

4-2010

# Mining Diversity on Networks

Lu LIU  
*Tsinghua University*

Feida ZHU  
*Singapore Management University, fdzhu@smu.edu.sg*

Chen CHEN  
*University of Illinois at Urbana-Champaign, USA*


Xifeng YAN  
*University of California at Santa Barbara, USA*

Jiawei HAN  
*University of Illinois at Urbana-Champaign*

*See next page for additional authors*

**DOI:** [https://doi.org/10.1007/978-3-642-12026-8\\_30](https://doi.org/10.1007/978-3-642-12026-8_30)

Follow this and additional works at: [https://ink.library.smu.edu.sg/sis\\_research](https://ink.library.smu.edu.sg/sis_research)

 Part of the [Databases and Information Systems Commons](#), and the [Numerical Analysis and Scientific Computing Commons](#)

---

## Citation

LIU, Lu; ZHU, Feida; CHEN, Chen; YAN, Xifeng; HAN, Jiawei; YU, Philip; and YANG, Shiqiang. Mining Diversity on Networks. (2010). *Database Systems for Advanced Applications: 15th International Conference, DASFAA 2010, Tsukuba, Japan, April 1-4, 2010, Proceedings, Part I*. 5981, 384-398. Research Collection School Of Information Systems.

**Available at:** [https://ink.library.smu.edu.sg/sis\\_research/509](https://ink.library.smu.edu.sg/sis_research/509)

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email [libIR@smu.edu.sg](mailto:libIR@smu.edu.sg).

---

**Author**

Lu LIU, Feida ZHU, Chen CHEN, Xifeng YAN, Jiawei HAN, Philip YU, and Shiqiang YANG

# Mining Diversity on Networks

Lu Liu<sup>1</sup>, Feida Zhu<sup>3</sup>, Chen Chen<sup>2</sup>, Xifeng Yan<sup>4</sup>,  
Jiawei Han<sup>2</sup>, Philip Yu<sup>5</sup>, and Shiqiang Yang<sup>1</sup>

<sup>1</sup> Tsinghua University

<sup>2</sup> University of Illinois at Urbana-Champaign

<sup>3</sup> Singapore Management University

<sup>4</sup> University of California at Santa Barbara

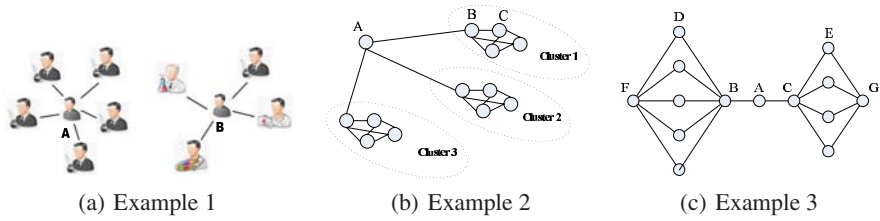
<sup>5</sup> University of Illinois at Chicago

**Abstract.** Despite the recent emergence of many large-scale networks in different application domains, an important measure that captures a participant's *diversity* in the network has been largely neglected in previous studies. Namely, diversity characterizes how diverse a given node connects with its peers. In this paper, we give a comprehensive study of this concept. We first lay out two criteria that capture the semantic meaning of diversity, and then propose a compliant definition which is simple enough to embed the idea. An efficient top-k diversity ranking algorithm is developed for computation on dynamic networks. Experiments on both synthetic and real datasets give interesting results, where individual nodes identified with high diversities are intuitive.

## 1 Introduction

Mining diversity is an important problem in various areas and finds many applications in real-life scenarios. For example, in information retrieval, people use information entropy to measure the diversity based on a certain distribution, e.g., one person's research interests diversity[12]. In social literature, diversity, which has been proposed under other terminologies like *bridging social capital*, proves its importance in many social phenomena. Putnam found that bridging social capital benefits societies, governments, individuals and communities[11]. In particular, bridging social capital helps reduce an individual's chance of catching certain diseases and the chance of dying, e.g., joining an organization cuts in half an individual's chance of dying within the next year, leading to the conclusion that "Network diversity is a predictor of lower mortality".

Mining diversity on network data is also critical for network analysis as network data emerge in abundance in many of today's real world applications. For example, advertisers may be very interested in the most diverse users in social network because they connect with users of many different types, which means "word of mouth" marketing on these users could reach potential customers of a much wider spectrum of varied tastes and budgets. In a research collaboration network of computer scientists, the diversity of a node could indicate the corresponding researcher's working style. A highly diverse researcher collaborates with colleagues from a wide range of institutions and communities, while a less diverse one might only work with a small group of people, e.g., his/her students. As such, an interesting query on such a network could be "Who



**Fig. 1.** Three Examples

are the top ten diversely-collaborating researchers in the data mining community?”. To illustrate the intuition of diversity on networks, let us look at an example.

**Example 1.** Consider a social network example in which nodes represent people and edges represent social connections between corresponding parties. Suppose we examine two nodes  $A$  and  $B$  in Fig.1(a) where  $A$  connects to 5 neighbors and  $B$  connects to 4 neighbors. However, the 5 neighbors of  $A$  are all from the same profession and the same community, while the 4 neighbors of  $B$  are from 4 different professions and/or communities. Here, although the neighborhood of  $B$  is smaller than that of  $A$ , it is obvious that  $B$  connects to a more diverse group of people, which could have important implications regarding the role he/she may play in the network, e.g., the profitability and impact if we are to choose a node to launch a marketing campaign.

Example 1 demonstrates that the diversity of a node on network is determined by the characteristics of its neighborhood. Greater difference between the neighbors translates into greater diversity of the node. In Example 1, the attributes or the labels are used to distinguish the neighbors. Then how can we measure the diversity if no attribute information is given? Example 2 illustrates another way to mine diversity which is based on the topological structure of the network.

**Example 2.** In Fig.1(b), comparing nodes  $A$  and  $C$  with the same degree of 3, it is easy to observe significant difference between the diversities of their neighborhoods.  $A$  connects to three neighbors, each of which belongs to a distinct community, while  $C$  connects to three closely connected neighbors that form a cohort. In many applications,  $A$  might be more interesting, because of its role of joining different persons together.

The two examples above give two different ways to measure diversity on networks. However regardless of using either neighborhood attributes or topology, certain common principles conveying the semantic meaning of diversity underlie any particular kind of computation or definition of diversity. In fact, it is our observation that there are two basic factors impacting the diversity measure on a network.

- *All else being equal, the greater the size of the neighborhood, the greater the diversity.*  
When all the neighbors are the same, in terms of both associated labels and neighborhood topology, more neighbors lead to a greater diversity.
- *The greater the differences among the neighbors, the greater the diversity.*

The neighbors can be distinguished either by their attributes and labels or by the topological information of the neighborhood. Whichever way, a larger difference should translate into a greater diversity.

The above two factors can also be treated as two criteria taken as the basis for proposing a reasonable definition for measuring diversity. In this paper, we focus on mining the diversity on network based on the topological structure. As pointed out in Section 2, existing measures like centrality can not accurately capture the notion of diversity in general, although certain degree of correlation between them can be observed for some data sets.

Our contributions can be summarized as follows.

- As far as we know, there has been no research work to investigate diversity on network structure data based on network characteristics. We are the first to propose the diversity concept on network and give two criteria that capture the semantic meaning of diversity.
- We investigate mining diversity based on topological information of a network, find a function which is simple enough to embed the two criteria and propose an efficient algorithms to obtain top-k diverse nodes on dynamic networks.
- Extensive experiment studies are conducted on synthetic and real data sets including DBLP. The results are interesting, where individual nodes identified with great diversities are highly intuitive.

The remaining of this paper is organized as follows. In Section 2, the related work is introduced and compared with our work. In Section 3, we propose a diversity definition based on topological information of network and develop an efficient top-k diversity ranking algorithm for dynamic networks in Section 4. The experiment results are reported in Section 5. Other kinds of diversity definition are discussed in Section 6. Section 7 concludes this study.

## 2 Related Work

As network data emerge in abundance in many of today's real world applications, many research work has been done on network analysis in recent literatures. Properties reflecting the overall characteristics of network, such as density, small world, hierarchical modularity and power law [15,5,2,10], have been observed for a long time. Compared to these, many measures that focus on individual components, e.g., degree, betweenness, closeness centrality, clustering coefficient, authority and etc, have also been proposed to distinguish the roles of nodes in network [13,9,14,7]. Besides, some other types of patterns, e.g., frequent subgraphs that focus more on local topologies [8,16], can be mined from the network.

However, all these measures are different from diversity and thus could not accurately capture the idea behind. Degree centrality, which is defined as the number of links for a given node, does not consider whether the neighbors are similar. Betweenness centrality assigns higher value to nodes appearing on the shortest paths of more node pairs. As we shall observe in the experiments, it might be correlated with diversity to some extent in particular data scenarios, but it is not a direct modeling of

diverseness and thus would not satisfy the two criteria we have proposed in general. Closeness centrality, which measures the average shortest-path length from a node to all other nodes in the network, has similar problems. Moreover, such shortest-path based measures require the global computation of all-pair shortest paths, which leads to the time-consuming measure calculations on a large network. The clustering coefficient value of a node corresponds to the number of edges among its neighbors normalized by the maximum number of such edges; intuitively, with higher clustering coefficient, the neighbors have more connections among them and thus are more similar to each other, which leads to lower diversity. However, clustering coefficient does not consider the scale of the neighborhood and only counts number of edges as the sole parameter, which is inevitably restricted. Interestingly, it can be treated as a degenerated version of our diversity definition when the latter is confined to a very special setting.

### 3 Diversity Definition

In this section, we will propose concrete diversity definitions based on nodes' neighborhood topology. First, a simple definition is given out and the calculation results on Example 2 illustrate that it matches our intuition of diversity. Then we will propose a general definition and show its calculation results on more examples, in which we analyze its parameters and compare it with centrality.

#### 3.1 Terminology and Representation

Let an undirected unweighted network be  $G = \{(V, E) \mid V \text{ is a set of nodes and } E \text{ is a set of edges, } E \in V \times V, \text{ an edge } e = (i, j) \text{ connects two nodes } i \text{ and } j, i, j \in V, e \in E\}$ .  $N(v)$  denotes the set of  $v$ 's neighbors.  $|N(v)|$  denotes the cardinality of  $N(v)$ , i.e., the number of neighbors.  $r$  is the radius of the neighborhood. If it is set to be 1,  $N(v)$  is the set of directly connected nodes and  $|N(v)|$  equals to the degree of node  $v$ .  $N_{-u}(v)$  denotes the set of  $v$ 's neighbors which excludes the nodes that become  $v$ 's neighbors through  $u$ . For example, when  $r = 1$ ,  $N_{-u}(v)$  is the set of the direct neighbors of  $v$  except  $u$  itself; when  $r = 2$ ,  $N_{-u}(v) = N(v) - \{x \mid \text{there is only one shortest path from } v \text{ to } x \text{ which is through } u\}$ .  $L(i, j)$  denotes the length of shortest path from node  $i$  to node  $j$ .

#### 3.2 A Simple Diversity Example

To illustrate the diversity measure, we first use a simple definition as below, which can get the intuitive results of Example 2 in Fig.1(b).

**Definition 1.** Given a network  $G$  and a node  $v \in V(G)$ , the diversity  $D(v)$  is defined as

$$D(v) = \sum_{u \in N(v)} \left( 1 - \frac{|N(v) \cap N(u)|}{|N(u)|} \right) \quad (1)$$

The underlying intuition of the definition is that, for a target node  $v$ , if a neighbor  $u$  has fewer connections with other neighbors of  $v$ ,  $u$  is considered to contribute more to

the diversity of  $v$ . Therefore the diversity of  $v$  is defined as the aggregation of every neighboring node  $u$ 's contribution which equals to the probability of leaving the direct neighborhood of  $v$  through  $u$  [7].

Based on this definition, we can get that the diversity values of  $A, B, C$  in Example 2 are 3, 2, 1.167 respectively. The relative values match our intuition of diversity ranking on this network.

### 3.3 Diversity: General Definition

While the previous definition based on direct common neighborhood is simple and intuitive in some cases, we need more flexibility and generality in the diversity definition for most applications to capture the measure more accurately. As we discussed above, the diversity in general grows in proportion with the size of the neighborhood. With this notion of each neighbor contributing to the diversity of the central node, we propose the general definition of diversity in an aggregate form as follows.

**Definition 2 [Diversity].** *The diversity of a node  $v$  is defined as an aggregation of each neighbor  $u$ 's contribution to  $v$ 's diversity.*

$$D(v) = \sum_{u \in N(v)} w_v(u) * F(u, v) \quad (2)$$

where  $F(u, v)$  is a function measuring the diversity introduced by  $u$ .  $w_v(u)$  is  $u$ 's weight in the aggregation.

According to our guiding principles, if a neighbor  $u$  is less similar to other neighbors of  $v$ ,  $u$  would contribute more to  $v$ 's diversity. Thus  $F(u, v)$  is a function evaluating the dissimilarity between  $u$  and other neighbors of  $v$  in the set radius  $r$ , i.e., the set  $N_{-u}(v)$ . In general,  $F(u, v)$  can be defined as a linear function of the similarity between  $u$  and  $N_{-u}(v)$  as

$$F(u, v) = 1 - \alpha * S(u, N_{-u}(v)) \quad (3)$$

$S(u, N_{-u}(v))$  is a function measuring the similarity between  $u$  and  $N_{-u}(v)$  up to a normalization.  $\alpha$  indicates its weight, which can be set empirically. We define  $S(u, N_{-u}(v))$  as the average similarity between  $u$  and each node  $x$  of  $N_{-u}(v)$ . There are various ways to measure the similarity between two nodes  $u$  and  $x$ , e.g., shortest path is a reasonable choice for many real-world scenario. However, computing shortest paths on a global scale is inefficient. Fortunately, since diversity is a local property defined on a neighborhood with a set radius, we can use the following definition based on local shortest path computation.

**Definition 3 [Similarity Between Node Pair].** *The similarity between two nodes  $u$  and  $x$  is defined as:*

$$S(u, x) = \begin{cases} \delta^{(l-1)}, & 0 < \delta < 1 \text{ if } L(u, x) = l \leq r \\ 0 & \text{otherwise} \end{cases}$$

**Table 1.** Computation Results for Example 2

Node	DC	BC	Diversity ( $\alpha = 0.8 \delta = 0.8$ )			
			r=1	r=2	r=3	r=4
A	3	48	3	5.208	5.208	5.208
B	4	27	1.6	2.763	4.147	4.245
C	3	0	0.867	1.767	2.962	4.489

If two nodes are too far apart, in the sense that their distance is larger than the neighborhood radius  $r$  of our interest, their similarity is considered to be zero; Otherwise, their similarity is inversely proportional to their distance.  $\delta$  is a damping factor to reflect the notion that nodes farther apart share less similarity. The effect of  $\delta$  is further explored in Section 3.4. With the similarity between a pair of nodes defined, we can give the definition of similarity between a node and a set of nodes.

**Definition 4 [Similarity Between Node and Node Set].** The similarity between a node  $u$  and a set of nodes  $N_{-u}(v)$  is defined as

$$S(u, N_{-u}(v)) = \frac{\sum_{x \in N_{-u}(v) \cap N_{-v}(u)} (w_v(x) * S(u, x))}{\sum_{x \in N_{-v}(u)} S(u, x)} \quad (4)$$

where  $w_v(x)$  is the weight of  $x$  in  $v$ 's neighborhood.

The purpose of setting weight, e.g.,  $w_v(u)$  and  $w_v(x)$ , is to prioritize all the nodes in  $v$ 's neighborhood. There are more than one possible ways to define the weights. In this paper, we define  $w_v(x) = S(v, x)$  based on the argument that distance-based similarity is an appropriate way to evaluate the priority of a node in  $v$ 's neighborhood when a radius larger than 1 is needed. Putting it together, we have

$$S(u, N_{-u}(v)) = \frac{\sum_{x \in N_{-u}(v) \cap N_{-v}(u)} (S(v, x) * S(u, x))}{\sum_{x \in N_{-v}(u)} S(u, x)} \quad (5)$$

It is easy to notice that the definition in Section 3.2 is a special case of this general definition.

### 3.4 Examples and Analysis

To illustrate the intuition of the diversity measure above and analyze the impact of its parameters, we get the computation results for Example 2 and 3 in Fig.1(b)(c) with changing parameters and show them in Table 1 and 2, where the computation results of degree and betweenness centrality are also listed<sup>1</sup>.

**Comparison with Degree and Betweenness.** Example 2 demonstrates that diversity does not equal to degree. E.g., A and C are with the same degree but their diversities differ a lot. In Example 3, as the neighbors of all the nodes are not directly connected with

<sup>1</sup> DC and BC denote degree and betweenness centrality for short respectively in this paper.



**Table 2.** Computation Results for Example 3

Node	DC	BC	Diversity ( $\alpha = 0.8, \delta = 0.5$ )						Diversity ( $\alpha = 0.8, \delta = 0.8$ )					
			r=1	r=2	r=3	r=4	r=5	r=6	r=1	r=2	r=3	r=4	r=5	r=6
A	2	42	2	4.70	4.74	4.74	4.74	4.74	2	5.31	4.97	4.97	4.97	4.97
B	6	47	6	3.19	3.92	3.99	3.99	3.99	6	3.04	4.37	4.39	4.39	4.39
C	5	43	5	2.98	3.90	3.96	3.96	3.96	5	2.85	4.50	4.51	4.51	4.51
D	2	1.6	2	2.39	2.69	3.19	3.24	3.24	2	2.33	2.96	4.25	4.38	4.37
E	2	2.25	2	2.16	2.48	3.10	3.15	3.15	2	2.14	2.82	4.41	4.51	4.51
F	5	5	5	2.34	2.73	3.15	3.39	3.41	5	2.13	3.01	4.11	5.06	5.18
G	4	3	4	2.08	2.47	2.90	3.19	3.21	4	1.92	2.83	3.94	5.13	5.25

each other, the value of diversity equals to degree when  $r = 1$ . But when  $r$  increases from 1 to 2, the diversity ranking changes. Example 3 demonstrates that diversity does not equal to betweenness centrality either. E.g., betweenness centrality of  $A$  and  $C$  in Fig.1(c) are roughly the same, but their diversities are obviously different.

**Radius of Neighborhood.** Table 1 and 2 show all the calculation results when  $r$  changes from 1 to the possible maximal value (it means that the neighborhood would no longer change when  $r$  increases more). It is found that a larger radius may lead to counter-intuitive ranking results. However, it is our belief and definition that diversity should measure an aspect of a node’s interaction with its local neighborhood. To judge a node’s diversity on a global scale (e.g., considering all the nodes as neighbors of the center node) is semantically controversial. On the other hand, it is discovered that “*small world*” phenomenon applies to a wide range of networks such as the Internet, the social networks like Facebook and the bio-gene networks, which means most nodes in these networks are found to be within a small number of hops from each other. In particular, the theory of “six degrees of separation” indicates that in social network most people can reach any other individuals through six persons. It follows that when  $r$  increases beyond a small number, a node’s diversity would be aggregated by nearly all the nodes’ contributions in the network, which deviates away from what diversity is meant to capture based on our previous discussion. Therefore, a small radius should be chosen in the computation. Furthermore, the results show that the top-k results in the diversity ranking become stable when  $r = 2$  or  $r = 3$  in most cases.

**Damping Factor.** The damping factor  $\delta$  controls a neighbor’s impact on the diversity measure in relation to its distance to the central node. Intuitively, neighbors far away should have smaller impact on the central node’s diversity. As we discussed above, diversity is influenced mainly by two factors: the size of the neighborhood and the difference among the neighbors. On real data sets, as the radius increases, the number of neighbors increases enormously, which makes the size of neighborhood be a dominating factor of diversity computation. This imbalance would sometimes distort the ranking result. Therefore an appropriate damping factor can be chosen to balance the two factors, e.g.,  $\delta = 0.5$  in Table 2 .

## 4 Top-K Diversity Ranking Algorithm

In real applications, top-k diversity ranking for query-based dynamic networks is often required in data scenarios. Still take the DBLP example. Suppose the original input network is the entire DBLP co-authorship network  $G$  generated by including papers from all the eligible conferences. If a user poses a query “Who are the most diverse researcher in Database community?”, it would result in the dropping of edges which correspond to papers published in non-database conferences. Diversity ranking is then computed on the resulting sub-network. The challenge for computing measures on dynamic networks is that it is no longer possible to compute once for all and answer all the queries by retrieving saved results. As such, the task is to develop efficient algorithms for top-k diversity measure on dynamic networks generated by user queries.

Our strategy is to find ways to quickly estimate an upper-bound of  $D(v)$  for each node  $v$  in the new sub-network. Meanwhile we store the smallest diversity value of top  $k$  candidates which is denoted as  $l\_bound$ . If the upper-bound of  $v$  is smaller than  $l\_bound$ , it can be tossed away to save computation. Otherwise we perform more costly computation to get the accurate measure value of  $D(v)$  and update  $l\_bound$ .

We obtain the upper-bound based on two scenarios. First, the diversity of a node should be smaller than the cardinality of its neighborhood. When all the neighbors have no connections, the diversity reaches the maximal value. On the other hand, as the query-based dynamic network is a subgraph of original network, one node’s neighborhood should be the sub-set of its original neighborhood. Thus two nodes’ similarity should be smaller than their similarity on the original network. By using the monotonicity property, we obtain the upper-bounds and propose an efficient top-k diversity ranking algorithm.

For any quantity  $W$  computed on a network  $G$ , we use  $W'$  to represent the same quantity computed on a sub-network  $G' \subseteq G$ . We use  $N_u(v)$  to denote the set of nodes in  $v$ ’s  $r$ -neighborhood which can only be reached by shortest paths passing through  $u$ , i.e.,  $N_u(v) = N(v) \setminus N_{-u}(v)$ .

**Lemma 1.** *For a network  $G$  and a node  $v \in V(G)$ ,  $D(v) \leq \sum_{u \in N(v)} w_v(u)$ .*

Lemma 1 is due to the fact that  $F(u, v) \leq 1$  by definition and  $F(u, v) = 1$  only when all the neighbors of  $v$  have no connections.

**Lemma 2.** *For a network  $G$  and a sub-network  $G' \subseteq G$ , for any two nodes  $u, v \in V(G)$ ,  $0 \leq S'(u, v) \leq S(u, v) \leq 1$ .*

Lemma 2 is due to the fact that the length of the shortest path  $L(u, v)$  for any two nodes  $u$  and  $v$  in  $G$  increases monotonically in sub-network  $G'$ .

We define some notations to simplify the formulas. We set  $C(v) = \sum_{u \in N(v)} w_v(u)$ . According to Lemma 1,  $C(v)$  is an upper bound of  $D(v)$ . Since in this paper we define  $w_v(u) = S(u, v)$ , we also have  $C(v) = \sum_{u \in N(v)} S(u, v)$ . Hence, for any sub-network  $G' \subseteq G$ ,  $C'(v) = \sum_{u \in N'(v)} S'(u, v)$ . We denote  $S = \sum_{x \in N_{-u}(v) \cap N_{-v}(u)} (S(v, x) * S(u, x))$  for short.

Input: Sub-network  $G'$  and  $K$   
Output: A set  $T$  of  $K$  nodes with top diversity  
1:  $Q \leftarrow$  Queue of  $V(G')$ , sorted by  $C'(v)$   
2:  $l\_bound \leftarrow 0$ ;  $T \leftarrow \emptyset$ ;  
3: Pop out the top node  $v$  in  $Q$   
4:   **if**  $C'(v) < l\_bound$  **return**  $T$ ;  
5:   **for each**  $u \in N'(v)$   
6:     Compute  $Upper(u, v)$ ;  
7:      $UP(v) \leftarrow UP(v) + \min\{1, Upper(u, v)\}$   
8:   **if**  $UP(v) < l\_bound$    **continue**;  
9:   **for each**  $u \in N'(v)$   
10:     Compute  $F'(u, v)$ ;  
11:      $D'(v) \leftarrow D'(v) + F'(u, v)$ ;  
12:   **if**  $D'(v) > l\_bound$    insert  $v$  into  $T$   
13:   **if**  $|T| > K$   
14:     remove the last node in  $T$ ;  
15:      $l\_bound \leftarrow$  smallest diversity in  $T$ ;  
16: **return**  $T$ ;

**Algorithm 1.** Top-K Diversity Ranking

Since  $0 \leq S(u, v), S'(v, x) \leq 1$  for any nodes  $u$  and  $v$ , we have for any node  $x$ ,

$$\begin{aligned} & S(v, x) - S'(v, x) + S(u, x) - S'(u, x) \\ & \geq (S(v, x) - S'(v, x)) * S(u, x) + (S(u, x) - S'(u, x)) * S'(v, x) \\ & = S(v, x) * S(u, x) - S'(u, x) * S'(v, x) \end{aligned}$$

If we sum up by  $x$  for the above inequality, since  $S(v, x) = 0$  for  $x \notin N(v)$  (resp. for  $S(u, x)$ ), and  $S(v, x) * S(u, x) = 0$  for  $x \notin (N(v) \cap N(u))$ , we have

$$C(v) - C'(v) + C(u) - C'(u) \geq S - S' + \sum_{x \in A} S(u, x) * S(v, x) - \sum_{x \in B} S'(u, x) * S'(v, x)$$

where  $A = N(u) \cap N(v) - N_{-v}(u) \cap N_{-u}(v)$ .  $B = N'(u) \cap N'(v) - N'_{-v}(u) \cap N'_{-u}(v)$ . As  $B \subseteq A$ ,  $S(u, x) \geq S'(u, x)$ ,  $\sum_{x \in A} S(u, x) * S(v, x) - \sum_{x \in B} S'(u, x) * S'(v, x) \geq 0$ . Therefore,

$$C(v) - C'(v) + C(u) - C'(u) \geq S - S'$$

So

$$\begin{aligned} F'(u, v) &= 1 - \alpha * \frac{S'}{\sum_{x \in N_{-v}(u)} S'(u, x)} \\ &\leq 1 - \alpha * \frac{(S - (C(u) - C'(u) + C(v) - C'(v)))}{\sum_{x \in N_{-v}(u)} S'(u, x)} \\ &\leq 1 - \alpha * \frac{(S - (C(u) - C'(u) + C(v) - C'(v)))}{C'(u)} \\ &= Upper(u, v) \end{aligned}$$

We thus derived another upper-bound  $Upper(u, v)$  for  $F'(u, v)$ . Thus  $F'(u, v) \leq \min\{1, Upper(u, v)\}$ .

To use this upper-bound, we compute  $S$  for each pair  $(u, v)$  which are each other's  $r$ -neighbors in the original network and store these values in the pre-computation stage. Likewise, we also compute and store  $C(v)$ . When the user inputs a query, we just need to compute  $C'(u)$  and  $C'(v)$  for the sub-network, which is simply a local neighbor checking, to get  $Upper(u, v)$ .

The top-k diversity ranking algorithm is as shown in Algorithm 1.

## 5 Experimental Results

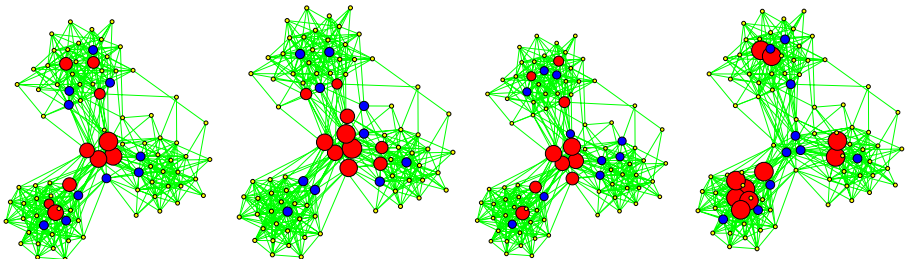
In this section, we did extensive experiments on both synthetic and real data and generated some interesting results. The most diverse nodes on different types of networks are highlighted to illustrate an intuition of diversity. We compare the results of diversity with two classical centrality measures – degree and betweenness centrality and show both the difference and the correlation between them. At last, we implemented our top-k ranking algorithm on dynamic network and demonstrate its efficiency.

### 5.1 Results on Synthetic Network

We first applied the algorithm to a synthetic network consisting of 92 nodes and 526 edges shown in Fig.2. The network was generated as following: first, we generated three clusters of nodes; in each cluster the nodes only connect with the nodes in the same cluster randomly; then we generated other 10 nodes connecting to any node arbitrarily.

Fig.2 shows the top 20 nodes ranked by degree, betweenness centrality and diversity respectively. The top 10 nodes are highlighted with red color and the sizes of nodes are linear with the ranking (The higher the rank, the larger the size). The second top 10 nodes are highlighted with blue color [1].

This figure demonstrates that the nodes which connect more nodes from different clusters tend to be more diverse. When  $r$  increases from 1 to 2, the diverse nodes will further move to the connection points of clusters. It seems that diversity is highly correlated with betweenness centrality on this network. Their correlation coefficients are



(a) Diversity when  $r = 1$  (b) Diversity when  $r = 2$  (c) Betweenness Centrality (d) Degree Centrality

Fig.2. Synthetic network results

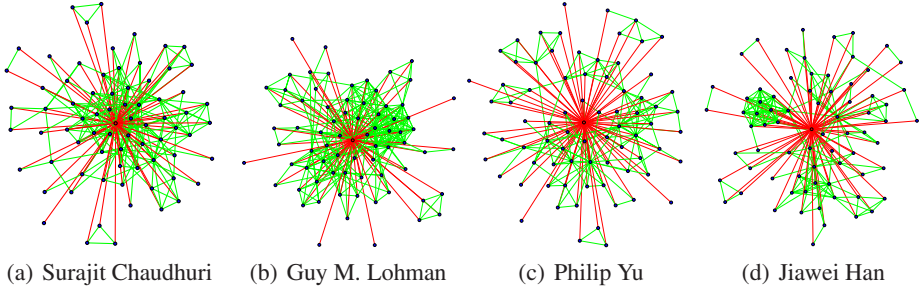


Fig. 3. Neighborhood of four authors

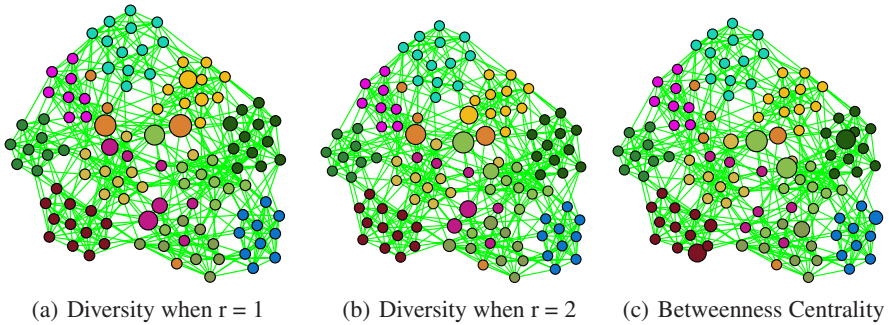


Fig. 4. Network of American football games

shown in Table 5<sup>2</sup>. This large correlation is caused by the characteristic of this network structure. As the network consists of three clusters and some other nodes connecting the clusters, the nodes with high betweenness centrality values also tend to locate on the connection points of clusters. However, diversity is different from betweenness centrality as we analyzed above. And we will show that they are lowly correlated on some networks with different structures.

## 5.2 Results on DBLP Network

We extracted the network of co-authorship on conference SIGMOD, VLDB and ICDE from DBLP data<sup>3</sup>, which means that if two authors cooperated a paper published on these conferences, an edge was generated to link them. Table 3 compares the top 20 author ranked by diversity and betweenness centrality. We set  $\alpha = 0.8$ ,  $\delta = 0.5$ . As it is proved that on an undirected network degree is consistent to authority (eigenvector centrality) obtained by PageRank [4], we can also treat degree as an authority value and compare it with diversity. Thus Table 3 demonstrates that diversity ranking is different from betweenness centrality ranking as well as authority (degree).

<sup>2</sup> SN denotes synthetic network for short.

<sup>3</sup> This network is called as "DB" for short in the remainder of the paper.

**Table 3.** Author Ranking Results on DB

Diversity when $r = 1$			Diversity when $r = 2$		Betweenness Centrality	
Author	DC	Value	Author	Value	Author	Value
Rakesh Agrawal	98	50.94	Rakesh Agrawal	450.84	Rakesh Agrawal	971048.8
David J. DeWitt	118	50.60	David J. DeWitt	434.77	Michael J. Carey	785089.9
Hector Garcia-Molina	98	48.20	Surajit Chaudhuri	402.93	Christos Faloutsos	747502.4
Divesh Srivastava	89	46.75	Michael J. Carey	386.85	David J. DeWitt	746523.0
Surajit Chaudhuri	73	45.53	Divesh Srivastava	373.34	Umeshwar Dayal	737304.2
Raghu Ramakrishnan	90	44.95	Jennifer Widom	367.29	Michael Stonebraker	705067.8
H. V. Jagadish	82	41.53	Hector Garcia-Molina	364.51	Hector Garcia-Molina	685955.0
Hamid Pirahesh	83	41.45	Raghu Ramakrishnan	360.98	Surajit Chaudhuri	631760.8
Michael J. Carey	115	41.05	Michael J. Franklin	360.09	Philip A. Bernstein	628037.5
Michael Stonebraker	113	40.93	Jeffrey F. Naughton	349.62	H. V. Jagadish	604977.7
Jennifer Widom	84	40.29	Hamid Pirahesh	343.99	Divesh Srivastava	562573.6
Christos Faloutsos	94	39.21	H. V. Jagadish	339.80	Raghu Ramakrishnan	555216.0
Jeffrey F. Naughton	95	38.86	Gerhard Weikum	333.76	Gerhard Weikum	540029.5
Guy M. Lohman	73	37.98	Umeshwar Dayal	330.88	Elisa Bertino	533129.3
Michael J. Franklin	76	37.42	Philip A. Bernstein	327.75	Dennis Shasha	526097.3
Nick Koudas	69	37.32	Michael Stonebraker	326.91	Jiawei Han	520527.3
C. Mohan	66	36.19	Abraham Silberschatz	326.70	Michael J. Franklin	518074.6
Gerhard Weikum	80	34.11	C. Mohan	322.23	Gio Wiederhold	517573.1
Philip A. Bernstein	61	33.45	Guy M. Lohman	320.67	Kian-Lee Tan	513349.0
Rajeev Rastogi	75	33.36	Bruce G. Lindsay	312.36	C. Mohan	509267.1

Table 3 demonstrates some interesting results. For example, although the difference between the degrees of R. Agrawal and D. DeWitt is as large as 20, their diversities are nearly the same. The reason should be that R. Agrawal is from industry area and has worked in many companies, e.g., Microsoft, IBM Almaden Research Center, Bell Laboratories, etc. Therefore, Agrawal’s cooperators are very diverse. We also compare the diversity of two authors, Surajit Chaudhuri and Guy M. Lohman, who have the same degree. Their neighborhoods as shown in Fig.3(a) and Fig.3(b) demonstrate that Lohman’s cooperators connect with each other more closely than Chaudhuri’s. Therefore the diversity of Chaudhuri is larger than Lohman as obtained in Table 3.

We can also get similar results on the co-author network of conference KDD and ICDM from DBLP data<sup>4</sup> as shown in Table 4. For example, although Philip S. Yu and Jiawei Han’s degrees are roughly the same, their diversities differ a lot, which can also be demonstrated from their neighborhoods as shown in Fig.3(c) and Fig.3(d). The reason should be that Philip S. Yu had worked in industry area and has cooperated with many different persons who have no close relationship. Thus his diversity value is much larger than Jiawei Han.

### 5.3 Results on Network of American Football Games

We obtained another social network of American football games between Division IA colleges during regular season Fall 2000 [6]. In this data, nodes represent teams and

<sup>4</sup> The network is called as “DM” for short in the remainder of the paper.

**Table 4.** Author Ranking Results on DM

Diversity when $r = 1$			Diversity when $r = 2$		Betweenness Centrality	
Author	DC	Value	Author	Value	Author	Value
Philip S. Yu	76	39.72	Philip S. Yu	160.82	Philip S. Yu	544203.3
Jiawei Han	73	26.25	Haixun Wang	107.15	Christos Faloutsos	335598.8
Christos Faloutsos	60	24.77	Jiawei Han	96.85	Heikki Mannila	179383.3
Jian Pei	51	20.37	Christos Faloutsos	93.26	Mohammed Javeed Zaki	158551.1
Haixun Wang	32	19.21	Ke Wang	92.37	Jiawei Han	132043.5
Ke Wang	36	17.30	Jian Pei	91.13	Eamonn J. Keogh	123389.1
Heikki Mannila	39	16.54	Ada Wai-Chee Fu	82.14	Padhraic Smyth	116926.1
Bing Liu	32	15.15	Jianyong Wang	75.56	Jian Pei	112538.7
Mohammed Javeed Zaki	30	14.50	Charu C. Aggarwal	74.11	Charu C. Aggarwal	107042.4
Eamonn J. Keogh	37	14.32	Wei Fan	73.63	Bing Liu	103081.9
Wei Fan	29	14.26	Wei Wang	71.52	Gregory Piatetsky-Shapiro	101267.2
Padhraic Smyth	32	13.89	Bing Liu	70.26	Srinivasan Parthasarathy	95692.4
Wei-Ying Ma	34	13.73	Spiros Papadimitriou	69.17	Ada Wai-Chee Fu	91889.1
Ada Wai-Chee Fu	25	13.70	Hong Cheng	69.14	Ke Wang	90909.1
Qiang Yang	41	13.68	Eamonn J. Keogh	67.69	Haixun Wang	88484.7
Vipin Kumar	29	13.21	Alexander Tuzhilin	64.71	Vipin Kumar	82333.2
Wei Wang	39	13.13	Jiong Yang	63.58	Rakesh Agrawal	80409.2
Hui Xiong	27	13.02	Hongjun Lu	62.50	Huan Liu	79472.5
Huan Liu	28	12.92	David W. Cheung	60.45	Spiros Papadimitriou	78784.6
Alexander Tuzhilin	17	12.16	Michail Vlachos	60.28	Prabhakar Raghavan	77359.7

edges denote that two teams had a game. Fig.4 shows the top 10 nodes with largest diversity and betweenness centrality, which are highlighted by the larger sizes of nodes. The degrees of all the nodes are roughly the same, with the range from 8 to 12. Thus we do not show the degree ranking results. The data also contain the node labels which indicate the conference that each team belongs to. We use different colors to distinguish the labels in the figure. Therefore the results illustrate that the diversity calculated based on network topology is consistent to the diversity based on node labels, which means that the nodes whose neighbors are from more clusters tend to be more diverse. Table 5<sup>5</sup> demonstrates that on this network the diversity is lowly correlated with degree and betweenness centrality.

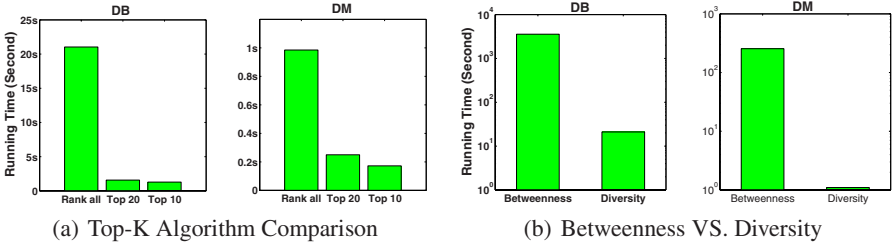
## 5.4 Performance Comparison

Fig.5(a) compares the running time of Top-K algorithm with the time of ranking all the nodes on DB and DM networks. It demonstrates that Top-K algorithm is much more efficient and can meet online query needs. We also implemented an efficient betweenness algorithm [3] and compared it with diversity. Fig.5(b) demonstrates that diversity calculation is much faster than betweenness calculation. The reason is that to some extent betweenness centrality is a global measure based on the shortest path calculation between all the pair-nodes which is very time consuming while the diversity measure only needs to count the local neighborhood.

<sup>5</sup> FN denotes the social network of American football games for short.

**Table 5.** Correlation Coefficients of Metrics

Network	#node	#edge	DC vs. BC	DC vs. Diversity		BC vs. Diversity	
				r = 1	r = 2	r = 1	r = 2
SN	92	526	0.470	0.874	0.399	0.709	0.828
FN	115	616	0.151	0.345	0.224	0.413	0.463
DB	7640	22309	0.810	0.881	0.819	0.829	0.716
DM	3405	6496	0.665	0.908	0.683	0.701	0.576

**Fig. 5.** Performance comparison

## 6 Discussion

As diversity is a highly subjective concept, we do not think there exists one optimal definition which is applicable for all scenarios. Rather than narrowing ourselves down to one specific definition, we are fully aware of other possible definitions that may be better geared for other applications. For example, a highly intuitive definition can be based on clustering, where nodes are first assigned labels by certain clustering algorithm and then diversity is computed by calculating the information entropy of the cluster distribution of neighbors. This kind of definition needs to at least solve the following issues: (i) The choice of the clustering algorithm dictates the resulting clusters, which in turn determines the diversity computation. The decision on clustering parameters becomes critical and difficult. (ii) The internal cohesion of clusters, which reflects the topology of network, is also an important component for diversity. The diversity of a node connected with a compact cluster should be different from the diversity of a node connected with a loose cluster. Therefore in general still lots of aspects and factors should be exploited for the clustering-based definition. In this paper, we propose a straightforward diversity definition based on the similarity between neighbors instead of solving these problems of clustering.

## 7 Conclusion

In this paper, we investigated the problem of mining diversity on networks. We gave two criteria to characterize the semantic meaning of diversity and to provide the basis of proposing a reasonable measure definition. Then we studied diversity measure based on network topology and picked a concrete definition to embed the idea. We



developed an efficient algorithm to find top-k diverse nodes on dynamic networks. Extensive experiment studies were conducted on synthetic and real data sets. The results are interesting, where individual nodes identified with high diversities are intuitive.

## Acknowledgements

The work was supported in part by the U.S. National Science Foundation grants IIS-08-42769 and IIS-09-05215, and the NASA grant NNX08AC35A, and 973 Program of China grant 2006CB303103, and the State Key Program of National Natural Science of China grant 60933013. Any opinions, findings, and conclusions expressed here are those of the authors and do not necessarily reflect the views of the funding agencies.

## References

1. <http://graphexploration.cond.org/index.html>
2. Barabasi, A.-L., Oltvai, Z.N.: Network biology: Understanding the cell's functional organization. *Nat. Rev. Genet.* 5(2), 101–113 (2004)
3. Brandes, U.: A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology* 25, 163–177 (2001)
4. Cover, T.M., Thomas, J.A.: *Elements of information theory*. John Wiley & Sons Inc., Chichester (2006)
5. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. In: *SIGCOMM*, pp. 251–262 (1999)
6. Girvan, M., Newman, M.E.J.: Community structure in social and biological networks. *Proceedings of the National Academy of Sciences* 99(12) (2002)
7. Hwang, W., Kim, T., Ramanathan, M., Zhang, A.: Bridging centrality: graph mining from element level to group level. In: *KDD*, pp. 336–344 (2008)
8. Kuramochi, M., Karypis, G.: Frequent subgraph discovery. In: *ICDM*, pp. 313–320 (2001)
9. Lawrence, P., Sergey, B., Motwani, R., Winograd, T.: The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University (1998)
10. Leskovec, J., Kleinberg, J.M., Faloutsos, C.: Graphs over time: densification laws, shrinking diameters and possible explanations. In: *KDD*, pp. 177–187 (2005)
11. Putnam, R.D.: Bowling Alone: America's Declining Social Capital. *Journal of Democracy* 6(1) (1995)
12. Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P.: The author-topic model for authors and documents. In: *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, Arlington, VA, USA, pp. 487–494. AUAI Press (2004)
13. Stephenson, K., Zelen, M.: Rethinking centrality: Methods and examples. *Social Networks* 11(1), 1–37 (1989)
14. Wasserman, S., Faust, K.: *Social Network Analysis, Methods and Applications*. Cambridge University Press, Cambridge (1994)
15. Watts, D.J., Strogatz, S.H.: Collective dynamics of 'small-world' networks. *Nature* 393(6684), 440–442 (1998)
16. Yan, X., Han, J.: gSpan: Graph-based substructure pattern mining. In: *ICDM*, pp. 721–724 (2002)