$$

3 模型的仿真

实验的硬件环境为:3.0 GHz Intel 奔腾 4 处理器,1 GB 内存,80 GB 硬盘。实验的软件环境为:Windows XP SP3, Matlab 7.0, Visual C++ 6.0。实验的默认设置为: $T_1 = T_2 = T_3 = T_4 = T_5 = T_6 = 10$, $\lambda = 100$, 实验迭代的次数为 500 次,即实验模拟的时间长度为 500 个时间单位。

3.1 联合污染条件下的 Kad 查询效果

受到联合污染时,对于方程(1)~(11),令 $\theta_4 = \theta_7 = \theta_8 = \theta_9 = 0.2$, $\omega_7 = \omega_9 = 0.6$, $P_c = P_{ic} = 0.3$, $P_f = P_{if} = 0.5$ 。仿真结果如图 2 所示。在图 2 中,“有污染查询成功”和“有污染

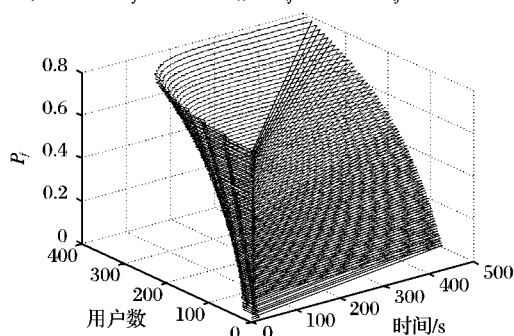
查询失败”分别是 $Y_s(t)$ 、 $Y_f(t)$ 随时间变化的值。对比“有污染查询失败”和“有污染查询成功”可知,在受到联合污染的影响时,Kad 网络中查询失败的用户数远大于查询成功的用户数。由图 2 还可知在联合污染模型中,Kad 网络内某个时刻查询成功的用户数与查询失败的用户数都将随着时间的增加趋向于稳定。对 $Y_s(t)$ 、 $Y_f(t)$ 作回归分析,得到的回归方程如下。

$$Y_s(t) = -0.0060t^2 + 0.6890t - 4.8764$$

$$Y_f(t) = 0.0004t^3 - 0.0631t^2 + 5.5480t - 18.3700$$

3.2 污染程度对联合污染效果的影响

以用户查询到虚假关键词信息的概率 P_f 和查询到虚假文件源信息的概率 P_{if} 来反映 Kad 网络中的污染程度。对于方程(1)~(11),令 $\theta_4 = \theta_7 = \theta_8 = \theta_9 = 0.2$, $\omega_7 = \omega_9 = 0.6$, $P_{ic} = 0.3$, $P_{if} = 0.5$, $P_c + P_f = 0.8$, P_f 的取值对联合污染效果的影响如图 3(a) 所示。令 $\theta_4 = \theta_7 = \theta_8 = \theta_9 = 0.2$, $\omega_7 = \omega_9 = 0.6$, $P_c = 0.3$, $P_f = 0.5$, $P_{ic} + P_{if} = 0.8$, P_{if} 的取值对联合污染效果的影响如图 3(b) 所示。



(a) 关键词污染的污染程度对联合污染效果的影响

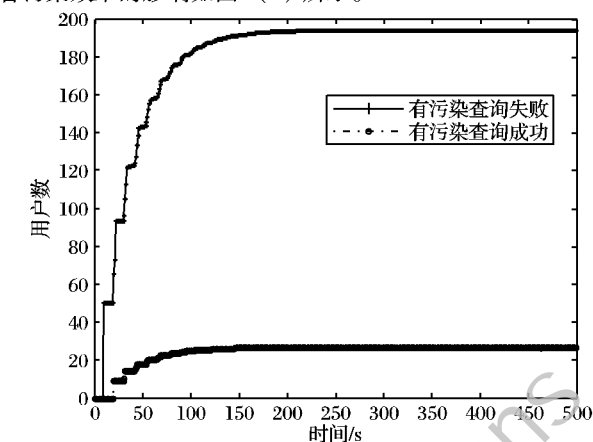
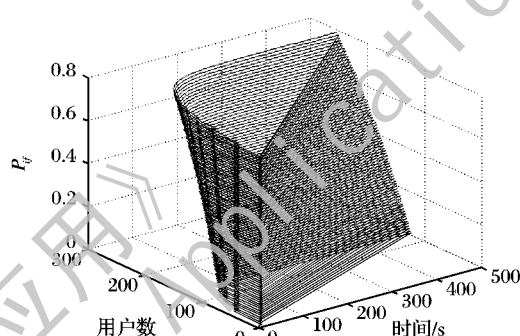


图2 联合污染模型对查询结果的影响



(b) 文件源污染的污染程度对联合污染效果的影响

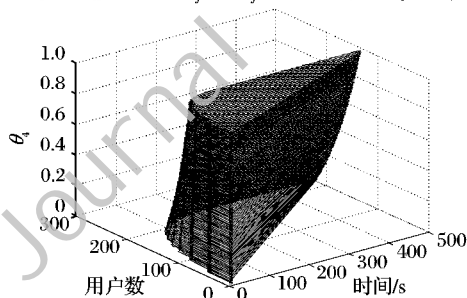
图3 污染程度对联合污染效果的影响

由图 3(a) 可知,当 P_{if} 取固定值时,查询失败的用户数趋向的稳定值随着 P_f 的取值的增加而快速增加;由图 3(b) 可知,当 P_f 取固定值时,查询失败的用户数趋向的稳定值也随着 P_{if} 的取值的增加而增加,但是增加的速率要低于图 3(a) 中 P_{if} 取固定值,以 P_f 为变量的情况。这说明相对于文件源污染,关键词污染的程度对联合污染的效果影响更大。

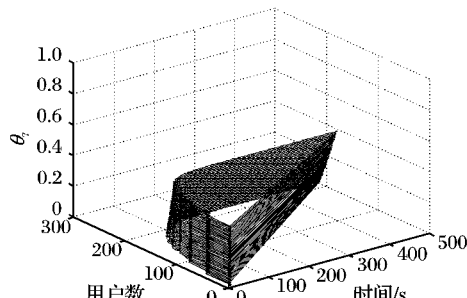
3.3 退出率对联合污染效果的影响

退出率即用户选择结束查询的概率,反映在联合污染模型中为 θ_4 、 θ_7 、 θ_8 、 θ_9 这四个参数。对于方程(1)~(11),令 $\omega_7 = \omega_9 = 0.6$, $P_c = P_{ic} = 0.3$, $P_f = P_{if} = 0.5$, $\theta_7 = \theta_8 = \theta_9 =$

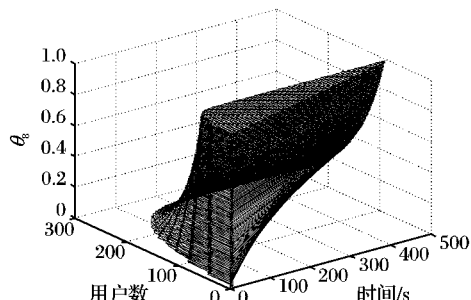
0.2, 则 θ_4 的取值对联合污染效果的影响如图 4(a) 所示。同理,当其他三个退出率的取值为 0.2 时, θ_7 、 θ_8 、 θ_9 的取值对联合污染效果的影响分别如图 4(b)~(d) 所示。由于分别受到 ω_7 与 ω_9 取值的影响, θ_7 与 θ_9 的取值在 0.01 到 0.39 之间。由图 4 可知,随着退出率的增加,查询失败的用户数趋向的稳定值减少,这是由于停留在 Kad 网络中的资源请求者数量减少,导致查询失败的用户数随之减少引起的。在四个退出率参数中, θ_8 的取值变化对联合污染效果的影响最大——即因为受到关键词污染,无法找到正确的文件源信息中间节点而退出 Kad 网络的节点的退出率对联合污染效果的影响最大。



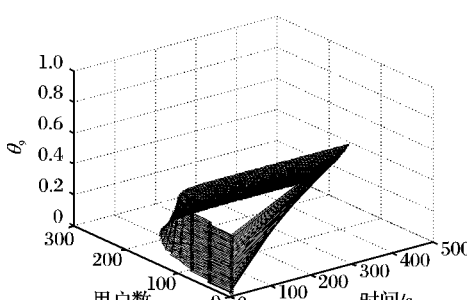
(a) θ_4 的取值对联合污染效果的影响



(b) θ_7 的取值对联合污染效果的影响



(c) θ_8 的取值对联合污染效果的影响



(d) θ_9 的取值对联合污染效果的影响

图4 退出率对联合污染效果的影响

3.4 等待率对联合污染效果的影响

等待率即用户选择停留在当前状态的概率,反映在联合污染模型中为参数 ω_7 和 ω_9 。对于方程(1)~(11),令 $P_c = P_{ic} = 0.3, P_f = P_{if} = 0.5, \theta_4 = \theta_7 = \theta_8 = \theta_9 = 0.2$ 。当 ω_9 为0.6时, ω_7 的取值对联合污染效果的影响如图5(a)所示;当 ω_7 为0.6时, ω_9 的取值对联合污染效果的影响如图5(b)所示。由图5可知,随着等待率的增加,Kad网络内查询失败的用户数趋向的稳定值减少,但是 ω_7 取值的变化对污染效果的影响非常小,几乎可以忽略不计。此外,对比图3、4可知, ω_7 和 ω_9 的取值对污染效果的影响要小于其他因素。因此,在联合污染模型中,等待率属于影响污染效果的次要因素。

3.5 仿真结论

实验表明,联合污染对Kad查询的控制从原理上是有效的,在影响联合污染效果的若干因素中,污染程度对联合污染的效果有决定性的影响,退出率的影响次之,等待率的影响最小。由此可见,提高污染效果的最好办法是提高关键词污染

和文件源污染的污染程度。

4 结语

本文介绍了针对Kad网络的P2P污染的基本原理,提出了一种Kad网络中的联合污染模型,该模型将关键词污染和文件源污染结合起来,综合评估两种污染对于Kad网络的污染效果。通过对模型的数学建模与仿真实验,证明了联合污染的有效性,并分析了模型中不同的因素对污染效果的影响。实验结果表明,Kad网络受到联合污染后,系统内能正常完成查询的节点数明显少于无法完成查询的节点数,从理论上讲联合污染是有效的。此外,在影响联合污染效果的若干因素中,污染程度对联合污染效果的影响最大。

在未来的工作中,联合污染模型将进行进一步完善,使其不仅能反映Kad网络的查询过程,也能反映节点之间的数据传输过程。此外,更加精确地描述模型中各个参数的取值范围也是有待进一步研究的内容。

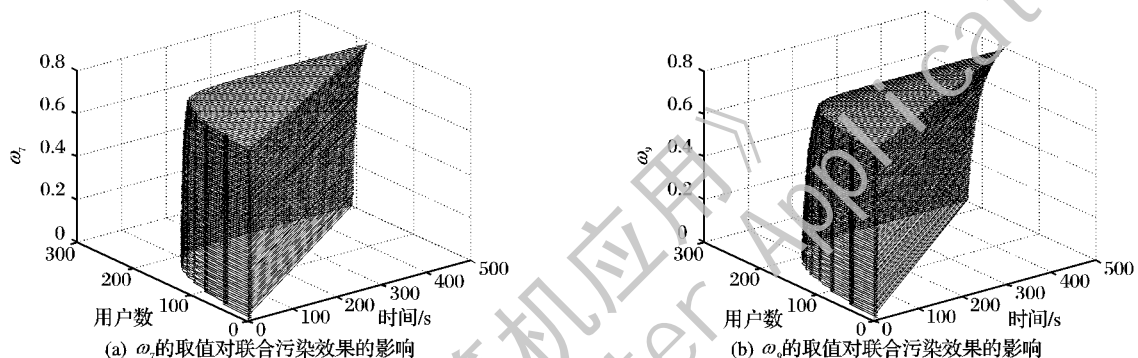


图5 等待率对联合污染效果的影响

参考文献:

- [1] MAYOUNKOV P, MAZIERES D. Kademlia: A peer-to-peer information system based on the XOR metric [C]// IPTPS'01: Proceedings of the First International Workshop on Peer-to-Peer Systems. Berlin: Springer-Verlag, 2001: 53-65.
- [2] LIANG J, KUMAR R, XI Y, et al. Pollution in P2P file sharing systems [C]// INFOCOM 2005: Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies. Piscataway, NJ: IEEE Press, 2005: 1174-1185.
- [3] DHUNGEL P, WU D, SCHONHORST B, et al. A measurement study of attacks on BitTorrent leechers [C]// IPTPS'08: Proceedings of the 7th International Conference on Peer-to-Peer Systems. Berkeley, CA: USENIX Association, 2008: 7-7.
- [4] LIANG J, NAOUMOV N, ROSS K W. The index poisoning attack in P2P file sharing systems [C]// INFOCOM 2006: Proceedings of the 25th IEEE International Conference on Computer Communications, Joint Conference of the IEEE Computer and Communications Societies. Piscataway, NJ: IEEE Press, 2006: 1-12.
- [5] 胡辛遥. 对等网络污染的建模[D]. 上海: 上海交通大学, 2008.
- [6] 左敏, 李建华, 蒋兴浩. P2P文件污染的建模与仿真分析[J]. 上海交通大学学报, 2008, 42(2): 239-243.
- [7] 方群, 吴国新, 于坤, 等. P2P文件污染的Markov生灭模型[J]. 东南大学学报: 自然科学版, 2008, 38(4): 593-597.
- [8] 方群. P2P文件污染随机模型[J]. 小型微型计算机系统, 2009, 30(10): 1980-1984.
- [9] 云燕. P2P文件污染的传播建模分析和防治策略研究[D]. 大连: 大连理工大学, 2008.
- [10] MAO JUNPENG, CUI YANLI, HUANG JIANHUA, et al. Analysis of pollution disseminating model of P2P network [C]// Proceedings of the Second International Symposium on Intelligent Information Technology Application 2008. Washington, DC: IEEE Computer Society, 2010: 790-794.

(上接第2151页)

- [3] GOSEVA-POPOSTOJANOVA K, WANG F, WANG R, et al. Characterizing intrusion tolerant systems using a state transition model [C]// DISCEX'01: Proceedings of the DARPA Information Survivability Conference and Exposition. Anaheim, CA: [s.n.], 2001: 211-221.
- [4] MEHTA V, BARTZIS C, ZHU H, et al. Ranking attack graphs [C]// RAID 2006: Proceedings of the International Symposium on the Recent Advances in Intrusion Detection. Berlin: Springer-Verlag, 2006: 127-144.
- [5] 黄光球, 乔坤, 朱华平. 基于FPN的模糊攻击图模型及生成算法研究[J]. 微电子学与计算机, 2007, 24(5): 162-165.
- [6] 黄光球, 任大勇. 基于双枝模糊决策与模糊Petri网的攻击模型[J]. 计算机应用, 2007, 27(11): 2689-2693.
- [7] 黄光球, 王金成. 基于双枝模糊集的一致性模糊变权Petri网攻击模型[J]. 计算机应用, 2009, 29(2): 529-533.
- [8] 黄光球, 李艳. 基于粗糙图的网络风险评估模型[J]. 计算机应用, 2010, 30(1): 190-195.
- [9] 尚大鹏, 张冰, 周渊, 等. 一种深度优先的攻击图生成方法[J]. 吉林大学学报: 工学版, 2009, 39(2): 447-451.
- [10] 赵芳芳, 陈秀真, 李建华. 基于权限提升的网络攻击图生成方法[J]. 计算机工程, 2008, 34(23): 158-160.
- [11] 姜伟, 方滨兴, 田志宏. 基于攻防博弈模型的网络安全测评和最优主动防御[J]. 计算机学报, 2009, 32(4): 817-825.
- [12] 王纯子, 黄光球. 基于脆弱性关联模型的网络威胁分析[J]. 计算机应用, 2010, 30(11): 3046-3050.