

Identifying Community Structure in Semantic Peer-to-Peer Networks*

Hanhua Chen, Hai Jin

Cluster and Grid Computing Lab

Huazhong University of Science and Technology, Wuhan, 430074, China

Email: hjin@hust.edu.cn

Abstract

*The research community has turned to create **Semantic Overlay Networks** for information retrieval in large scale peer-to-peer networks. Much research work in semantic overlay protocols and searching algorithms is done and their results indicate that semantic overlay is powerful for content searching in peer-to-peer network. However, very limited work has been done in analyzing and evaluating characteristic about semantic overlay network. In this paper we identify a natural property of semantic overlay network, community structure. We setup a mathematical model to evaluate the community structure property. A heuristic backtrace-greedy hybrid algorithm is also designed to optimize the evaluation value of a given overlay network. Using the evaluation model we compare the SemreX semantic overlay with the Gnutella network. We find that the SemreX overlay network has a distinct feature of community structure, while the Gnutella network does not have such feature.*

1. Introduction

Due to characteristics of low maintenance overhead, improved scalability and reliability, synergistic performance, and increased autonomy, peer-to-peer networks have shown a great potential to become an excellent information sharing tool.

Recently the research community has turned to create *Semantic Overlay Networks* (SON) to improve the efficiency of p2p information sharing [6]. Notable examples are Interest-based overlay [1], Edutella [2], Bibster [3], SSW [4]. In [1], Sripanidkulchai et al identify the natural principle in Gnutella and other p2p networks called interest-based locality, which posits that if a peer has a particular piece of content that one

is interested in, it is very likely that it will have other items that one is interested in as well. Based on the principle of interest-based locality, a self-organizing protocol, interest-based shortcuts, is proposed and implemented on top of Gnutella. Peers that share similar interests create shortcuts to one another. Peers then use these shortcuts to locate content. Using interest-based shortcuts, the flooding messages are greatly reduced.

Edutella [2] proposes RDF metadata models that standardize the way data is organized and enable routing and clustering strategies based on the metadata schemas, attributes and ontologies used. In Bibster [3], peers advertise the expertise of themselves, which contains a set of topics that the peer is an expert. Other peers may accept these advertisements or not, thus creating a semantic link to their neighbors. These semantic links form a semantic overlay, which is the basis for intelligent query routing. SSW [4] overlay network dynamically clusters peers with semantically similar data closer to each other and maps these clusters in a high-dimensional semantic space into a one-dimensional small world network, which has an attractive trade-off between search path length and maintenance costs.

In this paper we identify a property of semantic overlay, community structure [5], a distinct character of semantic overlay networks. Nodes with semantically similarity content are “clustered” together and network nodes are self-organized into groups within which the networks connections are dense, but between which they are sparser, shown in Fig. 1.

We identify this property in the SemreX¹, a semantic peer-to-peer system for reference exchange [15, 16]. We create a model to evaluate the community structure property of semantic overlay networks. A heuristic backtrace-greedy hybrid algorithm is designed to optimize the evaluation value of a given overlay network. By comparing the SemreX semantic overlay with the Gnutella like network, we find that the

* This paper is supported by the National 973 Key Basic Research Program under grant No.2003CB317003.

¹ <http://grid.hust.edu.cn/semrex/>

SemreX overlay network has much more distinct feature of community structure than Gnutella like network.

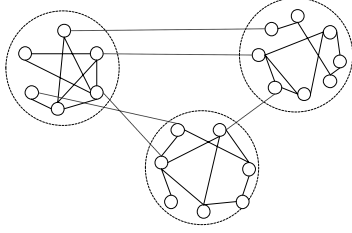


Figure 1. Community structure

The rest of this paper is organized as follows. In section 2, we give an introduction of the SemreX Semantic Overlay. We propose the community structure evaluation model in section 3. Section 4 introduces the heuristic algorithm for optimizing the evaluation value. In section 5, we compare the community structure of SemreX semantic overlay and the Gnutella network. We review some related works in section 6. Section 7 concludes the paper and describes our future work.

2. Semantic Peer-to-Peer Overlay

SemreX [15, 16] is a prototype peer-to-peer system developed by Cluster and Grid Computing Lab of Huazhong University of Science and Technology. The system architecture of SemreX, which can be viewed as four layers, includes the logical concept layer, the document object space layer, the underlying peer communication layer, and the semantic overlay layer.

2.1 Semantic overlay

Before introducing the community structure evaluation model, we give a formal definition of semantic similarity of peers in SemreX at first.

Definition 1. Semantic Similarity of Peers. In SemreX the similarity of peers is defined as the semantic similarity between the corresponding pair of sets of weighted topics, which are concept nodes on the ACM Topic IS-A concept structure.

$$Sim(P_1, P_2) = Sim(\{ \langle T_i, \lambda_i \rangle, i=1,2,\dots,m \}, \{ \langle T_j, \lambda_j \rangle, j=1,2,\dots,n \}) \quad (1)$$

where, P_1 and P_2 denote any pair of peer, T_i and T_j separately denote the topic from P_1 and P_2 , and λ_i quantifies the weight of the corresponded topic, that is the percent of document classified to T_i .

$$\lambda_i = \frac{N_i}{\sum_{j=1}^n N_j} \quad (2)$$

The study of semantic similarity between lexically expressed concepts has been a part of natural language processing for many years. A number of measuring methods have been developed, among which Yuhua's measure [7] gives the best results.

$$Sim(T_i, T_j) = f_1(l) \cdot f_2(h) = \begin{cases} e^{\alpha l} \cdot \frac{e^{\beta h} - e^{-\beta h}}{e^{\beta h} + e^{-\beta h}} & \text{if}(T_i \neq T_j) \\ 1 & \text{if}(T_i = T_j) \end{cases} \quad (3)$$

Here, l counts the shortest path length between T_1 and T_2 and h counts the depth from the subsumer of T_1 and T_2 to the top of the concept hierarchy. $\alpha > 0$ and $\beta > 0$ are parameters scaling the contribution of shortest path length and depth, respectively. The strongest correlation between equation (3) and human judgments is at 0.2 and 0.6.

Based on the method, we use the following equation to quantify the similarity between two peers:

$$Sim(P_1, P_2) = \sum_{j=1}^{|P_2|} \sum_{i=1}^{|P_1|} [Sim(T_i, T_j) \times (\lambda_i \times \lambda_j)] \quad (4)$$

Here, $|P_1|$ and $|P_2|$ are the topic numbers in the two peers. The similarity between the sets of ranked concepts of each other is calculated by summing up products of the similarity value between two topics separately selected from P_1 , P_2 and the ranks of both topics.

2.2 Community structure of SemreX

We identify the power principle, community structure, in semantic overlay, by simulating SemreX protocol on top of Gnutella-like network. Figure 2 gives a sub-graph of Gnutella-like topology. On top of the Gnutella-like sub topology, we simulate the population of documents as Zipf distribution [8]. Each document belongs to a topic on the ACM topic tree. After that we simulate the semantic link protocol on the Gnutella-like topologies and get SemreX overlay network. We identify strong property of community structure in SemreX overlay networks.

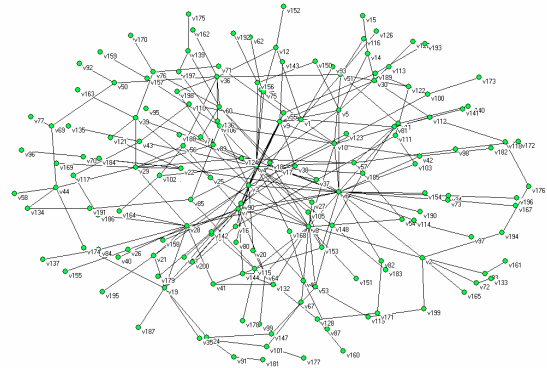


Figure 2. Gnutella-like overlay network

Figure 3 shows the semantic overlay network, where Gnutella nodes are re-organized by the semantic overlay protocol, into groups within which the connections are dense, but between which they are sparse. Figure 3 also displays the similarity value between the neighbor peers in the semantic overlay. Peers in the same group are much more similar than those in different groups, that is, inside the group the semantic link is strong with a higher similarity value, but semantic links between groups are much weaker.

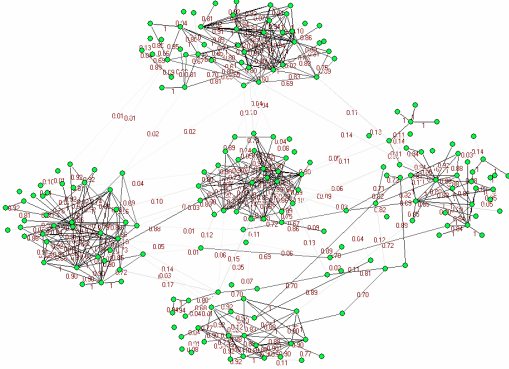


Figure 3. Community structure of SemreX semantic overlay network

3. Model for Evaluating Community Structure

We first give a formal definition of semantic overlay as follow.

Definition 2. Semantic Overlay. A semantic overlay network is defined as a network graph $G=(E, V)$, where E denotes the set of semantic links between peers, and V denote the set of peers.

We use the following adjacency matrix B with Boolean entries b_{ij} to represent the simple connection relation of any two vertices of the network.

$$b_{ij} = \begin{cases} 1 & \text{if } (e_{ij} \in E), \\ 0 & \text{if } (e_{ij} \notin E). \end{cases} \quad (5)$$

A semantic overlay network with n vertices can be quantified by an $n \times n$ adjacency matrix A with entries a_{ij} that are not simply zero or 1, but are equal to the weights on the edges, which are the semantic similarity values of pairs of peers. Here similarity values between two peers record their ‘‘similarity strength’’ relative to each another.

$$a_{ij} = \begin{cases} Sim(P_i, P_j) & \text{if } (e_{ij} \in E), \\ 0 & \text{if } (e_{ij} \notin E). \end{cases} \quad (6)$$

As aforementioned, we easily identify in semantic overlay that peers are self-organized into communities. The links inside a community is dense and with strong

strength, while the links between the communities are sparse and with weak strength. The aim of our model is to evaluate how strength the structure feature is.

Supposing the overlay network is partitioned into communities, let C_k denotes community k in this particular division. We use the following symmetrical matrix C with elements to represent the semantic strength between communities m and n , where c_{mn} denotes one-half of the fraction of weights of edges that connect vertices in community m to those in community n , so that the total fraction of such weights of edges is $c_{mm}+c_{nm}$.

$$c_{mn} = \frac{\sum_{v_i \in C_m} \sum_{v_j \in C_n} a_{ij}}{2 \times \sum_i \sum_j a_{ij}} \quad (7)$$

The only exception is the diagonal elements c_{mm} , which are equal to the fraction of weights of edges that fall within community m .

At the same time we use another symmetrical matrix C' with elements:

$$c'_{mn} = \frac{\sum_{v_i \in C_m} \sum_{v_j \in C_n} b_{ij}}{2 \times \sum_i \sum_j b_{ij}} \quad (8)$$

to quantify the connection relationship between communities n and m , where c'_{mn} quantities one-half of the fraction of edges that connect vertices in community m to those in community n , so that the total fraction of such edges is $c'_{mn} + c'_{nm}$. The only exception is the diagonal elements c'_{mm} , which are equal to the fraction of edges that fall within community m .

Let δ_k be the fraction of all ends of edges that are attached to vertices in group k . We can calculate δ_k directly by $\delta_k = \sum_i c'_{ki}$. If the ends of edges are connected together at random, the fraction/probability of the resulting edges that connect vertices within group k is δ_k^2 .

We define the *Community Structure Evaluation Model* as follow:

$$Q = \sum_k \left(c_{kk} - \delta_k^2 \times \frac{\bar{\omega}'_k}{\bar{\omega}_{total}} \right) \quad (9)$$

where $\bar{\omega}_{total}$ is the average weight (semantic similarity strength) of all the edges in the semantic overlay:

$$\bar{\omega}_{total} = \frac{\sum_j \omega_j}{|V|} \quad (10)$$

and $\bar{\omega}_k$ quantifies the average weight of the edges related to community k , which include the edges whose both ends are in the community k and the edges only one end of whose are in the community C_{kk} .

$$\bar{\omega}_k = \frac{\sum_{e_j \in C_k} \omega_j + \frac{1}{2} \sum_{e_i \in E_k} \omega_i}{|C_k| + \frac{1}{2} |E_k|} \quad (11)$$

Here, E_k is the set of edges that have only one end in community C_k , that is:

$$E_k = \{e = \langle v_i, v_j \rangle \mid (e \in E) \wedge ((v_i \in C_k \wedge v_j \notin C_k) \vee (v_i \notin C_k \wedge v_j \in C_k))\}$$

So, $\bar{\omega}_k / \bar{\omega}_{total}$ quantifies the relative weight of the edges within community k to that of all the edges of the overlay.

$(c_{kk} - \delta_k^2 \times \frac{\bar{\omega}_k}{\bar{\omega}_{total}})$ is the sum weight of the edges that

fall within community C_k , minus the expected value of the same quantity if edges fall at random without regard for the community structure.

We use Q to quantify the strength of community structure, which sums up the deviations of all the communities in the given community division. We can easily find that, if a particular division gives no within-community weights of edges than would be expected by random chance, this modularity will get the minimum value $Q=0$. Values other than 0 indicate deviations from randomness. So the aim of our model is to optimize Q .

4. Algorithms for Optimizing Q

As the model shows, a high value of Q represents a good community division of semantic overlay network. To find the strength of community structure of a given semantic overlay can be carried out simply by optimizing Q over all possible divisions to find the best one. However, the problem is that true optimization of Q is very costly. The number of ways to divide n vertices into m non-empty groups is at least exponentially in n . To carry out an exhaustive search of all possible divisions for the optimal value of Q would therefore take at least an exponential amount of time, and is in practice infeasible for systems larger than twenty or thirty vertices.

Generally speaking, various approximate optimization methods, simulated annealing, greedy, and so forth, are available for solving such kind of problems. However, in the community division problem we find these algorithms are not powerful for solving this problem. For example let us take a standard greedy optimization algorithm into

consideration. The algorithm will start with a state in which each vertex is the sole member of one of n communities. We can repeatedly join communities together in pairs, choosing at each step the join that results in the greatest increase in Q . Until the Q does not increase any more we get the best Q and the best community division. This algorithm may be fast to get a good Q , however it is obvious that the join operations will not be able to step back. So a bad start means a bad result when using the standard greedy algorithm.

In this paper we design a backtrack-greedy hybrid algorithm, shown in Figure 4. The algorithm starts with an initial state Γ in which each vertex is the sole member of one of n communities. The main difference between our algorithm and standard greedy algorithm is that we choose the next state from which we will achieve the greatest final increase in Q using greedy searching among all the states by joining two communities of current division instead of choosing the direct next state which has the greatest increase in Q among all the states by joining two communities of current division state.

```

Algorithm: Backtrace Search
/*  $\Gamma$  : the current state of the partition structure */
backtrace_search ( $\Gamma$ ) {
 $Q \leftarrow$  calculate  $Q(\Gamma)$ ;  $Best\_Q \leftarrow 0$ ;  $Best\_Gamma \leftarrow \Gamma$ ;
 $n \leftarrow$  number of communities of  $\Gamma$ ;
do
   $\Gamma^* \leftarrow \Gamma$ ;  $Q_{max} \leftarrow 0$ ;
  for ( each pair ( $C_i, C_j$ ) within  $\Gamma$ ) do
     $\Gamma \leftarrow$  join_community( $C_i, C_j$ );
     $Q_{ij} \leftarrow$  greedy_search ( $\Gamma$ );
    if ( $Q_{ij} > Q_{max}$ ) do
       $Q_{max} \leftarrow Q_{ij}$ ;
       $i^* \leftarrow i$ ;  $j^* \leftarrow j$ ;
    end if
     $\Gamma \leftarrow \Gamma^*$ ; /*backtrace to  $\Gamma^*$ */
  end for
   $\Gamma \leftarrow$  join_community( $C_{i^*}, C_{j^*}$ );
  calculate  $\Delta Q_{i^*j^*}$ ;
  /* the value by which  $Q$  increases if join  $C_{i^*}$  and
 $C_{j^*}$  to a single community*/
   $Q \leftarrow Q + \Delta Q_{i^*j^*}$ ;
  if ( $Q > Best\_Q$ ) do
     $Best\_Q \leftarrow Q$ ;  $Best\_Gamma \leftarrow \Gamma$ ;
  end if
   $n \leftarrow n - 1$ ;
until ( $n = 0$ )
return ( $Best\_Gamma$ );
}

```

Figure 4. Backtrack based greedy search algorithm for the best community division

Our algorithm will not stop if the Q does not

increase when we take a step. The best state of division is not decided until all the nodes are joined into a single community. Figure 5 give the algorithm for greedy search for every direct branch of the current state of community division.

```

Algorithm: Greedy Search
/*greedy search for every direct branch of current
state*/
/* $\Gamma$ : current state of the partition structure */
/*  $\Delta Q_{ij}$ : the value by which  $Q$  increase if  $C_i$  and  $C_j$ 
joined into a single community*/
greedy_search( $\Gamma$ ) {
 $Q \leftarrow \text{calculate } Q(\Gamma)$ ;
do
 $\Delta Q^* \leftarrow -1$ ;  $i^* \leftarrow -1$ ;  $j^* \leftarrow -1$ ;
for (each pair ( $C_i, C_j$ ) within  $\Gamma$ ) do
calculate  $\Delta Q_{ij}$ ;
if ( $\Delta Q_{ij} > \Delta Q^*$ ) do
 $\Delta Q^* \leftarrow \Delta Q_{ij}$ ;
 $i^* \leftarrow i$ ;  $j^* \leftarrow j$ ;
/*  $\Delta Q_{max} \leftarrow \max \{ \Delta Q_{ij} \}$  */;
end if
end for
if ( $\Delta Q^* > 0$ ) do
 $\Gamma \leftarrow \text{join\_community}(C_{i^*}, C_{j^*})$ ;
 $Q \leftarrow Q + \Delta Q^*$ ;
end if
until ( $\Delta Q^* < 0$ )
return ( $Q$ );
/*return the best  $Q$  of the structures start from  $\Gamma$  using
greedy search.*/

```

Figure 5. Greedy search from every direct branch of current state of community division

5. Performance Evaluation

On top of the Gnutella-like topology, we simulate the population of documents as Zipf distribution, as studies have shown that the file popularity distribution in Kaza follows Zipf's law [8]. Each document belongs to a topic on the ACM topic tree. We simulate the semantic link protocol proposed in [15] on the Gnutella topologies and get SemreX overlay networks. The scales of the semantic overlay networks vary from 100 to 1000. Semantic links are generated according the peer similarity threshold value 0.5. Table 1 summarizes the settings for the simulation of semantic overlay on top of Gnutella topology.

The metric of our simulation is the Q , the value of the community divisions. The best community structure will achieve the maximal value of Q . Figure 6 shows when the number of nodes varies from 100 to 1000, the maximal values of Q in semantic overlay network greatly exceed that of Gnutella networks in different scale.

Table 1. Settings for simulation semantic overlay

Parameter Descriptions		Values
N	Number of nodes in the network	100-1000
T	Number of ACM topics	30
D	Max number of documents per node	200
Tsd	Threshold of similarity	0.5
Q	Max number of queries each peer	200
Z(α, n)	Zipf distribution of documents	$\alpha=1.2$ $n=D$

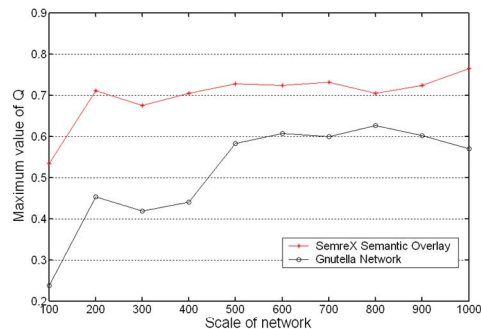


Figure 6. Comparison of maximal value of Q in Gnutella like network and semantic overlay network (number of node=100~1000)

Figure 7 compares the number of communities of the best divisions in Gnutella network and semantic overlay network. We find that the community structure of semantic overlay has much less number of communities than Gnutella networks. The experiment results show that semantic overlay networks have distinct property of community structure.

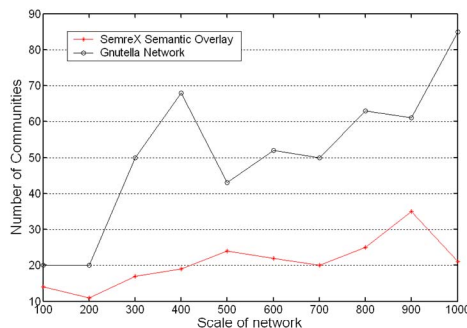


Figure 7. Comparison of number of communities of the best divisions in Gnutella network and semantic overlay network (number of node=100~1000)

6. Related Works

Community structure is a rather new property that has attracted considerable recent attention in the

research on complex network [5, 11]. Recent research shows many networks possess strong community structure, including World Wide Web [10], citation networks [11], social networks [12], and so on. In [13] Newman et al propose a community structure evaluation model for un-weighted network. However, we can not reasonably use this model to evaluate the community structure of semantic overlay network, a complex weighted network [14].

7. Concussion and Future Works

We propose a community structure evaluation model for the semantic overlay network, a complex weighted network. A backtrace-greedy hybrid algorithm is designed to solve the model without the problem of “bad start”. Using the evaluation model we compare the SemreX semantic overlay with the Gnutella network. We find that a SemreX overlay network has a distinct feature of community structure, while the Gnutella network does not have such feature. We expect our work to be useful for the work of developing more efficient search algorithms in semantic overlay networks.

References

- [1] K. Sripanidkulchai, B. Maggs, and H. Zhang, "Efficient content location using interest-based locality in peer-to-peer systems", *Proceedings of INFOCOM*, 2003.
- [2] W. Nejdl, M. Wolpers, W. Siberski, C. Schmitz, M. Schlosser, and I. Brunkhorst, "Super-peer-based routing and clustering strategies for RDF-based peer-to-peer networks", *Proceedings of WWW'03*, Budapest, Hungary, 2003.
- [3] P. Haase, J. Broekstra, M. Ehrig, M. Menken, P. Mika, M. Plechawski, P. Pyszlak, B. Schnizler, R. Siebes, S. Staab, and C. Tempich, "Bibster: a semantic-based bibliographic peer-to-peer system", *Proceedings of ISWC'04*, 2004.
- [4] M. Li, W.-C. Lee, and A. Sivasubramaniam, "Semantic small world: an overlay network for peer-to-peer search", *Proceedings of ICNP'04*, Berlin, Germany, 2004.
- [5] M. Newman, "Modularity and community structure in networks", *PNAS, USA*, Vol.103, No.23, June, pp.8577-8582, 2006.
- [6] H. Zhuge, X. Sun, J. Liu, E. Yao, and X. Chen, "A scalable p2p platform for the knowledge grid", *IEEE Transactions on Knowledge and Data Engineering*, Vol.17, No.12, pp.1721-1736, 2005.
- [7] L. Yuhua, Z. A. Bandar, and D. McLean, "An approach for measuring semantic similarity between words using multiple information sources", *IEEE Transactions on Knowledge and Data Engineering*, Vol.15, pp.871-882, 2003.
- [8] A. Iamnitchi, M. Ripeanu, and I. Foster, "Small-world file-sharing communities", *Proceedings of INFOCOM*, 2004.
- [9] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi, "Defining and identifying communities in networks", *PNAS, USA*, Vol.101, pp.2658-2663, 2004.
- [10] G. W. Flake, S. R. Lawrence, C. L. Giles, and F. M. Coetzee, "Self-organization and identification of web communities", *IEEE Computer*, Vol.35, pp.66-71, 2002.
- [11] M. E. J. Newman, "Coauthorship networks and patterns of scientific collaboration", *PNAS, USA*, Vol.101, pp.5200-5205, 2004.
- [12] D. Zhou, E. Manavoglu, J. Li, C. L. Giles, and H. Zha, "Probabilistic models for discovering E-Communities", *Proceedings of WWW'06*, Edinburgh, Scotland, 2006.
- [13] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks", *Physical Review E*, Vol.70, pp.0066111, 2004.
- [14] A. Barrat, M. Barthelemy, R. Pastor-Satorras, and A. Vespignani, "The architecture of complex weighted networks", *PNAS, USA*, Vol.101, pp.3747-3752, 2004.
- [15] H. Jin, H. Chen, and X. Ning, "SemreX: A semantic peer-to-peer system for literature documents retrieval", *Proceedings of ASWC'06*, Beijing, China, 2006.
- [16] H. Jin, X. Ning, and H. Chen, "Efficient search for peer-to-peer information retrieval using semantic small world", *Proceedings of WWW'06*, Edinburgh Scotland, 2006.