Multimed Tools Appl (2012) 56:179–205
DOI 10.1007/s11042-010-0568-1

Mining diversity on social media networks

Lu Liu · Feida Zhu · Meng Jiang · Jiawei Han · Lifeng Sun · Shiqiang Yang

Published online: 27 July 2010

Abstract The fast development of multimedia technology and increasing availability of network bandwidth has given rise to an abundance of network data as a result of all the ever-booming social media and social websites in recent years, e.g., Flickr, Youtube, MySpace, Facebook, etc. Social network analysis has therefore become a critical problem attracting enthusiasm from both academia and industry. However, 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. Based on the approach, we can measure not only a user's sociality and interest diversity but also a social media's user diversity. An efficient top-k diversity ranking algorithm is developed for computation on dynamic networks. Experiments on both synthetic and real social media datasets give interesting results, where individual nodes identified with high diversities are intuitive.

Keywords Social network · Mining · Diversity

L. Liu (✉)
Tsinghua University, Capital Medical University, Beijing, China
e-mail: lu-liu@mails.tsinghua.edu.cn

M. Jiang · L. Sun · S. Yang
Tsinghua University, Beijing, China

F. Zhu
Singapore Management University, 81 Victoria St., Singapore, Singapore

J. Han
University of Illinois at Urbana-Champaign, Urbana, IL, USA

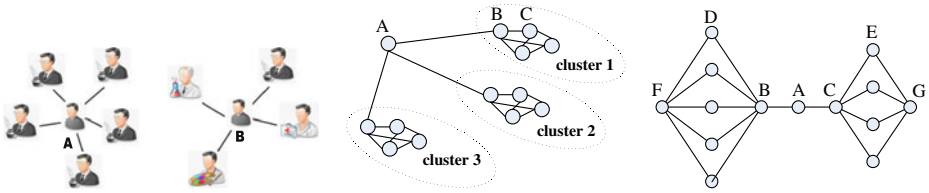
1 Introduction

Thanks to the fast advancement of multimedia technology and increasing availability of network bandwidth, social media and social websites (e.g., Flickr, Youtube, MySpace, Facebook, etc.) have emerged into proliferation in recent years. On these social websites, users can not only connect with their friends in real life but also create and share their social media with friends. For example, Renren is a Facebook-style website in China. Users can find and create links with their old friends with whom they have lost connections for a long time. Moreover, they could share their daily life, e.g., what are they doing now? how do they like a particular movie? They can also create or share social media, e.g., text, image, video, etc., among their friends. With the help of the social media, friends would know each other better and their relationship in real life would be strengthened.

Many of today's real-world applications, such as the social media websites, have generated an enormous amount of network data. Therefore, social network analysis has become a critical problem of focus for both academia and industry. To gain a deep understanding of the structures and functions of these network datasets, it is fundamental to investigate various properties of the network and its constituent components, i.e., nodes, edges, coherent subnetworks, etc. To this end, various kinds of network analysis have been proposed and conducted, offering useful insight into a great number of network-structured data. For example, small world phenomena and hierarchical modularity have been discovered for real-world network systems [2, 23], as well as inhomogeneous nodal degree distribution and the presence of a limited number of highly connected hubs as a result of power-law degree distribution [6]. While these global statistical properties provide interesting observations of the real-world networks as a whole, various network measures have also been introduced to characterize individual properties of the network components, e.g., a number of different centrality measures were proposed, including degree/betweenness/closeness centrality [22] and eigenvector centrality [14]. These measures, if used properly, can successfully capture the distinctiveness of different participants in the network.

However, as we shall argue in the following of this paper, one important measure, different from those defined in previous literature that focuses on distinctiveness, has so far been neglected. This measure, namely *diversity*, characterizes how diverse a node's connection to its neighborhood is. To illustrate the intuition behind, 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. 1a 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 want to choose a node to launch a marketing campaign.



(a) Example 1 (b) Example 2 (c) Example3

Fig. 1 Three examples

Example 1 demonstrates that the diversity of a node on a 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. 1b, 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.

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 [25]. 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 [18]. 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, which leads 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 blossom in many of today's real-world applications. For example, advertisers may be very interested in the most diverse users in social networks 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 are the top ten diversely-collaborating researchers in the

data mining community?". If this kind of diversity on homogeneous networks can be used to measure a person's sociality, the diversity measure on heterogenous networks, e.g., the bipartite network between users and social media, can also be used to distinguish users' interest diversity as well as the popularity diversity of social media, i.e., the problem of what kind of topic attracts more diverse users.

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.

These two factors can also been 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 networks 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 networks 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 and networks from social websites. 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

Properties reflecting the overall characteristics of networks, such as density, small world, hierarchical modularity and power law, have been observed for a long time [2, 6, 15, 23]. Compared to these, there are also measures that focus on individual components, cf. [20, 22]. Degree centrality, which is defined as the number of links for a given node, is often used to identify highly connected nodes which get exposure to whatever is flowing through the network (such as a virus). Betweenness centrality assigns higher values to nodes appearing on the shortest paths of more node pairs, with extensions considering paths that are not shortest as well. Closeness centrality is related to betweenness centrality, which measures the average shortest-path length from a node to all other nodes in the network. Centrality measures have also been enhanced for analysis performed on the level of node groups [9]. The clustering coefficient of a node assesses the local connectivity among its direct neighbors. Eigenvector centrality is a way to measure authority. It assigns relative scores to all nodes in the network based on the principle that a node that connects to high-scoring nodes should be assigned with high scores also. PageRank [14] is a variant of eigenvector centrality, which has a natural interpretation relating to random network surfing. Finally, some other types of patterns, e.g., frequent subgraphs that focus more on local topologies [13, 24], can also be mined from the network.

It is worth noting that all these measures are different from diversity and thus could not accurately capture the idea behind. From Example 2, it is obvious that degree centrality does not consider whether the neighbors of a given node are similar. As we shall observe in the experiments, betweenness centrality might be correlated with diversity to some extent in particular data scenarios, however, it is not a direct modeling of diverseness and thus would not satisfy the two criteria we have proposed in general. Similarly, closeness centrality has the same kind of problems; moreover, such shortest-path based measures require the global computation of all-pair shortest paths, which might be time-consuming on a large network. PageRank depicts the authority of a node in the network, which is an orthogonal dimension of measurement. 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 only considers number of edges as the sole parameter, which is inevitably restricted. Interestingly, it can be considered as a degenerated version of our diversity definition when the latter is confined to a very special setting.

The broad literature on graph clustering is also related, because we can always use clustering algorithms to first "classify" all nodes in the network, and then diversity can be easily computed. With characterization of the similarity among node pairs, e.g., simRank [10], various distance-based clustering algorithms can be applied on networks. Spectral clustering [16] conducts partitioning based on graph cut theory. In the maximum clique approach, clustering is performed by identifying fully connected subgraphs [19], and extensions have been proposed to overcome this relative stringency by considering quasi cliques and dense subnetworks [7, 17]. Recently, RankClus also integrates authority ranking information into the clustering procedure [21]. Still, scalability is a big concern here, as clustering requires us to

perform global processing over the whole network, while our diversity definition might only involve a small local neighborhood.

As for potential applications, the influence of a node has been examined in terms of how large or how quick the spread of influence is, given a node as the starting point. Many algorithms have been proposed in recent years to obtain these measures effectively and efficiently. Kossinets et al. [12] tries to extract the "backbone" of a social network, i.e., the subgraph that consists of edges on which information has the potential to flow the quickest. In [11], the spreading of influence through a social network has been studied, in which the goal was to maximize the number of nodes that could be reached. In real advertising scenarios, it is very likely that the advertiser would aim for not only an audience of large size, but also the one with great diversity as well.

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 $G = \{(V, E) \mid V$ is a set of nodes and E is a set of edges, $E \in V \times V$, an edge $e = (i, j)$ connects two nodes i 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 $r = 2$, $N_{-u}(v) = N(v)$ - $\{x|$there is only one shortest path from v to x 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. 1b.

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

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

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 [9].

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.

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

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

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

$S(u, N_{-u}(v))$ is a function measuring the similarity between u and $N_{-u}(v)$ up to a normalization. α 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:

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

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. δ is a damping

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

factor to reflect the notion that nodes farther apart share less similarity. The effect of δ 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

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

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

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

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 Examples 2 and 3 in Fig. 1b and c with changing parameters and show them in Tables 1 and 2, where the computation results of degree and betweenness centrality are also listed.[1]

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 each other, the values of diversity equal 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.1c are roughly the same, but their diversities are obviously different.

[1] 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

Radius of neighborhood Tables 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 δ 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, the top-k diverse nodes are more interesting and meaningful for users than all nodes' diversity values. Furthermore, the diversity of nodes should be associated with topics. As the network structure changes with topics, nodes' diversity values vary on different topics. Thus, top-k diversity ranking for topic-based dynamic

networks is often required in data scenarios. For example, a user may pose a query "Who are the most diverse researchers in Database community?".

The topic-based dynamic networks are sub-networks of original input network. 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. The query "Who are the most diverse researchers in Database community?" actually results 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 subset 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 $v \in V(G)$, $D(v) \leq \sum_{u \in N(v)} w_v(u)$.*

According to (3), we can get that $F(u, v) \leq 1$ as $S(u, N_{-u}(v)) \geq 0$. And only when all the neighbors of v have no connections, $F(u, v) = 1$. Thus Lemma 1 is easy to be proved.

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 $S = \sum_{x \in N_{-u}(v) \cap N_{-v}(u)} (S(v, x) \times S(u, x))$ for short.

Algorithm 1 Top-K diversity ranking

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** $C'(v) < l\_boundQ$ **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;

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

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

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 $S(v, x) \times S(u, x) = 0$ for $x \notin \big(N(v) \bigcap N(u)\big)$, we have

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

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 $B \subseteq A$, $S(u, x) \ge S'(u, x)$, $\sum_{x \in A} S(u, x) \times S(v, x) - \sum_{x \in B} S'(u, x) \times S'(v, x) \ge 0$.

Therefore,

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

And

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

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 Experiments

In this section, we do extensive experiments on both synthetic and real data which generate some interesting results. The most diverse nodes on different genres 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 implement our top-K ranking algorithm on dynamic network and demonstrate its efficiency.

5.1 Data description

Four genres of network data are used in the experiments.

- *Synthetic Network*
  A synthetic network consisting of 92 nodes and 526 edges is generated to get the intuitive impression of diversity. The synthetic network is generated as below: first, we generate three clusters of nodes; in each cluster the nodes only connect with the nodes in the same cluster randomly; then we generate other 10 nodes connecting to any node arbitrarily.
- *Network of Co-authorship on DBLP*
  We extract the networks of co-authorship on two areas: database ("DB") and data mining ("DM"). The former co-authorship network is obtained from the conference SIGMOD, VLDB and ICDE from DBLP data, which means that if two authors cooperated a paper published on these conferences, an edge is generated to link them. And the latter co-authorship network is obtained from the conference KDD, ICDM of DBLP data in the same way.
- *Network of American Football Games*
  We obtain another social network of American football games between Division IA colleges during regular season Fall 2000 [8]. In this data, nodes represent teams and edges denote that two teams had a game.
- *Social Network from Renren Website*
  We crawl the real social network data from the website Renren. The data contain two parts: one is the relationship between users; the other is the social media data shared by users. We have crawled about 5,000 users and their relationship as well as about 300,000 social media shared by these users. On average, each user shares about 100 social media (including duplicate copies).

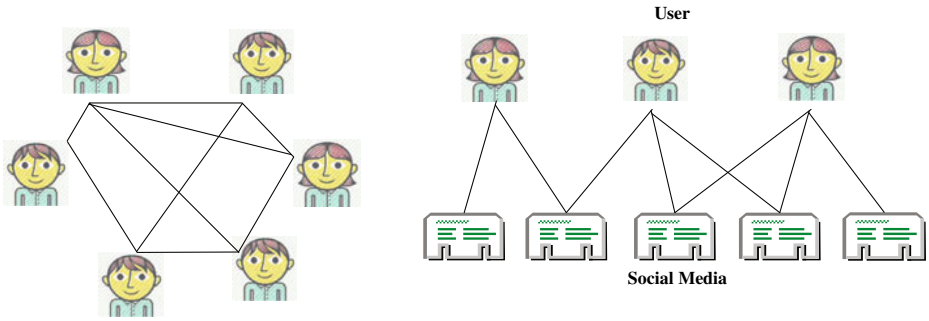


Fig. 2 Social network from Renren website

Based on the data, two types of social network can be built as shown in Fig. 2:

- the social network among the users: the nodes denote users while the edges represent their real-life relationships;
- the bipartite network between users and social media: the nodes represent users and social media while the edges represent the interactions between them.

5.2 Results illustrations

5.2.1 Synthetic network

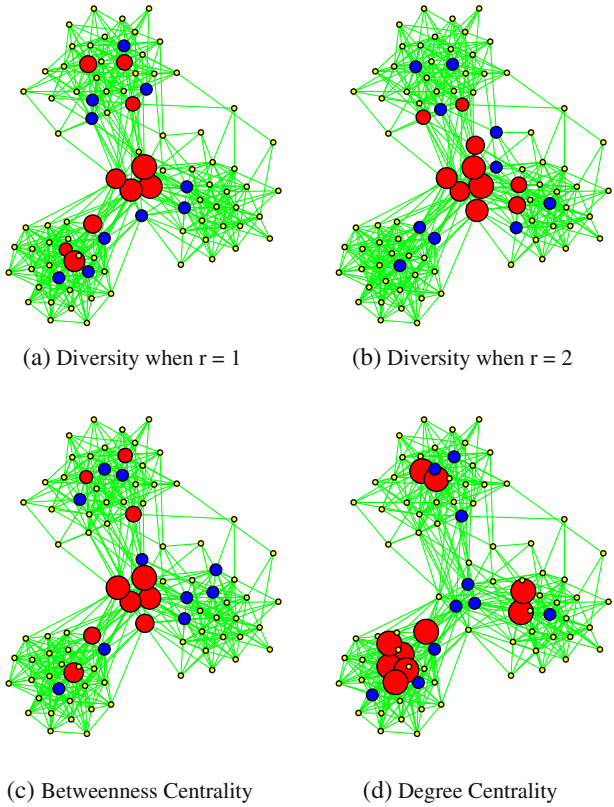
Figure 3 shows the results on the synthetic network. The top 20 nodes ranked by degree, betweenness centrality and diversity respectively are highlighted. The top 10 nodes are colored in red and their sizes are linear with the ranking (The higher the rank, the larger the size). The second top 10 nodes 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 shown in Table 3.[2] 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.

5.2.2 DBLP network

Table 4 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

[2]SN denotes synthetic network for short.



(a) Diversity when r = 1

(b) Diversity when r = 2

(c) Betweenness Centrality

(d) Degree Centrality

Fig. 3 Synthetic network results

consistent to authority (eigenvector centrality) obtained by PageRank [5], we can also treat degree as an authority value and compare it with diversity. Thus Table 4 demonstrates that diversity ranking is different from betweenness centrality ranking as well as authority (degree).

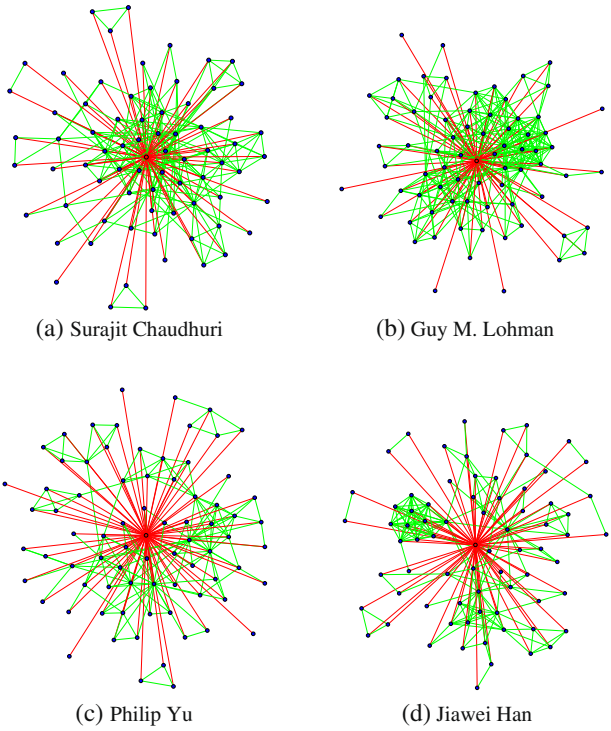
Table 4 demonstrates some interesting results. For example, although the difference between the degrees of R. Agrawal and D. DeWitt is as large as 20, their diversities are nearly the same. The reason should be that R. Agrawal is from industry area and has worked in many companies, e.g., Microsoft, IBM Almaden Research Center, Bell Laboratories, etc. Therefore, Agrawal's cooperators are very diverse. We also compare the diversities of two authors, Surajit Chaudhuri and Guy

Table 3 Correlation coefficients of different measures

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	7,640	22,309	0.810	0.881	0.819	0.829	0.716
DM	3,405	6,496	0.665	0.908	0.683	0.701	0.576

Table 4 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	971,048.8
David J. DeWitt	118	50.60	David J. DeWitt	434.77	Michael J. Carey	785,089.9
Hector Garcia-Molina	98	48.20	Surajit Chaudhuri	402.93	Christos Faloutsos	747,502.4
Divesh Srivastava	89	46.75	Michael J. Carey	386.85	David J. DeWitt	746,523.0
Surajit Chaudhuri	73	45.53	Divesh Srivastava	373.34	Umeshwar Dayal	737,304.2
Raghu Ramakrishnan	90	44.95	Jennifer Widom	367.29	Michael Stonebraker	705,067.8
H. V. Jagadish	82	41.53	Hector Garcia-Molina	364.51	Hector Garcia-Molina	685,955.0
Hamid Pirahesh	83	41.45	Raghu Ramakrishnan	360.98	Surajit Chaudhuri	631,760.8
Michael J. Carey	115	41.05	Michael J. Franklin	360.09	Philip A. Bernstein	628,037.5
Michael Stonebraker	113	40.93	Jeffrey F. Naughton	349.62	H. V. Jagadish	604,977.7
Jennifer Widom	84	40.29	Hamid Pirahesh	343.99	Divesh Srivastava	562,573.6
Christos Faloutsos	94	39.21	H. V. Jagadish	339.80	Raghu Ramakrishnan	555,216.0
Jeffrey F. Naughton	95	38.86	Gerhard Weikum	333.76	Gerhard Weikum	540,029.5
Guy M. Lohman	73	37.98	Umeshwar Dayal	330.88	Elisa Bertino	533,129.3
Michael J. Franklin	76	37.42	Philip A. Bernstein	327.75	Dennis Shasha	526,097.3
Nick Koudas	69	37.32	Michael Stonebraker	326.91	Jiawei Han	520,527.3
C. Mohan	66	36.19	Abraham Silberschatz	326.70	Michael J. Franklin	518,074.6
Gerhard Weikum	80	34.11	C. Mohan	322.23	Gio Wiederhold	517,573.1
Philip A. Bernstein	61	33.45	Guy M. Lohman	320.67	Kian-Lee Tan	513,349.0
Rajeev Rastogi	75	33.36	Bruce G. Lindsay	312.36	C. Mohan	509,267.1



(a) Surajit Chaudhuri (b) Guy M. Lohman

(c) Philip Yu (d) Jiawei Han

Fig. 4 Neighborhoods of four authors

M. Lohman, who have the same degree. Their neighborhoods as shown in Fig. 4 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 4.

We can also get similar results on the co-authorship network on "DM" as shown in Table 5. 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. 4. 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's.

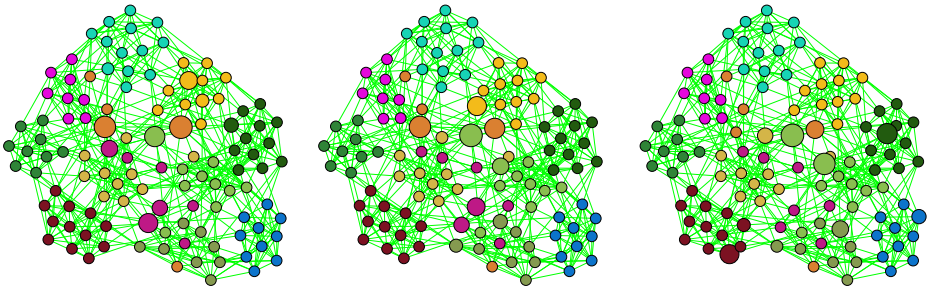
5.2.3 Network of American football games

Figure 5 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 3[3]

[3]FN denotes the social network of American football games for short.

Table 5 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	544,203.3
Jiawei Han	73	26.25	Haixun Wang	107.15	Christos Faloutsos	335,598.8
Christos Faloutsos	60	24.77	Jiawei Han	96.85	Heikki Mannila	179,383.3
Jian Pei	51	20.37	Christos Faloutsos	93.26	Mohammed Javeed Zaki	158,551.1
Haixun Wang	32	19.21	Ke Wang	92.37	Jiawei Han	132,043.5
Ke Wang	36	17.30	Jian Pei	91.13	Eamonn J. Keogh	123,389.1
Heikki Mannila	39	16.54	Ada Wai-Chee Fu	82.14	Padhraic Smyth	116,926.1
Bing Liu	32	15.15	Jianyong Wang	75.56	Jian Pei	112,538.7
Mohammed Javeed Zaki	30	14.50	Charu C. Aggarwal	74.11	Charu C. Aggarwal	107,042.4
Eamonn J. Keogh	37	14.32	Wei Fan	73.63	Bing Liu	103,081.9
Wei Fan	29	14.26	Wei Wang	71.52	Gregory Piatetsky-Shapiro	101,267.2
Padhraic Smyth	32	13.89	Bing Liu	70.26	Srinivasan Parthasarathy	95,692.4
Wei-Ying Ma	34	13.73	Spiros Papadimitriou	69.17	Ada Wai-Chee Fu	91,889.1
Ada Wai-Chee Fu	25	13.70	Hong Cheng	69.14	Ke Wang	90,909.1
Qiang Yang	41	13.68	Eamonn J. Keogh	67.69	Haixun Wang	88,484.7
Vipin Kumar	29	13.21	Alexander Tuzhilin	64.71	Vipin Kumar	82,333.2
Wei Wang	39	13.13	Jiong Yang	63.58	Rakesh Agrawal	80,409.2
Hui Xiong	27	13.02	Hongjun Lu	62.50	Huan Liu	79,472.5
Huan Liu	28	12.92	David W. Cheung	60.45	Spiros Papadimitriou	78,784.6
Alexander Tuzhilin	17	12.16	Michail Vlachos	60.28	Prabhakar Raghavan	77,359.7



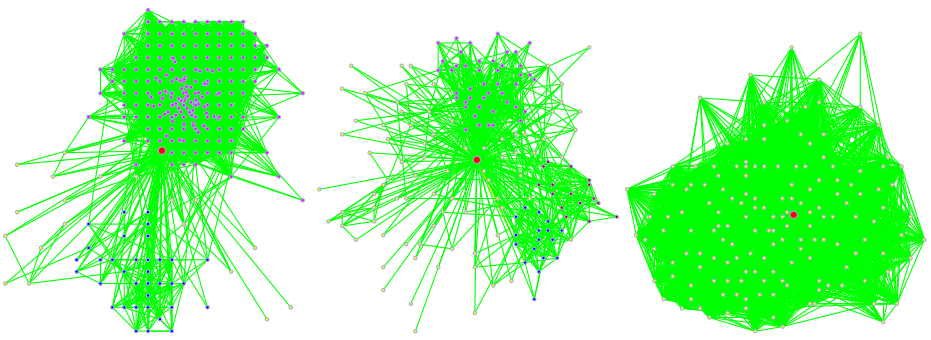
(a) Diversity when r = 1 (b) Diversity when r = 2 (c) Betweenness Centrality

Fig. 5 Network of American football games

demonstrates that on this network the diversity is lowly correlated with degree and betweenness centrality.

5.2.4 Social network of Renren website

- **Social diversity patterns**
  If we apply the diversity measure on the social network consisting of users' connections, we can measure the social diversity of users. For example, if one person's friends all know each other, his/her diversity should be small. Contrarily, if one person connects with more different clusters of persons, his/her diversity should be larger. We test the diversity on the real social network from Renren website. Figure 6 shows three persons' different social diversity patterns. The node with the red color represents the person we studied. Other nodes represent his/her friends. The edges denote their social connections.
  It is obvious that Person A connects mainly to two communities, which are distinguished by different colors. In fact, the two communities are consisting of his/her school mates from the same university and the same middle school respectively. Moreover, the community of the university (denoted by the purple



(a) Person A (b) Person B (c) Person C

Fig. 6 Three persons' neighborhoods on Renren website

Table 6 Social diversity of users on Renren website

Measure	Person A	Person B	Person C
Degree	255	158	158
Diversity	144.844	129.175	74.564
Ratio	0.568	0.818	0.472

color) is much bigger than the other one. Person *B* mainly lives in three communities which are represented in three different colors. Person *C* nearly knows the persons in only one community. Table 6 shows their degree, diversity values and the ratio of diversity to degree. It demonstrates that although Person *B* and Person *C* have the same number of friends, their diversity values differ a lot. Furthermore, Person *B*'s ratio of diversity to degree is the largest. The calculation results are consistent to the intuition.

- **Audience diversity of social media**
  Besides the social diversity of users, the audience diversity of social media can also be calculated by the our proposed measure based on the bipartite graph between users and social media. The semantic meaning is that if the audience diversity value of one social media is large, it denotes that the social media attracts users from more diverse communities.
  In order to test the intuition of social media's audience diversity, we select the top-20 most popular social media which are shared by more than 400 persons and calculate their diversity values. Table 7 shows the content of social media which have the largest and smallest diversity values. It is interesting to find that although the social media which is talking about the Ftp of Tsinghua University has been shared by as many people as other social media, its diversity is the smallest. It means that the users who access this social media are from a small community as the content is on a narrow topic.
  We classify the social media into 4 clusters by their content, which are "learning", "going abroad", "working" and "living" respectively. We calculate the average and variance values of the diversity of each cluster, which are shown in Fig. 7. The dash line in the left figure represents the average value of all the social media and the real line denotes the average value of each cluster. Thus we can tell that the diversity value of the cluster "working" and "living" are over the average level while the cluster of "learning" and "going abroad" are not. Furthermore, the variance value of the cluster "living" is much larger than others. All the results are easy to understand and explain, which prove that the diversity value of social

Table 7 Examples of social media's diversity on Renren website

Rank	Keywords	Diversity	Degree	Diversity/degree
1	All over the world, beautiful, landscape	475.2	580	0.819
2	Office, printing, electronic, typing	458.0	558	0.821
3	TOEFL, language, summary, New Oriental School	417.6	516	0.809
18	Resume, Oreal, High-level	329.2	442	0.745
19	Colleague, reference book, solution, textbook	316.6	399	0.793
20	Tsinghua, ftp, useful, intramural	155.8	447	0.349

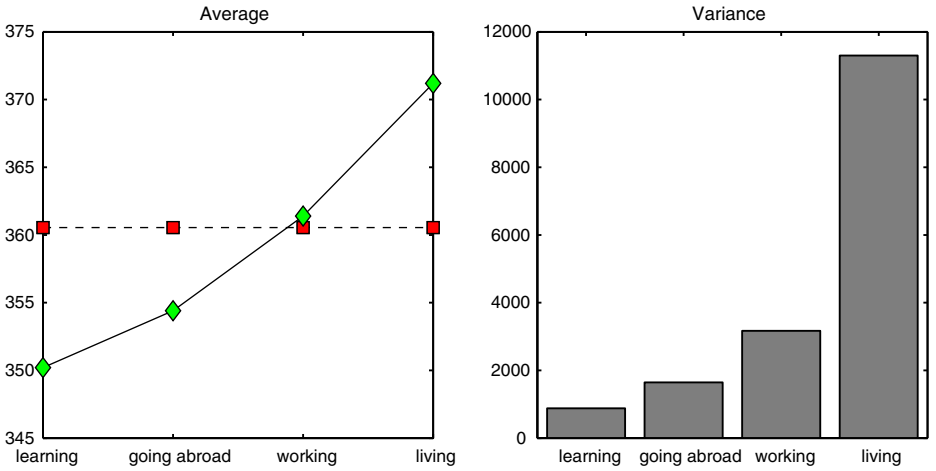


Fig. 7 The average and variance values of four clusters' diversities of social media

media reflects the diversity of the related topic, e.g., "living" is a broad topic which attracts more diverse users.

- **User interest diversity**
  Based on the bipartite graph between users and social media, user interest diversity can also be measured. If a user shares many social media which have different content, we can say that the user has a diverse interest.
  Table 8 shows the content of two users' shared social media. Their interest diversity values are also shown in the first row. We can find that the interest diversity of User 2 is much smaller than User 1 although both of them shared five social media. The reason can be found from the table: User 2 shared the social media all related to campus life while the social media shared by User 1 are on more topics.
  Information entropy (IE) is also used to measure the researchers' interest diversity in some literature [25]. However, the way to measure the diversity by our approach is very different. Information entropy needs to get the topic distribution of users by applying clustering or a topic model to social media content but our measure only uses the network structure instead of content information. In real applications, the content of some social media is ambiguous, which may result in an inaccurate user interest distribution. Our approach is a straightforward measurement which does not rely on the content information,

Table 8 The example of the social media accessed by two users

User 1: 4.035	User 2: 0.061
How to restore what you have just deleted.	The most beautiful landscape over the world.
A girl is bargaining to buy pants.	The interesting things happened in the campus.
Thirteen pieces of advice to impact your life.	A girl is bargaining to buy pants.
The test for civil servant in 2010.	The love story of a famous hostess.
Yu Minhong taught us how to memorize English words.	The love songs for single persons of twelve constellation.

Table 9 Interest diversity of two users

User	# Work	# Life	# Study	# Love	Degree	IE	Diversity
1	4	15	21	3	43	0.488	32.43
2	0	17	7	19	43	0.444	23.85

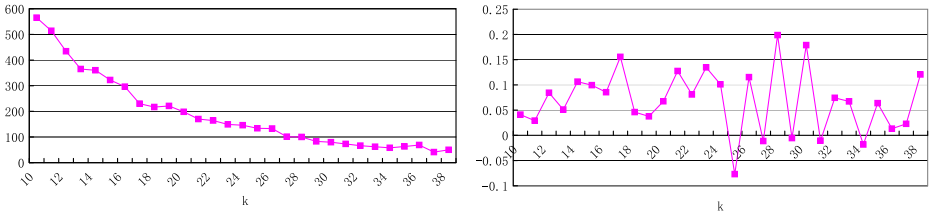
thus it can overcome this problem. Although these two measures are proposed in different ways, we study their correlation in the paper.

Table 9 shows an example of two users who both shared 43 social media. We manually classify these social media into four clusters: "work", "life", "study", "love" and count the number of social media belonging to each cluster, based on which we can calculate the information entropy. It is interesting that our diversity shown in Table 9 is consistent with information entropy.

Furthermore, we analyze the correlation coefficient between information entropy and diversity. First, all the social media are assigned a topic label by the topic model LDA [3]. As our diversity value is correlated to the size of neighborhood, in order to compare the two measures fairly, we select out the users who access the same number of social media. Suppose the number of social media is k. The curve of user numbers changing with k is shown in Fig. 8a. We calculate users' diversity values as well as the information entropy based on their topic distributions. The correlation coefficients of these two measures are calculated as below:

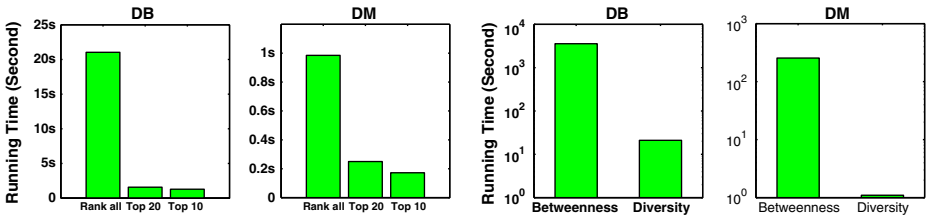
$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}} \tag{6}$$

where x_i and y_i denote the diversity value and information entropy of one user respectively. n is the number of users. Thus when $r_{xy} > 0$, it represents that these two measures are positive correlated; otherwise they are negative correlated. The curve of the correlation coefficients according to each list of users w.r.t. k is shown in Fig. 8b. Therefore the results demonstrate that the user interest diversity obtained by our approach is positive correlated to the information entropy in most cases.



(a) The number of users changing with k (b) The correlation coefficients changing with k

Fig. 8 The comparison between diversity and information entropy



(a) Top-K algorithm comparison (b) Betweenness VS. Diversity

Fig. 9 Performance comparison

5.3 Performance comparison

Figure 9a compares the running time of Top-K algorithm with the time of ranking all the nodes on DB and DM networks, which contain thousands of nodes. The algorithms were run on an Intel Core 2 T7200 (2 cores, 2 GHZ) processor with 2GB DDR2 RAM. The results demonstrate that more than 50% nodes can be pruned by Top-K diversity ranking algorithm, which is much more efficient and can meet online query needs. We also implemented an efficient betweenness algorithm [4] and compared it with diversity. Figure 9b 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.

6 Discussion

As diversity is a highly subjective concept defying any optimal definition applicable for all scenarios, we propose two basic principles which should convey the intrinsic rules behind any reasonable diversity definition. Guided by these principles, we studied one such definition in detail to illustrate our proposed concept.

However, rather than narrowing ourselves down to one specific definition, we are fully aware of other possible definitions that may better geared for other applications. For example, a highly intuitive definition can be based on clustering, where the network is first assigned labels by certain clustering algorithm and diversity is then computed by checking the pre-computed cluster labels of the neighbors. This kind of definition needs to at least solve the following issues:

- 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.
- A node may connect to a varied number of nodes in one cluster, which indicates the connection strength between the node and the cluster. It is also an important factor of diversity which should be taken into consideration.
- The internal cohesion of clusters, which reflects network topology, is also an important component for diversity.

Therefore, still lots of aspects and factors should be exploited for the clustering-based definition. Besides, the link weights are also important components for diversity computation. But we aim at illustrating the main idea of diversity in this paper. Thus we neglect this factor and would incorporate it into the computation in our future work.

We should also notice that similarity definition is an essential factor on diversity computation, as it directly determines how closely the neighbors connect with each other. Previous work proposed various similarity definitions which are suitable for different situations. In this paper, similarity definition is not our emphasis, thus we chose a simple one as a starting point. For other kinds of diversity definition, different similarity function could be exploited.

7 Conclusion

With the prevalence of social media websites, social network data emerge in abundance consequentially. In this paper, we investigated the problem of mining diversity on social media 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.

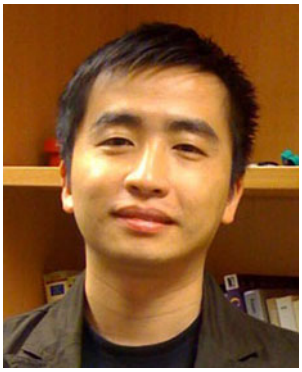
Acknowledgements The work was supported in part by the U.S. National Science Foundation grants IIS-08-42769 and IIS-09-05215, and the NASA grant NNX08AC35A, and 973 Program of China grant 2006CB303103, and the State Key Program of National Natural Science of China grant 60933013. Any opinions, findings, 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 AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization. Nat Rev Genet 5(2):101–113
3. Blei D, Ng A, Jordan M (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
4. Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25:163–177
5. Cover TM, Thomas JA (2006) Elements of information theory. Wiley
6. Faloutsos M, Faloutsos P, Faloutsos C (1999) On power-law relationships of the internet topology. In: SIGCOMM, pp 251–262
7. Gibson D, Kumar R, Tomkins A (2005) Discovering large dense subgraphs in massive graphs. In: VLDB, pp 721–732
8. Girvan M, Newman MEJ (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826
9. Hwang W, Kim T, Ramanathan M, Zhang A (2008) Bridging centrality: graph mining from element level to group level. In: KDD, pp 336–344
10. Jeh G, Widom J (2002) Simrank: a measure of structural-context similarity. In: KDD, pp 538–543
11. Kempe D, Kleinberg J, Tardos E (2003) Maximizing the spread of influence through a social network. In: KDD, pp 137–146

12. Kossinets G, Kleinberg JM, Watts DJ (2008) The structure of information pathways in a social communication network. In: KDD, pp 435–443
13. Kuramochi M, Karypis G (2001) Frequent subgraph discovery. In: ICDM, pp 313–320
14. Lawrence P, Sergey B, Motwani R, Winograd T (1998) The pagerank citation ranking: bringing order to the web. Technical report, Stanford University
15. Leskovec J, Kleinberg JM, Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In: KDD, pp 177–187
16. Ng AY, Jordan MI, Weiss Y (2001) On spectral clustering: analysis and an algorithm. In: NIPS, pp 849–856
17. Pei J, Jiang D, Zhang A (2005) On mining cross-graph quasi-cliques. In: KDD, pp 228–238
18. Putnam RD (1995) Bowling alone: America's declining social capital. J Democr 6(1):65–78
19. Spirin V, Mirny LA (2003) Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci USA 100(21):12123–12128
20. Stephenson K, Zelen M (1989) Rethinking centrality: methods and examples. Soc Netw 11(1): 1–37
21. Sun Y, Han J, Zhao P, Yin Z, Cheng H, Wu T (2009) Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In: EDBT, pp 565–576
22. Wasserman S, Faust K (1994) Social network analysis, methods and applications. Cambridge University Press
23. Watts DJ, Strogatz SH (1998) Collective dynamics of 'small-world' networks. Nature 393(6684):440–442
24. Yan X, Han J (2002) gSpan: graph-based substructure pattern mining. In: ICDM, pp 721–724
25. Zvi MR, Griffiths T, Steyvers M, Smyth P (2004) The author-topic model for authors and documents. In: UAI, pp 487–494

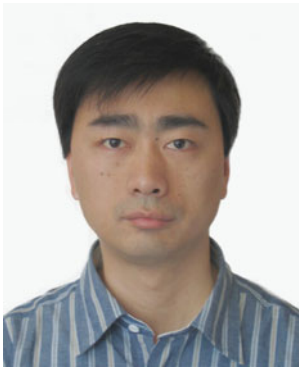
Lu Liu received her B.E. Degree and Ph.D. Degree in the Department of Computer Science and Technology of Tsinghua University in 2005 and 2010 respectively. She is now an Assistant Professor at Capital Medical University, Beijing, China. Her research interests include multimedia analysis, information retrieval, social network mining, generative graphical model, and etc.



Feida Zhu is currently an Assistant Professor at the School of Information Systems in Singapore Management University. His research interests are data mining, web mining, algorithms and complexity analysis for data mining and database problems. He got his Ph.D. in Computer Science from University of Illinois at Urbana-Champaign under the supervision of Dr. Jiawei Han in 2009. During his Ph.D. study, he has won two Best Student Paper Awards from ICDE (International Conference on Data Engineering Conference) 2007 and PAKDD (The Pacific-Asia Conference on Knowledge Discovery and Data Mining) 2007 respectively.

Meng Jiang was born in 1989. He received the B.S. degree in Computer Science and Technology from Tsinghua University in 2010. He is currently pursuing his M.S. and Ph.D. degrees in Computer Science and Technology in Tsinghua University, with his research interests in data mining on social web, etc.

Jiawei Han is a Professor in the Department of Computer Science at the University of Illinois. He has served on many program committees of the major international conferences in the fields of data mining and database systems, and also served or is serving on the editorial boards for Data Mining and Knowledge Discovery, IEEE Transactions on Knowledge and Data Engineering, Journal of Computer Science and Technology, and Journal of Intelligent Information Systems. He is the founding Editor-in-Chief of ACM Transactions on Knowledge Discovery from Data (TKDD). Jiawei has received IBM Faculty Awards, HP Innovation Award, the Outstanding Contribution Award at the International Conference on Data Mining (2002), ACM SIGKDD Innovation Award (2004), IEEE Computer Society Technical Achievement Award (2005), and IEEE CS W. Wallace McDowell Award (2009). He is a Fellow of ACM and IEEE. He is currently the Director of Information Network Academic Research Center (INARC) supported by the Network Science-Collaborative Technology Alliance (NS-CTA) program of U.S. Army Research Lab. His book "Data Mining: Concepts and Techniques" (Morgan Kaufmann) has been used worldwide as a textbook.



Lifeng Sun received his B.S. Degree and Ph.D. Degrees in System Engineering in 1995 and 2000 separately from National University of Defense Technology, Changsha, Hunan, China. He is now an associated professor of the Department of Computer Science and Technology at Tsinghua University. Dr. Sun's professional interests lie in the areas of interactive multi-view video, video sensor network, peer-to-peer streaming, distributed video coding.

Shiqiang Yang graduated from the Department of Computer Science and Technology, Tsinghua University in 1977 and received the M.E. degree in 1983. He is now a professor at Tsinghua University. His research interests include multimedia technology and systems, video compression and streaming, content-based retrieval and semantics for multimedia information, embedded multimedia systems. He has published more than 100 papers in the international conference and journals. Mr. Yang is currently the President of the Multimedia Committee of the China Computer Federation. He is a senior member of IEEE.