A complex network model based on the Gnutella protocol

Xiaoguang Qi [a,c], Guang Yue [b,c], Liang Zhang [c], Mingfeng He [c,d,*]

[a] School of Energy and Power Engineering, Dalian University of Technology, Dalian 116024, China
[b] School of Physics and Optoelectronic Technology, Dalian University of Technology, Dalian 116024, China
[c] School of Innovation Experiment, Dalian University of Technology, Dalian 116024, China
[d] School of Mathematical Science, Dalian University of Technology, Dalian 116024, China

ARTICLE INFO

ABSTRACT

Gnutella is one of the basic protocols for P2P software. In this paper, a novel network model based on Gnutella is introduced. The mechanism of this network is based on resource occupancy and search activities of peers. As for the structure, the power-law exponent of in-degree $\gamma_{in} \approx 4.2$, the length of the average shortest path $\langle l \rangle = 57.74$, and the diameter of the network is 156; these topological properties of the proposed structure differ from known results.

© 2009 Elsevier B.V. All rights reserved.

1. Introduction

In recent years, there is a considerable number of complex networks [1] in communication systems such as the Internet, World Wide Web [2], and peer-to-peer networks. In these networks users are represented by nodes, and communication sessions correspond to edges. The whole network can be modelled by a graph, which constitutes an active research area in the study of complex networks [3].

P2P (peer-to-peer) technology has been a hot topic in Internet development in recent years. A number of P2P networks are based on the Gnutella protocol [4]. The prominent characteristic of the Gnutella protocol system is its decentralization, meaning the equality of each user in the network. Moreover, users are linked by functions in the protocol; for more details, see Ref. [5]. A number of studies on Gnutella and models derived from it have been carried out, including topological properties of P2P networks based on measurements of real implementations [6–8], models and analyses based on dynamic activities of peers such as file sharing and data researching [9–13], and discussions of modified protocols and relative network structures with special characteristics [14–17]. Moreover, scale-free phenomena have been seen to emerge in these networks [18,19].

These discussions do not include simplified mechanisms of resource occupancy, search or dynamics. The objective of this paper is to constitute a network structure under representative rules derived from Gnutella protocol v0.4, and we made these rules as concise as possible, meanwhile we focused on the topology of this network structure in this work. We thought that in this way our result may generally reflect the influence of Gnutella protocol on network structure. "flooding" mechanism

is the core concept in message searching of Gnutella protocol which has been applied to P2P softwares widely, and these softwares have additional technical details in their operations. Here, to give prominence to Gnutella, we have ignored other technical details. It is significant to note that we cannot construct a network only with Gnutella protocol, hence we need to introduce definite rules as we described in our paper. The two principles are: Nodes in the network select their new neighbors that have the highest data transfer rate; each node has limited resource storage. We employ complex network theory to analyze this model. The topological properties of this structure are different from earlier results.

2. Structure based on the Gnutella protocol

Gnutella is a typical P2P protocol in which terminals in the network engage in reciprocal activity under decentralization. This protocol is significant for the evolution of current P2P networks; therefore, models based on it are appropriate for simulating a P2P environment.

One cannot construct a network only with this protocol. Considering the following background: (1) Clients link to neighbors depend on partiality to certain resource, transfer efficiency has been taken into consideration as well. (2) Clients delete old files in order to storage new resource that has large capacity. Hence, we give the two rules based on Gnutella in order to construct a network: Nodes in the network select their new neighbors that have the highest data transfer rate; each node has limited resource storage.

According to the above rules, nodes in this network need to consider transfer efficiency as well as required resource while selecting their neighbors. We established a two-dimensional plane, made position of each node randomly. And then simplify all the factors that have effects on transfer efficiency, regarded linear distance between nodes as transfer rate. This is a reflection of real network, nodes are distributed in the network environment and in our case, and we assumed that transfer rate became lower when distance is longer although the transfer efficiency is not determined by the distance completely in real condition.

The reader is referred to Gnutella v0.4 [5] for further details on this directed structure.

Definition of parameters
N Number of terminals in the network
L Number of neighbors to which each terminal connects
K_i In-degree of terminal i, meaning number of connections to this terminal
(x_i, y_i) Relative position of terminal i
TTL Time To Live, the lifetime of a data request
R Resource storage capacity for each terminal
M Number of existing resource categories in the network.
It should be noted that $M > R$, $L \times R > M$.

Establishing this network requires four steps, of which steps 1 and 2 involve initialization:

(1) Using a random technique [20], obtain N pairs of coordinates as coordinates for N nodes in the system, where the terms "node" and "terminal" are equivalent. These coordinates are kept constant to fix nodes on the two-dimensional coordinate plane.

(2) Let each node connect to L nodes randomly. As nodes build up directed connections to others, their neighbor lists will record their neighborhoods. Meanwhile, neighbors are listed in order of technical length. Technical length is defined as coordinate distance, expressed as $\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$, between node i and j. This parameter is introduced to simulate differences in connection efficiency because of various factors such as the Internet environment. Next, allow each node to have resources that are random both in quantity and in category. These resources are represented by integers between 1 and M.

Steps 3 and 4 constitute the dynamic mechanism:

(3) Each node will send a data request to all its neighbors to search randomly for one category of resource that is not available in its own storage. Each data request includes resource information (an integer from 1 to M) and TTL (lifetime of this data request; TTL counts down after being sent from one terminal to another, and the request will cease when TTL becomes zero. Hence, a request from one node will theoretically transfer to L^{TTL} nodes in this case.)

(4) While a data request sent by a node is stopped, those receiving the request will feed back the desired information if they possess it, and the requester will select nodes among them according to their technical lengths toward the requester. Then this requester covers its neighbor list by selecting nodes with minimum technical lengths. Meanwhile, it will switch the initial connections to these new neighbors. On the other hand, the node will save this retrieved resource, eliminating an existing one randomly if its storage is full. Now this node has finished its search process.

The definition of a time step in this instance denotes accomplishment of search activity by N nodes seriatim.

3. Results and analysis

In this simulation, the following parameter values are used:

$N = 10000$, $L = 5$, $0 \leq x \leq 100$, $0 \leq y \leq 100$, $R = 5$, $M = 16$, $TTL = 3$.
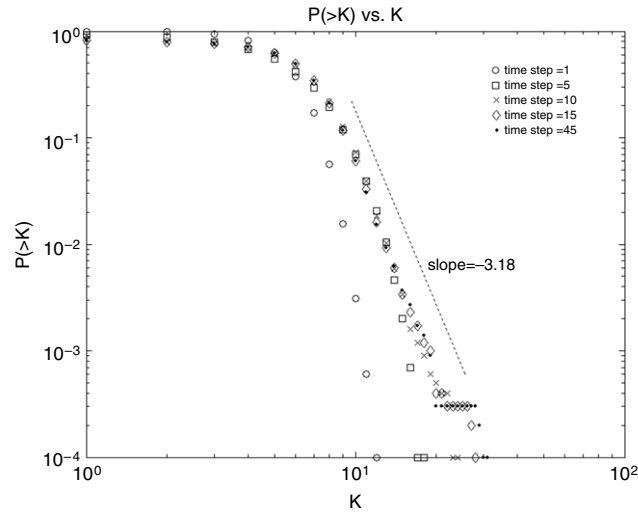
Fig. 1. Cumulative degree distributions at selected time steps.
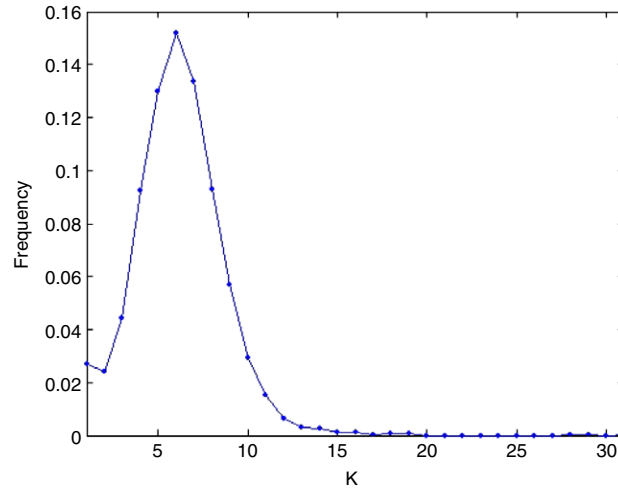

Fig. 2. Degree frequency distribution.

As mentioned above, this is a directed network, and the out-degree of each node is fixed while the in-degree varies according to a dynamic procedure. Hence, the in-degree encapsulates the connectedness of the terminals. Consequently, "degree" means in-degree except when otherwise noted in the following discussion.

Set $L \times R > M$; this means that one terminal is able to obtain all categories of resources through its neighbors, and this will result in stability of the structure after a certain number of time steps. With the parameter values used, the structure attains stability at an average of 45 time steps. It is noteworthy that fewer than 19% of the nodes become inefficient nodes, meaning that the in-degree of these nodes is zero, when the system has reached steady state.

The cumulative degree distribution $p(>K)$ is the probability that a randomly selected node has an in-degree greater than K and is shown in Fig. 1.

At steady state, this curve clearly, except for its flat head, obeys a power-law function such as $\sim K^{-\tau}$ with $\tau = 3.18 \pm 0.01$. One can immediately obtain the degree distribution $p(K) \sim K^{-\gamma_{in}}$ with $\gamma_{in} = \tau + 1 = 4.18$. The value of γ_{in} varies in real-world examples such as social networks and Internet website linking networks. According to previous work [1], the degree exponent normally falls in the range $2 < \gamma_{in} < 3$. Obviously, the structure proposed here has an exponent that is outside of this range.

The following analyses are based on a steady-state condition. The degree frequency distribution is shown in Fig. 2.

From Fig. 2, it is clear that most nodes have an in-degree between 3 and 10. The average in-degree $\langle K \rangle = 5$ can be calculated and is definitely equal to L. Typical scale-free models use a probability based on the degree of each node to construct connections [1], while in the present case, the probability is based on resource occupancy state.
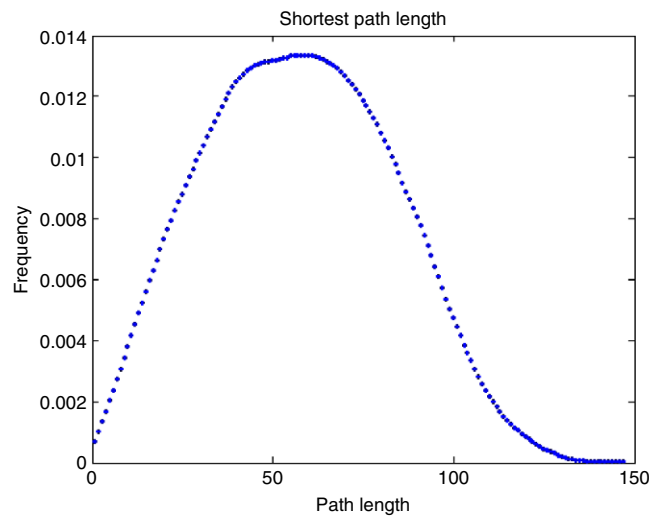
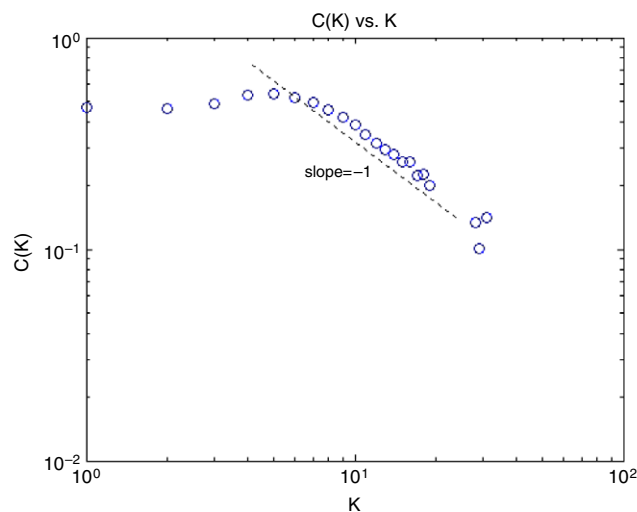**Fig. 3.** Shortest-path-length distribution.

Fig. 4. Local clustering coefficient vs. in-degree distribution; the dashed line has slope −1.

The shortest-path-length distribution is shown in Fig. 3. The shortest path length is defined as the minimum number of connections between two nodes that have at least one path connecting them.

The length of the average shortest path $\langle l \rangle$, which is the mean of shortest path length, can be calculated. The result obtained here is 57.74, and the diameter is 156, the maximum length.

Previous studies on directed Internet structures propose values for $\langle l \rangle$ between 10 and 20 [1,21]. The $\langle l \rangle$ value obtained here is outside this range, which can be explained by the fact that the structure proposed here is based on a resource-searching process. Thanks to the equal probability of resource selection and the fact that $N \gg M$, there is a considerable number of nodes which contain the same type of resource. Therefore, the terminal will select nearby neighbors instead of distant neighbors. A pair of nodes can be connected through a large number of neighbors if their technical length is not short. This phenomenon gives rise to the emergence of clusters and homogenizes the structure. In addition, it will influence the risk of being attacked [22].

The clustering coefficient is illustrated in Fig. 4. The clustering coefficient of node i is defined as $C_i = \frac{2E_i}{K_i(K_i-1)}$, that is, the ratio between the number E_i of edges that actually exist between the K_i neighbors of node i and the total number of possible edges, $K_i(K_i - 1)$, when $K_i = 1$, $C_i = 0$. The clustering coefficient of the whole structure, C, is the average of all individual C_i's. The network was made undirected so that the clustering coefficient could be measured [23].

The power-law form of the clustering coefficient is evident in this case. Its form is similar to that of the semantic web of synonyms in the Merriam-Webster dictionary [24]. Here, $C = 0.3467$, and the distribution clearly shows the small-world property [25], with prominent clusters. Fig. 5 shows snapshots of the network process which represent this phenomenon.
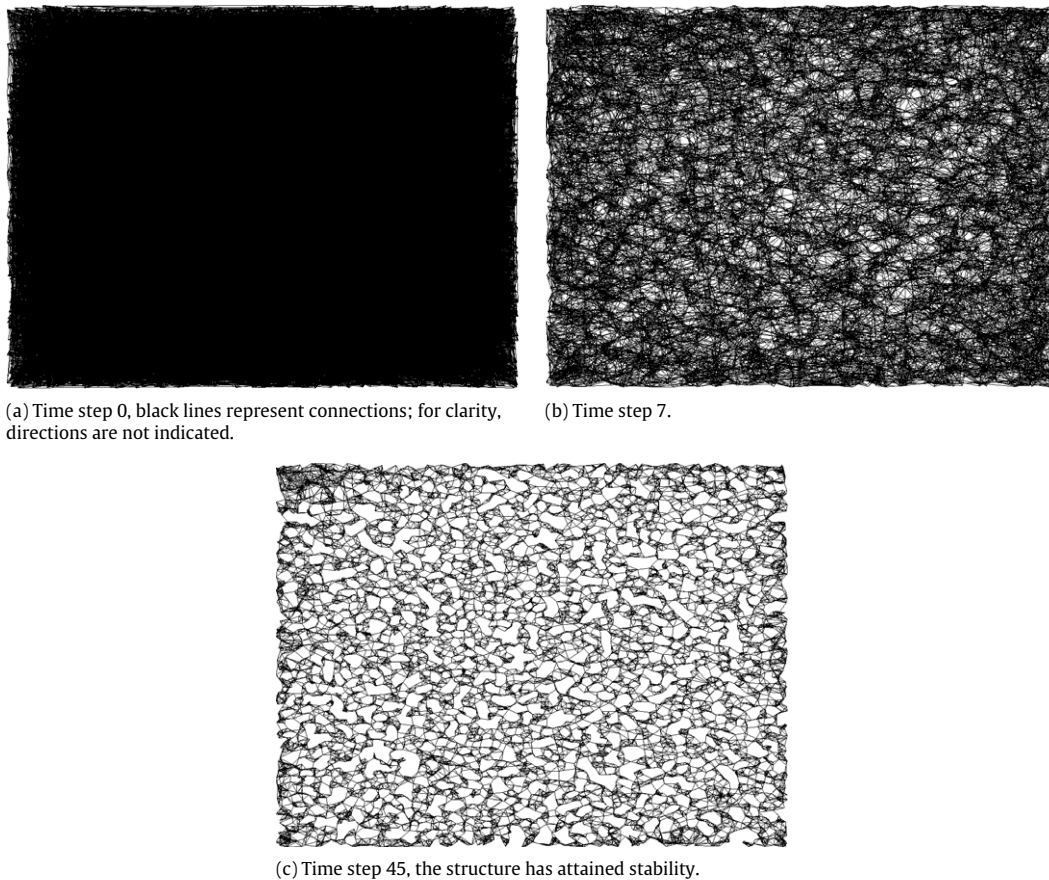
(a) Time step 0, black lines represent connections; for clarity, directions are not indicated.

(b) Time step 7.

(c) Time step 45, the structure has attained stability.

Fig. 5. Snapshots of the network process.

Through visualizations, it is possible to observe that the dynamic process generates clusters and attains stability gradually after a visible initial transformation.

4. Conclusions

In the structure proposed here, a dynamic mechanism has been implemented based on a resource-searching operation derived from the Gnutella protocol. This structure has the scale-free property, and the measurements made in this study of the power-law exponent, the average shortest path length, and the diameter of the network show particularity when compared with previous work. This research provides a novel perspective on the construction and evolution of real P2P networks.

Acknowledgment

References

 [1] R. Albert, A.L. Barabási, Reviews of Modern Physics 74 (2002) 47.
 [2] F. Fu, L.H. Liu, L. Wang, Physica A 387 (2008) 675–684.
 [3] Y. Liu, J. Yuan, X.M. Shan, Y. Ren, Z.X. Ma, Physica A 387 (2008) 2145–2154.
 [4] Gnutella, 2000. http://www.Gnutella.com.
 [5] Gnutella Protocol Specification v0.4, 2007. http://www.clip2.com.
 [6] C. Xie, et al., Computer Communications 31 (2008) 190.
 [7] D. Stutzbach, R. Rejaie, S. Sen, IEEE-ACM Transactions on Networking 16 (2008) 267.
 [8] M. Ripeanu, A. Lamnitchi, I. Foster, IEEE Internet Computing 6 (2002) 50.
 [9] S. Lee, S.-H. Yook, Y. Kim, Physica A 385 (2007) 743–749.
[10] S. Blanas, V. Samoladas, Further Generation Computer Systems 25 (2009) 100.
[11] F. Wang, Y.R. Sun, Computational Intelligence 24 (2008) 213.

[12] R. Gaeta, M. Sereno, Concurrency and Computation—Practice & Experience 20 (2008) 713.
[13] E. Pournaras, G. Exarchakos, N. Antonopoulos, Computer Communications 31 (2008) 3030.
[14] Z.Z. Zhu, P. Kalnis, S. Bakiras, IEEE Transactions on Parallel and Distributed Systems 3 (2008) 363.
[15] M. Karakaya, I. Korpeoglu, O. Ulusoy, Computer Networks 52 (2008) 675.
[16] G. Pandurangan, P. Raghavan, E. Upfal, IEEE Journal on Selected Areas in Communications 21 (2003) 995.
[17] A. Beygelzimer, G. Grinstein, R. Linsker, I. Rish, Physica A 357 (2005) 593–612.
[18] L.A. Adamic, R.M. Lukose, A.R. Puniyani, et al., Physical Review E 64 (2001) 046135.
[19] A.L. Barabási, et al., Science 286 (1999) 509.
[20] D.P. Landau, K. Binder, A Guide to Monte Carlo Simulations in Statistical Physics, Cambridge University Press, Cambridge, UK, 2000.
[21] A. Broder, et al., Computer Networks 33 (2000) 309.
[22] E. Ben-Naim, H. Frauenfelder, Z. Toroczkai (Eds.), Complex Networks, Springer Press, Berlin, 2004.
[23] L.A. Adamic, B.A. Huberman, Nature 401 (1999) 131.
[24] E. Ravasz, A.L. Barabási, Physical Review E 67 (2003) 026112.
[25] S.H. Strogatz, Nature 410 (2001) 268.