Physica A 390 (2011) 4627–4635

Contents lists available at SciVerse ScienceDirect

Physica A

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

Diversity of social ties in scientific collaboration networks

Quan Shi [a,c], Bo Xu [b], Xiaomin Xu [b], Yanghua Xiao [b,*], Wei Wang [b], Hengshan Wang [c]

[a] School of Computer Science and Technology, Nantong University, Nantong 226019, PR China
[b] School of Computer Science, Fudan University, Shanghai 201203, PR China
[c] School of Management, University of Shanghai for Science & Technology, Shanghai 200090, PR China

ARTICLE INFO

Article history:
Received 26 November 2010
Received in revised form 23 June 2011
Available online 8 July 2011

Keywords:
Social network
Diversity
DBLP
Scientific collaboration network

ABSTRACT

Diversity is one of the important perspectives to characterize behaviors of individuals in social networks. It is intuitively believed that diversity of social ties accounts for competition advantage and idea innovation. However, quantitative evidences in a real large social network can be rarely found in the previous research. Thanks to the availability of scientific publication records on WWW; now we can construct a large scientific collaboration network, which provides us a chance to gain insight into the diversity of relationships in a real social network through statistical analysis. In this article, we dedicate our efforts to perform empirical analysis on a scientific collaboration network extracted from DBLP, an online bibliographic database in computer science, in a systematical way, finding the following: distributions of diversity indices tend to decay in an exponential or Gaussian way; diversity indices are not trivially correlated to existing vertex importance measures; authors of diverse social ties tend to connect to each other and these authors are generally more competitive than others.

1. Introduction

Social networks form the backbone of our social and economic life [1]. With more and more social network data being available to researchers, social network analysis (SNA) has become a valuable tool for us to understand the social structure as well as the behaviors of individuals [2–12]. Among various SNA tasks, characterizing the *diversity* of individuals' social ties in a social network is attracting increasing interest both in academic and industries.

Intuitively, an individual's social ties are *diverse* if he or she maintains connections to different communities or groups. More heterogeneous the linking targets are, more diverse the individual's social ties are. For example, in the hypothetical social network shown in Fig. 1, user *A*'s links are more diverse than user *B*, since user *A* is connected to three different groups while user *B*'s connections come from the same group. As a result, user *A* generally has more chance to collect information or synthesize idea from three diverse groups than others. By serving as an intermediary with a position bridging different groups in a social network, user *A* becomes more competitive or owns more opportunities than others in the network.

Recently, diversity has attracted increasing research interests. It was shown that heterogeneous social ties may generate more opportunities from a range of diverse contacts and are responsible for innovations of ideas or techniques [3,13]. A quite recent research showed that the diversity of individuals' relationships is strongly correlated with the economic development of communities [1], suggesting that diversity is a potential powerful measure of community development.

Diversity analysis in social networks also promises to be valuable in a variety of real applications. For example, ranking diversities of users in an online social network such as Facebook[1] LinkedIn[2] can help us to identify those users that will bring

* Corresponding author.
E-mail address: shawyh@fudan.edu.cn (Y. Xiao).

1 http://www.facebook.com.

2 http://www.linkedin.com.

0378-4371/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2011.06.072

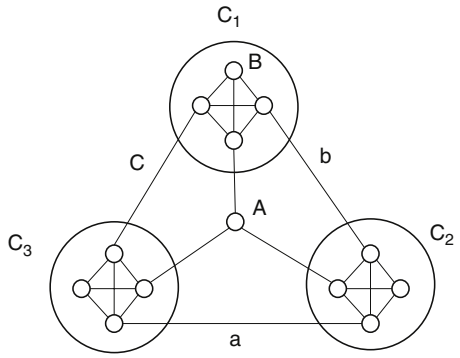


Fig. 1. Illustration of diversity in a hypothetical social network. It is evident that there exist three communities C_1, C_2, C_3 in the network.

us opportunities of new jobs. As a vertex ranking mechanism, diversity also can be incorporated into other ranking mechanisms so that we can rank people in a social network appropriately, which is the underlying technique to search persons from online social networks.

However, up until now, little is known about the statistical properties of diversity in a social network. Many interesting problems about diversity in social networks remain to be unsolved. For example, what is the distribution of diversity? Is there any correspondence between an individual's diversity and his/her status in the social network? Is diversity trivially correlated to the existing measures of vertex? Are the individuals of significantly diverse social ties tend to connect to each other? Are the individuals with diverse social ties more competitive than others? Answers to these questions are the foundations to utilize diversity analysis in real applications.

Scientific collaboration networks (SCN) [14,15] with vertex representing scientists and edges representing collaborations among them are typical social networks. Many publication datasets available on WWW, such as DBLP (Digital Bibliography & Library Project),[3] provide us a chance to extract an SCN and explore diversity-related properties of this network. In this article, we construct a large SCN with hundreds of thousands of authors from the public DBLP dataset. We systematically investigate the properties of diversity in this SCN through empirical analysis. Our major findings include the following.

1. Diversity indices tend to decay in an exponential or Gaussian way due to the high costs of maintaining connections to authors of different communities.
2. Diversity is a novel perspective to characterize vertex importance and cannot be replaced by existing vertex importance measures.
3. Authors of diverse social ties tend to connect to each other.
4. Authors of diverse social ties are generally quite competitive.

In the following texts, we first give a short overview about diversity and the dataset that will be used in this paper in Section 2. In Section 3, we present our major results of empirical analysis. Finally, we close the paper with a brief conclusion in Section 4.

2. Diversity of social networks

2.1. Diversity and structural holes

One of the concepts that are closely related to but essentially different from diversity is *structural hole*, which has been extensively studied in the previous research in sociology. A structural hole is a separation between non-redundant contacts [16]. Contacts to the same group or community that are strongly knit are usually regarded as redundant since information can be obtained from any one in the group. A structural hole exists for two groups, if no links exist between two groups. At its heart, structural hole theory argues that individuals benefit when they serve as intermediaries or bridges between others who are not directly connected [16]. As an example, if links a, b, c are removed from the network shown in Fig. 1, structural holes will form among groups C_1, C_2, C_3. By bridging gaps among these groups, user A gains power by brokering the flow of information between different parts of the network.

It is clear to see that when structural holes exist between groups, an individual with diverse links to these different groups is competitive in the sense of bridging more structural holes. However, in many cases where structural holes do not exist, individuals of diverse social ties still have much chance to be successful especially in those less competitive environments. For example, in a scientific collaboration network, researchers of diverse links have more chance to collaborate with researchers from different communities and synthesize ideas from different communities.

[3] http://www.informatik.uni-trier.de/ley/db/.

2.2. SCN extracted from DBLP

DBLP is a computer science bibliography website hosted at Universität Trier, in Germany. It was originally a database and logic programming bibliography site, and has existed at least since the 1980s. DBLP listed more than 1.3 million articles on computer science in January 2010. The DBLP dataset presents information on computer science publications. The dataset used in this paper was derived from a snapshot of the bibliography in May, 2010.

We only keep the publications on conferences since conferences have limited and relatively fixed topics, which can help us accurately associate a research area to each conference. We use the conference classification criteria given in Microsoft Academic Search[4] that is widely accepted in computer science to classify all conferences into 24 major categories. Among all publications, papers on conferences of lower ranks in general are of low quality and can be considered as noises in our study. Hence, among all conferences (overall 2691), we only keep the top-5 conferences of each category. Conference rank given in Microsoft Academic Search is used to filter out conferences of low rank. After preprocessing, we have overall 155 442 papers published on 120 conferences, covering 18.8% of the entire publications (overall 826 506) in the original DBLP dataset. There are overall 146 343 authors and 805 324 coauthoring relationships in the preprocessed network.

Now we can model the scientific collaboration network after preprocessing as a vertex-labeled graph $G(V, E, \ell)$, where ℓ is a labeling function that assigns a unique label of L to each vertex in V. If any two authors v_i, v_j have ever coauthored at least one paper, an edge $e_{ij} = (v_i, v_j)$ is added to E. For simplicity of description, we will use the following notations in subsequent sections. $Neg(v)$: the neighbor set of v; $Inc(v)$: the set of edges that are incident with v; $\ell(Neg(v))$: the set of labels of vertex in $Neg(v)$; $deg(v)$: the degree of a vertex, i.e., the number of incident edges of v.

2.3. Quantifying diversity

A straightforward measure of diversity of an author's relationships is the number of research areas of his or her collaborators in the SCN, as given in Eq. (1). We refer to this measure as *global diversity* (GD), and denote it by $d_G(v)$. Clearly, $d_G(v)$ has an upper bound: $\min\{deg(v), |L|\}$, which is reachable when none of his/her two collaborators have the same label or the maximal number of labels $|L|$ is reached. An author has a larger $d_G(v)$ implies that his collaborations are more diverse

$$d_G(v) = |\ell(Neg(v))|. \tag{1}$$

Note that GD is subject to ℓ, i.e., the way we construct community structure of the network. In general, for an SCN, there may exist multiple choices to specify ℓ. The 24 major categories naturally imply non-overlapping communities, which will be used in this paper. For each author, if he publishes papers on conferences of different categories, we will use the category that the largest number of his publications belong to as his label. As a result, each author is assigned a unique label, which ensures to produce non-overlapping communities. However, it is also possible to specify ℓ by other means. For example, we can use purely the structure information to mine the communities such that the vertex within a community are closely connected and vertex between communities are less closely connected. Then, use the resulting communities to define ℓ.

Linking strategy will be a major concern of an author, when the author has only limited resources to maintain the collaboration relationships with other authors. Some of them may tend to maintain the relationships with authors in the same community, whereas others may tend to diversify his collaborations. To distinguish between authors of different degree of relationship diversity but with the same number of collaborations, we need another diversity measures of authors: *local diversity* (LD), which is defined in Eq. (2) with each p_a being the fraction of v's neighbors with label a among $Neg(v)$. The definition of $d_L(v)$ follows the framework to define entropy. It holds that $d_L(v)$ lies in $[0, upp]$, where $upp = \ln d_G(v)$. When collaborations of v are equally distributed among $d_G(v)$ communities, we have $d_L(v) = upp$. When all collaborations come from the same community, that is $d_G(v) = 1$, the minimal value $d_L(v) = 0$ is reached

$$d_L(v) = -\sum_{a \in \ell(Neg(v))} p_a \ln(p_a). \tag{2}$$

Thus, we can establish relationship between local diversity and global diversity as follows: for any vertex v, we have

$$d_G(v) \geq \exp(d_L(v)). \tag{3}$$

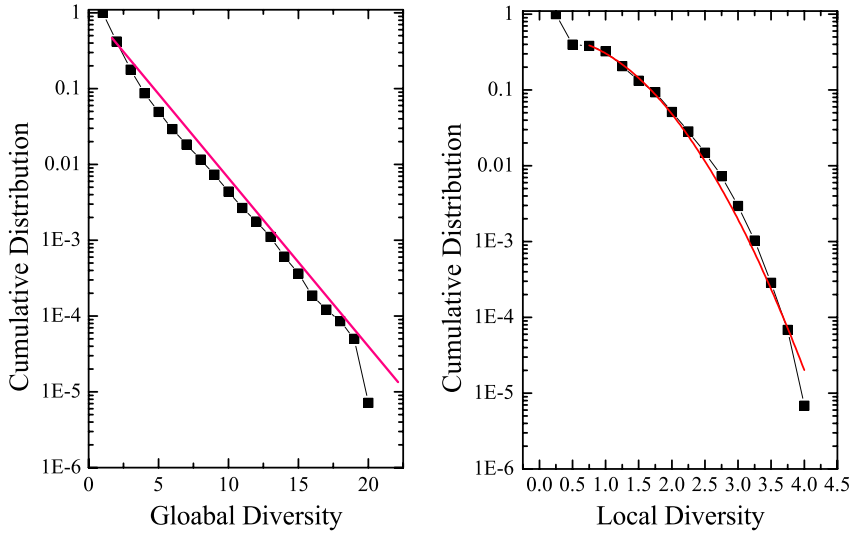
3. Empirical analysis of diversity

In this section, we will present the analysis result of diversity in scientific collaboration network extracted from DBLP. Our evaluation focuses on (a) distribution of global diversity and local diversity, (b) correlations between diversity indices and other vertex importance measures, (c) assortatively mixing by diversity, (d) correlation between diversity and structural holes and (e) top-k analysis by diversity.

3.1. Distribution

The cumulative distributions of global diversity and local diversity are shown in Fig. 2. It is visually apparent that data of global diversity is consistent with an exponential distribution $P(k) \sim \exp(-bk)$ with $b = 0.53 \pm 0.007$ from the linear-log

[4] http://academic.research.microsoft.com/.



(a) Global diversity. (b) Local diversity.

Fig. 2. (Color online) Cumulative distributions of global diversity and local diversity. (a) shows the linear-log plot of global diversity distribution. The full red line is the fit to the cumulative distribution of an exponential distribution $P(k) \sim \exp(-bk)$ with $b = 0.53 \pm 0.007$, where k is the global diversity. (b) shows the linear-log plot of the local diversity distribution. The full red line is the fit to the cumulative distribution of a Gaussian distribution $P(k) \sim \exp(-2(\frac{k-\mu}{\omega})^2)$ with $\mu = -0.15 \pm 0.196$ and $\omega = 1.71 \pm 0.16$.

plot of Fig. 2(a), where $P(k)$ is the probability that an author has GD as k. Although majority of authors have small GDs (for example, 58% of authors have $d_G(v) = 1$ and 24% of authors have $d_G = 2$), there still exist authors of high global diversity. There are overall 611 authors with global diversity ≥ 10, covering about 0.4% of all authors. Among them, one author has the maximum value $d_G(v) = 20$ (We will further discuss this author in Section 3.5).

The data of global diversity distribution seems to fall faster in the tail than they would for an exponential decay, suggesting a Gaussian decay in the tail. Compared to power-law decay, exponential decay or even faster decay in the tail of SCN decays much faster, sufficiently suggesting that *diversity of an author's relationships is limited by the high costs of maintaining diverse relationships*. Physical costs of adding links have successfully explained the cut-off of power-law distribution in an airline transportation network [17]. Similar to airline networks, it will cost much time and resources for authors in SCN to maintain relationships to authors in different communities. Hence, it is quite impossible for an author to maintain connections to too many authors in different domains. In general, when $d_G(v)$ exceeds a threshold ξ, a faster decay, such as Gaussian decay, will be significant. In SCN, ξ is approximately to be 19, which can be visually observed from Fig. 2(a). Hence, in general distribution of global diversity can be described by $P(k) \sim \exp(-bk)f(k/\xi)$, where ξ is the typical size that Gaussian decay becomes significant and $f(k/\xi)$ introduces a sharp cutoff for $k > \xi$.

The data of cumulative distribution of local diversity is consistent with a Gaussian distribution with exceptions when $d_L(v) = 1$, suggesting that it is of quite low possibility for the existence of authors with large local diversity. Clearly, for a fixed number of connections, it is quite difficult for an author to maintain his relationships if they are linked to diverse communities. Hence, the Gaussian decay of local diversity can also be attributed to the high costs of maintaining the connections to authors in different communities.

3.2. Correlation

In this subsection, we will investigate the following question: *Is diversity a novel measure that cannot be replaced by existing measures of vertex in graphs?* To answer this question, we need to summarize the correlation between diversity and existing measures of vertex in a graph. We are especially interested in measures that quantify the importance of individuals in a social network, such as SCN. In general, an author's importance can be measured from different aspects [18]. Among them, the widely used include degree, Page Rank and centrality and betweenness.

Degree quantifies the number of connections of an author, which is a straightforward measure of vertex importance. Page Rank [19] is widely used for ranking web pages that are returned as the answers to a query on search engine. The principle behind Page Rank is that every link is a vote and votes from important webs or authors have significant contributions to the importance of this web or author. Hence, the more links from important authors to an author, the higher Page Rank it is assigned to the author. An author's importance ranked by Page Rank is given by following equation:

$$PR(u) = (1 - d) + d \sum_{v \in Neg(u)} \frac{PR(v)}{deg(v)} \tag{4}$$

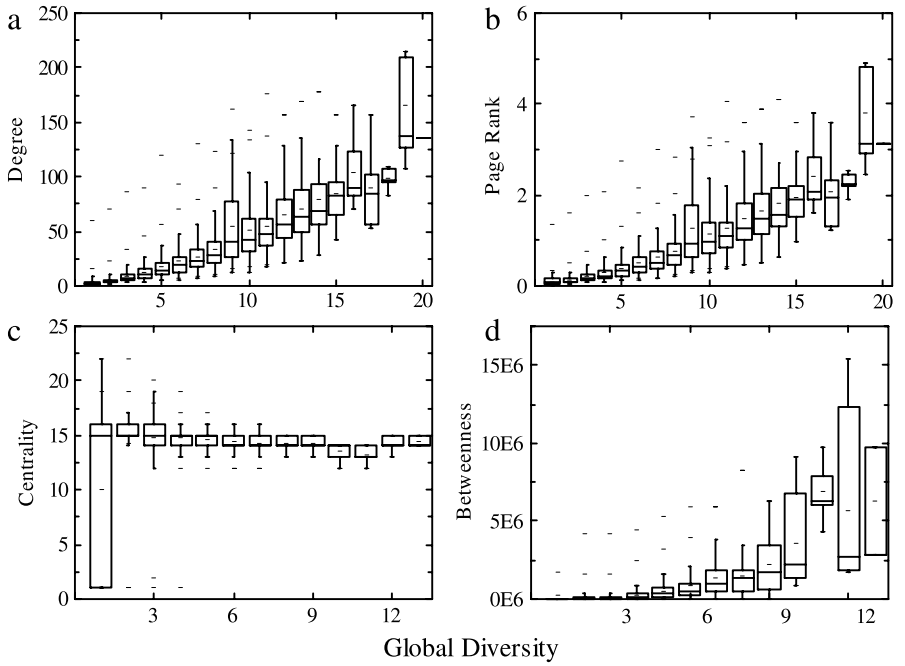


Fig. 3. Correlations between global diversity and other measures of vertex importance. (a) shows the correlation between GD and degree, the correlation coefficient is 0.7162. (b) shows the correlation between GD and Page Rank, the correlation coefficient is 0.6799. (c) shows the correlation between GD and centrality, the correlation coefficient is −0.2834. (d) shows the correlation between GD and betweenness, the correlation coefficient is 0.5296. (c) and (d) are summarized from an SCN of small size due to the unaffordable computation complexity of centrality and betweenness. Specifically, for each category, we only keep the papers published on the top-one conference. Then, we use these papers to construct SCN and summarize all statistics from this simplified SCN. The simplified SCN contains 30 729 papers covering 20% of overall papers. There are overall 39 677 authors and 92 318 links in the simplified SCN.

where $PR(u)$ is the Page Rank of author u and d is a damping factor which can be set between 0 and 1. The computation of Page Rank simulates the process that an imaginary surfer randomly clicks on links and eventually stops clicking. Damping factor d is the probability that an imaginary surfer will continue clicking a link. It is generally assumed that the damping factor is set around 0.85.

Centrality measures the importance of vertex by locations of vertex. For example, in SCN, if an author is located in the center of the network, he is quite influential since information can fast propagate from him to other authors. Thus, an author's centrality (denoted by $C_E(u)$) can be quantified by the inverse of the maximum distance from u to any other vertex v in the network, i.e.,

$$C_E(u) = \frac{1}{\max\{dist(u, v) : v \in V\}} \tag{5}$$

where $dist(u, v)$ is the length of a shortest path from u to v. This measure is consistent with our general notion of vertex centrality, since $C_E(u)$ grows if the maximal distance from u to other vertex decreases. An author of maximal $C_E(u)$ is located at the center of the network. $C_E(u)$ is larger implies that u is closer to the center and u is more important.

In general, information or traffic flows along the shortest path in a network. Hence, in SCN, if many shortest paths pass through an author, the author is regarded as important. Betweenness [20] measures the importance of an author by the number of shortest paths passing through him. For a graph $G = (V, E)$ with n vertices, the betweenness ($C_B(v)$) of $v \in V$ given by

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{6}$$

where σ_{st} is the number of shortest paths between s and t, and $\sigma_{st}(v)$ is the number of shortest paths passing through v among σ_{st}.

The correlations between global diversity and above important measures are shown in Fig. 3 (where box charts are plotted). It is apparent from the figure that the correlations between global diversity and the tested vertex importance measures are weak. In general, authors of the same global diversity may exhibit diverse behaviors under the measurement of other measures, as can be seen from the large variance of the box chart under each possible value of global diversity. This fact suggests that diversity is a novel perspective to measure vertex and the information captured by diversity in general cannot be expressed by existing measures. Note that, it seems that an author of large GD tends to have a large minimal

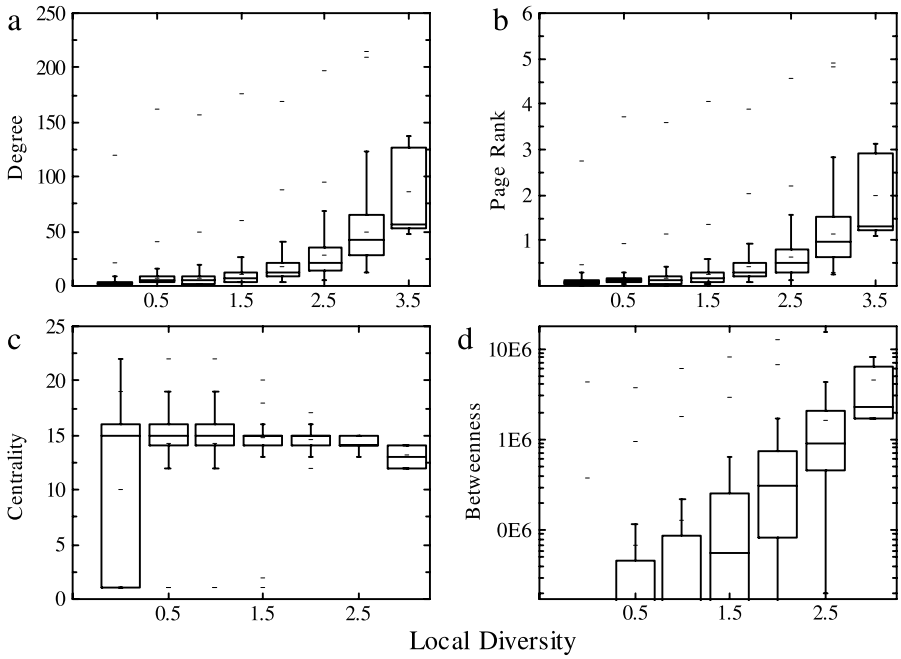


Fig. 4. Correlations between local diversity and other measures of vertex importance. (a) shows the correlation between LD and degree, the correlation coefficient is 0.4351. (b) shows the correlation between LD and Page Rank, the correlation coefficient is 0.3816. (c) shows the correlation between LD and centrality, the correlation coefficient is −0.32488. (d) shows the correlation between LD and betweenness, the correlation coefficient is 0.32448. (c) and (d) are also summarized from the simplified SCN the same as that used in Fig. 3.

degree and Page Rank, which can be attributed to the fact that $deg(v) \geq d_G(v)$ and degree is positively correlated to Page Rank to a certain degree.

Similar to global diversity, local diversity shows no correlations to tested vertex importance measures. The results are shown in Fig. 4 in the form of box chart. From the figure, we can see that authors of the same value quantified by either one of degree, Page Rank, centrality and betweenness exhibit diverse behaviors when measured by LD, suggesting that LD is another perspective to characterize individual's behavior in a social network.

3.3. Assortative mixing

In a typical social network, large degree individuals preferentially attach to large degree individuals, which is known as *assortatively mixing by degree* [21]. Another property that is closely related to assortative mixing and is also shared across social networks is *rich club* [11,12,22,23], i.e., the individuals of rich connections are well connected to each other. It was shown that scientific collaboration networks generally have rich clubs [22,23].

Since diversity is the major concern of this paper, we may wonder whether authors in SCN are *assortatively mixed by diversity*, which is equivalent to answer the following question: *Are authors of diverse social ties tend to connect to each other?* To address this issue, we will summarize the statistics about average GD of the nearest neighbors of a vertex with GD as k, which is used to measure GD–GD correlation in SCN. Since global diversity and local diversity display similar behaviors, in the following analysis, we will only focus on global diversity.

We follow the form of average degree of nearest neighbors of a k-degree vertex [24] to define average GD of the nearest neighbors of a vertex with GD as k, which is defined as

$$\bar{k}_{nn}(k) = \sum_i k_i \frac{E_{kk_i}}{kN_k} \tag{7}$$

where k_i is the GD of vertex i, N_k is the number of vertex with GD as k, E_{kk_i} counts the connections between vertex i and vertex with GD as k. In general, if authors with large GD have large $\bar{k}_{nn}(k)$, these authors tend to connect to each other.

The behavior of $\bar{k}_{nn}(k)$ is shown in Fig. 5. The result confirms that authors of large GD tend to coauthor with each other. $\bar{k}_{nn}(k)$ generally linearly increases with the growth of GD with some exceptions when GD $= 9$. We will have a closer look at these exceptions in the following subsection.

3.4. Structural holes

It was shown that in a social network, an individual of many structural holes is more competitive than others [16]. Thus, we may wonder *whether authors of diverse social ties are more competitive than others?* One way to explore this problem is to

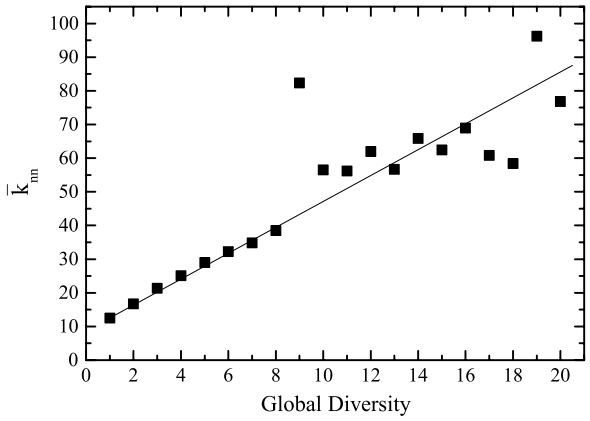


Fig. 5. The dependence on global diversity of $\bar{k}_{nn}(k)$. The correlation coefficient is 0.87431.

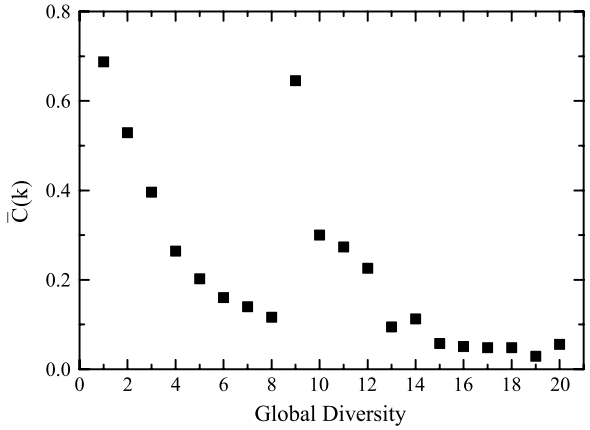


Fig. 6. The dependence on global diversity of $\bar{C}(k)$.

examine the correlation between diversity and abundance of structural holes. For this purpose, we summarize the average clustering coefficient for a vertex with GD as k, which is defined as

$$\bar{C}(k) = 2\frac{\sum\limits_{i\in V_k} E_i}{\sum\limits_{i\in V_k} k_i(k_i-1)} \tag{8}$$

where E_i sums up the number of edges between the neighbors of vertex i, V_k is the set of vertex with GD as k, and k_i is the degree of i. In general, if authors of large GD tend to have smaller $\bar{C}(k)$, more structural holes can be found from these authors' neighborhoods and we can confidently come to the conclusion: *diverse social ties lead to competition advantages*.

The result about $\bar{C}(k)$ is shown in Fig. 6. It is evident from the figure that in general $\bar{C}(k)$ decreases with the growth of GD with a few exceptions when GD $= 9, 10, 11, 12$. Authors of larger GD (larger than 12) generally tend to have smaller $\bar{C}(k)$, suggesting that abundant structural holes exit in their neighborhoods. Superlinear decrease of $\bar{C}(k)$ also implies that structural holes are quite sensitive to the increase of global diversity.

Now we have a closer look at the exceptions when GD $= 9, 10, 11, 12$, especially when GD $= 9$. These exceptions can be attributed to the close connections among authors with these GD values. Note that Fig. 5 has shown that authors of GD $= 9$ have an extremely large $\bar{k}_{nn}(k)$, suggesting that they tend to coauthor with each other. To provide more evidence, we also summarize the distribution of coauthoring relationships, which is shown in Fig. 7. It is obvious that there exist significant number of edges among authors with GD varying from 9 to 12.

Furthermore, we give some statistics about authors of GD $= 9$ to gain deep insights into behaviors of authors of large diversity. We denote the set of all authors of GD $= 9$ by V_9. There are overall 417 authors in V_9, and 3754 links among them. The average degree of subgraph induced by V_9 is 18, which is significantly larger than the average degree of the entire network, that is 5.76. On average, for any author v in V_9, it is expected that there exist 8[5] links among v's neighbors in V_9.

[5] Let $G[V_9]$ be the induced subgraph of V_9. A spanning tree of $G[V_9]$ consumes 416 edges. Then remaining $3754-416 = 3338$ edges need to be distributed in the neighbors of authors. Hence, for an author $v \in V_9$ there exist $3338/417 \approx 8$ edges linking v's neighbors in V_9 on average.

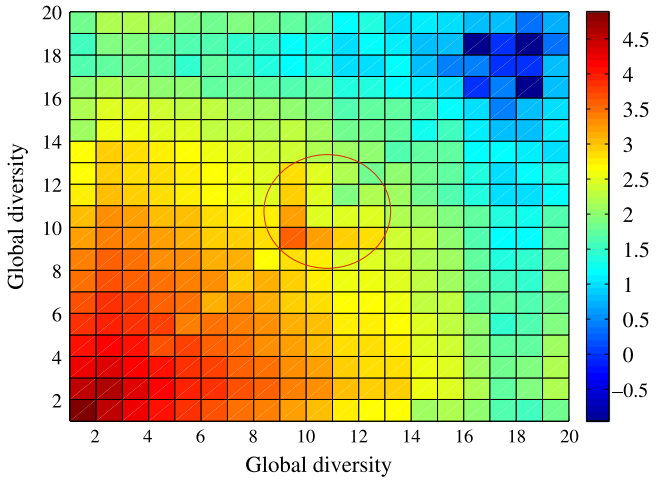


Fig. 7. (Color online) The color representation of edge distribution matrix. Each element of the matrix is log μ_{ij}, where μ_{ij} is the number of edges with two ends representing authors with GD as i, j, respectively. Color runs from blue for minimal μ_{ij} to red for maximal μ_{ij}. The region marked by the circle implies that authors of GD $= 9, 10, 11, 12$ tend to coauthor with each other.

Table 1
Author list of top-10 global diversity. As a comparison, degree of authors and corresponding ranks are also given.

Author name	GD rank	GD	Degree rank	Degree
Wei Li	1	20	23	136
Jiawei Han	2	19	1	214
Philip S. Yu	2	19	2	209
Christos Faloutsos	2	19	3	198
Xin Li	2	19	21	137
Ming Li	2	19	38	127
H.V. Jagadish	2	19	161	107
Li Zhang	8	18	156	110
Michael I. Jordan	8	18	161	107
Elke A. Rundensteiner	8	18	182	97

Hence, neighbors of an author in V_9 are expected to be closely connected. Consequently, for an author in V_9, structural holes tend to miss in the author's neighborhoods.

3.5. Top-k analysis

Finally, we will provide more evidence to show that authors of diverse social ties are really competitive through top-k analysis. Table 1 gives the authors of top-10 global diversity.

A quick observation on the table indicates that the top-10 author list include some very prolific authors. For example, Jiawei Han, one of the most prestigious authors in data mining, is a professor of department of computer science at university of Illinois at Urbana-Champaign. Han has published 560 publications in top conferences and journals, and his publications are cited more than 15 000. From 1973 to 2010, he has collaborated with 448 authors. He is ranked as the top-one expert in data mining by Arnetminer.[6] Due to his contribution, he won the 2009 Wallace McDowell Award. Similar to Han, most of other authors in the top-10 list, such as Philip Yu, Jagadish, Christos Faloutsos, Michael Jordan, and Elke Rundensteiner are also prestigious authors in their respective research fields.

From Table 1, we can also find some outliers. They are not very famous, but their global diversity is very high. These outliers can be attributed to the fact that there are many authors with identical names and these authors cannot be easily distinguished from each other in DBLP dataset. For example, 'Wei Li', the one with the largest GD actually are 73 authors with the same name.[7] These authors come from diverse research fields. Hence, if we regard them as one person, this person will have quite large global diversity. Identifying duplicate names is still a challenging problem. However, it seems that diversity provides us a new alternative to solve this problem. From the result of the above case analysis and fast decay of diversity distribution, one truth is self-evident: *it is of quite low probability for a single person to own a high diversity value*, which may be employed to help us accurately identify duplicate names.

6 http://arnetminer.org/, an online analysis platform for DBLP data.

7 This data is available from Arnetminer.

4. Conclusion

In this article, we have systematically investigate diversity of author's relationships in a real SCN through statistical analysis. Our quantitative results confirm that diverse ties can help an author own more competition advantage. We also find authors of large diversity tend to connect to each other. Our results also suggest that diversity maybe an important ranking mechanism that has been ignored in the previous research.

As a new perspective to explore behaviors of individuals in social networks, research about diversity is still quite preliminary and can be extended from many aspects, including empirical analysis of diversity in typical online social networks, building network models that produce desired diversity behavior, and so on.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under grant Nos 61003001, 71071098 and 60703093; Specialized Research Fund for the Doctoral Program of Higher Education No. 20100071120032; the Natural Science Foundation of Jiangsu Province (BK2009153, BK2010280); Key Program of National Natural Science Foundation of China under grant No. 61033010; National Science and Technology Major Project of the Ministry of Science and Technology of China under grant No. 2010ZX01042-003-004.

References

[1] N. Eagle, M. Macy, R. Claxton, Science 328 (2010) 1029.
[2] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann et al., Science 323 (2009) 721.
[3] M.E.J. Newman, SIAM Rev. 45 (2003) 167.
[4] J. Scott, Social Network Analysis: A Handbook, SAGE Publications, ISBN: 9780761963394, 2000.
[5] S. Wasserman, K. Faust, Social network analysis: methods and applications, in: Structural Analysis in the Social Sciences, Cambridge University Press, ISBN: 9780521387071, 1995.
[6] G. Kossinets, D.J. Watts, Science 311 (2006) 88.
[7] D. Centola, Science 329 (2010) 1194.
[8] J.H. Fowler, N.A. Christakis, Proc. Natl. Acad. Sci. 107 (2010).
[9] J. Zhao, J. Wu, K. Xu, Phys. Rev. E 82 (2010) 016105.
[10] N. Masuda, Y. Kawamura, H. Kori, Phys. Rev. E 80 (2009) 046114.
[11] S. Zhou, R.J. Mondragón, Commun. Lett., IEEE 8 (2003).
[12] N. Masuda, N. Konno, Soci. Netw. 28 (2005) 15.
[13] S.E. Page, The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies, Princeton University Press, ISBN: 0691128383, 2007.
[14] M.E.J. Newman, Phys. Rev. E 64 (2001) 016131.
[15] M.E.J. Newman, Phys. Rev. E 64 (2001) 016132.
[16] R. Burt, Structural Holes: The Social Structure of Competition, Harvard University Press, ISBN: 9780674843714, 1995.
[17] L.A.N. Amaral, A. Scala, M. Barthélémy, H.E. Stanley, Proc. Natl. Acad. Sci. (2000).
[18] P. Csermely, Trends in Biochemical Sciences 33 (2008) 10.
[19] L. Page, S. Brin, R. Motwani, T. Winograd, Proceedings of the 7th International World Wide Web Conference, Brisbane, Australia, 1998, pp. 161–172.
[20] U. Brandes, J. Math. Sociol. 25 (2001) 163.
[21] M.E.J. Newman, Phys. Rev. Lett. 89 (2002) 208701.
[22] V. Colizza, A. Flammini, M.A. Serrano, A. Vespignani, Nat. Phys. 2 (2006) 110.
[23] T. Opsahl, V. Colizza, P. Panzarasa, J.J. Ramasco, Phys. Rev. Lett. 101 (2008) 168702.
[24] M. Boguñá, R. Pastor-Satorras, Phys. Rev. E 68 (2003) 036112.