

Received December 17, 2019, accepted January 13, 2020, date of publication January 29, 2020, date of current version February 7, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2969992

Revisiting the Feld's Friendship Paradox in Online Social Networks

XIAOPING ZHOU¹, (Member, IEEE), XUN LIANG¹², (Senior Member, IEEE), JICHAO ZHAO³, HAIYAN ZHANG⁴, AND YANG XUE⁵

¹Beijing Key Laboratory of Intelligent Processing for Building Big Data, Beijing University of Civil Engineering and Architecture, Beijing 100044, China

²School of Information Science, Qufu Normal University, Shandong 276826, China

³CNPC Managers Training Institute, Beijing 100096, China

⁴School of Information Engineering, Ningxia University, Yinchuan 750021, China

⁵School of Information, Renmin University of China, Beijing 100872, China

Corresponding author: Xun Liang (xunliangruc@163.com)


This work was supported in part by the National Social Science Foundation of China under Grant 18ZDA309, in part by the National Natural Science Foundation of China under Grant 71531012, Grant 71601013, and Grant 71271211, in part by the Natural Science Foundation of Beijing under Grant 4202017 and Grant 4172032, in part by the Youth Talent Support Program of Beijing Municipal Education Commission under Grant CIT&TCD201904050, and in part by the Youth Talent Project of Beijing University of Civil Engineering and Architecture.

ABSTRACT Feld's friendship paradox is a widely accepted observation that states that the friends of any particular individual tend to have more friends on average than the individual has. Due to its implications for information transmission, the friendship paradox has become a generalized paradigm in many disciplines. However, how many of an individual's friends have more friends than the individual does is still unknown, yet interesting. In this paper, we revisited the Feld's friendship paradox and we found that only a limited number of a person's friends have more friends than the person herself has. This conclusion was reached from empirical studies using real-world networks and is true regardless of the number of one's neighbours, which contradicts the intuitive deduction from the friendship paradox. For one thing, if a person is unpopular, the number of her friends that have more friends than she does grows with the number of friends that she makes. This observation crystallizes the tenable margin of the friendship paradox. In another case, if a person is popular, as she acquires more friends, fewer of her friends have more friends than she does. This finding suggests an observation bias in the friendship paradox, which makes individuals feel less popular than their friends. Besides enriching our knowledge of the friendship paradox in psychological science, the findings reported here are also beneficial for technical areas such as large-scale triangle discovery. Although the friendship paradox was proposed from the perspective of the number of neighbours, the findings reported here can shed light on theories and applications from different disciplines like information, happiness, obesity, and extroversion, to name a few.

INDEX TERMS Computer science, Feld's friendship paradox, social computing, social network.

I. INTRODUCTION

The "friendship paradox" is the observation that the friends of individuals tend to have more friends on average than the individuals themselves [1]. This notion was first articulated for social networks by the sociologist Scott L. Feld in 1991 and has become a generalized paradigm for many areas [2], [3], e.g., information [4]–[6], interaction [7], wealth [8], happiness [9], [10], obesity [11], interesting personality [12], and extroversion [13], to name a few.

The associate editor coordinating the review of this manuscript and approving it for publication was Aneel Rahim .

This paradox has many interesting applications. For example, when an individual knows that her sex partner has more partners than she does or that her friends are much more popular, wealthy, and happy than her, this creates distortions in psychological well-being. Additionally, the friendship paradox has many applications [4]–[16] due to its implications for information transmission. Nonetheless, the friendship paradox does not tell us how many of one's friends have more friends than oneself. This is another interesting yet unanswered question. Exploring this issue can enrich our understanding of the friendship paradox.

An intuitive answer to this question would be that a large portion of one’s friends have more friends than oneself. Zuckerman and Jost [32] also argued that, from the subjective aspect, most people thought they had more friends than their friends do. However, what is the real answer of this question from the objective aspect? If the objective aspect was consistent with the subjective aspect, the finding might mitigate the distortions in psychological well-being and set theoretical basis for many technical problems, e.g., triangle discovery. This study investigates this question by analyzing many online social networks.

Our empirical studies on online social networks revealed the interesting phenomenon that only a limited number of one’s friends, as a matter of fact, have more friends than oneself, regardless of the number of one’s friends. It has also been proven theoretically that the expected number of friends with more friends than a given person in a real-world network is $O(1)$. Furthermore, removing the ultra-popular people with tens of millions of friends from the study group can mitigate the friendship paradox. In other words, most people’s friendship groups are over populated with ultra-popular people, making them feel upset about being less popular than their friends.

These findings are quite contradictory to the intuitive deduction from the friendship paradox and reveal its associated observation bias. Concretely, this study directly demonstrates that only a few of one’s friends are more popular than oneself. This indicates that the common perception of the friendship paradox is just a trick of the way social networks form. When one knows that only a few of one’s friends are more popular, wealthy, and happy than oneself, this provides a more intelligent viewpoint on the friendship paradox. Besides its significance for psychological science, this finding is useful in various technical areas. We illustrated the power of our finding in a technical area by developing a novel parallel triangle discovery algorithm that outperforms state-of-the-art solutions. In summary, our findings reveal the observation bias in the friendship paradox, and advance our understanding of the friendship paradox.

II. RELETED WORKS OF FRIENDSHIP PARADOX

A. FRIENDSHIP PARADOX

The friendship paradox is the phenomenon that most people have fewer friends than their friends do, on average [1]. Consider an undirected network $G = (V, E)$ with $n = |V|$ nodes and $m = |E|$ edges, where V and E represent the sets of nodes and edges respectively. Intuitively, (u, v) represents the edge between nodes u and v . Denote by $N(v) = \{u \in V, (u, v) \in E\}$ the set of neighbours of u and by $d(v) = |N(v)|$ the degree of v . The average number of friends of the friends of individual v is defined as:

$$f(v) = \frac{\sum_{u \in N(v)} d(u)}{d(v)}.$$

The friendship paradox holds that $d(v) < f(v)$ for most individuals in a social network. Figure 1 presents an

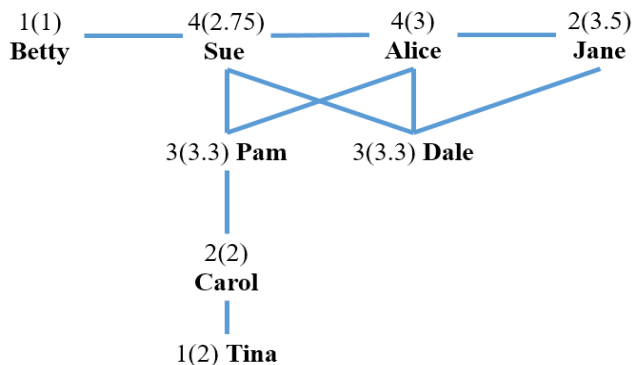


FIGURE 1. Example of friendship paradox from a sub-network of Marketville high school. The number beside each name is the number of friends that the individual has, and the number in parentheses beside each name is the mean number of friends that her friends have. Obviously, the friendship paradox holds for 5 out of 8 individuals, including Betty, Pam, Tina, Dale, and Jane.

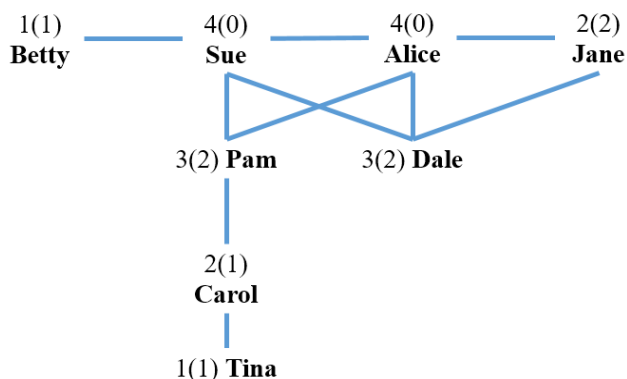


FIGURE 2. How many friends have more friends than she has: example from a sub-network of Marketville high school. The number beside each name is the number of friends of the individual, while the name in parentheses beside each name is the number of friends with more friends than she has. Obviously, the friendship paradox holds for 5 out of 8 individuals, including Betty, Pam, Tina, Dale and Jane.

example from a sub-network of the Marketville High School dataset [1]. The sub-network has eight students. The number beside each name is the number of friends that the individual has, and the number in parentheses beside each name is the mean number of friends that her friends have. Obviously, the friendship paradox holds for five of the eight individuals, including Betty, Pam, Tina, Dale, and Jane. Feld also noted that this finding holds for the whole Marketville High School network. This phenomenon has also been observed over extensive areas.

Let $\zeta(v)$ denote by the higher degree of node v , the number of nodes in $N(v)$ with degree larger than $d(v)$. One may argue intuitively from the friendship paradox that $\zeta(v)$ occupies a large portion of $d(v)$. Figure 2 presents an example from the sub-network of Marketville High School. Six out of eight individuals, with Sue and Alice as the exceptions, thought that a large portion of their friends possessed more friends than they had. In other words, 75% of individuals held that $\zeta(v) / d(v) \geq 50\%$ in the sub-network. This intuitive deduction appears valid in this scenario. However, is this deduction tenable in large-scale real-world networks, and will $\zeta(v)$ increase

TABLE 1. Datasets used for evaluation.

Dataset	Number of nodes	Number of edges	Maximum degree	Network type
Enron email	3.7×10^4	1.8×10^5	1383	Email network
web-BerkStan	6.9×10^5	6.6×10^6	84230	Web network
as-Skitter	1.7×10^6	1.1×10^7	35455	Network topology network
LiveJournal	4.8×10^6	4.3×10^7	20333	Social network

as $d(v)$ or as n ? These questions are also interesting, yet unanswered to the best of our knowledge.

B. BACKGROUND AND RELATED WORKS

The friendship paradox [1] reveals that most individuals have fewer friends than do their friends. Interestingly, Zuckerman and Jost [32] found that most people tended to believe that they had more friends than their friends.

In the past decades, the friendship paradox has been extended to other attributes. Hodas *et al.* [12] confirmed the friendship paradox in Twitter and discovered two new paradoxes, namely virality paradox and activity paradox. The virality paradox states that your friends receive more viral content than you on average, and the activity paradox conveys that your friends are more active than you on average. Eom *et al.* [2] generalized the friendship paradox from a case study from scientific collaboration, while Fotouhi *et al.* [33] studied the generalized friendship paradox in an analytical approach. Bollen *et al.* [9] found the friendship paradox in happiness, termed happiness paradox. Munzel *et al.* [10] broadened the friendship paradox from happiness to human well-being. Kramer *et al.* [34] investigated multistep friendship paradox, where the friends' friends were considered. Higham [35] developed the centrality-friendship paradoxes to answer the question when our friends are more important than us. Momeni and Rabbat [36] measured the generalized friendship paradox in networks with quality-dependent connectivity.

Besides being an interesting phenomenon, the applications of friendship paradox also can be found in many areas. Momeni *et al.* [3] explored the qualities and inequalities in online social networks using the generalized friendship paradox. Christakis and Fowler [6] developed a social network sensor system for early detection of contagious outbreaks using the friendship paradox. Golder *et al.* [7] studied the messaging within a massive online network utilizing the friendship paradox. Christakis and Fowler [11] investigated the spread of obesity in a large social network on top of the friendship paradox. Han and Srinivasan [14] identified influential mobile users through random-walk sampling, and found that your friends have more friends than you do. Netasinghe and Krishnamurthy [37] proposed three efficient polling methods for networks using friendship paradox.

The friendship paradox has been studied extensively and applied to some areas. However, the friendship paradox does not tell us how many of one's friends have more friends than oneself. This study aims to answer this question through a data analytics approach on many online social networks.

III. EMPIRICAL STUDY

A. A LIMITED NUMBER OF ONE'S FRIENDS HAVE MORE FRIENDS THAN ONESELF

To study the number of friends with more friends than oneself, we conducted experiments on nine real-world networks, including email, the Web, Internet topologies, and social networks. These networks are released publicly at the Stanford SNAP [17] and provide possibility to show applicability of our findings to a wide variety of different online networks. Four typical networks, namely Enron email, web-BerkStan, as-Skitter, and LiveJournal networks, were selected and discussed in this subsection, and the experimental results of the other five networks are available in Appendices. Usually, a friendship usually refers to a mutual-following relationship in online social networks [38]. Since not all networks in this study are undirected, we considered directed edges as undirected in directed networks. Their descriptions and concrete meanings of an edge can also be found at the Stanford SNAP. Table 1 presents the number of nodes, the number of edges, and the maximum degree of the four networks. The numbers of nodes in these networks vary from 3.7×10^4 to 4.8×10^6 and the maximum degree from 1383 to 84230. Like most real-world networks, the degrees of these networks follow a power-law distribution (see Appendix B).

We computed the higher degree, $\zeta(v)$, of each node in these networks. Interestingly, the maximum values of higher degree were no greater than 70, 130, 231 and 142 in Enron, web-BerkStan, as-Skitter, and LiveJournal respectively. As the number of nodes increases, the maximum value of higher degree increases slightly. Although the LiveJournal network has more nodes than the as-Skitter network, the maximum value of higher degree in the LiveJournal network is smaller than in the as-Skitter network. Although the number of nodes is 4.8×10^6 in the LiveJournal network and the maximum degree in the web-Berkstan network is 84230, the maximum values of higher degree are 142 and 130 in LiveJournal and web-BerkStan, respectively.

Figure 3 presents the distribution of higher degree. Notably, the higher degrees also follow a power-law distribution. More interestingly, we observed that most values of $\zeta(v)$ are less than 20 in all four datasets, regardless of the maximum degree of the network or the number of friends that an individual has. In other words, only a limited number of one's friends have more friends than oneself. This observation is quite contrary to the intuitive deduction from the friendship paradox.

We also explored the relationship between degree and higher degree. Figure 4 shows scatter diagrams of degree and

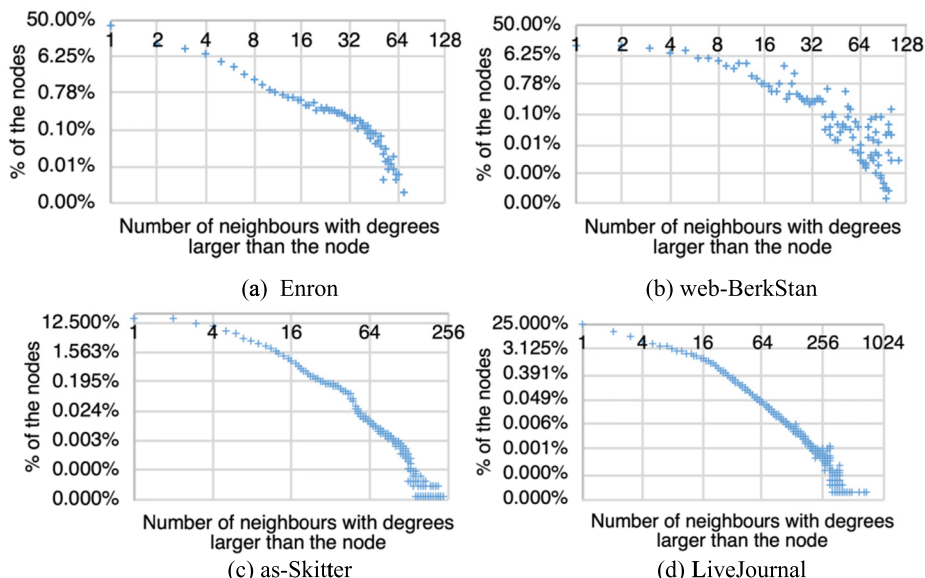


FIGURE 3. Distributions of number of neighbours with degrees larger than the node. The distributions of the number of neighbours with degrees larger than the node approximately follow a power law. Most of the nodes have less than 20 friends with more friends than they have in all four datasets.

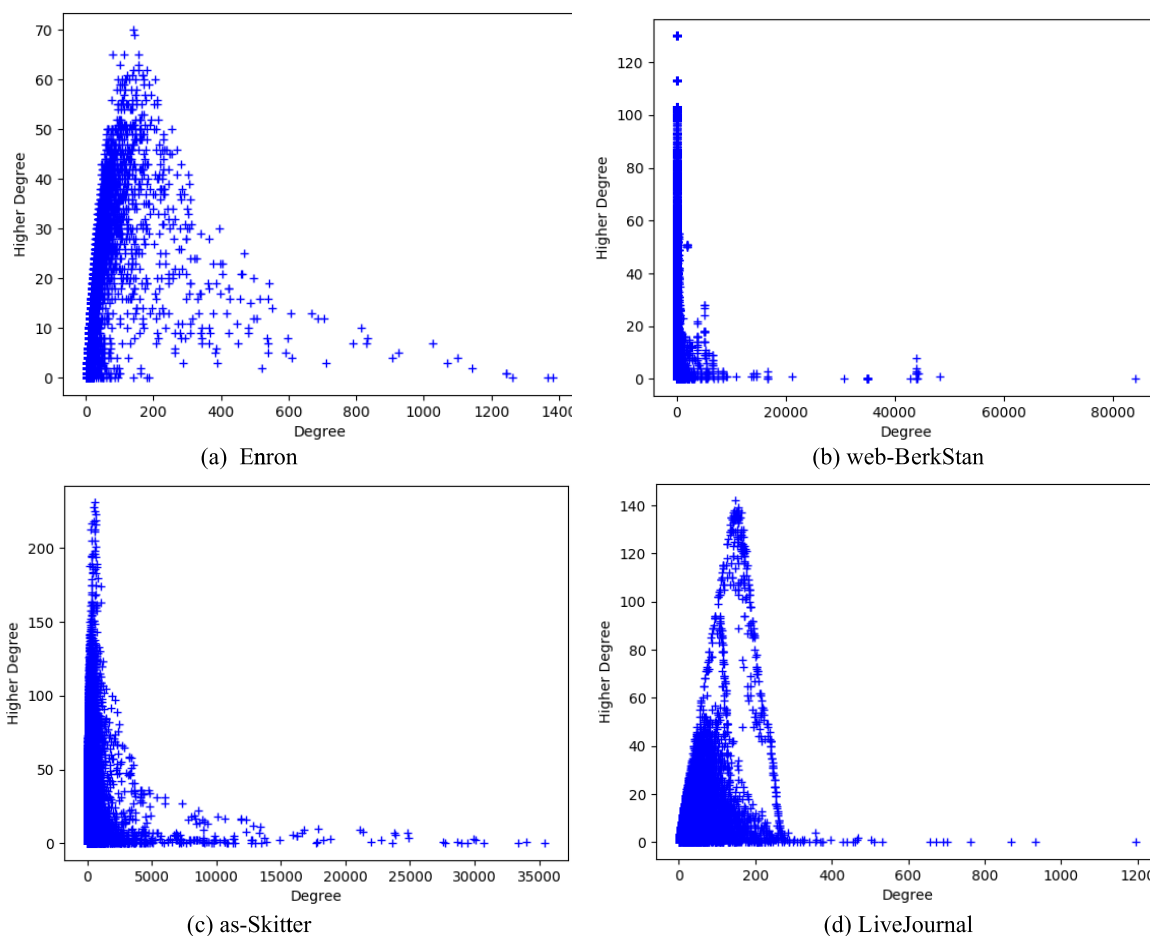


FIGURE 4. Relationship between degree and higher degree. The higher degree increases with the degree when the nodes are unpopular, but decreases sharply otherwise.

higher degree for each node in the four real-world networks. The nodes with low degree can be considered as unpopular nodes and the nodes with high degree as popular nodes.

Obviously, the higher degree grows along with the degree for most of the unpopular nodes, but the higher degree drops sharply with degree for the popular nodes in all the real-world

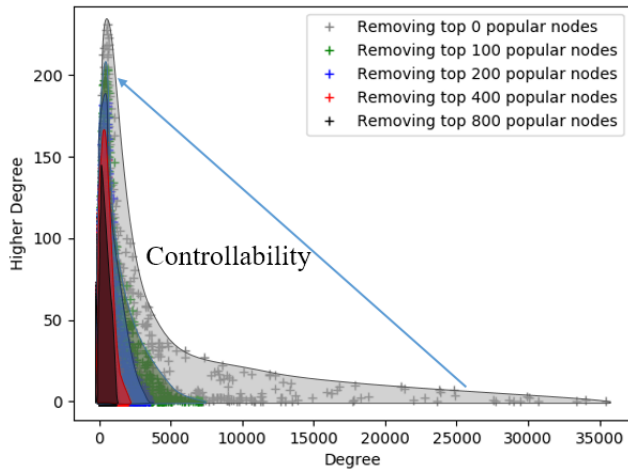


FIGURE 5. Controllability of the higher degree. Ultra-popular nodes have the ability to control the higher degree. When more ultra-popular nodes are removed, the higher degree decreases.

datasets. The higher degrees reach peak values of 70, 130, 231 and 142 when the degrees are 141, 160, 604 and 149 in the Enron email, web-BerkStan, as-Skitter, and LiveJournal networks respectively. This observation suggests that Feld's friendship paradox holds for scenarios where nodes have a limited number of neighbours or where individuals are unpopular. In scenarios where nodes are popular, things are the other way around. In other words, the higher the degree of a node, the smaller the number of her friends who have more friends than the individual herself. Subsequently, no matter how many friends an individual has, only a limited number of her friends have more friends than she has. This finding suggests an observation bias of individuals when they feel less popular, wealthy, or happy than their friends.

We also studied the controllability of higher degrees. It is commonly accepted that ultra-popular online users with thousands of friends have a significant influence on the higher degree. We confirmed this assertion by removing the top-ranking popular nodes. Figure 5 presents the experimental result on the as-Skitter dataset. The more top-ranking nodes are removed, the lower the value of the maximum higher degree. In other words, the top-ranking nodes can in fact control the higher degree. This phenomenon shows that individuals often worry about the friendship paradox when many of their friends are ultra-popular people. In this scenario, leaving aside the ultra-popular friends can mitigate their depressed feeling.

B. THEORETICAL ANALYSIS OF FINDINGS

Currently, a massive number of large-scale, real-world networks have been verified as scale-free [18], meaning that the networks follow a power-law degree distribution. As revealed in extensive real-world datasets, the exponent k of the power-law degree distribution typically lies in the range $2 < k < 3$ [19]. Here, we prove theoretically that $\zeta(v), v \in V$ is expected to be a constant with a value no greater than $n^{1/k}$ in a network following a power-law degree distribution.

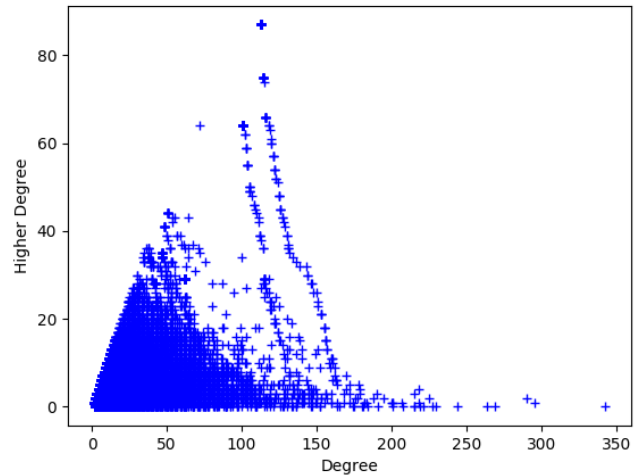


FIGURE 6. Relationship between degree and higher degree in Dbp network. The peak points is (113, 87).

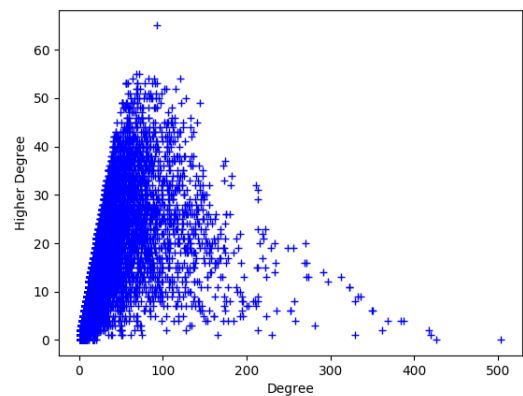


FIGURE 7. Relationship between degree and higher degree in AstroPh network. The peak points is (93, 65).

Lemma 1: For any node u in a network, $\mathbf{E}(\zeta(u)) = m/n$, where $\mathbf{E}(\cdot)$ represents the expected value. $\mathbf{E}(\zeta(u)) = O(1)$ in a network following a power-law distribution with exponent $k > 2 + \epsilon$, where $\epsilon > 0$ is a small number.

Proof: Let $S = \sum_{v \in V} \zeta(v)$. For any edge (u, v) in a network, either $u > v$ or $v > u$. Hence, each edge contributes one point to S , and $S = m$. Subsequently, the expected value of $\zeta(u)$, $\mathbf{E}(\zeta(u))$, is m/n .

Because $p_x \sim x^{-k}$, it follows that $p_x = cx^{-k}$, where c is the normalization constant. Note that $\sum_{x=1}^{\infty} p_x = 1$. It can be concluded that $\sum_{x=1}^{\infty} p_x \approx \int_1^{\infty} cx^{-k} dx = c/(k-1) = 1$. Hence, $c = k-1$. Because $\sum_{v \in V} d(v) = 2m$, $\sum_{v \in V} d(v) = \sum_{x=1}^{\infty} xnp_x \approx \int_1^{\infty} xncx^{-k} dx = cn \int_1^{\infty} x^{1-k} dx$. When $k > 2$, $cn \int_1^{\infty} x^{1-k} dx = cn/(k-2) = n(k-1)/(k-2) = 2m$. Consequently, $\mathbf{E}(\zeta(u)) = m/n = 0.5(k-1)/(k-2)$ with $k > 2$. Let $k = 2 + \epsilon$. It must be true that $\mathbf{E}(\zeta(\cdot)) = 0.5 + 0.5/\epsilon$. Hence, even a small value of ϵ can result in a constant value of $\mathbf{E}(\zeta(u))$. \square

Intuitively, when $\epsilon = 0.0005$, $\mathbf{E}(\zeta(\cdot)) = 1000$. Lemma 1 conveys that $\mathbf{E}(\zeta(u))$ can be considered as a constant with $k > 2^+$ for any node u . This conclusion can also be deduced effortlessly using the Barabási and Albert (BA) model [18], the most celebrated model for scale-free

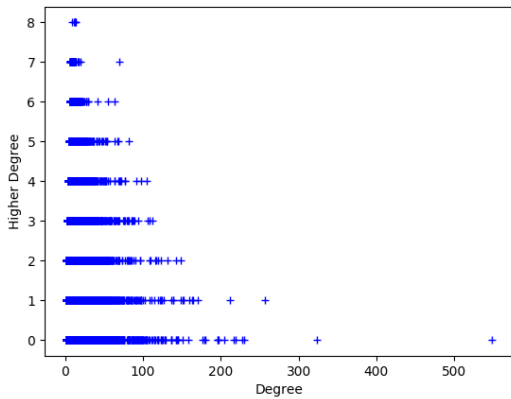


FIGURE 8. Relationship between degree and higher degree in Amazon network. The peak points is (8, 8).

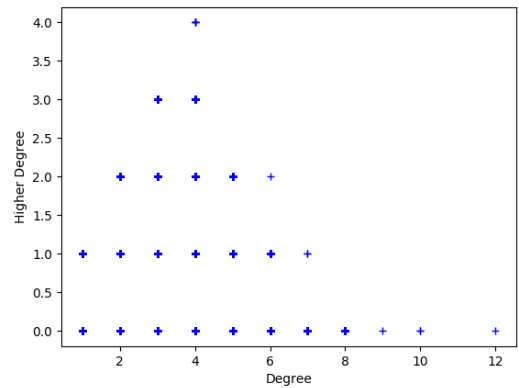


FIGURE 10. Relationship between degree and higher degree in road network. The peak points is (4, 4).

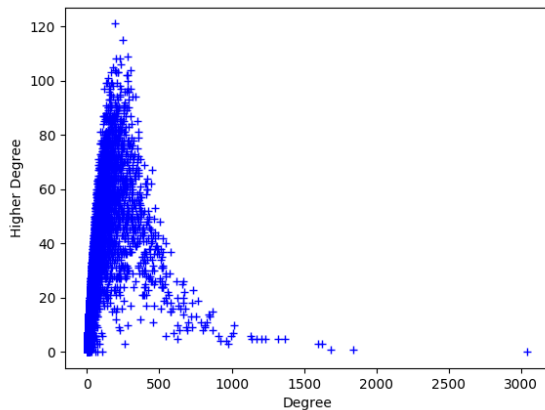


FIGURE 9. Relationship between degree and higher degree in Epinions network. The peak points is (194, 121).

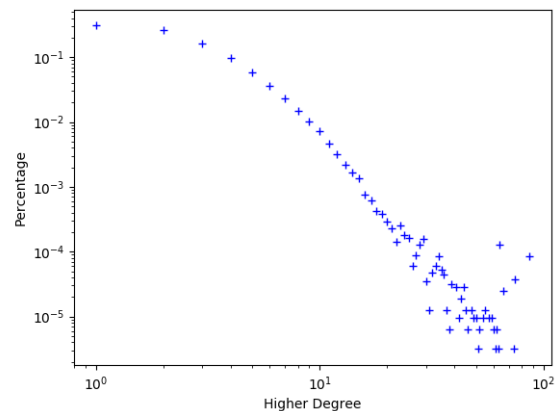


FIGURE 11. Higher degree distribution in Dblp network.

networks in which $k = 3$. Actually, in the BA model, the expected value of $\zeta(v)$, $v \in V$ is the number of links established when a node is added to the network.

Lemma 2: For any node u in a network, $\zeta(u) \leq (2m)^{1/2}$, and $\zeta(u) \leq n^{1/k}$ in a network following a power-law distribution with exponent k .

Proof: Let $\rho = (2m)^{1/2}$, and let $|H(\rho)|$ be no more than $2m/\rho = (2m)^{1/2}$ because $\sum_{v \in H(\rho)} d(v) \leq 2m$. For any $v \in H(\rho)$, $\zeta(v) \leq |H(\rho)| \leq (2m)^{1/2}$; for any $v \in L(\rho)$, $\zeta(v) \leq d(v) < (2m)^{1/2}$.

As discussed in Lemma 1, $c = k - 1$ in any network following a power-law distribution with exponent k . Here, the proof uses continuous distributions. Then, $p_x \approx \int_x^{x+1} cx^{-k} dx = x^{-k+1} - (x+1)^{-k+1}$. Finally, $|H(\rho)| = n\rho^{-k+1}$. When $\rho = n^{1/k}$, $|H(\rho)| = n^{1/k}$. Similarly, it can be concluded that $\zeta(u) \leq n^{1/k}$ for $u \in V$. \square

Lemma 1 conveys that only a limited number of friends have more friends than oneself, whereas Lemma 2 defines the upper bound of the number of friends that have more friends than oneself. Subsequently, it can be concluded that the higher degree is not scale-free, although the higher degree follows a power-law distribution.

C. APPLICATION OF TRIANGLE DISCOVERY

This exercise illustrates the power of the findings of this study by developing an approach for triangle discovery tailored to

oversize real-world networks using MapReduce [20], the *de facto* standard framework for parallel computing.

A triangle in an undirected network is a set of three nodes with edges between any two of them. Triangles are elementary structures in networks and have been used to define a variety of metrics related to information retrieval in a network, including clustering coefficients [21], the transitivity ratio [22], and triangular connectivity [23]. In addition, the power of triangles has been examined in a host of applications, e.g., network centrality measures [24]. Hence, exploring the triangles in a network lays a foundation for further studies on networks.

For decades, triangle discovery has been studied extensively. Although many algorithms exist to discover all the triangles in networks, these traditional solutions are mainly based on internal memory [25] and cannot be applied to large-scale real-world networks. Although some traditional algorithms have been upgraded to be scalable using a parallel framework such as MapReduce [26], [27], they need to load all the neighbours of each node into internal memory. This requirement causes the ‘‘curse of the last reducer’’ [27] and hinders the use of these algorithms in scenarios where some nodes have too many neighbours to be loaded into internal memory. For instance, almost 0.2 billion users have connections with the official account of SinaWeibo [28], and

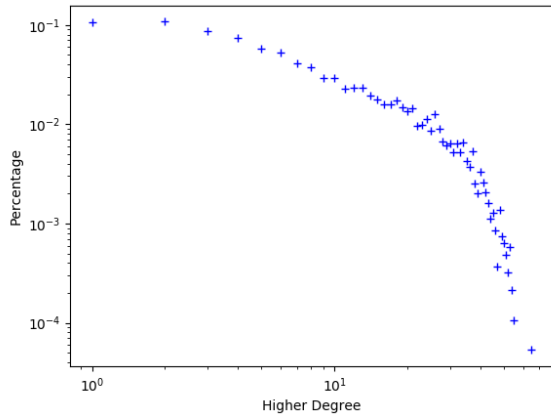


FIGURE 12. Higher degree distribution in AstroPh network.

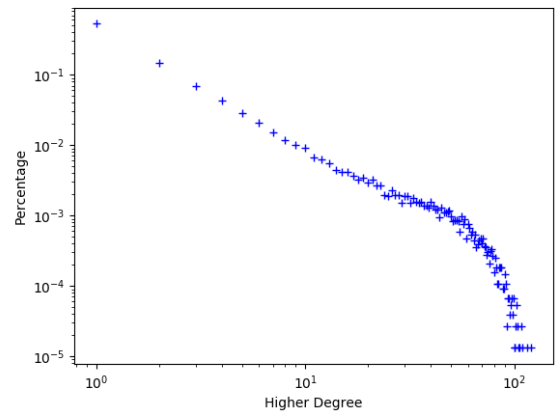


FIGURE 14. Higher degree distribution in Epinions network.

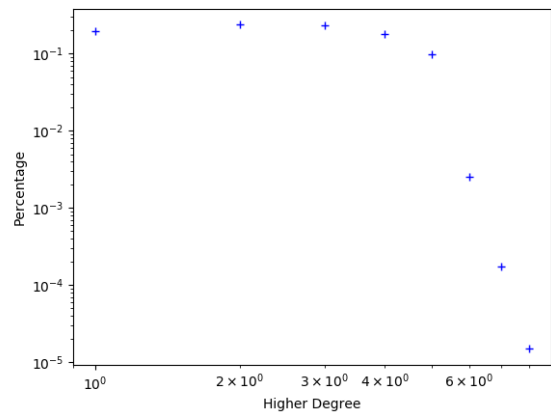


FIGURE 13. Higher degree distribution in Amazonnetwork.

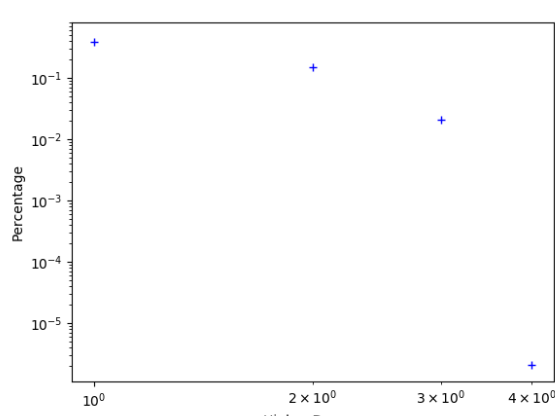


FIGURE 15. Higher degree distribution in roadnetwork.

more than 0.1 billion comments have been observed for a single tweet in Sina Weibo [29]. Obviously, these numbers are still increasing every minute. Clearly, innovative technical approaches are needed to discover all the triangles in social and other large-scale networks.

Because only a limited number of nodes have more neighbours than any given node in a real-world network, the “curse of the last reducer” can be resolved if only the neighbours with more neighbours than the node in question have to be loaded into internal memory in any reducer. With this motivation, we developed a novel triangle discovery algorithm based on degree partition, called DePart.

The following discussion presents an empirical evaluation of DePart, which we implemented in Hadoop on a cluster server with 40 nodes. Each node has 2G memory size, 500 G disk space, and a 2.8 GHz CPU. Because graph partition(GP) [27], triangle type partition(TTP) [30], and coloured triangle type partition (CTTP) [31] are the scalable triangle discovery algorithms that use MapReduce, we took them as the baseline for our algorithm.

To verify the effectiveness of DePart on real-world datasets, we conducted experiments on the datasets listed in Table 1. In GP and TTP, $\lambda = 30$, where λ is a control parameter given in advance. In CTTP, the number of colours is 200, and the number of iterations of MapReduce is 30.

For experiments with GP, TTP, and CTTP, node iteration was used as the algorithm in the last reducer. To reduce the internal memory requirement of GP, TTP, and CTTP, a hash table was used for storage, and the triangles were counted in the Reduce step. DePart was called MR-EI (MapReduce with Edge-Iteration) when $\rho = d_{max}$, where ρ is a control parameter in DePart and d_{max} is the maximum value of the degree. We tested the performance of DePart using the Enron, web-BerkStan, as-Skitter, and LiveJournal datasets with $\rho = 300, 15000, 23000$ and 10000 respectively.

Table 2 shows the results of GP, TTP, CTTP, MR-EI, and DePart in terms of running time, internal memory, and the number of key-value pairs generated. MR-EI, on which DePart is based, outperformed GP, TTP, CTTP, and DePart in running time and total disk space (except LiveJournal) in the four datasets. However, MR-EI required more internal memory than DePart. DePart outperformed GP, TTP, and CTTP in running time and performed the best in internal memory on all five datasets. Hence, DePart required much less internal memory than the other algorithms and appeared much more efficient in running time, which is consistent with the theoretical analysis (see Appendix C). Although DePart cost more in total disk space, increasing ρ can also reduce the total disk space in DePart.

TABLE 2. Empirical comparison among GP, TTP, CTPP, MR-EI, and depart algorithms.

Dataset	Running Time (minutes)					Maximum Nodes/Edges Loaded into Memory ($\times 10^9$)					Total Disk Space (Total Shuffled Nodes/Edges) ($\times 10^6$)				
	DePart	MR-EI	GP	TTP	CTTP	DePart	MR-EI	GP	TTP	CTTP	DePart	MR-EI	GP	TTP	CTTP
Enron	0.02	0.01	0.04	0.04	0.05	0.3	1.4	4.8	3.4	1.5	38.5	5.2	7.3	5.3	33.3
web-BerkStan	0.46	0.39	1.74	1.61	1.74	16.7	84.2	167.2	118.7	59.8	225.4	161.5	265.9	192.8	1203.6
as-Skitter	0.98	0.71	2.85	2.64	3.06	23.6	35.5	232.7	155.7	55.6	346.4	299.9	450.2	321.7	2008.5
LiveJournal	12.32	11.27	14.38	15.75	12.91	12.4	20.3	875.1	589.9	175.9	2064.7	2049.6	1736	1242.7	7756.5

To illustrate DePart's performance further, Fig. 6 shows the detailed running-time ratios, maximum number of nodes loaded into memory, and number of key-value pairs generated in the triangle discovery process. As shown in Fig. 6(a), the running times of GP, TTP, and CTPP were more than 3.5 times that of DePart in web-BerkStan. DePart showed unremarkable improvement in running time in LiveJournal compared to GP, TTP, and CTPP, but reduced the internal memory requirement in the Reduce step by 70 times, 47 times, and 15 times respectively. Furthermore, GP generated around 1.3 times as many key-value pairs in the Map step as DePart. Figure 6(c) shows that although CTPP reduces the disk space requirement for each iteration of MapReduce, the total amount of data to be shuffled increases dramatically. Although increasing λ in GP and TTP or running iterations of MapReduce in CTPP can decrease the internal memory requirement, both the running time and the total amount of data to be shuffled increase simultaneously.

IV. CONCLUSIONS

The friendship paradox suggests that most people have fewer friends than their friends have, on average. Due to the implications for information transmission, this observation has been widely applied in studying social networks, happiness, obesity, finance, and other fields. Nonetheless, how many of one's friends have more friends than oneself is a still unanswered, yet interesting question to the best of our knowledge. An intuitive deduction from the friendship paradox would be that a large portion of one's friends have more friends than oneself. In contradiction to this intuitive deduction, this research found, by analysing Big Data from real-world networks, that only a limited number of one's friends have more friends than oneself. For most unpopular individuals, the number of their friends with more friends than they have increases as the person makes more friends. However, this phenomenon works the other way around for popular individuals. In other words, empirical studies show that only a limited number of one's friends have more friends than oneself. A theoretical analysis for these findings has been presented here. Besides the contributions of this work to psychological science, it is also relevant to various technical areas. The power of these findings in a technical area has also been illustrated by developing a triangle discovery algorithm.

These findings crystallize the tenable margin of the friendship paradox, which is suitable for studies of unpopular

TABLE 3. Datasets.

Dataset	Number of nodes	Number of edges	Maximal degree	Network Type
Enron email	3.7×10^4	1.8×10^5	1383	Email network
web-BerkStan	6.9×10^5	6.6×10^6	84230	Web network
as-Skitter	1.7×10^6	1.1×10^7	35455	Internet topology network
LiveJournal	4.8×10^6	4.3×10^7	20333	Social network
Dblp	3.1×10^5	1.0×10^6	343	Co-authorship network
AstroPh	1.9×10^4	2.0×10^5	504	Collaboration network
Epinions	7.6×10^4	5.1×10^5	3044	Trust network
Road	2.0×10^6	2.8×10^6	12	Road network
Amazon	3.3×10^5	9.3×10^5	549	Product network

Note: The datasets are obtained from the Stanford SNAP datasets.

individuals. Undoubtedly, innovative techniques and solutions should be developed for popular individuals. Hence, the findings presented here advance our understanding of Feld's friendship paradox and offer benefits to more challenging tasks in social science, medical science, psychology, and computer science.

APPENDIX A

See Table 3.

APPENDIX B FIGURES OF RELATIONSHIPS

See Figures 6–15.

APPENDIX C ANALYSIS OF DePart

This appendix verifies the feasibility of DePart.

Theorem 1: InDePart, the size of the input to any reduce instance is no more than $O(\max(\rho, 2m/\rho))$, and the total disk space used is $O(m \times \max(\rho, |H(\rho)|) + \sum_{v \in H(\rho)} |L(\rho)|d(v))$. The optimal internal memory requirement is $O(m^{0.5})$, and the total disk space used is $O(m^{1.5} + \sum_{v \in H(\rho)} |L(\rho)|d(v))$.

Proof: For any node $u \in L(\rho)$, Reduce 1 receives $d(u) < \rho$ nodes. Since the size of $H(\rho)$ is no more than $2m/\rho$, for any node $u \in H(\rho)$, Reduce 1 receives no more than $|H(\rho)|$ nodes. Consequently, the internal memory requirement in Reduce 1 is no more than $O(\max(\rho, m/\rho))$. For any possible edge (u, v) , if (u, v) does not exist in the original edge list or $v \in H(\rho)$, Reduce 2 receives $M(v) = \{w \in N(v), w > v\}$, which is no more than $|H(\rho)|$ nodes. Otherwise, Reduce 2 receives $C(u, v) = \{w \in N(u), w > v\} + \{w \in N(v), w > v\}$, which is no more than $2 \times \max(\rho, |H(\rho)|)$ nodes. Therefore, the size of the input to any Reduce instance is no more than $O(\max(\rho, |H(\rho)|)) = O(\max(\rho, m/\rho))$.

For the edges existing in the original edge list (i.e. "real edges," others are called "fake edges"), Reduce 1 emits $m \times (2 \times \max(\rho, |H(\rho)|))$ nodes. For any node $u \in H(\rho)$, DePart generates no more than $\min(n - d(u), |L(\rho)|) \leq |L(\rho)|$ fake edges and is no more than $\min(d(u), |H(\rho)|) \leq d(u)$ nodes for any fake edge. Thus, all the fake edges occupy no more than $\sum_{v \in H(\rho)} |L(\rho)|d(v)$ units of disk space. So the total disk space used is $O(m \times \max(\rho, 2m/\rho) + \sum_{v \in H(\rho)} |L(\rho)|d(v))$.

When $\rho = 2m/\rho$, or $\rho = (2m)^{0.5}$, the internal memory requirement is optimal at $O(m^{0.5})$. In this case, the total disk space costs $2\sqrt{2}m^{1.5} + \sum_{v \in H(\rho)} |L(\rho)|d(v)$, that is $O(m^{1.5} + \sum_{v \in H(\rho)} |L(\rho)|d(v))$.

When $\rho = d_{max}$, DP turns into MR-EI; in this case, $|H(\rho)| = 0$, $|L(\rho)| = n$ and the internal memory complexity and disk space complexity are $O(d_{max}) = O(n)$ and $O(m^{1.5})$, respectively.

Lemma 1: In a scale-free network with exponent k , the size of the input to any Reduce round is $O(\max(\rho, n\rho^{-k+1}))$.

Proof: $|H(\rho)| = n\rho^{-k+1}$ in scale-free network with exponent k . According to Theorem 1, the maximal input to any Reduce round is $O(\max(\rho, n\rho^{-k+1}))$.

We design DePart to mitigate problematic nodes with too many neighbors to be loaded into internal memory; ρ is larger than $n^{1/k}$ in practice. Accordingly, the maximal internal memory requirement in DePart is $O(\rho)$.

Lemma 2: In a scale-free network with exponent k , the expected total disk space used in DePart is $O(m + n^2\rho^{-k+1})$.

Proof: Note that only the nodes larger than both u and v are received in Reduce 2 for any real edge (u, v) . According to Lemma 1, the expected number of nodes loaded into internal memory in Reduce 2 is $O(1)$. Thus, the expected total disk space occupied by the real edges is $O(m)$. For any node $u \in H(\rho)$ with degree d DePart generates no more than $|L(\rho)| = n - n\rho^{-k+1}$ fake edges and for any fake edge no more than $O(1)$ nodes. So the total disk space occupied by fake edges is no more than $O(L(\rho)H(\rho)) = O(n^2(1 - \rho^{-k+1})\rho^{-k+1}) < O(n^2\rho^{-k+1})$. Subsequently, the expected total disk space is $O(m + n^2\rho^{-k+1})$.

Theorem 2: In DePart, the total amount of the work performed by all the parallel servers is $O(m \times \max(\rho, |H(\rho)|) + \sum_{v \in H(\rho)} |L(\rho)|d(v))$.

Proof: Because there are m edges in the network, Map 1 takes $O(m)$ time. According to Theorem 2, Reduce 1 generates $O(m \times \max(\rho, |H(\rho)|) + \sum_{v \in H(\rho)} |L(\rho)|d(v))$ nodes in total, which means that the total amount of work performed in all Reduce 1 rounds is $O((m \times \max(\rho, 2m/\rho) + \sum_{v \in H(\rho)} |L(\rho)|d(v)))$ in time complexity. In addition, all the Reduce 2 rounds take the same amount of work, $O((m \times \max(\rho, 2m/\rho) + \sum_{v \in H(\rho)} |L(\rho)|d(v)))$, to output the triangles – thus, the total amount of work performed by all the parallel servers in DePart is $O((m \times \max(\rho, 2m/\rho) + \sum_{v \in H(\rho)} |L(\rho)|d(v)))$.

We can easily deduce that MR-EI ($\rho = d_{max}$) costs $O(m \times \max(\rho, 2m/\rho))$ time. When $\rho = 2m/\rho$ and $\rho = (2m)^{0.5}$, the running time is optimal with $O(m^{1.5})$.

Lemma 3: In a scale-free network with exponent k , the expected total amount of the work performed by all parallel servers is $O(m + n^2\rho^{-k+1})$.

Proof: In the first round of MapReduce in DePart, m edges are traversed, which costs $O(m)$ in running time. For each edge (u, v) , Reduce 2 costs $O(1)$ in time to output all the triangles including nodes u and v . According to Lemma 2, the total number of nodes generated in Reduce 1 is $O(m + n^2\rho^{-k+1})$. Thus, the expected running time of the second round of MapReduce in DePart is $O(m + n^2\rho^{-k+1})$, and the expected total amount of work performed by all parallel servers is $O(m + nm/\rho^2)$.

Obviously, when internal memory complexity is optimal, or $\rho = n^{1/k}$, the total amount of the work performed by all parallel servers is $O(m + n^{1+1/k})$. Since $k > 2$, $O(m + n^{1+1/k}) < O(m + n^{1.5}) < O(m^{1.5})$.

Basically, Lemma 3 clearly reveals that DePart discovers all the triangles more efficiently than GP, TTP or CTPP especially in scale-free networks, at $O(m^{1.5})$ in time complexity.

Based on the above, DePart uses $O(m + n^2\rho^{-k+1})$ in total disk space and $O(\rho)$ in internal memory when $\rho > n^{1/k}$. In other words, compared to MR-EI, DePart reduces the internal memory requirement by a factor of n/ρ by creating an $O(n^2\rho^{-k+1})$ increase in total disk space. As illustrated through real-world datasets, DePart has an adequate tradeoff between internal memory and total disk space in practice, which is quite favorable considering RAM is typically much more expensive than disk space.

REFERENCES

- [1] S. L. Feld, "Why your friends have more friends than you do," *Amer. J. Sociology*, vol. 96, no. 6, pp. 1464–1477, 1991.
- [2] Y. H. Eom and H. H. Jo, "Generalized friendship paradox in complex networks: The case of scientific collaboration," *Sci. Rep.*, vol. 4, Apr. 2014, Art. no. 04603.
- [3] N. Momeni and M. Rabbat, "Qualities and inequalities in online social networks through the lens of the generalized friendship paradox," *PLoS ONE*, vol. 11, no. 2, Feb. 2016, Art. no. e0143633.
- [4] B. Huberman, D. M. Romero, and F. Wu, "Social networks that matter: Twitter under the microscope," Dec. 2008, *arXiv:0812.1045*. [Online]. Available: <https://arxiv.org/abs/0812.1045>
- [5] J. P. Bagrow, C. M. Danforth, and L. Mitchell, "Which friends are more popular than you? Contact strength and the friendship paradox in social networks," Mar. 2017, *arXiv:1703.06361*. [Online]. Available: <https://arxiv.org/abs/1703.06361>

- [6] N. A. Christakis and J. H. Fowler, "Social network sensors for early detection of contagious outbreaks," *PLoS ONE*, vol. 5, no. 9, Sep. 2010, Art. no. e12948.
- [7] S. A. Golder, D. M. Wilkinson, and B. A. Huberman, "Rhythms of social interaction: Messaging within a massive online network," in *Communities and Technologies*. London, U.K.: Springer, 2007, pp. 41–66.
- [8] P. Gai and S. Kapadia, "Contagion in financial networks," *Proc. Roy. Soc. London A, Math. Phys. Eng. Sci.*, vol. 466, no. 2120, pp. 2401–2423, Aug. 2010.
- [9] J. Bollen, B. Gonçalves, and I. G. van de Leemput Ruan, "The happiness paradox: Your friends are happier than you," *EPJ Data Sci.*, vol. 6, no. 1, p. 4, 2017.
- [10] A. Munzel, J.-P. Galan, and L. Meyer-Waarden, "Getting by or getting ahead on social networking sites? the role of social capital in happiness and well-being," *Int. J. Electron. Commerce*, vol. 22, no. 2, pp. 232–257, Apr. 2018.
- [11] J. Stockman, "The spread of obesity in a large social network over 32 years," *Yearbook Pediatrics*, vol. 2009, pp. 464–466, Jan. 2009.
- [12] N. O. Hodas, F. Kooti, and K. Lerman, "Friendship paradox redux: Your friends are more interesting than you," in *Proc. ICWSM*, 2013, pp. 8–10.
- [13] D. C. Feiler and A. M. Kleinbaum, "Popularity, similarity, and the network extraversion bias," *Psychol. Sci.*, vol. 26, no. 5, pp. 593–603, 2015.
- [14] B. Han, J. Li, and A. Srinivasan, "Your friends have more friends than you do: Identifying influential mobile users through random-walk sampling," *IEEE/ACM Trans. Netw.*, vol. 22, no. 5, pp. 1389–1400, Oct. 2014.
- [15] K. N. Hampton, L. S. Goulet, C. Marlow, and L. Rainie, "Why most facebook users get more than they give," *Pew Internet Amer. Life Project*, vol. 3, pp. 1–40, Feb. 2012.
- [16] S. A. Golder and S. Yardi, "Structural predictors of tie formation in twitter: Transitivity and mutuality," in *Proc. IEEE 2nd Int. Conf. Social Comput.*, Aug. 2010, pp. 88–95.
- [17] J. Leskovec and A. Krevl. (Jun. 2014). *SNAP Datasets: Stanford Large Network Dataset Collection*. [Online]. Available: <http://snap.stanford.edu/data>
- [18] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, no. 5439, pp. 509–512, Oct. 1999.
- [19] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Rev.*, vol. 51, no. 4, pp. 661–703, Nov. 2009.
- [20] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," *Commun. ACM*, vol. 51, no. 1, p. 107, Jan. 2008.
- [21] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, p. 440, 1998.
- [22] F. Harary and H. H. Paper, "Toward a general calculus of phonemic distribution," *Language*, vol. 33, no. 2, p. 143, Apr. 1957.
- [23] Z. E. Roth and Y. Baram, "Multidimensional density shaping by sigmoids," *IEEE Trans. Neural Netw.*, vol. 7, no. 5, pp. 1291–1298, Sep. 1996.
- [24] X. Zhou, X. Liang, J. Zhao, and S. Zhang, "Cycle based network centrality," *Sci. Rep.*, vol. 8, no. 1, 2018, Art. no. 11749.
- [25] M. Latapy, "Main-memory triangle computations for very large (sparse (power-law)) graphs," *Theor. Comput. Sci.*, vol. 407, nos. 1–3, pp. 458–473, Nov. 2008.
- [26] J. George, C.-A. Chen, R. Stoleru, and G. G. Xie, "Hadoop MapReduce for Mobile Clouds," *IEEE Trans. Cloud Comput.*, vol. 7, no. 1, pp. 224–236, Jan. 2019.
- [27] S. Suri and S. Vassilvitskii, "Counting triangles and the curse of the last reducer," in *Proc. 20th Int. Conf. World Wide Web*, 2011, pp. 607–614.
- [28] (Jun. 2017). *SinaT*. [Online]. Available: <http://weibo.com/sinat>
- [29] L. Han. (Jun. 2017). *That's Why I Love Manchester United*. [Online]. Available: <http://weibo.com/1537790411/yB8xz28NI>
- [30] H.-M. Park, F. Silvestri, U. Kang, and R. Pagh, "Mapreduce triangle enumeration with guarantees," in *Proc. 23rd ACM Int. Conf. Conf. Inf. Knowl. Manage. - CIKM '14*, 2014, pp. 1739–1748.
- [31] H.-M. Park and C.-W. Chung, "An efficient mapreduce algorithm for counting triangles in a very large graph," in *Proc. 22nd ACM Int. Conf. Conf. Inf. Knowl. Manage. - CIKM '13*, 2013, pp. 539–548.
- [32] E. W. Zuckerman and J. T. Jost, "What makes you think you're so popular? self-evaluation maintenance and the subjective side of the 'friendship paradox'," *Social Psychol. Quart.*, vol. 64, no. 3, p. 207, Sep. 2001.
- [33] B. Fotouhi, N. Momeni, and M. G. Rabbat (November), "Generalized friendship paradox: An analytical approach," in *Proc. Int. Conf. Social Inform.*, Springer, Cham, 2014, pp. 339–352.
- [34] J. Brown Kramer, J. Cutler, and A. J. Radcliffe, "The multistep friendship paradox," *The Amer. Math. Monthly*, vol. 123, no. 9, p. 900, 2016.
- [35] D. J. Higham, "Centrality-friendship paradoxes: When our friends are more important than us," *J. Complex Netw.*, vol. 7, no. 4, pp. 515–528, Aug. 2019.
- [36] N. Momeni and M. G. Rabbat, "Measuring the generalized friendship paradox in networks with quality-dependent connectivity," in *Complex Networks VI*. Cham, Switzerland: Springer, 2015, pp. 45–55.
- [37] B. Nettasinghe and V. Krishnamurthy (2019), "What do your friends think?": Efficient polling methods for networks using friendship paradox," *IEEE Trans. Knowl. Data Eng.*, to be published.
- [38] X. Zhou, X. Liang, X. Du, and J. Zhao, "Structure based user identification across social networks," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 6, pp. 1178–1191, Jun. 2018.



XIAOPING ZHOU (Member, IEEE) received the B.E. and M.E. degrees from Beijing Information Science and Technology University, in 2006 and 2009, respectively, and the Ph.D. degree from the Renmin University of China, in 2018. He is currently an Associate Professor with the Beijing University of Civil Engineering and Architecture. His research interests include data mining, artificial intelligence, and building information modeling (BIM).



XUN LIANG (Senior Member, IEEE) received the B.Sc. and Ph.D. degrees in computer engineering from Tsinghua University, Beijing, China, in 1989 and 1993, respectively, and the M.Sc. degree in economics and operations research from Stanford University, Stanford, CA, USA, in 1999. He worked as a Postdoctoral Fellow with the Institute of Computer Science and Technology, Peking University, from 1993 to 1995, and the Department of Computer Engineering, University

of New Brunswick, from 1995 to 1997. He has worked as a Software Architect or CTO leading over ten intelligent information products in California, from 2000 to 2007, and an Associate Professor with the Institute of Computer Science and Technology, Peking University, from 2005 to 2009. He is currently a Distinguished Professor with the School of Information Science, Qufu Normal University, Shandong, China. He focused his research on machine learning and social computing, hosted several projects from the NSFC, and led the Business Intelligence Laboratory, Renmin University of China, from 2005 and 2019. His research interests include machine learning, web mining, and social computing.



JICHAO ZHAO received the B.E. degree in information system from Northwest A&F University, in 2015, and the M.E. degree in management science from the Renmin University of China, in 2018. She is currently a Lecturer with the CNPC Managers Training Institute. Her research interests include management information system and data mining.

HAIYAN ZHANG is currently an Associate Professor with Ningxia University. Her research interests are data mining and artificial intelligence.

YANG XUE is currently a Post Graduate Student with the Renmin University of China. His research interest includes data mining.

• • •