AAAS
Research
Volume 2021, Article ID 9831621, 10 pages
https://doi.org/10.34133/2021/9831621

Research
A SCIENCE PARTNER JOURNAL

*Research Article*

# The Strength of Structural Diversity in Online Social Networks

**Yafei Zhang,[1,2] Lin Wang,[1] Jonathan J. H. Zhu,[2] Xiaofan Wang,[1,3] and Alex 'Sandy' Pentland[4]**

[1]*Department of Automation, Shanghai Jiao Tong University and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China*
[2]*Department of Media and Communication and School of Data Science, City University of Hong Kong, Hong Kong S.A.R., China*
[3]*Department of Automation, Shanghai University, Shanghai 200444, China*
[4]*Media Lab, Massachusetts Institute of Technology, Cambridge, MA 02139, USA*

Correspondence should be addressed to Jonathan J. H. Zhu; j.zhu@cityu.edu.hk and Xiaofan Wang; xfwang@sjtu.edu.cn

Received 20 January 2021; Accepted 29 April 2021; Published 26 May 2021



Understanding the way individuals are interconnected in social networks is of prime significance to predict their collective outcomes. Leveraging a large-scale dataset from a knowledge-sharing website, this paper presents an exploratory investigation of the way to depict structural diversity in directed networks and how it can be utilized to predict one's online social reputation. To capture the structural diversity of an individual, we first consider the number of weakly and strongly connected components in one's contact neighborhood and further take the coexposure network of social neighbors into consideration. We show empirical evidence that the structural diversity of an individual is able to provide valuable insights to predict personal online social reputation, and the inclusion of a coexposure network provides an additional ingredient to achieve that goal. After synthetically controlling several possible confounding factors through matching experiments, structural diversity still plays a nonnegligible role in the prediction of personal online social reputation. Our work constitutes one of the first attempts to empirically study structural diversity in directed networks and has practical implications for a range of domains, such as social influence and collective intelligence studies.

## 1. Introduction

Recent years have witnessed the emergence and rapid proliferation of many social applications and media platforms. As the backbone of so many online social systems, network structure is becoming a complex and subtle force that drives the dynamics of a wide variety of social processes. In some cases, we seek to leverage social networks to maintain social capital, encourage the adoption of new products, or promote positive behaviors like cooperation and physical exercise [1–10], while in others to eliminate the spread of infectious diseases and fake news or change negative behaviors like conflict and unhealthy eating [11–15].

A wealth of studies suggests that the socioeconomic characteristics of individuals or communities are closely related to their network locations [16–23]. As evidenced in the literature, connected individuals generally show similar patterns in terms of diverse social interests and activities [24–28]. However, the information redundancy and prevalence of similarity in one's social realm may limit his/her potential of exposure to diverse information and interaction with people from different backgrounds, thereby reducing the efficiency of social networks, preventing the diffusion of innovative ideas, weakening the power of social influence and undermining the wisdom of crowds [29–33].

The recent availability of vast and fine-grained data of human activities in online social networks provides an unprecedented opportunity to investigate the nuanced or subtle social effects induced by social context diversity. Structural diversity, which is aimed at quantifying the diversity of one's social context from the view of component count among social neighbors, has been shown useful in predicting specific social processes in undirected networks [2, 6], while for directed networks, the strength of structural diversity still remains an open question [34]. In directed networks, the network connectivity patterns

become more complicated, and network ties may function quite differently according to the direction [28, 35]. Moreover, exposure to similar users could have a better chance to result in the similarity between two individuals even without an explicit social tie between them [36], which is largely overlooked by previous studies on structural diversity.

Online social reputation is the consensus public opinion of an individual or entity based on the ratings from members in a social network [37–39]. It helps to foster trust among online users and is one of the most valuable assets in our online social lives [40, 41]. Using data from an online knowledge-sharing platform, this paper addresses the problem of how to quantify the structural diversity of an individual in directed networks and the role of it in empirically predicting personal social reputation. We first consider the weak and strong connectivity patterns among one's social contacts and find that personal social reputation is positively correlated with the number of weakly or strongly connected components in one's contact neighborhood. Conditional on how many followers one has, individuals whose followers come from more diverse social backgrounds tend to have higher social reputations on the platform. Regression analysis indicates that structural diversity measures via weak and strong connectivity among one's neighbors yield better predictions of personal social reputation than the number of followers one has. We further take the coexposure network of one's social neighbors into consideration, which goes beyond the sheer number of social contacts or connected social components in one's contact neighborhood and provides an additional ingredient to predict personal social reputation.

To eliminate the effects of confounders from structural diversity in the prediction of personal social reputation, we conduct a series of matching experiments. After synthetically controlling possible confounding factors, we present compelling evidence that for individuals with an equal number of followers, same gender, and similar activity-related patterns, those whose followers come from more diverse social backgrounds (measured by structural diversity) are more likely to have higher social reputations. Our work presents one of the first attempts to empirically study structural diversity in directed networks and demonstrates the potential utility of structural diversity in predicting personal social outcomes and could also shed significant insights into the study of a range of social processes such as the diffusion of innovations, the spread of infectious diseases, and the influence maximization problem.

## 2. Results

*2.1. Network Data and Social Reputation Index.* We collect data from Zhihu, a Chinese knowledge-sharing website which claims to have more than 200 million registered users. On this platform, social ties are created when users choose to follow other accounts. Starting from a randomly selected user, we collect follower and followee lists of 234,834 users in a snowball sampling manner. These 234,834 users are hereafter referred to as *ego* users since we know their complete followers and followees. In addition, we also collect the followee lists of the ego users' followers (user accounts two hops away from the ego users). The whole social network is then constructed based on the explicit social ties between user accounts. In total, the constructed network covers more than 10 million user accounts and 300 million directed social ties. For these ~230,000 ego users, we further collect their popularity data, including how many upvotes, thanks, and favorites they have received, which indicate their social reputations on the platform. Other kinds of informative data are also collected, such as how many questions they have asked and answered, self-reported gender, followed topics, and questions. Note that all the collected data are based on public information on the platform and do not include any users with privacy restrictions.

The distributions of the number of received upvotes, thanks, and favorites are illustrated in Figure 1, where each point in the panel indicates the fraction or relative frequency $P(k)$ of users with a specific quantity which equals $k$. The dashed grey line in each panel shows the power-law fitting [42, 43] for each distribution, where the power-law exponents are 1.342, 1.473, and 1.400, respectively. As the distributions span several orders of magnitude, the inequality of personal popularity is very striking. Considering the fact that these three popularity measures are highly correlated and very sparse, we adopt nonnegative matrix factorization (NMF) [44–47]—a widely used dimensionality reduction technique—to collapse them into a single measure which we term the *social reputation index* (see Methods for details on the construction of the social reputation index).

*2.2. Weak and Strong Connectivity.* To quantify the structural diversity of a given node in directed networks, we first consider the weak and strong connectivity between the neighboring nodes. Note that for any two nodes $u$ and $v$ in a directed network, $u$ and $v$ are said to be weakly connected as long as there is a path linking $u$ and $v$ regardless of the direction of the path and strongly connected if and only if there is at least a directed path from $u$ to $v$ as well as a directed path from $v$ to $u$. Therefore, according to the weak or strong connectivity patterns between nodes, a directed network can be decomposed into several social components.

There may exist multiple approaches to quantify the structural diversity of individuals, but the most simple and straightforward way would be the number of connected social components in one's contact neighborhood [2, 34]. Taking networks with three nodes as examples, Figure 2(a) shows how directed networks are projected into different numbers of social components based on the weak or strong connectivity patterns. For a constructed ego network with the ego node/user (we use the terms node and user interchangeably throughout the paper) located at the hub of the wheel, Figure 2(b) further illustrates how the corresponding structural diversity measures are computed compared with indegree (number of followers). For the given ego user in Figure 2(b), (i) indegree is equal to the number of followers, which is 9 in the given example; (ii) weak diversity measure is equal to the number of weakly connected components in

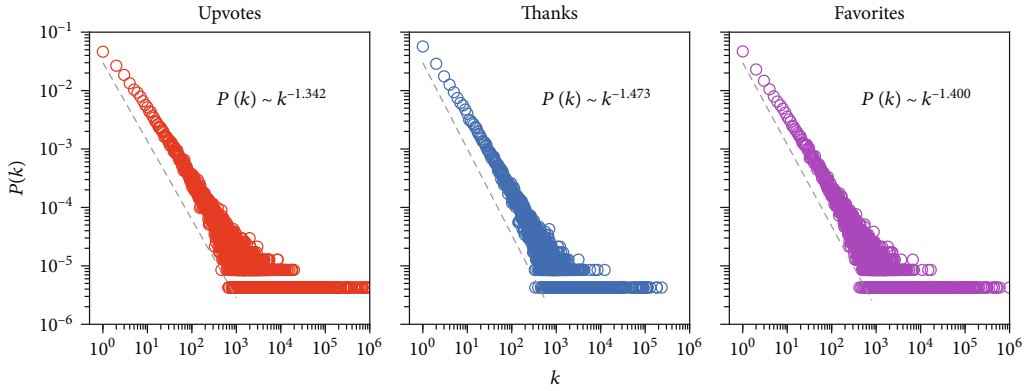


Figure 1: Distributions of the number of received upvotes, thanks, and favorites. For ease of visualization, quantities of zeros are not shown. The dashed grey line in each panel shows the power-law fitting for each distribution.

the contact neighborhood of the ego user, which is 4 in the given example; and (iii) strong diversity measure is equal to the number of strongly connected components in the contact neighborhood, which is 6 in the given example. In fact, we have the property that weak diversity measure ≤ strong diversity measure ≤ indegree for any given ego user. For these ~230,000 ego users, more than 91% of them have less than 20 followers, and only 1,442 of them have more than 1,000 followers (Figure 2(c), c1). As shown in Figure 2(c), the distributions of their weak and strong diversity measures are also very highly skewed. The dashed grey line in each panel shows the power-law fitting [42, 43] for each distribution, where the power-law exponents are 1.641, 1.690, and 1.659, respectively. As expected, the social reputation index increases with all three metrics, with Spearman's rank correlation coefficients as 0.646 for indegree ($p < 0.001$), 0.648 for weak diversity measure ($p < 0.001$), and 0.647 for strong diversity measure ($p < 0.001$).

With a closer look at individuals with an equal number of followers (indegree), we find that the social reputation index generally grows monotonically with the increase of weak and strong diversity measures (Figure 3). In other words, for individuals with the same number of followers, those whose followers come from more diverse social backgrounds (measured by weak or strong diversity measure) are likely to have higher social reputations. Figure 4(a) summarises the prediction accuracy ($R^2$) of social reputation by indegree, weak diversity measure, and strong diversity measure through three separate ordinary least square (OLS) regressions. In each regression, the social reputation index is set as the dependent variable and each measure (log-transformed) is set as the sole independent variable. As shown in the figure, structural diversity measures yield better predictions of online social reputation than indegree with $R^2$ values equal to 0.691, 0.699, and 0.698 for indegree, weak diversity measure, and strong diversity measure, respectively.

To directly compare these measures, we further adopt $L_1$ -regularized linear regression, which is also called least absolute shrinkage and selection operator (LASSO) regression. LASSO regression is a standard model in sparse regression and has been widely used for simultaneous estimation and variable selection [48–52] (see Methods for more information on LASSO regression). Note that, with the increase of the regularization level, LASSO would continuously shrink the coefficients of less important features to be zero [49, 50]. Figure 4(b) shows the regularization path obtained from LASSO regression where indegree, weak diversity measure, and strong diversity measure (log-transformed) are set as predictors in the prediction of social reputation. With the decrease of the regularization level, weak and strong diversity measures are first selected before indegree, which may imply that it is not how many followers one has but rather the structural diversity of one's followers that matters in capturing one's social reputation.

*2.3. Social Bridges.* To quantify one's social context diversity from the view of component count, previous studies have focused exclusively on explicit and direct social ties in one's contact neighborhood. However, as social neighbors provide a reliable way to infer personal characteristics [24, 27, 28, 53–55], it is reasonable to speculate that two individuals who are exposed to a similar set of users will tend to be similar with each other, even without a direct social tie between them. Therefore, the common followees between them could act as *social bridges* that implicitly "link" two unconnected individuals or social components in the network [36].

Figure 5(a) illustrates the process of how we apply social bridges to capture the implicit structural diversity of ego nodes. Figure 5(b) gives another example with all the followers (denoted by black circles) isolated (i.e., no direct social ties between the followers). However, after social bridges are taken into consideration, user a and user b are likely to form one connected component (see Figures S3 and S4 of the Supplementary Materials for more examples). Specifically, for two followers $i$ and $j$ of an ego user, we adopt Jaccard similarity of their followee sets to determine whether there is a "bridged connection" between them: $\text{JaccardSim}(i, j) = |F_i \cap F_j|/|F_i \cup F_j|$, where $F_i$ and $F_j$ denote the followee sets of $i$ and $j$, respectively; $\cap$ and $\cup$ denote the intersection and union operators of two sets, respectively; and $|\cdot|$ denotes the size of a given set. A bridged connection exists between $i$ and $j$ when $\text{JaccardSim}(i, j)$ is larger than a given

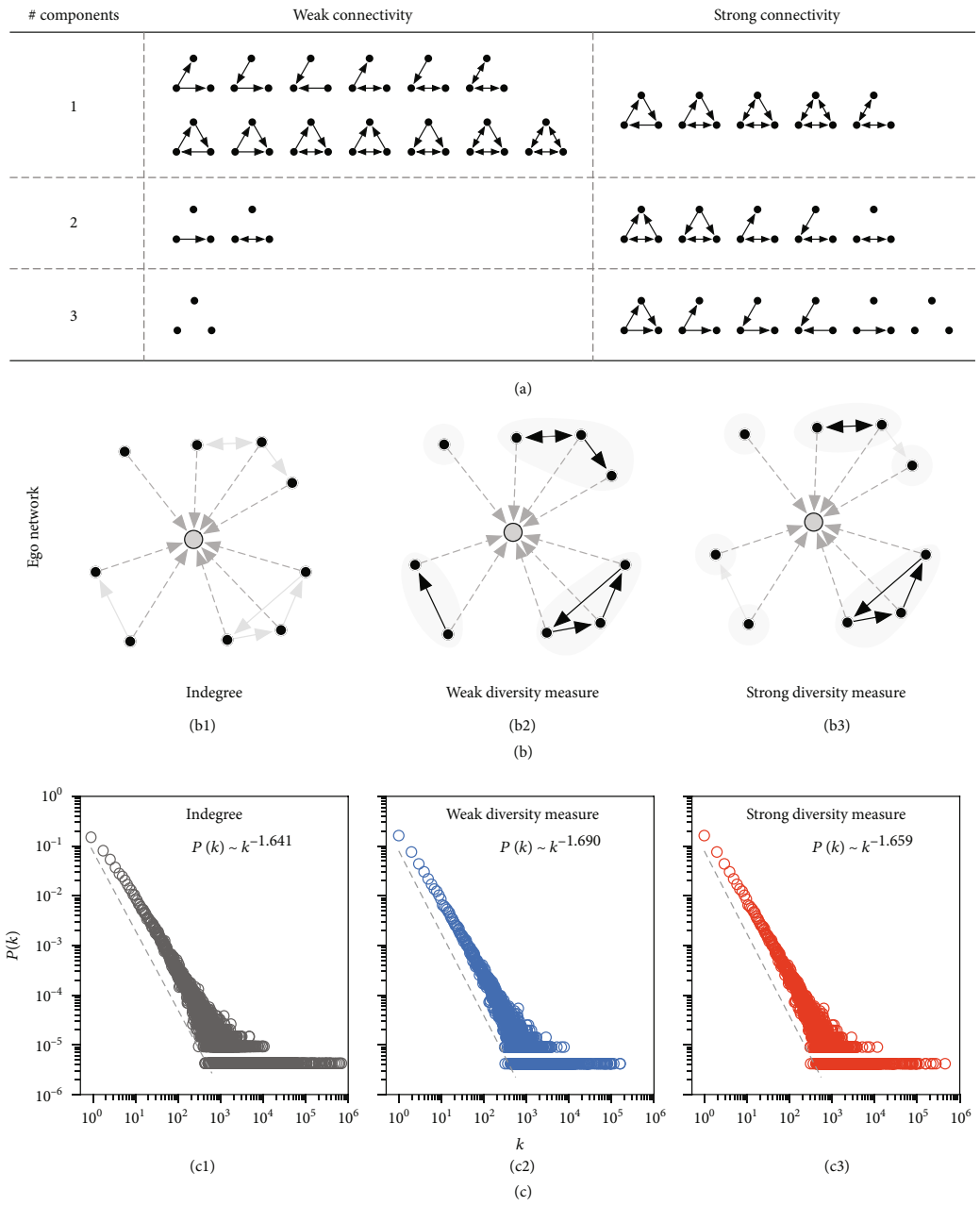


Figure 2: Weak and strong connectivity. (a) Projecting three-node directed networks into the number of weakly or strongly connected components depending on the connectivity patterns between nodes. The first, second, and third rows correspond to situations where the number of weakly or strongly connected components is exactly one, two, and three, respectively. (b) Illustration of how the structural diversity measures are quantified compared with indegree. (b1) For the given ego network, the ego user and his/her followers are shown in grey and black dots, respectively, and the incoming links of the ego user are shown in dashed lines while the links among followers are faded in grey. (b2, b3) Illustration of how weak and strong diversity measures are computed, with the informative links highlighted in black and weakly or strongly connected social components shown in shaded areas. For the given example, the indegree of the ego user is 9 as he/she has 9 followers; the weak diversity measure is 4 since there are only 4 weakly connected components formed by these 9 followers; and the strong diversity measure is 6 since there are 6 strongly connected components formed by these 9 followers. (c) Distributions of indegree, weak diversity measure, and strong diversity measure. For ease of visualization, quantities of zeros are not shown. The dashed grey line in each panel shows the power-law fitting for each distribution.

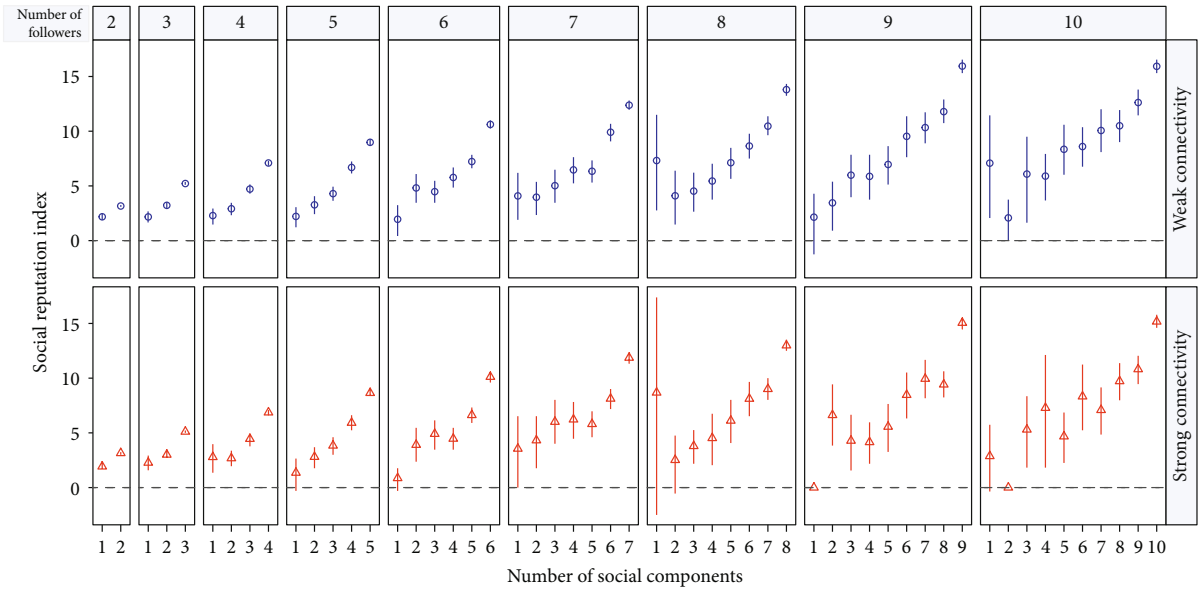


FIGURE 3: Social reputation index versus structural diversity. Social reputation index for two- to ten-follower ego networks in terms of weak and strong diversity measures stratified by the number of followers. Error bars are 95% confidence intervals.

threshold. A small threshold indicates that a bridged connection exists between two individuals as long as they share a small fraction of followees, while a large threshold requires two individuals sharing a large proportion of followees to determine the existence of a bridged connection.

Note that the bridged connection meets the requirement of strong connectivity between two nodes. Therefore, if two nodes share a bridged connection, they are also weakly and strongly connected. After social bridges are considered, the number of weakly and strongly connected components is termed the enhanced diversity measures via social bridges, denoted as "Weak connectivity+social bridges" and "Strong connectivity+social bridges" respectively. For ease of computation, ego users with more than 30,000 followers (i.e., 88 out of 234,834 ego users) are omitted in the analysis. Figure 5(c) shows the prediction accuracy ($R^2$) of social reputation by diversity measures via social bridges with the change of the Jaccard similarity threshold. Note that each $R^2$ value is obtained via OLS regression with the social reputation index set as the dependent variable and diversity measure (log-transformed) under each condition set as the sole independent variable. For weak connectivity, a threshold around 0.25 achieves the best performance; for strong connectivity, a threshold around 0.2 achieves the best performance.

In Figure 5(d), we summarise the best prediction accuracy ($R^2$) achieved by each diversity measure via different approaches, including two approaches that are proposed for undirected networks—$k$-core decomposition and $k$-brace decomposition [2]. Similar to the above, to facilitate the computation and make these diversity measures comparable, only ego users with no more than 30,000 followers are included in the analysis. We also make some slight changes to make $k$-core and $k$-brace decomposition applicable in directed networks (see Methods for details). Specifically, $k$-core decomposition and $k$-brace decomposition achieve the best prediction accuracy ($R^2$) when $k$ is set to be 2 and 1, respectively. As we can see from Figure 5(d), social bridges provide the most additional ingredient to predict personal online social reputation, and diversity measure via weak connectivity and social bridges yields the best prediction performance. Furthermore, Figure 5(e) shows the regularization path of the $L_1$-regularized linear regression (LASSO regression) model with several diversity measures (log-transformed) set as predictors. As shown in the figure, with the decrease of the regularization level, diversity measures via social bridges are selected before other diversity measures, which further suggest the effectiveness of social bridges in the prediction of online social reputation.

*2.4. Robustness Analysis.* In the previous sections, we have shown that individuals with higher levels of structural diversity tend to have higher online social reputations. However, the positive correlation between structural diversity and online social reputation may be biased by other factors. For example, as shown in Figure 6(a), the answer count is also positively correlated with online social reputation (Spearman's $r_s = 0.858$, $p < 0.001$). In other words, the more answers a user has contributed to the knowledge-sharing community, the higher the social reputations he/she tends to have on the platform. Therefore, personal online social reputation may also be induced by contributed answers rather than merely the structural diversity of individuals. We also find that gender could be another potentially confounding factor, as male users tend to receive higher social reputations than female users and users of unknown gender on the platform (one-way

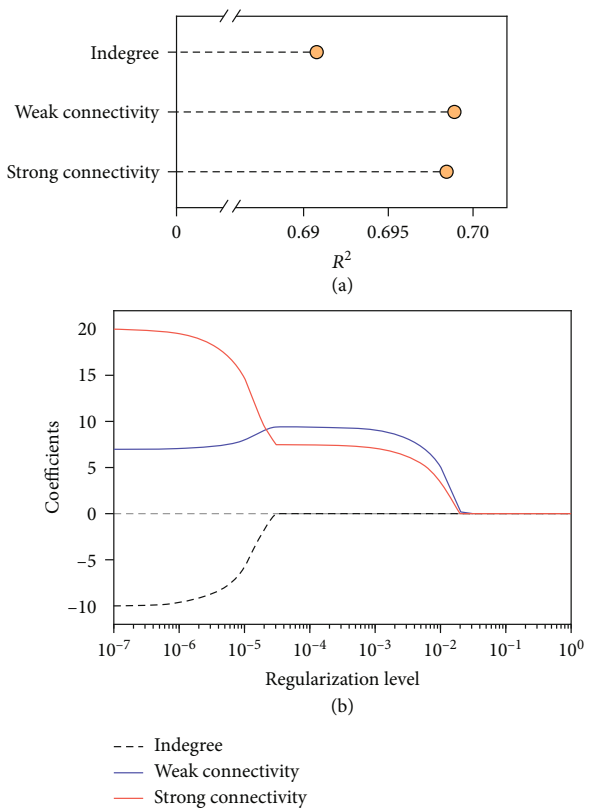


FIGURE 4: Prediction accuracy and regularization path. (a) Prediction accuracy ($R^2$) of social reputation by indegree and diversity measures (log-transformed) obtained via weak and strong connectivity. (b) Regularization path of the $L_1$-regularized linear regression model with indegree, weak diversity measure, and strong diversity measure (log-transformed) set as predictors. The log transformation is done by $\log_{10}(x+1)$ in each regression.

ANOVA, $F(2, 234831) = 6636.8$, $p < 0.001$) (Figure 6(b)). Combining with several other activity-related factors, such as question count and article count, a series of matching experiments are then conducted to distinguish the effects of structural diversity from these possible confounding factors in capturing personal social reputation.

Propensity score matching (PSM) is a widely used method for matching experiments in the literature [25, 56–61]. The key intuition of PSM is to match the treatment group with a control group whose members do not receive the treatment but are statistically indistinguishable or at least only marginally different (within a reasonable limit) from the treatment group on all observable covariates. We use the quantified diversity measure via weak connectivity and social bridges to depict the structural diversity of users in the network. Here, in our setting, a user is said to be treated (treatment group) if his/her diversity measure is larger than or equal to $m$ ($m$ is a given integer), otherwise untreated (control group). We do exact matching on indegree and gender (i.e., matched pairs have equal indegree and the same gender) and propensity score matching on other covariates (see Table S6 of the Supplementary Materials for detailed covariates controlled in the matching experiments). The difference of social reputation index in each matched pair (treatment-control) indicates the relative social reputation induced by the increase of structural diversity.

After matching, we obtain a well-balanced dataset with all standardized mean differences between the treated and untreated groups being less than 0.25. Figure 6(c) shows the differences of social reputation index between matched pairs (for space constraint and simplicity, we only present results when $m$ is set to be in the range $[2, 10]$). As shown in the figure, after ruling out several potentially confounding factors, users with higher levels of structural diversity (assigned as treated users in the matching experiments) would still tend to have higher online social reputations (paired $t$-test, $p < 0.001$ for all matching experiments). Taken together, after rigorously controlling possible confounders, matching experiments provide compelling evidence for the role of structural diversity in elevating personal online social reputation.

## 3. Discussion

The advent of social networking sites and knowledge-sharing platforms has radically shifted the way we consume information, acquire knowledge, and exchange ideas. As social ties among individuals provide the primary pathways along which interactions occur, the way we are connected and embedded in social networks is thought to affect various personal social outcomes, ranging from personal health to socioeconomic characteristics.

Taking advantage of network data which cover more than 10 million users (including more than 230,000 ego users) from an online knowledge-sharing platform, our study highlights the importance of structural diversity in online social networks and suggests an alternate perspective for people to accumulate their social capital, for policy makers to make appropriate interventions and for market operators to set up effective campaigns. Our findings are also of prime significance to understand why network structure matters in a range of social and economic domains. For example, individuals who are located in a diversified social context are generally accessible to novel information and ideas, which has important implications for viral marketing and fake news research. Moreover, as we live in such a connected world of overloaded information, how we aggregate opinions around us (e.g., adopt information from diverse backgrounds) to arrive at a reliable and accurate estimation is not only crucial to make better decisions but also important to improve the collective intelligence.

Our work is subject to a number of limitations. Our results are the product of one study based on the collected data from an online knowledge-sharing platform. Therefore, additional studies are needed to validate our findings in other kinds of social applications or domains. In this paper, we use snowball sampling during the data collection, but this approach may over- or undersample some data and induce bias to the results. To quantify the structural diversity of individuals, this work mainly considers the number of connected social components in one's social realm, but other network factors, such as the weight of ties and other structural

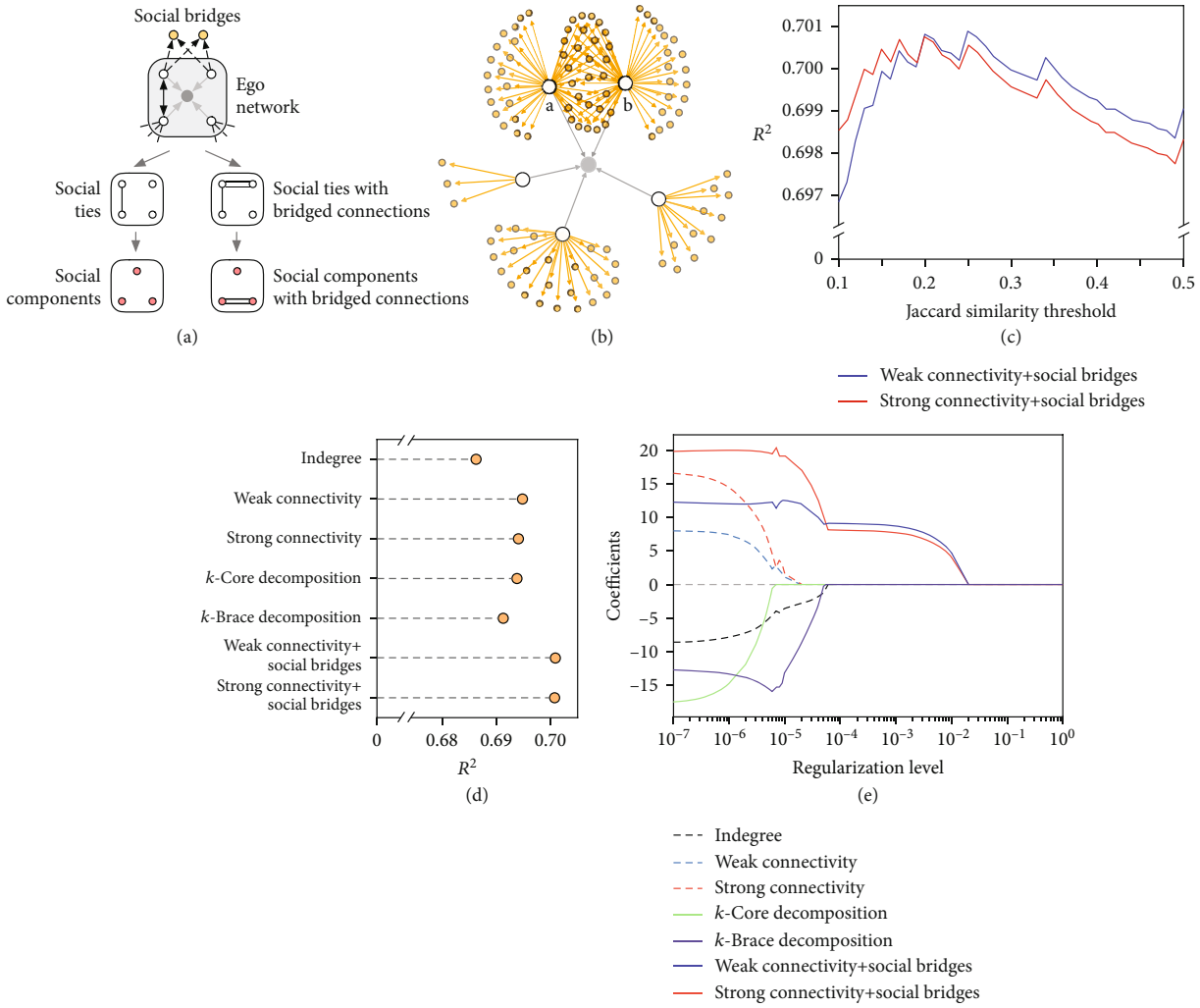


FIGURE 5: Social bridges. (a) Illustration of an ego user (grey dot) with four followers (black circles) and the projection from the original ego network to social components. Direct social ties between followers are highlighted in black, while social connections between followers and their followees are shown in dashed lines. In the given example, the top two followers have three followees (act as social bridges) in common: one is the ego user and the other two are colored in orange; as such, a "bridged connection" (double solid line) between the two followers is induced due to the function of social bridges. Direct social ties lead a four-node connected neighborhood to three social components, whereas social bridges link two of the three social components. In other words, the number of connected components becomes 2 after social bridges are considered. (b) An example ego network from the data with five isolated followers. In this example, social bridges provide an enhanced ability to depict structural diversity as node a and node b share a large proportion of followees. (c) Prediction accuracy ($R^2$) of social reputation with the changing of the Jaccard similarity threshold in the identification of bridged connections. (d) Prediction accuracy ($R^2$) of social reputation by diversity measures (log-transformed) obtained via different approaches. (e) Regularization path of the $L_1$ -regularized linear regression model with several diversity measures (log-transformed) set as predictors. To facilitate the computation and make these measures comparable, we focus on ego users with no more than 30,000 followers. Diversity measures are log-transformed by $\log_{10}(x+1)$ in each regression.

properties, may help to achieve that goal and are also worth exploring in future works. Although social bridges help to capture the implicit structural diversity of individuals, the computation will require a large amount of computing time and memory resources, especially for users with a large number of followers. In our study, the matching experiments have already accounted for several observed characteristics of users, but due to the limitation of data availability, the estimation results may still be biased without properly controlling those unobserved or unmeasured confounding factors. We emphasize that our results are built upon correlation analysis on observational data, thus not implying causality sufficiently. The current study mainly considers social ties induced by following relationships, but social connections induced by behavioral changes, such as comment or retweet, may provide another way to infer the interrelationships

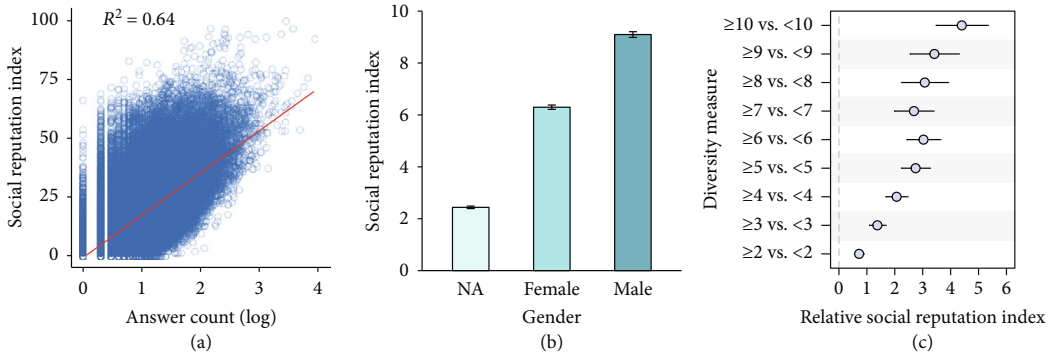


FIGURE 6: Robustness analysis. (a) Scatter plot of social reputation index versus answer count, where red line indicates the linear fit. (b) Gender differences in terms of social reputation index, where NA refers to users whose gender is not disclosed. (c) Matching results where the treatment assignment is set as whether the quantified structural diversity measure is larger than or equal to $m$ ($m \in [2, 10]$). Error bars are 95% confidence intervals.

between individuals and thus are worth exploring in future studies.

## 4. Methods

*4.1. Construction of Social Reputation Index.* The current study adopts nonnegative matrix factorization (NMF) to construct the social reputation index from the original data. NMF is an unsupervised approach for dimensionality reduction with the constraints that the input and output matrices do not contain any negative elements [44–47]. Specifically, for a given nonnegative data matrix $V$, NMF attempts to find an approximate factorization: $V \approx W \cdot H$ such that the original matrix $V$ can be decomposed into two nonnegative submatrices $W$ and $H$, with the goal of minimizing the reconstruction error between $V$ and $W \cdot H$.

In our setting, the decomposition of an original matrix with $n$ users and $m$ dimensions of features can be decoded as $V_{n\times m} \approx W_{n\times r} \cdot H_{r\times m}$, where $r$ is a given parameter prior to the matrix decomposition and indicates the expected dimension after NMF. Here in our study, $V_{n\times m}$ is the original data matrix while $n$ equals 234,834 (i.e., the number of ego users in the sample) and $m$ equals 3 as there are three types of popularity measures (i.e., number of received upvotes, thanks, and favorites), whereas $r$ is set to be 1 as we aim to obtain a single measure of social reputation. After NMF, the reduced submatrix $W_{n\times r}$ is denoted as the social reputation index.

In practice, these three popularity measures are first log-transformed by $\log_{10}(x + 1)$ and then fed into NMF simultaneously. Once the social reputation index (the first submatrix $W$ after NMF decomposition) is obtained, we normalize it to the range [0, 100] as follows: given a vector $W$ indicating the social reputation index, every element of $W$ is normalized as $W_i = 100 \times (W_i - W_{\min})/(W_{\max} - W_{\min})$, where $W_{\max}$ and $W_{\min}$ are the maximum and minimum values of $W$, respectively.

*4.2. LASSO Regression.* Generally speaking, OLS regression yields nonzero estimates to the coefficients of all the features. But LASSO adds the $L_1$ penalty of the coefficients (i.e., the sum of the absolute value of the estimated coefficients) as the regularization term to the loss function of OLS regression (see Refs. [49–51] for further technical details). Therefore, when the regularization level is set to be zero, LASSO regression is equivalent to OLS regression. With the increase of the regularization level, LASSO would gradually force the coefficients of less important features to be zero. When the regularization level becomes sufficiently large, all the estimated coefficients will be zero.

*4.3. k-Core and k-Brace Decomposition.* $k$-Core decomposition and $k$-brace decomposition have been shown useful in depicting structural diversity in undirected networks [2]. Considering the fact that users with completely isolated followers tend to have higher social reputations than their counterparts, node removal in $k$-core and $k$-brace decomposition would not work here. In this regard, we make some slight changes to make $k$-core and $k$-brace decomposition applicable in the current scenario. In the contact neighborhood formed by the followers of an ego user, only edges that are affiliated to specific nodes are removed (instead of node removal). For $k$-core decomposition, we use the degree (i.e., indegree+outdegree) of nodes to do the decomposition process: edges that are affiliated with nodes whose degrees are less than $k$ are repeatedly removed. For $k$-brace decomposition, we transform the directed networks to undirected ones and then implement the decomposition process: edges whose two endpoints share less than $k$ neighbors are repeatedly removed. After decomposition, the number of (weakly) connected components in the decomposed contact neighborhood is named the diversity measure via $k$-core or $k$-brace decomposition.

## Data Availability

The relevant data that support the findings of this paper are available from the corresponding authors upon reasonable request.

## Conflicts of Interest

The authors declare no competing interest.

## Authors' Contributions

Y.Z., L.W., J.J.H.Z., X.W., and A.S.P. conceived the present study. Y.Z. collected and processed the data. Y.Z., J.J.H.Z., and X.W. analyzed the data. Y.Z. wrote the paper with input from L.W., J.J.H.Z., X.W., and A.S.P.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant Nos. 61773255 and 61873167), Hong Kong RGC (GRF 11505119), and City University of Hong Kong (CCR 9360120 and HKIDS 9360163). The authors thank Shuyang Shi for assistance with data collection; Xiang Meng and Hai Liang for critical reading of the manuscript; and Liya Dou, Chuang Deng, Bo Liang, and Qi Zhang for helpful discussions.

## Supplementary Materials

Figure S1: data collection procedure of an ego network. Figure S2: gender distribution. Figure S3: illustration of the function of social bridges in an example ego network. Figure S4: examples of social bridges in ego networks. Table S1: data descriptions of three popularity measures. Table S2: correlation coefficients between three popularity measures. Table S3: data descriptions of other activity-related factors. Table S4: data descriptions of indegree, weak diversity measure, and strong diversity measure. Table S5: correlation coefficients between indegree, weak diversity measure, and strong diversity measure. Table S6: covariates accounted for in propensity score matching. *(Supplementary Materials)*

## References

[1] N. B. Ellison, C. Steinfield, and C. Lampe, "The benefits of Facebook "friends:" social capital and college students' use of online social network sites," *Journal of Computer-Mediated Communication*, vol. 12, no. 4, pp. 1143–1168, 2007.

[2] J. Ugander, L. Backstrom, C. Marlow, and J. Kleinberg, "Structural diversity in social contagion," *Proceedings of the National Academy of Sciences*, vol. 109, no. 16, pp. 5962–5966, 2012.

[3] A. Banerjee, A. G. Chandrasekhar, E. Duo, and M. O. Jackson, "The diffusion of microfinance," *Science*, vol. 341, no. 6144, 2013.

[4] A. Mani, I. Rahwan, and A. Pentland, "Inducing peer pressure to promote cooperation," *Scientific Reports*, vol. 3, no. 1, p. 1735, 2013.

[5] D. Centola, "The spread of behavior in an online social network experiment," *Science*, vol. 329, no. 5996, pp. 1194–1197, 2010.

[6] S. Aral and C. Nicolaides, "Exercise contagion in a global social network," *Nature Communications*, vol. 8, no. 1, p. 14753, 2017.

[7] T. Althoff, P. Jindal, and J. Leskovec, "Online actions with offline impact: how online social networks influence online and offline user behavior," in *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pp. 537–546, Cambridge, UK, 2017.

[8] A. Proestakis, E. P. di Sorrentino, H. E. Brown et al., "Network interventions for changing physical activity behaviour in preadolescents," *Nature Human Behaviour*, vol. 2, no. 10, pp. 778–787, 2018.

[9] Z. C. Steinert-Threlkeld, D. Mocanu, A. Vespignani, and J. Fowler, "Online social networks and offline protest," *EPJ Data Science*, vol. 4, no. 1, p. 19, 2015.

[10] C. Jin, C. Song, J. Bjelland, G. Canright, and D. Wang, "Emergence of scaling in complex substitutive systems," *Nature Human Behaviour*, vol. 3, pp. 837–846, 2019.

[11] N. A. Christakis and J. H. Fowler, "The spread of obesity in a large social network over 32 years," *New England Journal of Medicine*, vol. 357, no. 4, pp. 370–379, 2007.

[12] A. Madan, S. T. Moturu, D. Lazer, and A. Pentland, "Social sensing: obesity, unhealthy eating and exercise in face-to-face networks," in *Wireless Health 2010*, pp. 104–110, ACM, San Diego, CA, USA, 2010.

[13] R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani, "Epidemic processes in complex networks," *Reviews of Modern Physics*, vol. 87, no. 3, pp. 925–979, 2015.

[14] E. L. Paluck, H. Shepherd, and P. M. Aronow, "Changing climates of conflict: a social network experiment in 56 schools," *Proceedings of the National Academy of Sciences*, vol. 113, no. 3, pp. 566–571, 2016.

[15] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, 2018.

[16] N. Lin, "Building a network theory of social capital," *Connections*, vol. 22, no. 1, pp. 28–51, 1999.

[17] M. M. Wasko and S. Faraj, "Why should I share? Examining social capital and knowledge contribution in electronic networks of practice," *MIS Quarterly*, vol. 29, no. 1, pp. 35–57, 2005.

[18] N. Eagle, M. Macy, and R. Claxton, "Network diversity and economic development," *Science*, vol. 328, no. 5981, pp. 1029–1031, 2010.

[19] S. Luo, F. Morone, C. Sarraute, M. Travizano, and H. A. Makse, "Inferring personal economic status from social network location," *Nature Communications*, vol. 8, no. 1, p. 15227, 2017.

[20] Y. Leo, E. Fleury, J. I. Alvarez-Hamelin, C. Sarraute, and M. Karsai, "Socioeconomic correlations and stratification in social-communication networks," *Journal of the Royal Society Interface*, vol. 13, no. 125, article 20160598, 2016.

[21] J. Bollen, B. Goncalves, I. van de Leemput, and G. Ruan, "The happiness paradox: your friends are happier than you," *EPJ Data Science*, vol. 6, p. 4, 2017.

[22] J. Blumenstock, G. Cadamuro, and R. On, "Predicting poverty and wealth from mobile phone metadata," *Science*, vol. 350, no. 6264, pp. 1073–1076, 2015.

[23] J. Gao, Y.-C. Zhang, and T. Zhou, "Computational socioeconomics," *Physics Reports*, vol. 817, pp. 1–104, 2019.

[24] M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a feather: homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[25] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proceedings of the National Academy of Sciences*, vol. 106, no. 51, pp. 21544–21549, 2009.

[26] D. Centola, "An experimental study of homophily in the adoption of health behavior," *Science*, vol. 334, no. 6060, pp. 1269–1272, 2011.

[27] N. A. Christakis and J. H. Fowler, "Social contagion theory: examining dynamic social networks and human behavior," *Statistics in Medicine*, vol. 32, no. 4, pp. 556–577, 2013.

[28] H. Liang and F. Shen, "Birds of a schedule flock together: social networks, peer influence, and digital activity cycles," *Computers in Human Behavior*, vol. 82, pp. 167–176, 2018.

[29] M. S. Granovetter, "The strength of weak ties," *American Journal of Sociology*, vol. 78, no. 6, pp. 1360–1380, 1973.

[30] A. W. Woolley, C. F. Chabris, A. Pentland, N. Hashmi, and T. W. Malone, "Evidence for a collective intelligence factor in the performance of human groups," *Science*, vol. 330, no. 6004, pp. 686–688, 2010.

[31] L. Muchnik, S. Aral, and S. J. Taylor, "Social influence bias: a randomized experiment," *Science*, vol. 341, no. 6146, pp. 647–651, 2013.

[32] P. S. Park, J. E. Blumenstock, and M. W. Macy, "The strength of long-range ties in population-scale social networks," *Science*, vol. 362, no. 6421, pp. 1410–1413, 2018.

[33] J. M. Moore, M. Small, and G. Yan, "Inclusivity enhances robustness and efficiency of social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 563, p. 125490, 2021.

[34] J. Su, K. Kamath, A. Sharma, J. Ugander, and S. Goel, "An experimental study of structural diversity in social networks," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, pp. 661–670, Atlanta, GA, USA, 2020.

[35] A. Almaatouq, L. Radaelli, A. Pentland, and E. Shmueli, "Are you your friends' friend? Poor perception of friendship ties limits the ability to promote behavioral change," *PLOS ONE*, vol. 11, no. 3, article e0151588, 2016.

[36] X. Dong, Y. Suhara, B. Bozkaya, V. K. Singh, B. Lepri, and A. Pentland, "Social bridges in urban purchase behavior," *ACM Transactions on Intelligent Systems and Technology*, vol. 9, no. 3, p. 33, 2018.

[37] X.-L. Liu, J.-G. Liu, K. Yang, Q. Guo, and J.-T. Han, "Identifying online user reputation of user–object bipartite networks," *Physica A: Statistical Mechanics and its Applications*, vol. 467, pp. 508–516, 2017.

[38] A. Jøsang, R. Ismail, and C. Boyd, "A survey of trust and reputation systems for online service provision," *Decision Support Systems*, vol. 43, no. 2, pp. 618–644, 2007.

[39] J. Gao, Y.-W. Dong, M.-S. Shang, S.-M. Cai, and T. Zhou, "Group-based ranking method for online rating systems with spamming attacks," *Eeurophysics Letters*, vol. 110, no. 2, p. 28003, 2015.

[40] J. Gao and T. Zhou, "Evaluating user reputation in online rating systems via an iterative group-based ranking method," *Physica A: Statistical Mechanics and its Applications*, vol. 473, pp. 546–560, 2017.

[41] P. Resnick, K. Kuwabara, R. Zeckhauser, and E. Friedman, "Reputation systems," *Communications of the ACM*, vol. 43, no. 12, pp. 45–48, 2000.

[42] A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM Review*, vol. 51, no. 4, pp. 661–703, 2009.

[43] J. Alstott, E. Bullmore, and D. Plenz, "powerlaw: a python package for analysis of heavy-tailed distributions," *PLOS ONE*, vol. 9, no. 1, article e85777, 2014.

[44] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[45] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, pp. 535–541, Denver, CO, USA, 2000.

[46] A. Cichocki and A.-H. Phan, "Fast local algorithms for large scale nonnegative matrix and tensor factorizations," *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, vol. 92, no. 3, pp. 708–721, 2009.

[47] C. Févotte and J. Idier, "Algorithms for nonnegative matrix factorization with the $\beta$-divergence," *Neural Computation*, vol. 23, no. 9, pp. 2421–2456, 2011.

[48] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[49] H. Zou, "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1418–1429, 2006.

[50] P. Zhao and B. Yu, "On model selection consistency of lasso," *The Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[51] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of Statistical Software*, vol. 33, no. 1, pp. 1–22, 2010.

[52] H. Yin, A. R. Benson, and J. Ugander, "Measuring directed triadic closure with closure coefficients," *Network Science*, vol. 8, no. 4, pp. 551–573, 2020.

[53] F. Al Zamal, W. Liu, and D. Ruths, "Homophily and latent attribute inference: inferring latent attributes of twitter users from neighbors," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 6, pp. 387–390, Dublin, Ireland, 2012.

[54] D. Liben-Nowell and J. Kleinberg, "The link-prediction problem for social networks," *Journal of the American Society for Information Science and Technology*, vol. 58, no. 7, pp. 1019–1031, 2007.

[55] L. Lü and T. Zhou, "Link prediction in complex networks: a survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[56] D. B. Rubin, "Using propensity scores to help design observational studies: application to the tobacco litigation," *Health Services and Outcomes Research Methodology*, vol. 2, pp. 169–188, 2001.

[57] E. A. Stuart, "Matching methods for causal inference: a review and a look forward," *Statistical Science*, vol. 25, no. 1, pp. 1–21, 2010.

[58] L. M. Aiello, R. Schifanella, M. Redi, S. Svetlichnaya, F. Liu, and S. Osindero, "Beautiful and damned. Combined effect of content quality and social ties on user engagement," *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2682–2695, 2017.

[59] S. Kumar, W. L. Hamilton, J. Leskovec, and D. Jurafsky, "Community interaction and conflict on the web," in *Proceedings of the 2018 World Wide Web Conference*, pp. 933–943, Lyon, France, 2018.

[60] S. F. Way, A. C. Morgan, D. B. Larremore, and A. Clauset, "Productivity, prominence, and the effects of academic environment," *Proceedings of the National Academy of Sciences*, vol. 116, no. 22, pp. 10729–10733, 2019.

[61] D. E. Ho, K. Imai, G. King, and E. A. Stuart, "MatchIt: nonparametric preprocessing for parametric causal inference," *Journal of Statistical Software*, vol. 42, no. 8, pp. 1–28, 2011.