TSINGHUA SCIENCE AND TECHNOLOGY
ISSN 1007-0214 07/19 pp465-469
Volume 14, Number 4, August 2009

Locating Highly Connected Nodes in P2P Networks with Heterogeneous Structures*

ZHANG Haoxiang (张浩翔), ZHANG Lin (张 林)**, SHAN Xiuming (山秀明), Victor O. K. LI (李安国)†

Department of Electronic Engineering, Tsinghua University, Beijing 100084, China;
† Department of Electrical and Electronic Engineering, The University of Hong Kong, Hong Kong, China

Abstract: Peer-to-peer (P2P) networks aggregate enormous storage and processing resources while minimizing entry and scaling costs. Gnutella-like P2P networks are complex heterogeneous networks, in which the underlying overlay topology has a power-law node degree distribution. While scale-free networks have great robustness against random failures, they are vulnerable to deliberate attacks where highly connected nodes are eliminated. Since high degree nodes play an important role in maintaining the connectivity, this paper presents an algorithm based on random walks to locate high degree nodes in P2P networks. Simulations demonstrate that the algorithm performs well in various scenarios and that heterogeneous P2P networks are very sensitive to deliberate attacks.

Key words: random walks; scale-free; peer-to-peer; largest component; deliberate attack

Introduction

Large, complex networks are an essential part of the modern world. A complex network consists of a set of nodes and a set of edges (directed or undirected) connecting the nodes. Complex networks can be divided into two major classes based on their connectivity distribution p_k[1], which gives the probability that a randomly chosen node in the network is connected to k other nodes. The first class of networks has a homogeneous structure, in which each node has approximately the same number of links. However, many networks, such as the World Wide Web (WWW)[2], the Internet, and other large networks[3], belong to the second class of heterogeneous networks. One important class of the heterogeneous networks is the scale-free (SF) networks characterized by power-law degree connectivity distributions[2]. The connectivity distribution follows $p_k \sim k^{-\gamma}$, where γ is the power-law exponent.

In recent years, peer-to-peer (P2P) systems have emerged as a significant social and technological phenomenon. This new breed of distributed systems creates application-level virtual networks that aggregate enormous storage and processing resources while minimizing entry and scaling costs[4]. As most P2P applications operate on unstructured overlay networks, which exhibit self-organized characteristics, the topology of these overlay networks significantly impacts the application's properties. Recent measurements of Gnutella and simulated FREENET networks, both very popular P2P file-sharing systems, show that the underlying overlay topology has a power-law node degree distribution[5]. The node degrees exhibit a high variance, with a few nodes having very high degrees while many others have very low degrees. This study indicates that Gnutella-like P2P file-sharing networks belong to the class of heterogeneous complex

Received: 2008-04-03; revised: 2008-10-25

∗ Supported by the National Natural Science Foundation of China (Nos. 60672107, 60674048, 60603068, and 60772053), the National High-Tech Research and Development (863) Program of China (No. 2006AA10Z261), and the National Key Basic Research and Development (973) Program of China (No. 2007CB307105)

∗∗ To whom correspondence should be addressed.
E-mail: linzhang@tsinghua.edu.cn; Tel: 86-10-62781382

networks.

Due to the network's ad hoc nature, file locating becomes one of the most challenging issues in these P2P file-sharing systems. Random walk-based search strategies have been proposed to tackle the scalability problem[6], in which a query message referred to as a walker is forwarded to a randomly chosen neighbor until the search succeeds or the time to live (TTL) reaches 0. In homogeneous networks, random walks achieve load balancing, while in heterogeneous topologies with power-law degree connectivity distributions, walkers naturally gravitate towards the high degree nodes[5].

This paper describes a random walk-based search algorithm for high degree node localization. Such highly connected nodes play an important role in SF networks with the total system dependability highly dependent on these nodes. Simulations are used to investigate the performance of the localization algorithm and how the network connectivity varies when the highly connected nodes are removed.

1 Random Walks in Heterogeneous Networks

Several related works have focused on the properties of random walks in hierarchical networks to show that the high degree nodes will be encountered more often by the walker.

Since the random walk is a fundamental stochastic process, Noh and Rieger[7] mapped the random walk onto a dynamic Ising spin chain system to investigate the characteristic relaxation time in heterogeneous networks. Noh and Rieger[8] focused on the mean first passage time (MFPT) of the random walk between two arbitrary nodes.

Consider a finite undirected network with N nodes. The connectivity of the network is represented by the adjacency matrix $\boldsymbol{A}$ whose elements $A_{ij}=1$ or 0 depending on whether there is a link between nodes i and j. Since the network is undirected, $A_{ij}=A_{ji}$. The degree of each node i is then denoted by $d_i=\sum_j A_{ij}$.

Suppose that the walker starts at node i at time $t=0$. In discrete time, the stochastic process of the random walk is described by the master equation,

$$P_{ij}(t+1)=\sum_k \frac{A_{kj}}{d_k}P_{ik}(t) \tag{1}$$

where $P_{ij}(t)$ denotes the probability of finding the walker at node j at time t.

If the network contains an odd number of loops, the infinite time limit of the probability converges to the stationary distribution $P_j^{\infty}=\lim_{t\to\infty}P_{ij}(t)$ given by

$$P_j^{\infty}=\frac{d_j}{D} \tag{2}$$

with $D=\sum_i d_i$ [8]. The stationary distribution is equal to the node's normalized degree. Thus, those nodes with a very high degree will be visited by a random walker more often. In finite networks, the MFPT of the random walker between two nodes is

$$\langle T_{ij}\rangle=\begin{cases}\dfrac{D}{d_i}, & j=i;\\ \dfrac{D}{d_j}[R_{jj}^{(0)}-R_{ij}^{(0)}], & j\neq i\end{cases} \tag{3}$$

with $R_{ij}^{(n)}=\sum_{t=0}^{\infty}t^n(P_{ij}(t)-P_j^{\infty})$. The average recurrence time $\langle T_{ii}\rangle$ is determined only by the total number of links and the degrees of the nodes. In a heterogeneous network with power-law degree distribution $p_k\sim k^{-\gamma}$, the MFPT will be much smaller for the high degree nodes than that for the low ones.

The random walk centrality (RWC)[8] is a measure of the effectiveness of communication between nodes. The definition of the RWC for node i is

$$C_i=\frac{P_i^{\infty}}{\tau_i} \tag{4}$$

where τ_i is the relaxation time of node i, $\tau_i=R_{ii}^{(0)}$.

Since the relaxation time, τ_i, is presumed to have weak node dependence, the RWC of node i explicitly depends on the node degree. Thus, for random walks in heterogeneous networks, nodes with larger RWC will typically be visited earlier by the random walkers than nodes with smaller RWC by the random walkers.

2 High Degree Node Detection in P2P Networks

With the power-law nature of the node degree distribution, a very small percentage of peers have a very high degree and the total system dependability relies greatly on the operation of such peers[4]. These peers can be located by a random walk-based search algorithm for high node degree localization.

In a general random walk search in P2P networks, each query message contains a descriptor header. The intermediate node, upon receiving the walker, discards the message if its descriptor ID is the same as the one it has received before. The unique descriptor ID avoids routing loops, which reduces the search bandwidth. While the routing loops are avoided in the search algorithms, they are fully utilized in the algorithm for high degree node localization.

A power-law graph with exponential cutoff is generated using the algorithm given by Newman et al.[9] The default graph has $N = 10\,000$ nodes, with the degree distribution $p_k = \dfrac{\kappa^{-\gamma} \mathrm{e}^{-k/\kappa}}{\mathrm{Li}_\gamma(\mathrm{e}^{-1/\kappa})}$, where $\mathrm{Li}_n(x) = \sum_{k=1}^{\infty} \dfrac{x^k}{k^n}$ is the n-th polylogarithm of x, γ is the power-law exponent, and κ is the cutoff. Each peer initiates a random walker to locate the high degree nodes. When the TTL reaches 0, the walker sends back the list of peers encountered along the reverse path. The initiating nodes choose the k nodes appearing most frequently in the list as the high degree node candidates, thus forming an $N \times k$ matrix.

Figure 1 shows the average rate of the three highest degree nodes appearing in the matrix. As expected, the nodes with larger degrees are visited more often by a random walker. Thus, these nodes have a high probability of being listed. The order of the node appearance rate is the same as the order of the node degree. Also, as the TTL increases, the random walk process converges gradually to a stationary distribution. Thus, the highly connected nodes will be visited more often by a walker, resulting in steady growth of the appearance rates for the highest degree nodes.

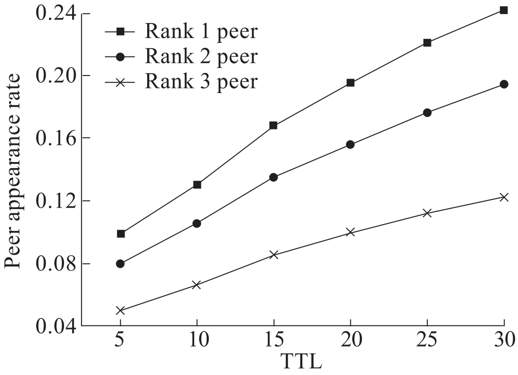

Fig. 1 Appearance rate of the highest degree nodes (γ= 2.041 289, κ= 1000, N= 10 000, k = 3)

Since the scale of P2P networks can be quite large and may cover a wide geographical area, it is more realistic to label and control a small number of the nodes belonging to the P2P network. Suppose that we have a small number of the total nodes under control. The problem is then to locate the most highly connected nodes in this small part of the already known nodes.

The algorithm developed for high degree node detection is quite straight forward with two steps. S is a subset of the peers in the already known P2P system. The first step is the same with each node, i, in S initiating a random walker with a large TTL. Since the subset S contains N_s nodes, each node chooses the k nodes appearing most frequently in the list to form an $N_s \times k$ matrix. The second step is more collaborative with all the nodes in S gathering the lists to one head node, which can be selected randomly. The head node then chooses and locates those k peers which appear most often in the matrix. Therefore, if N_s is small, the overhead would be reduced, as the algorithm would generate $N_s \times \mathrm{TTL}$ messages.

Figure 2 shows how well the three nodes chosen after two rounds of filtration match the three largest degree nodes in the network. The metric is the absolute detection rate, i.e., the rate that the k nodes chosen based on the random walks exactly match the k largest degree nodes. Since the TTL and the ratio of the participating nodes become larger, the detection rate increases. The algorithm achieves detection rates

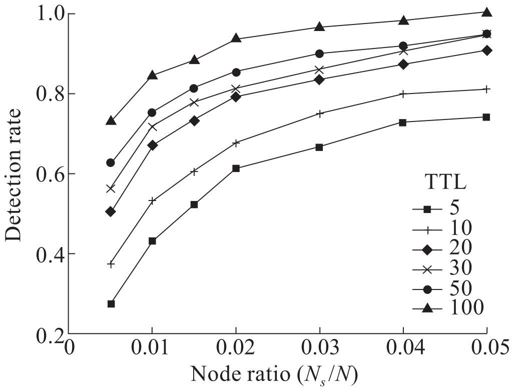

Fig. 2 Absolute detection rate for different node ratios (γ= 2.041 289, κ= 1000, N= 10 000, k = 3)

greater than 80% even if 2% of the total nodes initiate a random walk with TTL larger than 20.

Since the difference in the degree number between rank k and rank $k+1$ nodes decreases with increasing k, it gets more difficult to distinguish the two nodes based on random walks. Also, as k increases, the differences in the roles played by the two nodes in the network become less significant. Define a parameter η, called the weighted detection rate, as a new performance metric, where η is the weighted average success rate ranked by the node degree.

$$\eta = \frac{\sum_{i=1}^{k} R_i \gamma^{-(i-1)}}{\sum_{i=1}^{k} \gamma^{-(i-1)}} \tag{5}$$

where γ is the power-law exponent and R_i denotes the success rate in locating the i-th largest degree node. The power-law exponent γ determines the role difference that the nodes play in scale-free networks, which affects the weighted detection rate.

Figure 3 shows the performance of the largest degree node localization based on this metric. The weighted detection rate is much larger than the absolute detection rate shown in Fig. 2. With the limited number of nodes in the network, the degree of the third most connected node is close to that of the less connected ones. Random walks can easily distinguish the top two nodes, but sometimes have difficulty in locating the third highest when only a small fraction of the total nodes is used.

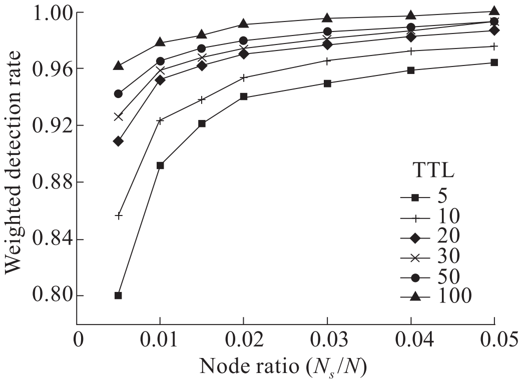

Fig. 3 Weighted detection rate for different node ratios (γ=2.041 289, κ=1000, N=10 000, k=3)

High degree nodes play an important role in maintaining network connectivity. Albert et al.[1] illustrate that though scale-free networks display a surprisingly high degree of tolerance against random failures, they are extremely vulnerable to attacks. A deliberate attack will not randomly eliminate nodes, but will preferentially target the most connected node in the SF networks.

The damage caused by the attack/removal of the high degree nodes can be quantified in terms of the relative size, G, of the largest connected component.

$$G = \frac{N'}{N} \tag{6}$$

where N and N' are the numbers of nodes in the largest connected component before and after the removal of the highest degree nodes.

Figure 4 shows the network fragmentation under a malicious deliberate attack. After the removal of the three most connected nodes, the largest connected component in the topology consists of only 86.54% of the remaining nodes. As the number of the most connected nodes being removed increases, the overlay will be shattered into a large number of disconnected components. For example, the relative size of the largest connected component drops below 75% when only 0.1% of the nodes (the most connected ones) are removed from the network.

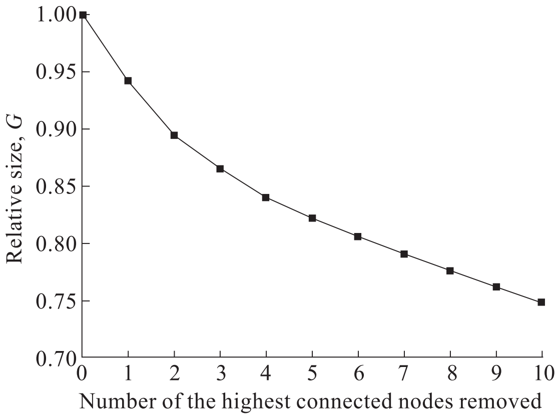

Fig. 4 Relative size of the largest connected component (γ=2.041 289, κ=1000, N=10 000)

The random walk algorithm may be utilized by malicious attackers to locate the high degree nodes and reduce the system failure tolerance. As a precautionary measure, the status of these highly connected nodes should be periodically examined.

3 P2P Botnets

Recently, Botnets have been identified as one of the biggest threats to the security of the Internet[10]. A Botnet is a collection of a large number of bots, which are programs automatically performing user-centric tasks without any interaction from a user[11]. The traditional hierarchical architectures with a centralized command and control (C&C) master are susceptible to attack; therefore, more resilient C&C architectures have to be developed. For example, P2P-based Botnets are becoming increasing popular and may become more widespread than traditional architectures[11]. The P2P's decentralized feature will avoid single point failures, which makes the Botnet a much greater threat. Phatbot[12] infected hosts by utilizing Gnutella cache servers to find other peers. A highly-connected node localization algorithm based on random walks can then be implemented together with other metrics[13] to detect P2P Botnets.

4 Conclusions

A localization algorithm was developed to find high degree nodes in P2P networks based on random walk searches. Simulations verify that nodes with very high connectivity degrees will be visited more often by a random walker. The simulations also show that the power-law degree distribution is not so resilient to attacks as previously thought. Although SF networks display a high degree of tolerance against random failures, they are much more vulnerable to deliberate attacks which target the easily identified highly connected nodes. Thus, the heterogeneous connectivity distribution feature of P2P overlays is vulnerable to deliberate attacks and new overlay topologies should be developed.

References

[1] Albert R, Jeong H, Barabasi A. Error and attack tolerance of complex networks. *Nature*, 2000, **406**(6794): 378-382.

[2] Albert R, Jeong H, Barabasi A. Diameter of the World Wide Web. *Nature*, 1999, **401**: 130-131.

[3] Albert R, Jeong H, Barabasi A. Statistical mechanics of complex networks. *Rev. Mod. Phys.*, 2002, **74**: 47-97.

[4] Risson J, Moors T. Survey of research towards robust peer-to-peer networks: Search methods. *Computer Networks*, 2006, **50**(17): 3485-3521.

[5] Adamic L, Lukose R, Puniyani A, et al. Search in power-law networks. *Physical Review E*, 2001, **64**(4): 046135.

[6] Lv Q, Cao P, Cohen E, et al. Search and replication in unstructured peer-to-peer networks. In: Proceedings of the 16th International Conference on Supercomputing. New York, USA, 2002: 84-95.

[7] Noh J, Rieger H. Constrained spin-dynamics description of random walks on hierarchical scale-free networks. *Physical Review E*, 2004, **69**(3): 036111.

[8] Noh J, Rieger H. Random walks on complex networks. *Physical Review Letters*, 2004, **92**(11): 118701.

[9] Newman M, Strogatz S, Watts D. Random graphs with arbitrary degree distributions and their applications. *Physical Review E*, 2001, **64**(2): 026118.

[10] Schiller C A, Binkley J, Harley D, et al. Botnets: The Killer Web App. Rockland, MA, USA: Syngress Publishing, 2007.

[11] Grizzard J, Sharma V, Nunnery C, et al. Peer-to-peer Botnets: Overview and case study. In: Proceedings of 1st Workshop on Hot Topics in Understanding Botnets. Berkeley, USA, 2007.

[12] Joe S. Phatbot Trojan analysis. http://www.secureworks.com/research/threats/phatbot/?threat=phatbot. March 15, 2004.

[13] Akiyama M, Kawamoto T, Shimamura M, et al. A proposal of metrics for Botnet detection based on its cooperative behavior. In: Proceedings of the 2007 International Symposium on Applications and the Internet Workshops. Hiroshima, Japan, 2007.