The Parallel Rings Topology in Semantic Peer-to-Peer Networks

Bertalan Forstner, Dr. Hassan Charaf

Budapest University of Technology and Economics
Goldmann György tér 3, H-1111 Budapest, Hungary
{bertalan.forstner, hassan.charaf}@aut.bme.hu

Abstract: The Peer-to-Peer networks are very popular in decentralized information retrieval; however, decreasing the network traffic generated by the broadcast messages is still an open question. There are some semantic approaches that implements metadata-based routing or node selection as a new layer on existing protocols. However, the high connectedness of such networks could decrease the number of successful queries.
In this paper we present a solution for the clustering of semantic peer-to-peer networks by altering a semantic protocol to construct a topology that better serves for the intelligent neighbor selection. The topology eliminates the counterproductive links from the network, hence the number of hops of the messages required for a succesful query can be decreased and the communication costs can be kept low.

Keywords: Peer-to-Peer Networks, Protocols, Topology, Modeling

1 Introduction

As in Peer-to-Peer (P2P) networks the required network bandwidth is very important, more and more attempts were made to develop more efficient and scalable protocols. After examining some existing systems, we developed a new protocol, the SemPeer, which was introduced in [15]. SemPeer is a new layer on existing protocols that utilizes the semantic information available by the stored documents to transform the P2P network to be able to benefit from the locality in interest. We need a mathematical model to examine the theoretical capabilities of the new protocol. After examining different P2P models, we found that it is reasonable to prepare a new mathematical model that captures the aspects of the different fields of interest related to the nodes (users) in the system. The new model also describes the effect of clustering in the small-world network.

Selecting nodes with similar fields of interest to connect is not sufficient in self to increase the number of query hits. The protocol should ensure the construction of proper topology to avoid undesired side effects. In case of SemPeer the ideal

topology has to satisfy the following three criteria. First, the topology must benefit from the connections to similar nodes. Then, it should eliminate the effect of the clustering. And third, the topology should be constructed without using significant resources or nodes with special roles. We describe how our models helped us to develop a suitable topology for networks where network traffic should be decreased.

2 The SemPeer Protocol

Some initiatives are launched to make the Internet semantic, namely, provide it with metadata. Ontology-based information retrieval makes the search more intelligent than string matching alone [2]. We already mentioned WordNet, and another good example is the Dublin Core Metadata Initiative [3]. This project is dedicated to facilitate the widespread adoption of interoperable metadata standards and to develop specialized metadata vocabularies for describing resources that enable more intelligent information discovery systems. The viability and benefit of this initiative has been proven by the numerous projects built on it [3].

The Peer-to-Peer approach enables to make information retrieval more efficient, using a model well-known from everyday life. In the real world, working relationship is established among the people with a labor of the same topic. For example, if one's job is connected to the 19th century French literature, one's associates will have the same field of interest and probably have experience, books (documents), that is, relevant information, on the topic. If some related information needs to be found, then probably nobody would start with asking random people, but the mentioned experienced colleagues. This phenomenon is detailed in Section VI.

The Internet and Peer-to-Peer makes it possible to contact those people with whom we cannot enter into relations, because of geographical or other barriers. In the basic peer-to-peer protocols, the mentioned circumstances do not play any role in the selection of the adjacent nodes, thus search for the documents starts with querying the randomly selected neighbors. However, there are some methods elaborated to acquire ontology from documents [7], [8]. Then SemPeer creates profiles for the nodes that will describe the owner's fields of interests, for example with the appropriate weighting of the semantic categories provided by Dublin Core, or using the WordNet taxonomy. This profile creation can be fully automated. Thus, the construction of the peer network is not random, but we fundamentally consider that the fields of interest, namely, the profiles of the connecting nodes overlap as much as possible. As the individual nodes select their neighbors this way, we can assume that nodes in distance of two or more hops (the neighbors of our neighbors) also have a similar profile. This has the benefit of making the information retrieval more efficient, as the nodes reached during the

lifetime of the request (TTL) have relevant information with greater probability than selecting the neighbors in a random way. In a mobile P2P network, it is essential to decrease the network traffic with a good TTL strategy. The TTL parameter can be decreased when the most of the queries are answered by the nodes in only few hops distance.

There are other approaches that try to organize the nodes into clusters for better performance (for example [19]), but they request all the nodes to use the advanced P2P algorithm. A summary of the challenges with respect to these types of clustered P2P networks can be found in [20].

3 Peer-to-Peer Network Models

Considerable research effort has been involved in the examination of the performance of networks with client-server architecture [21]. There are some models elaborated to analyze the throughput, response time, and other parameters of the network. However, there are only very few papers concerning these issues of P2P networks. The main research directions can be characterized by the following types of network models.

The aspects of connection distribution of the large-scale P2P networks are modeled in [23]. This work describes the measures that affect the quality of service of the network, such as network latency or the short-circuit effect. However, it does not answer such questions such as the probability of success or the influencing parameters.

We found a quite useful model in [22]. The main goal of this model was to capture network throughput for three different classes of P2P networks. The one that describes the P2P architecture of distributed indexing with flooding architecture is suitable to obtain probabilistic results for Gnutella networks. However, it can hardly be transformed to use with clustered SemPeer networks, but we can use it to validate new models in extreme cases, as we will do it later in this paper.

After examining the available models we found that we should elaborate a new one to fully describe the novel SemPeer protocol.

4 The Clustering Problem

To be able to describe the connectivity of a graph in a formal way, we use a modified version of the clustering coefficient graph measure introduced by Watts and Strogatz [18].

Because of the nature of the Gnutella-based protocols, the high connectivity of the nodes with similar semantic profiles could lead to a very high clustering coefficient. This results in a query to arrive multiple times in different ways to some of the nodes in the group. Because of the connectedness, fewer nodes can be reached by a query, and also unnecessary computational resources are required. This can be described in a more formal manner as follows.

Consider a set of nodes, where the clustering coefficient is zero, i.e. no neighbors are connected with each other (Figure 1.a). In this case the number of nodes a query can reach is written as

$$M = \sum_{i=1}^{TTL} k^i \tag{1}$$

In Eq. (1), *TTL* represents the Time-To-Live parameter: a query should be propagated through *TTL* hops, and *k* the number of connections per node. Now we consider the worst case, when the clustering coefficient is 1. In this case the neighboring nodes form a fully connected directed graph, thus, the number of nodes reached by a query are decreased to *k* (Figure 1.b)

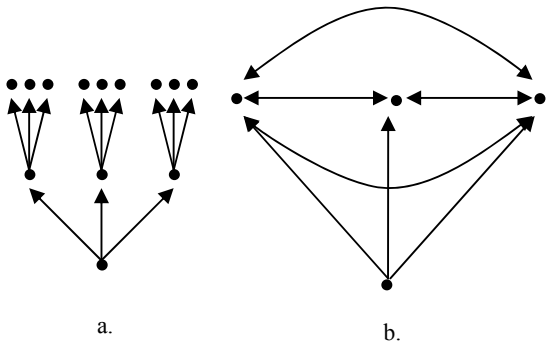

Figure 1
Directed graphs with extreme clustering coefficients, k=3, TTL=2, a. C=0, b. C=1

In a standard Gnutella network the coefficient is nearly zero as the graph can be regarded as a random mesh. In case of SemPeer, this measure can be quite high, depending on the popularity of the given group.

In the case of SemPeer not only the high connectedness of the neighbors causes a problem, but also some other kinds of links: first, the links backward in the propagation tree; second, links between nodes in the same level (siblings, in our wording); and third, links to neighbors of a sibling node. The first type decreases the nodes reached by a query in an obvious manner. The second and third types

can also cause ineffective query propagation, because when a query is issued by a node, it can be propagated back with high probability to a node that has already received it. These three types of connections should be avoided. They can be seen on the graph representation marked with dotted lines in Figure 3.

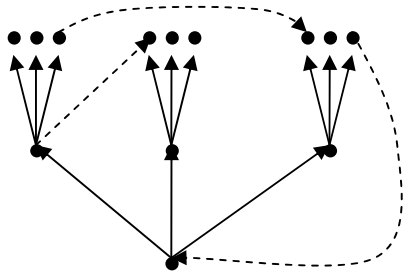

Figure 3
The dotted links decreases the number of reached nodes

To be able to measure this kind of connectedness we introduce a modified clustering coefficient. This measure has to be 0 if the nodes reached by a query constitute a tree, and it approaches 1 as the number of the counterproductive links increases.

Let $\{E_r^*\}$ stand for the set of counterproductive links in the propagation tree of a query initiated by the node v_r. For the sake of simplicity, we assume that each node has the same number of neighbors. In that case the modified clustering coefficient for the node r is

$$C_r = \frac{\left|\{E_r^*\}\right|}{\sum_{m=1}^{TTL}\left[k^m\left(\sum_{n=0}^{m-1}\left(k^n\right)+k^m-1\right)\right]+\sum_{m=1}^{TTL-1}k^m\left(k^{m+1}-k\right)} \tag{2}$$

Recall that Eq. 1 gives the number of nodes reached by a query. The defined number cannot be reached because of the effect of clustering: the same message is delivered to a node more than once. In [23] there is an experiment with the Gnutella protocol, where the proportion of the practical and theoretical number of reached nodes is computed, based on snapshots of the Gnutella network. The reach is measured with different TTL values. We were able to reproduce these results with the GXS Simulator [14].

5 Modeling the Networks

A mathematical model is required to prove that the theoretical efficiency of a P2P network can be extended with the advanced SemPeer protocol. Our goal is to increase the query hit in mobile networks with low network traffic, thus, we need a simple model to approximate the probability of a successful query in the standard Gnutella as well as in the SemPeer network. This new model should take clustering into account to be able to examine its effect with different topologies.

We regard the P2P network as a directed graph. Consider that the fields of interest of a user (represented by a node) can be determined with a single topic. We assume t different topics, the same number of nodes (V_t) and documents (D_t) with each topic, and every node obtaining D_n documents. The documents and the initiating links between the nodes are selected randomly with uniform distribution.

We can approximate the probability of a successful query in case of Gnutella as shown below:

$$P_{Success,Gnutella} = 1 - \left(1 - \frac{1}{tD_t}\right)^{D_n E_q} \tag{3}$$

In this formula, we compute the probability of not finding the requested document at any reached node, and subtract it from 1. The expression in bracket is the probability of selecting any disinterested document from all the documents that exist in the network. E_q is the number of reached nodes:

$$E_q = \sum_{i=1}^{TTL} \left[(1-C)k\right]^i \tag{4}$$

where C, the modified clustering coefficient is the average of all the C_r s.

C can be approximated by the following fraction:

$$C = \frac{\sum_{m=1}^{TTL}\left[k^m\left(\sum_{n=0}^{m-1}\left(k^n\right)+k^m-1\right)\right]+\sum_{m=1}^{TTL-1}k^m\left(k^{m+1}-k\right)}{tV_t\sum_{i=0}^{TTL}k^i} \tag{5}$$

In the case of SemPeer, we regard the steady state when nodes are connected only to other nodes from the same cluster. Therefore a search happens only in the set of documents related to only one topic, but also the clustering coefficient rises because the multiplier t in the denominator of Eq. 5 decreases to 1. The approximate probability of a successful query will be

$$P_{Success,SemPeer} = 1 - \left(1 - \frac{1}{D_t}\right)^{D_n E_q} \tag{6}$$

To validate our results and examine the behavior of our protocol, we have evaluated a series of simulations on the GXS Peer-to-Peer Simulator [14]. The model, supported by the results of the simulations, showed us that the response probability is quite low because of the clustering. The network has attributes which are characteristic for the "small worlds".

6 Small Worlds

The small-world phenomenon in the context of a worldwide social network refers to a widely accepted belief that we are all connected by a short chain of intermediate acquaintances. One of the first experimental studies of this phenomenon was conducted by Stanley Milgram in the late 1960s. The same phenomenon can be observed in computer networks with certain topologies. In case of Peer-to-Peer networks the task is also a search for short routes: the query issuer has to find a short route to the one that obtains the searched document. A couple of researches are involved in the investigation of the small world phenomenon [23] and the relations in such P2P networks.

We used the work of Faloutsos [24] as the basis for our research. He stated that certain power laws are valid in small world P2P networks. The most important for us is the one that introduces the hop-plot exponent. It states that the total number of pairs of nodes, *P(h)* within *h* hops is proportional to the number of hops to the power of a constant *H, the hop-plot exponent*:

$$P(h) \propto h^H \tag{7}$$

We consider the intuition behind the number of pairs of nodes *P(h)*. For $h=0$, we only have the self-pairs: $P(0)=N$. For the diameter of the graph δ, $h=\delta$, we have the self-pairs plus all the other possible pairs: $P(\delta)=N^2$, which is the maximum possible number of pairs. For a hypothetical ring topology, we have $P(h) \propto h^1$, and, for a 2-dimensional grid, we have $P(h) \propto h^2$, for $h<\delta$. [24] refines this approximation by calculating its proportionality constant as follows: The number of pairs within *h* hops is

$$P(h) = \begin{cases} c\,h^H, & h << \delta \\ N^2, & h \geq \delta \end{cases}, \tag{8}$$

where $c=N+2E$ satisfies the initial conditions. In the case of Peer-to-Peer searching, we have to reach a target node without knowing its exact position, thus selecting the TTL parameter of the broadcast is an issue. The effective diameter measure of a graph introduced by [24] helped us to approximate the TTL value which is satisfactory to reach a sufficiently large part of the network to have positive results with high probability. It is obvious that the intersection of the horizontal asymptote at level N^2 and the other one with the slope H gives us a promising solution. We can calculate the intersection point by the help of Power-Law 3. Given a graph with N nodes, E edges and H hop-plot exponent, we define the effective diameter, δ_{ef} as:

$$\delta_{ef} = \left(\frac{N^2}{N+2E} \right)^{1/H} \tag{9}$$

We will consider in the next section how this measure decreases with the unclustered SemPeer network.

7 The Parallel Rings Topology

Recall the clustering problem of the P2P networks. The task defined in Section IV is to ensure that a query does not arrive to a node more than once in different ways because of the high clustering. However, the nodes with similar fields of interest should reach each other. We found that clustering can be decreased when defining smaller rings, loops in the topology. As the nodes with similar semantic profile will connect to each other, there will exist more „parallel" rings in the network.

To find an optimal graph structure, we first define a minimum size for the rings in the SemPeer layer, that is, the number of nodes in a loop cannot be less than a predefined value. It can easily be seen regarding Eq. 2 that if this value is not less than TTL+1, we eliminate the backward links ($n<m$ case).

In the advanced SemPeer protocol, we define partitions for the nodes in the system, and each node in a partition can only be connected to a node in the next partition. This also eliminates the connections between nodes on the same level (the $n=m$ case). Each node has to identify the partition that it belongs to. To achieve this, we form a number from the address of each node with modulo division to define the corresponding partition. So far we have not found an optimal distributed strategy to eliminate the third type of counterproductive links. As the parallel rings network topology eliminates two types of counterproductive links, we can give a maximum for the modified clustering coefficient:

$$C_r = \frac{\sum_{m=1}^{TTL-1} k^m (k^{m+1} - k)}{\sum_{m=1}^{TTL} \left[k^m \left(\sum_{n=0}^{m-1} (k^n) + k^m - 1 \right) \right] + \sum_{m=1}^{TTL-1} k^m (k^{m+1} - k)} \tag{10}$$

The coefficient tends to zero as k tends to infinity, independent of the value of TTL. With reasonable parameters the worst case value of the coefficient is quite low. We can also observe a decrease in the diameter of the network. The reason is that the nodes with similar fields of interest are organized in a ring, thus, in most cases, only a subset of the nodes is involved in the search. We can analytically calculate the effective diameter from Eq. 9, when supposing that the number of nodes (and connections) is in average divided by the number of fields of interests (topics, *t*). In that case the effective diameter decreases as follows:

$$\delta_{eff,t} = t^{-\frac{1}{H}} \delta_{eff} \tag{11}$$

A series of simulations supported this analytical result. With the SemPeer protocol, we can achieve the same result as Gnutella does, but with a decreased TTL value. The practical meaning behind the symbols is that the number of messages required to perform a successful query can be decreased. This significantly decreases the network traffic, which is always an issue in mobile Peer-to-Peer networks. Regarding the typical simulation case described in the Section IV with a TTL parameter of 5, the average number of messages required for a successful query also decreases significantly with time (Figure 7).

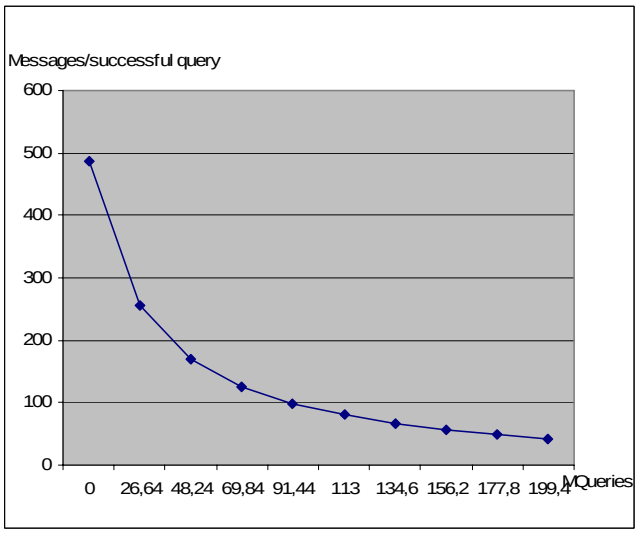

Figure 7
Number of messages required for a successful query

Conclusions

The Peer-to-Peer information retrieval systems have many advantages over the centralized networks. However, to be able to utilize them in mobile environment, the generated network traffic should be decreased. In this paper we examined the advanced SemPeer protocol and proposed a new topology that is more optimal for mobile networks. We constructed a mathematical model that captures the clustering effect in the small world networks. We used this model to test our new topology and to determine a maximum value for the clustering coefficient. The results are more then promising, however, more optimization to the topology could be made at the organization of connections as described in the paper.

References

[1] Oram, A. (edited by), Peer-to-Peer: Harnessing the benefits of a distruptive technology (O'Reilly & Associates, Inc., 2001)

[2] OLeary, D., "Using ai in knowledge management. Knowledge bases and ontologies", IEEE Intelligent Systems 13 (1998) pp. 34-39

[3] The Dublin Core homepage, http://dublincore.org/. (Projects built on the Dublin Core, (http://dublincore.org/projects/))

[4] The Gnutella homepage, http://gnutella.wego.com

[5] Resnik, P., "Semantic similarity in taxonomy: An information-based measure and its application problems of ambiguity in natural language", Journal of Artificial Intelligence Research 11 (1999) pp. 95-130

[6] Csúcs, G. et al., "Peer to Peer Evaluation in Topologies Resembling Wireless Networks. An Experiment with Gnutella Query Engine", ICON2003: The 11th IEEE International Conference on Networks (Sydney, 2003) pp. 673

[7] H. Assadi, "Construction of a Regional Ontology from Text and Its Use within a Documentary System International Conference on Formal Ontology and Information Systems", FOIS-98, IOS Press, Amsterdam (WebDB-2000), Springer-Verlag, Berlin, 2000, pp. 60–71

[8] J.-U. Kietz, A. Maedche and R. Volz, "Semi-Automatic Ontology Acquisition from a Corporate Intranet", Proc. Learning Language in Logic Workshop (LLL-2000), ACL, New Brunswick, N.J., 2000, pp. 31–43

[9] The Napster homepage, http://www.napster.com

[10] K. Sripanidkulchai, B. Maggs, H.Zhang, "Efficient content location using interest-based locality in peer-to-peer systems", Infocom, 2003

[11] Joseph S., "P2P MetaData Search Layers", Second International Workshop on Agents and Peer-to-Peer Computing (AP2PC 2003)

[12] Marcello W Barbosa, Mellssa M Costa, Jussara M Almeida, Virgilio A P Alameida, “Using Locality of reference to improve performance of peer-to-peer applications”. WOSP'04 & ACM SIGSOFT Software Engineering Notes V29n1(Jan 2004), pp. 216-227

[13] The WordNet project homepage, http://www.cogsci.princeton.edu/~wn/

[14] Bertalan Forstner, Gergely Csúcs, Kálmán Marossy, “Evaluating performance of peer-to-peer protocols with an advanced simulator”, Parallel and Distributed Computing and Networks, 2005, Innsbruck, Austria

[15] Bertalan Forstner, Hassan Charaf, “Neighbor Selection in Peer-to-Peer Networks Using Semantic Relations”, WSEAS Transactions on Information Science & Applications, 2(2), February 2005, ISSN 1790-0832, pp. 239-244

[16] Paul Erdős, Alfred Rényi, “On the Strength of Connectedness of a Random Graph”, Acta Math. Acad. Sci. Hungary, 12, 1961, pp. 261-267

[17] Mark S. Granovetter, “The Strength of Weak Ties”, American Journal of Sociology, 78 (1973), pp. 1360-1380

[18] D. J. Watts, S. H. Strogatz, “Collective Dynamics of ’Small-World’ Networks”, Nature, 393 (1998), pp. 440-442

[19] Wolfgang Nejdl et al, “Super-peer-based routing and clustering strategies for RDF-based peer-to-peer networks”, 20th International Conference on World Wide Web, Budapest, Hungary, May 2003, ISBN: 1-58113-680-3, pp. 536-543

[20] Wolfgang Nejdl, Wolf Siberski, Michael Sintek, “Design issues and challenges for RDF- and schema-based peer-to-peer systems”, ACM SIGMOD Record, 32(3), September 2003

[21] D. A. Menasc´e, V. A. F. Almeida, and L. W. Dowdy, “Capacity Planning for Web Services: metrics, models, and methods”. Prentice Hall, 2001

[22] Z. Ge, D. R. Figueiredo, S. Jaiswal, J. Kurose, and D. Towsley, “Modeling Peer-Peer File Sharing System” In Proceedings of INFOCOM 2003, San Francisco, USA, Apr 2003

[23] Mihajlo A. Jovanovic, “Modeling Large-scale Peer-to-Peer Networks and a Case Study of Gnutella”, MsC Thesis

[24] M. Faloutsos, P. Faloutsos, and C. Faloutsos, “On power-law relationships of the Internet topology”, ACM SIGCOMM, Sep 1-3, Cambridge MA, 1999, pp. 251–262