Physica A 388 (2009) 3955–3960

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

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

Article history:
Received 2 March 2009
Received in revised form 10 May 2009
Available online 6 June 2009

PACS:
89.75.-k
89.75.Da

Keywords:
Peer-to-peer
Gnutella
Complex networks
Resource searching
Scale-free

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.

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

* Corresponding author at: School of Mathematical Science, Dalian University of Technology, Dalian 116024, China. Tel.: +86 411 8470 6093.
E-mail address: mfhe@dlut.edu.cn (M. He).

0378-4371/$ – see front matter
doi:10.1016/j.physa.2009.06.007

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:

$$N = 10000, \quad L = 5, \quad 0 \leq x \leq 100, \quad 0 \leq y \leq 100, \quad R = 5, \quad M = 16, \quad TTL = 3.$$