Weibo, and a Tale of Two Worlds

Wentao Han, Xiaowei Zhu, Ziyan Zhu, Wenguang Chen, Weimin Zheng
Department of Computer Science and Technology
Tsinghua University
Beijing, China
{hwt04, zhuxw13}@mails.tsinghua.edu.cn
2011213022@bupt.edu.cn, {cwg,zwm-dcs}@tsinghua.edu.cn

Jianguo Lu
School of Computer Science
University of Windsor
Windsor, Canada
jlu@uwindsor.ca

***Abstract*—Weibo is the Twitter counterpart in China that has attracted hundreds of millions of users. We crawled an almost complete Weibo user network that contains 222 million users and 27 billion links in 2013. This paper analyzes the structural properties of this network, and compares it with a Twitter user network. The topological properties we studied include the degree distributions, reciprocity, clustering coefficient, PageRank centrality, and degree assortativity. We find that Weibo users have a higher diversity index, higher Gini index, but a lower reciprocity and clustering coefficient for most of the nodes. A surprising observation is that the reciprocity of Weibo is only about a quarter of the reciprocity of the Twitter user network. We also show that Weibo adoption rate correlates with economic development positively, and Weibo network can be used to quantify the connections between provinces and regions in China. In particular, point-wise mutual information is shown to be accurate in quantifying the strength of connections.**

I. Introduction

Sina Weibo, the Chinese equivalent of Twitter, has attracted hundreds of millions of users. Despite its immense impact on society, Weibo user network has not been systematically studied except for a few brief summaries using small sample data [10] [29] [13] [11]. Sample data can only infer a limited number of simple properties, and the inference may not be accurate. We crawled 222 million Weibo users from November 2012 to February 2013. To our knowledge, this paper is the first attempt to give an overall view of Weibo based on an almost complete user network.

Social network of this magnitude was studied within companies who own the data. For instance, Myers et al. studied 175 million active Twitter users in 2012 [22]; Uganer et al. characterized the entire Facebook user network of 721 million active users in May 2011 [27]; Leskovec et al. studied the communication network that consists of 240 million users of Microsoft instant messenger in 2006 [16]. Online social networks available for public study are often small [15] and incomplete [30][19][28][24]. Our Weibo network is the largest that is available for research independent of data providers. Social networks are evolving rapidly into a complex platform and playing a profound role in our daily life in many ways. An independent study of the networks is of utter importance.

The topological properties we studied include the degree distributions, reciprocity, clustering coefficient, PageRank centrality, degree assortativity, degree of separation. The ground truths of these properties are instrumental for understanding the formation and evolution of the network and information diffusion over the network.

The comparison between Weibo and Twitter is particularly interesting. On the one hand, the users are mostly disjoint, representing the online social network users of China and the rest of the world. On the other hand, the networking platforms are almost the same. Both are directed networks, allowing unlimited inbound links; both impose a default 2000 up-limit for the outbound links for each user; both attracted hundreds of millions of users. Such similar but isolated platforms provide an opportunity to study structural difference between China and the rest of the world. We find that the way people interacting in Weibo is fundamentally different, with very low reciprocity.

The ultimate goal of studying an online social network is to link it to the real world [12]. Some pioneering studies show that the way people socialize can have a connection with economic development. For instance, Eagle et al. established the connection between the diversity of a phone network and the level of the economic development in that region [8]. This paper demonstrates that Weibo penetration rate and clustering coefficients relate positively to economic development. More importantly, we use Weibo to quantify the connections between provinces and regions, and find that point-wise mutual information can reflect connection strength accurately.

II. The Data

We crawled a snapshot of the Weibo user network from November 2012 to February 2013. At that time, Weibo API was not as restrictive as today. Thus, an almost complete user network was obtained, containing most users who have followers. We used the breadth-first crawling strategy by following the out-links of the collected nodes (users). At the end of the crawling process, most retrieved nodes are old ones that have been crawled before. On average, 689 duplicates are retrieved in order to harvest one new node. Since the chance is very slim for spotting a new node, we stopped the crawling process. Considering that the user network was dynamic and expanding, there would always be new nodes if the crawling had continued. What we can get is a snapshot of the network within a certain time frame. Overall, the crawled graph contains 282 million distinct users. This number is substantially lower than the officially announced number of

registered users, which is 503 million by the end of 2012 [21].

To understand the discrepancy between the crawled graph and the original graph, we should be aware of the limitation of the crawling strategy. Note that the same strategy is used by the Twitter 2009 data [15] that will be compared with in this paper. The crawling follows the out-bound links, meaning that the nodes without any followers can not be collected. To find out the population of such unfollowed nodes, we run a random sampling that consists of one million uniform random users by probing the ID space. Among them, 48% are never followed for Weibo, and 40% for Twitter. Thus, our estimation of the total Weibo population in early 2013 is $282/0.52 \approx 542$ million, which is close to 503 million, the official number by the end of 2012. This confirms that our Weibo graph contains most of the followed users.

Our next question is the impact of the lack of those unfollowed nodes on the network properties to be studied. First, several properties we studied, such as clustering coefficient and the connections between regions, use a subgraph that consists of bi-directional (mutual) edges only. Those mutual relations are the same regardless the exclusion of these unfollowed nodes. The unfollowed nodes are not involved in any mutual relations, and we do not need to crawl them to find it out.

Secondly, our random sampling indicates that most of unfollowed nodes are isolated, not connecting with any other nodes. Although their number is large, accounting for almost half of the population, the proportion of the edges they contribute to the entire graph is small (5%). Thus, the impact on the overall structure of the graph is limited. The most direct influence is the distribution of component sizes, i.e., the sizes of the weakly and strongly connected components. The way we collect the data ignored huge number of isolated components. Therefore, we exclude this property from our study. For degree distributions and reciprocity, the influence is limited given the small proportion of these edges. In addition, the comparative study is worthwhile because the same crawling method is employed for the two networks. Particularly, both data sets do contain some unfollowed users that occur in the seed set, 1.7 million for Weibo and 1.5 million for Twitter. The number is not small, enabling a partial exploration of the unfollowed nodes. Because the data sets are crawled exhaustively, the selections of the seed sets are ad hoc, for instance from their occurrences in tweets. The seed selection causes little difference to the resulting graph except for the unfollowed nodes.

Statistics of the Weibo network is tabulated in Table I, along with the corresponding data from the Twitter 2009 network [15] for comparison. Among 282 million distinct users, we have explored all the out-links from 222 million users, in total 27 billion edges. In the following discussions, we will focus on the 222 million users because the remaining 60 million nodes have their out-bound links missing. They are newly added from the crawling process, indicating that they have low in-bound edges since they were not spotted earlier. For completeness, we list the average path length and SCC. SCC is the size of the largest strongly connected component. It accounts for 92% of the nodes for Weibo, and 80% Twitter. As we noted previously, this topology may be influenced by the way we collect he data, not necessarily the SCC structure of the entire networks. The average path length is calculated on the large component ignoring the directions of the graph. Weibo has a smaller degree of separations when unfollowed nodes are not counted. As a reference, Twitter 2014 is 4.12 [22] and Facebook is 4.7 [27].

Twitter 2009 data is selected for our comparative study because it is the largest Twitter user network available. In addition, the crawling method is the same as what we used. Incidentally, both networks are three year old after their launches. Weibo has a similar platform as Twitter, while the user groups are disjoint. Before diving into the structural properties of the networks, we give an bird's eye view of their users activities in Fig. 1. Clearly, these two networks complement each other geographically, highlighting the necessity of a comparative study. The world map is drawn from 37.5 million coordinates of postings from Weibo users, and 42.6 million from Twitter users. We can see that the earth is roughly divided into two worlds, the red one that is created by Weibo users, and the green one that consists of Twitter users. Brightness indicates the relative number of users in the area. Since Twitter service is not available in China, Weibo users dominate China, although there are occasional Twitter users distributed mainly in large cities.

All the analyses were carried on a server with four Xeon CPUs and 1TB memory. The memory is big enough to load the entire graph. Intermediate data and code are downloadable from [4].

III. Degrees

The first step to examine a network is the study of its degree distribution. Since Weibo user network is a directed graph, we plot both the in- and out-degree distributions in Fig. 2. Panels A and C are in-degree distributions; B and D are out-degree distributions. Panels A and B plot the frequency as a function of degrees. These plots are good for lower ends of the degrees. But the frequencies of the popular bloggers are not discernible. So we plot the degree as a function of its rank to focus the popular bloggers in Panels C and D.

The in-degree distributions resemble a power law for the degrees in the middle section of the data, with exponent -1 for degree-rank plot, and -2 for frequency-degree plot. Note that the exponent in the degree-rank plot is greater by one than that of the frequency-degree plot as expected [6].

Surprisingly, Weibo and Twitter have a similar slope, and that slope is close to the Zipf's law that characterizes the frequency of words in natural languages [32]. Secondly, the top bloggers have a much smaller exponent than the rest of the data. Such data can be better modelled using Mandelbrot law [20] instead of a simple power law.

Out-degrees are more influenced by restrictions imposed by service providers. Both Weibo and Twitter have a default 2000 limit for the out-bound links. The difference is that Weibo has

	#Nodes ($\times 10^6$)	#Links ($\times 10^9$)	Mean degree	Max ($\times 10^6$)	Std	CV	Simpson ($\times 10^{-4}$)	Gini	Path length	SCC (%)
Weibo	222	27	121	25	9028	74	0.25	0.88	3.44	92
Twitter	41	1.4	35	2.9	2419	69	1.13	0.83	4.12	80

TABLE I

STATISTICS OF WEIBO AND TWITTER 2009 DATA. MAX, STANDARD DEVIATION (STD), COEFFICIENT OF VARIATION (CV), SIMPSON INDEX AND GINI INDEX ARE FOR IN-DEGREES.

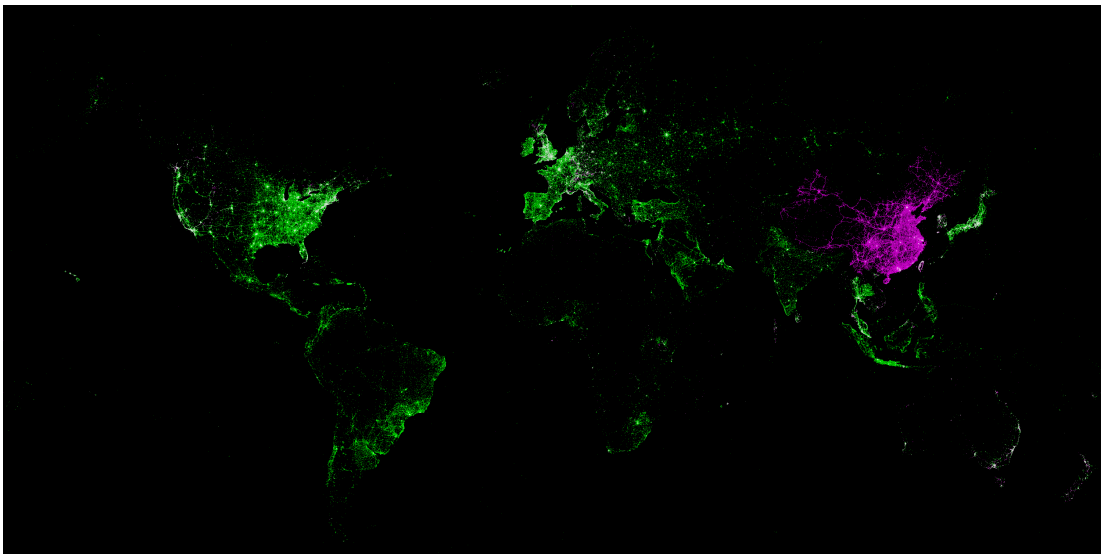

Fig. 1. The map of Weibo and Twitter user locations three years after their launches. Best viewed in colour and zoomed in online at [5]. Red: Weibo; green: Twitter; white: both Weibo and Twitter.

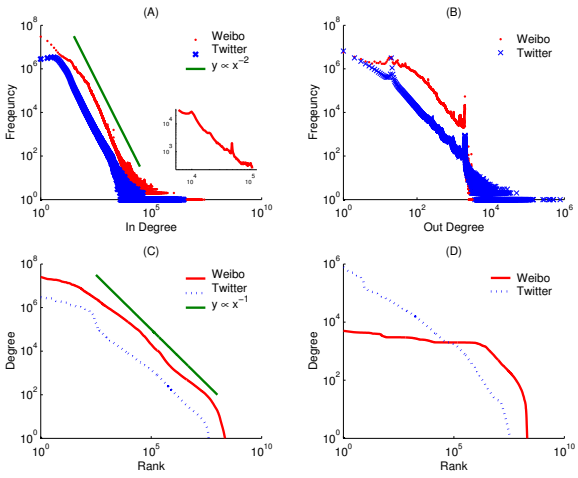

Fig. 2. Degree distributions of Weibo, and its comparison with Twitter. The inset in Panel A zooms in a segment of the Weibo data.

only a few exceptions that are slightly greater than 2000, while Twitter has a large number of privileged users who can have millions of out-bound links. The 2000 up-limit explains the spikes around 2000 in Panel B. These users maximize their presence by using up their quota. A spike around degree 20 for Twitter network is due to the automated recommended users.

Degree distribution alone can lead to interesting discoveries. The inset in Panel A focuses on a segment of in-degrees. There are apparent spikes around 10k and 50k. One possible explanation is that people typically buy fake followers in the amount of 10k and 50k.

Such highly engineered data are difficult to be modelled precisely using mathematical formulas, just the same as the Web graph that cannot be modelled by a simple power law [18]. Instead, we give several metrics that measures the variation of the degrees in Table I. We focus on in-degrees only, because its size is unrestricted by both systems. Twitter allows for many large out-degrees, hence the comparison is not meaningful. 1) Coefficient of variation (CV): it is the standard deviation normalized by the average degree. One of its intuitive interpretation is to measure the number of friends of your friends. In every social network, your friends have more friends than you do [9]. $CV^2 + 1$ measures how many times more [17]. In a hypothetical homogeneous network where every user has the same degree, $CV = 0$. Only in that case your friends have the same number of friends as you have. In Weibo, $CV^2 + 1 = 5567$, i.e., on average your friends have 5567 $times$ more friends than you do. This is considerably larger than that of Twitter (4761). 2) Simpson diversity index: It measures the evenness of the degrees [26]. It can be interpreted as the probability of following the same person when two random links are selected. Lower index means lower probability of collision, thus higher diversity. The Simpson index for Weibo is 0.25×10^{-4}, meaning that approximately $40,000$ random links are needed so that a

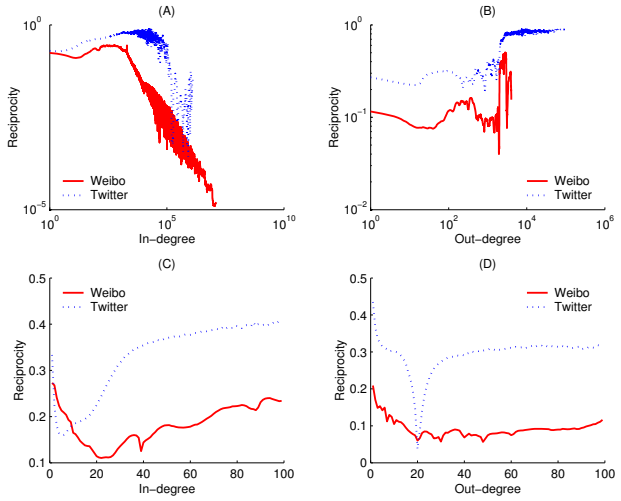

Fig. 3. Average reciprocity as a function of in- and out-degrees.

person could be followed twice. On the other hand, Twitter only needs less than $10,000$ random links. Hence, the diversity of Weibo is higher. 3) Gini coefficient: it is used to measure the inequality of the degrees. It is 0.88 for Weibo, and 0.83 for Twitter. The top 1% of Weibo users possess 62% followers (56% for Twitter). The top 0.1% Weibo accounts own 48% of followers (37% for Twitter).

Conclusion: Although their in-degree slopes are similar, Weibo has a higher inequality and higher diversity than Twitter.

IV. Reciprocity and Transitivity

The often posed question for Weibo or Twitter is whether it is a media or a social networking platform. The answer is not as clear as Facebook, because many people in Weibo and Twitter follow celebrities only, giving the impression that the networking functionality is weak on such platforms. One of the major tasks in the studies of the Twitter user networks in 2009 [15] and 2012 [22] is to answer this question. Reciprocity and transitivity (clustering coefficient) are used as the principal metrics.

A. Reciprocity

We find that among all the links of Weibo users, 10% are reciprocated. This reciprocity is substantially lower than that of Twitter 2009 (0.36) and Twitter 2012 (0.42). The huge difference is startling, and prompts us to look into it more closely.

First, there are alternative methods to calculate the reciprocity, notably the arch method and dyad method [14], that result in different values. The $arch$ method calculates the proportion of the arches that are reciprocated. In the example graph in Fig. 4, there are three arches (A,B), (B,A), and (B,C). Among them, two are reciprocated. Thus, reciprocity $r = 2/3 \approx 0.66$. The $dyad$ method, on the other hand, calculates the proportion of the dyads that are reciprocated. There are two dyads (A,B) and (B,C), one is reciprocated, the other is not. Thus, reciprocity $r = 0.5$. In the literature the method of calculation is usually not indicated explicitly. [23] adopted the arch method, while [15] used the dyad approach. Our first calculation of reciprocity uses the arch method. If we use the dyad method, the reciprocity is 0.05 for Weibo and 0.22 for Twitter. In other words, for every pair of nodes that has a relationship between them, the probability of that relationship being bi-directional is 0.05 for Weibo, 0.22 for Twitter. The difference between the two networks is even larger.

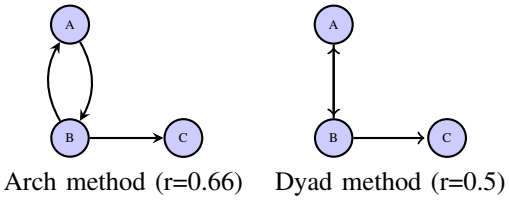

Fig. 4. Two kinds of reciprocities.

Next we examine the reciprocity for each type of nodes. Fig. 3 plots the average reciprocity as a function of in- and out-degrees in Panels A and B, respectively. The plots are smoothed with a window size 20. Panels C and D give the corresponding un-smoothed plots for the first 100 degrees. First, for most degrees, reciprocity of Twitter is consistently higher than that of Weibo, for both in- and out- degrees. Second, for in-degrees, reciprocity increases up to the 2000 limit, then decreases with the growth of the degree size. For large accounts who has many followers, they cannot reciprocate many followers because of the limit of out-bound link, thus their reciprocity becomes lower with the increase of in-degree. We see almost a monotonic decrease of the reciprocity of Weibo popular accounts as expected. However, Twitter has some large accounts whose reciprocities are very high. This is due to the existence of a large amount of privileged users who can have more out-bound links than the default two thousand limit. They are mostly business accounts, having the tendency of high reciprocity. Fig. 3 Panel B indicates that for those privileged users, almost all of them have a reciprocity that is close to one.

The large number of out-links given by privileged users may have boosted the overall reciprocity of the Twitter network. Thus, we discount all the out-links emanating from these privileged users. This brings down its arch reciprocity from 0.36 to 0.22, which is still significantly larger than that of Weibo (0.10).

Despite its low reciprocity, Weibo has the same average number of mutual friends as Twitter 2009. On average, Weibo has 12.9 mutual followers, while Twitter 2009 has 12.7, both after three years of evolution. One explanation is that there are two aspects of Weibo/Twitter: the social networking aspect and the information dissemination aspect. The networking aspect is similar for Weibo and Twitter, by developing the same number of mutual friends over three years. Weibo is more active in information dissemination by having a much higher average degree (121 vs. 35). Among these average links, about 13 go to friends, the rest (108 vs. 22) go to celebrities. Over time,

the number of mutual friends increases, as Twitter 2012 has an average of 48 for active users [22].

B. Clustering Coefficient (CC)

For clustering coefficient, we consider the mutual graphs in Weibo and Twitter, where each link is reciprocated, so that there is no direction considered when calculating the clustering coefficient. Since the arch reciprocity is 0.10, this subgraph contains 10% of the edges (2.7 billion) of the original graph.

We observe that for *most nodes*, CC of Weibo is smaller than that of Twitter. Fig. 5 plots the average CC as a function of degrees. Panel A compares all the nodes in Weibo and Twitter. It demonstrates that the trend is consistent. Overall, the *average* CC is 0.10 for Weibo and 0.12 for Twitter. Panel B focuses the degrees below 150. This is the Dunbar's number that is the cognitive limit to the number of relationships people can handle [7]. CCs within this range are more important since they are more likely accounts of ordinary people instead business accounts.

One lesson we learnt is that the *global* CC is a misleading metrics that should not be used to measure the rate that a friend of friend is still a friend. Despite the obvious trend, the global CC of Weibo is 0.0758, which more than doubles that of Twitter (0.0283). To understand such striking discrepancy between the global CC value and visual check on most degrees, we need to note that the global CC measures the portion of the triplets that are closed. A node with degree d has $\binom{d}{2}$ triplets. Thus, the CCs of large nodes dominate the value of global CC. Mutual graph of Weibo has a maximal degree 2000. On the other hand, Twitter mutual graph contains nodes with much larger degrees because of the existence of privileged users. These large accounts have a huge number of triplets that are not closed. In other words, a network with higher degree will most probably have a lower global CC, regradless the CC values for most nodes.

Another observation we made is that CC is positively correlated with economic development. Panels C and D show the CC of HongKong, Beijing, and Sichuan. Panel C is for all the degrees, Panel D focuses on the degrees below 150. We can see an apparent pattern, suggesting that economically more development regions have a higher CC.

Conclusion: The reciprocity and transitivity of most Weibo users are significantly lower than those of Twitter, indicating that Weibo is more used as a news media than a social network platform.

V. PageRank Centrality

The importance of the bloggers can be measured by various social network centralities. Two of them are degree centrality and PageRank centrality. The top 20 most followed users and their PageRanks are plotted in Fig. 6. PageRank is calculated using the classic power-iteration method with damping factor 0.85 [25]. The power method iterates for 50 times.

One perplexing phenomenon is the occurrence of several Weibo service accounts in the top list, for both follower number and PageRank centrality. As a comparison, none of

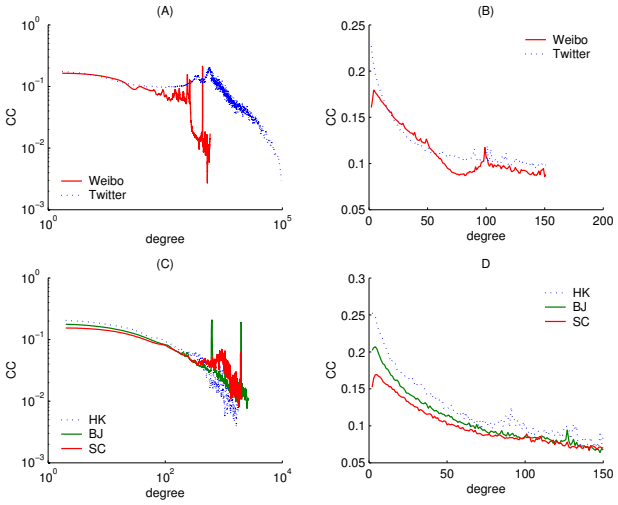

Fig. 5. Average clustering coefficient as a function of degree for mutual graphs in Weibo and Twitter. Panels A and C: smoothed with window size 20; B and D: degrees up to 150.

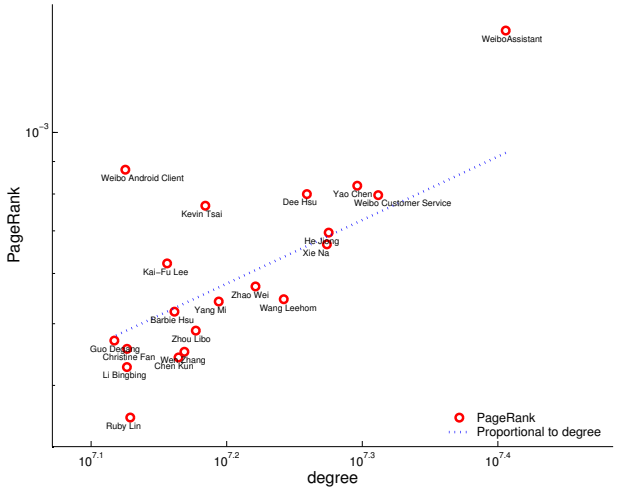

Fig. 6. Top 20 Weibo users with highest in-degrees and their corresponding PageRank values.

the Twitter service accounts come up to the top [15]. What is more startling is that these Weibo official accounts crop up even higher in terms of PageRank than in-degree. Three Weibo accounts are among the top 20 in-degree list, and five are among the top 20 PageRank list. Those official Weibo accounts attract not only huge amounts of followers, but also "important" ones. A detailed inspection of these accounts reveals that they share many followers. For instance, we find that the top two accounts (Weibo Assistant and Weibo Customer Service) share 43% of their followers, as shown in our web page [3].

The second observation is the correlation between in-degree and PageRank shown in Fig. 7. Panels A and B demonstrate that Weibo has a weaker correlation. Indeed the Pearson correlation coefficient is 0.82 for Weibo, and 0.94 for Twitter for the top 100 most followed users. This can be inferred from

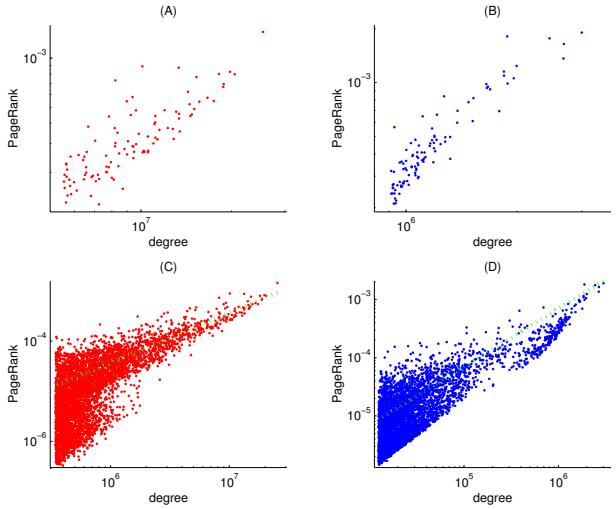

Fig. 7. PageRank as a function of degrees of the top 100 and 5000 most followed users. Left column: Weibo; right column: Twitter. Line: proportional to in-degree.

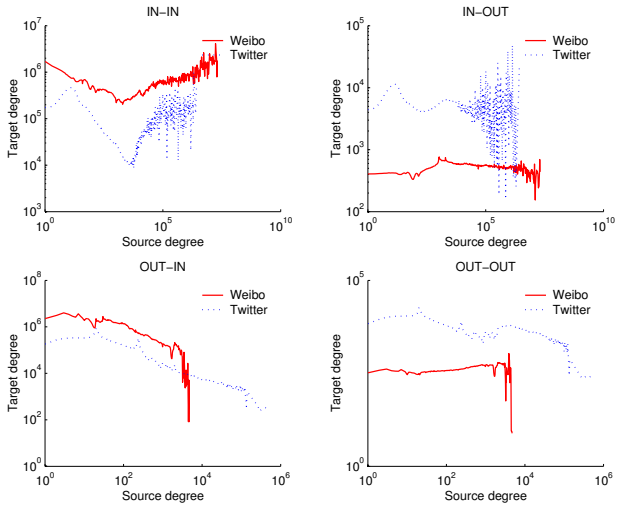

Fig. 8. Assortative mixing by degrees.

the fact that Twitter has a higher reciprocity. In an extreme case when every link were reciprocated, the graph would be undirected, and each PageRank value would be proportional to the degree.

Panels C and D draw lines where the PageRank is proportional to the degree. Nodes above the line attract "important" followers, while the ones below contain sub-quality followers. We can observe clusters of nodes whose PageRanks are substantially below the line, indicating possible spamming activities. Such anomaly happens for the top 200 Twitter accounts. For Weibo, those accounts mostly have less than one million followers.

Conclusion: The correlation between in-degree and PageRank is weaker in Weibo than Twitter. This can be derived from the high reciprocity of Twitter.

VI. Assortativity

Conventional wisdom tells us that celebrities socialize with celebrities. Such tendency can be measured by assortative mixing by degrees [23]. It was reported in [24] that social networks, such as citation networks, demonstrate positive assortative mixing, confirming the common sense widely perceived.

The question is whether such assortativity is carried on in large online networking sites such as Weibo and Twitter. Recently, [22] reported a surprising observation that popular users tend to follow unpopular users in Twitter user network, contradicting the common sense. This conclusion was drawn based on the negative Pearson correlation coefficient between logged in-degrees of the following relationship. We conducted the experiments on our two data sets, and observed negative coefficients as well. However, this coefficient is meaningful only for measuring linear correlations. To reveal the details of their relations, we plot the average in-(out-) degree of the target nodes as a function of the in-(out-) degrees of the source nodes in Fig. 8. For each in- or out-degree (x-axis), we collect all the source nodes that have that degree. Then we obtain all the target nodes that are linked from the source nodes. From those target nodes, we calculate the average in- and out-degrees (y-axis). Altogether there are four combinations, denoted by IN-IN, IN-OUT, OUT-IN and OUT-OUT in Fig. 8

The IN-IN plot shows that indeed there is no linear relation between the in-degree of a node and the average in-degree of the nodes it follows for both Weibo and Twitter. Instead, they exhibit an apparent V shape. The turning point is around 2000. Before that, smaller accounts tend to follow more popular bloggers. After receiving more than 2000 followers, popular bloggers do follow popular bloggers in general. This can be explained by the evolution of the network: users start by following celebrities, inducing higher average degree of their targets. When users accumulate more experience, they tend to connect with friends in their real life and people in their communities. This will drag down the average degree of the people they follow. This trend continues until the in-degree is around 2000. After that, there is a strong positive assortative mixing, i.e., the more popular you are, the more popular the people you follow are.

In addition to IN-IN correlations, we also studied other combinations as shown in Fig. 8. OUT-IN: the average target in-degree decreases almost monotonically with the out-degree of the source nodes, for both Weibo and Twitter. It reveals that the more people you follow, the less popular those people are. IN-OUT: the average target out-degree does not change with the in-degree of the source nodes. Regardless of the popularity of a blogger, the variance of their target out-degree is small. This is particularly true for Weibo, because it has a strict 2000 limit for the out-links. OUT-OUT: The more people you follow, the fewer targets these people follow.

Conclusion: In online social networks, celebrities tend to connect with celebrities more often in both Weibo and Twitter. This confirms our common sense in the real world,

and corrects the conclusion made in [22].

VII. Make the Links

The ultimate goal of studying an online social network is to link it to the real world. This paper gives two examples of linking Weibo to the real world.

A. Connections between Provinces

The connection between people in different provinces has never been studied quantitatively on population level. Thanks to the digitalization of the user relations, we can quantify the strength of the connection between provinces based on the Weibo user network. Mutual information (MI) is used to quantify such relation. Simpler similarity metrics such as the number of links or normalized versions such as in [27] could have been used, but they are hard to justify. Jaccard similarity and Dice index favour large provinces that have more Weibo users. MI measures the deviation from the independence of two random variables. It is a proved successful measurement text classification [31]. We use the following (normalized) MI to quantify the connection between provinces x and y:

$$MI(x,y) = \frac{\log \frac{P_{xy}}{P_x P_y}}{-\log P_{xy}}, \quad (1)$$

where the probabilities P_x and P_{xy} are estimated using the observed links as follows. Let n_{xy} be the number of links between provinces x and y, $n_x = \sum_y n_{xy}$ the number of links from x, and $N = \sum n_x$ the total number of links. Then $\hat{P}_x = n_x/N$, and $\hat{P}_{xy} = n_{xy}/N$. MI has an intuitive interpretation. It takes a positive value if x and y share more links than we expect by chance, and a negative one if they share less. $-\log P_{xy}$ is added to normalize the value so that the minimal (when they have no connection at all) is -1 and maximal (when they overlap completely) is 1.

We calculate the MIs of 34 provinces and regions on 72 million users whose profiles are crawled. Each profile contains self-claimed area/province data. Accounts with empty province data are excluded. The total number of cross province edges is $N = 5.32\times10^8$. Among them, $\max(n_x) = 8.65\times10^7$ (GD), $\max(n_{xy}) = 1.17 \times 10^7$ (GD and AB), $\min(n_x) = 930,000$ (QH), $\min(n_{xy}) = 1963$ (MC and QH). Due to the huge size of these statistics, the estimated MIs are regarded as accurate.

The pairwise MIs are plotted in Fig. 9. Provinces are sorted using hierarchical agglomerative clustering using average linkage. The results of single and complete linkages are similar. To highlight the accuracy of the clustering, we cut the dendrogram into six clusters, and plot them using different colours on the map of China. We want to emphasize that the clustering is accurate not only because the adjacent provinces are clustered together. More importantly, each cluster reflects the bondage in culture, dialect, and tradition. For instance, the closest pairs (the red pairs in the plot) of provinces/regions are (AB, GD), (CQ, SC), (HK, GD), and (XZ, QH). AB (abroad) represents oversea users. CQ and SC were in the same province only a few years ago; HK and GD are adjacent and speak the same dialect, the same for XZ and QH.

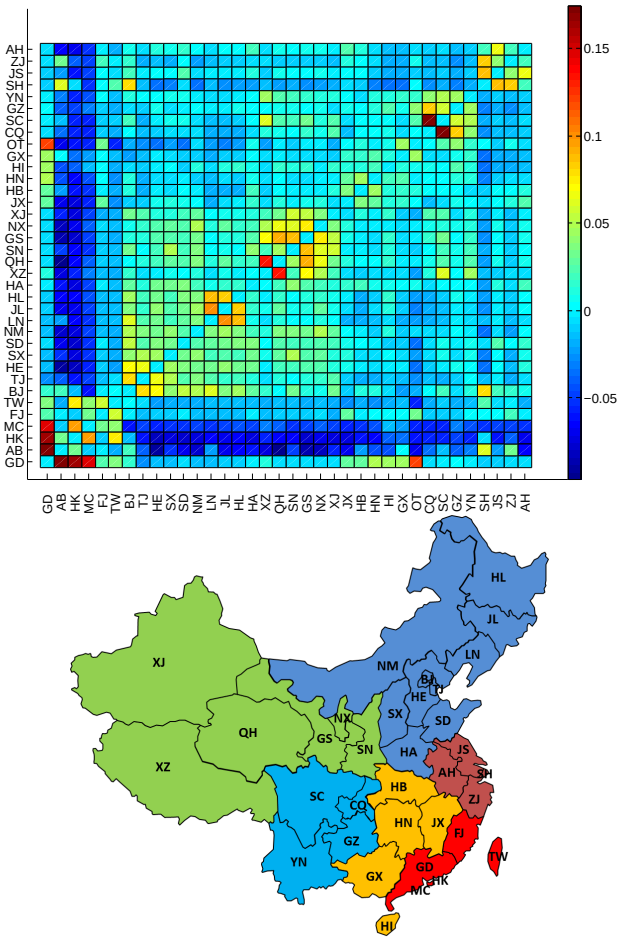

Fig. 9. MIs between provinces in China, and the corresponding six clusters in the map of China. Refer to [2] for province code.

Of all the links, slightly more than half (56.91%) are within provinces. This is in sharp contrast to countries, where 84% of edges are within countries [27]. The percentage varies widely across provinces. The highest is GD (78.55%), and the lower end includes XZ (10.56%), QH (13.74%), and TW (18.38%).

When all the users are grouped into the 34 regions, the modularity is 0.44. This is a value much smaller than the modularity grouped by countries (0.7486) [27]. It indicates that a much stronger bond exists between provinces than countries.

B. GDP vs. Weibo Penetration Rate

Regions that are willing to adopt a new technology should be a region technologically more advanced, henceforth economically more developed. We show that Weibo penetration rate is correlated with the economic development measured by GDP per capita. Provincial GDP in 2013 is from Statistics China [1]. Fig. 10 plots the provincial GDP per capita as a function of Weibo penetration rate. It demonstrates that there is a strong correlation, with Pearson correlation 0.76. Beijing, Shanghai, and Guangdong are the leading regions in Weibo

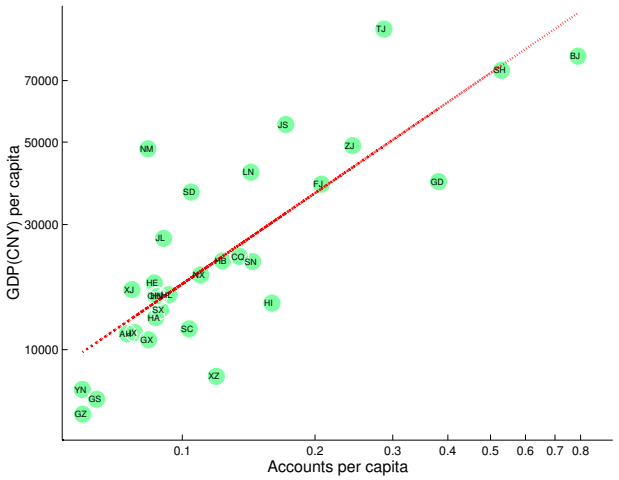

Fig. 10. GDP per capita vs. number of accounts per capita in log-log scale.

penetration rate.

VIII. Discussions and Conclusions

Weibo and Twitter are two of the major online social network sites. They share many similarities, in size, structure, and influence. Yet they barely overlap geographically. As the first comparative study of these two services, we find that Weibo is fundamentally different from Twitter in the way people interact. The arch reciprocity is smaller by a factor of 3.6 for Twitter 2009, and 4.2 for Twitter 2012. The clustering coefficient is also significantly smaller for most accounts. Overall, the interaction between people in Weibo is substantially weaker than that of Twitter.

The inequality of followers is also greater in Weibo. It is even higher than the inequality of wealth in any country in the world in terms of Gini index. 0.1% of the top bloggers attract almost 50% of the followers in Weibo. On average, a friend has 5567 (CV^2+1) times more followers than you have.

Anomalies are found in both Weibo and Twitter networks. Five Weibo service accounts are among the top 20 PageRank list. There are large accounts who share most of the followers. There are cliques containing thousands of nodes. Many Twitter privileged users reciprocate almost every follower. These anomalies highlight the necessity for independent studies of online social networks.

Thanks to the digitalization of human relationships, we can quantify the strength of connections between different provinces and regions in China. We find that mutual information is accurate in reflecting such bondage. Of particular interests is that Hong Kong has the lowest MI with most other provinces, even weaker than Taiwan, coinciding with its tenuous relationship with mainland China.

IX. Acknowledgements

This work is supported by NSERC (Natural Sciences and Engineering Research Council of Canada) Discovery grant (RGPIN-2014-04463), National Grand Fundamental Research 973 Program of China under Grant No. 2014CB340402, and NSFC No. 61433008, U1435216.

References

[1] http://data.stats.gov.cn/.
[2] http://en.wikipedia.org/wiki/Provinces_of_China.
[3] http://jlu.myweb.cs.uwindsor.ca/spammer/view_node-1642909335.
[4] http://pacman.cs.tsinghua.edu.cn/~hanwentao/weibo/.
[5] http://pacman.cs.tsinghua.edu.cn/~hanwentao/weibo/world.png.
[6] A. Clauset, C. R. Shalizi, and M. E. Newman. Power-law distributions in empirical data. *SIAM review*, 51(4):661–703, 2009.
[7] R. I. Dunbar. Neocortex size and group size in primates: a test of the hypothesis. *Journal of Human Evolution*, 28(3):287–296, 1995.
[8] N. Eagle, M. Macy, and R. Claxton. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.
[9] S. Feld. Why your friends have more friends than you do. *American Journal of Sociology*, pages 1464–1477, 1991.
[10] K.-w. Fu and M. Chau. Reality check for the chinese microblog space: a random sampling approach. *PLOS ONE*, 8(3):e58356, 2013.
[11] Q. Gao, F. Abel, G.-J. Houben, and Y. Yu. A comparative study of users? microblogging behavior on sina weibo and twitter. In *User modeling, adaptation, and personalization*, pages 88–101. Springer, 2012.
[12] J. Giles. Making the links. *Nature*, 488(7412):448–450, 2012.
[13] Z. Guo, Z. Li, and H. Tu. Sina microblog: an information-driven online social network. In *Cyberworlds (CW), 2011 International Conference on*, pages 160–167. IEEE, 2011.
[14] R. A. Hanneman and M. Riddle. Introduction to social network methods, 2005.
[15] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, pages 591–600. ACM, 2010.
[16] J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. In *Proceedings of the 17th international conference on World Wide Web*, pages 915–924. ACM, 2008.
[17] J. Lu and D. Li. Sampling online social networks by random walk. In *ACM SIGKDD Workshop on Hot Topics in Online Social Networks*, pages 33–40. ACM, 2012.
[18] R. Meusel, S. Vigna, O. Lehmberg, and C. Bizer. Graph structure in the web—revisited: a trick of the heavy tail. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 427–432. International World Wide Web Conferences Steering Committee, 2014.
[19] A. Mislove, M. Marcon, K. Gummadi, P. Druschel, and B. Bhattacharjee. Measurement and analysis of online social networks. In *SIGCOMM*, pages 29–42. ACM, 2007.
[20] M. Montemurro. Beyond the zipf–mandelbrot law in quantitative linguistics. *Physica A: Statistical Mechanics and its Applications*, 300(3):567–578, 2001.
[21] P. Mozur. How many people really use sina weibo?, 2013.
[22] S. A. Myers, A. Sharma, P. Gupta, and J. Lin. Information network or social network?: the structure of the twitter follow graph. In *WWW*, pages 493–498, 2014.
[23] M. Newman. *Networks: an introduction*. Oxford University Press, Inc., 2010.
[24] M. Newman and J. Park. Why social networks are different from other types of networks. *Physical Review E*, 68(3):036122, 2003.
[25] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. 1999.
[26] E. H. Simpson. Measurement of diversity. *Nature*, 1949.
[27] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.
[28] B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, August 2009.
[29] H. Wang and J. Lu. Detect inflated follower numbers in osn using star sampling. *The IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pages 127–133, 2013.
[30] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. Acm, 2009.
[31] Y. Yang and J. O. Pedersen. A comparative study on feature selection in text categorization. In *ICML*, volume 97, pages 412–420, 1997.
[32] G. Zipf. Human behavior and the principle of least effort. 1949.