SAMPLING WEIBO

Ronald S. Burt
Booth School of Business
University of Chicago

Hong Huang
Institute of Computer Science
University of Goettingen

Jie Tang and Jing Zhang
Department of Computer Science
Tsinghua University

ACKNOWLEDGEMENT: This is a research note written in preparation for substantive analysis of the Zhang et al. (2015) Weibo network data. Professor Burt and Hong Huang are grateful to the University of Chicago Booth School of Business for financial support during the work reported here. Professor Jie Tang and Jing Zhang provided the Weibo data. The current manuscript can be downloaded from http://faculty.chicagobooth.edu/ronald.burt/research/files/SW.pdf.

SAMPLING WEIBO

We search a large Weibo database to identify parameters that can productively guide sampling strategies for network analyses of specific substantive questions. Two parameters stand out, and are probably significant for network analyses of social media more generally: time and clustering. The most consequential sampling parameter is social time, the distinctions between early, bandwagon, and late adopters. User networks are different in size, growth, decay, and composition by the period during Weibo's diffusion when a user opened a Weibo account, and there are well-known substantive differences between early and late adopters. Fortunately, the Weibo data provide a robust distinction between early and late adopters. Second, the networks around individual users are often differentiated into distinct clusters. Sample data imply there are two or three clusters in most Weibo networks. Clustering has implications for estimating interpersonal influence and network dynamics, so modularity scores could be useful to hold constant clustering differences between users. Also significant are kind of user (common, star, celebrity, organization), and geographic location (Beijing, Guangdong, Shanghai, other China, and overseas). At minimum, models of making or breaking connections in Weibo should be tested for consistency across the sampling parameters.

Our data come from Weibo, a Chinese social media site combining elements of Twitter and Facebook. Users post short messages, re-post messages, and follow other users to have their posts and other user comments appear in the user's own timeline. Posts can be in simplified Chinese, traditional Chinese, and English. In addition, posts can contain user images, music, and video files. With half a billion registered users at the time our data were gathered,[1] and 100 million messages posted daily, Weibo is one of the world's largest micro-blogging websites, benefiting in part from Chinese restrictions on Western social media, and in part from strategic recruiting of Chinese celebrities to the platform at its launch.[2]

More specifically, our data come from a large snowball sample of Weibo users created by Zhang et al. (2015). One hundred seed users selected at random were

---

[1]As of December, 2012, Weibo had 503 million registered users (from Wikipedia entry for "weibo," retrieved June 17, 2015). In comparison, Twitter had 185 million registered users at the end of 2012 (from Statistica website, retrieved March 23, 2016).

[2]Weibo was launched by Sina Corporation in August 2009, right after microblogging services outside China – such as Twitter, Facebook, Plurk, and Fanfou [Sina's precursor to Weibo] – were blocked to Chinese users in July 2009 following the Ürümqi riots.

traced to others who the seed users followed, and then to still others who the followees followed. The three sampling steps provided 413,503,687 "follow" connections within the snowball sample as of August 28, 2012. Users in the initial sample were then tracked for the next month — the observation period — to see who they added to their networks with "follow" citations, and who they deleted from their networks with "unfollow" citations. On average, each day during the observation period provided 330,110 follow citations and 349,338 unfollow citations. The additional citations brought additional users into the snowball sample, for a total of 1,787,443 users. The web was also crawled for user profiles which define kind of user (common user, star user, celebrity, organization), gender, date when user account was created, city, province, cumulative number of messages posted by the user, number of contacts the user follows in all of Weibo, number of contacts in all of Weibo who follow the user, and number of mutual ties between the user and other Weibo users (user follows other and other follows user).

Illustrative Weibo networks are displayed in Figures 1 and 2 as sociograms of the ego-networks around two users. Dots in each sociogram represent contacts (other users followed by the focal user, ego). Lines in the sociograms indicate how contacts are connected with each other (ego is deleted to clearly display the network structure among contacts). Two contacts appear close together to the extent that the contacts are strongly connected to each other and to the same other contacts (spring embedding algorithm in NetDraw, Borgatti, 2002). A social-media connection is created when one user registers to follow the activity of another user. A disconnection is created when the user registers to stop following the other user. Thick lines in the sociograms indicate contacts who follow one another. Thin lines indicate a connection in which only one contact follows the other. The absence of a line between two contacts indicates that neither follows the other (they are both in the sociogram because both are followed by the focal user, ego).

——— Figure 1 and Figure 2 About Here ———

The sociograms are for two Weibo users, one who adopted late and the other early. Figure 1 describes a median-sized network for a male in Guangdong who

adopted Weibo after the service was well established (Weibo's diffusion is discussed below). The sociogram and network statistics in the figure show predominantly local contacts (79% Guangdong) connected by strong relations (13.6% of possible connections are present [density, aka "local clustering"], and 50% of connections are mutual), anchored on a single dense cluster (e.g., Burt, Kilduff, and Tasselli, 2013, for details on the network statistics). In contrast, Figure 2 describes a median-sized network for a Guangdong male who began using Weibo soon after the service became available. Figure 2 looks more like the broker network expected of an opinion leader (Burt, 1999; 2005:84-86): The sociogram shows three clusters of contacts rather than one, there are many contacts outside the user's city (83% not Guangdong), and connections between contacts are typically not strong (91.7% of possible connections are missing [8.3% density], and 78% of the connections present are not mutual).

It would be an extravagant use of time to search through the 414 million relations to build networks around each of the nearly two million users — extravagant in the amount of time needed, and extravagant in that reliable inferences can be drawn from sample networks without having to analyze every network on which data are available. For example, the software used to obtain network metrics for this note assembles a user's network by searching through the snowball data to extract a user's contacts and connections between the contacts, then analyzing the assembled network. It took 10 minutes to process the network in Figure 1 and 12 minutes to process the network in Figure 2. At a rate of 11 minutes per network, running all users in the snowball data would require a little more than 37 years. The rate is made practical by removing snowball data not relevant to a sample of users. We drew a sample of 1500 users to study network decay. Network assembly can be limited to snowball data on the sample users and their contacts and their contact's connections. The 11-minute per network rate drops to just under a minute, which means computations for the 1500 sample users can be run overnight. Our purpose in this note is to find population parameters that can productively guide the selection of sample users for network analyses of specific substantive questions.

SNOWBALL REFLECTS POPULATION

The snowball sampling did well in capturing the relative frequency with which sample users were connected more broadly in Weibo. Figure 3 plots mean numbers of connections for the sample users. Celebrities and organizations are excluded from Figure 3. Sample users are ordered on the horizontal axis by the date at which they created their Weibo account. To the far right, there are users who signed up with Weibo on the first day the service was available, August 14, 2009 (slightly more than 36 months before the snowball-sample observation period). To the far left, there are users who joined Weibo less than two months before the observation period. There are only a few hundred users at the right and left extremes of Figure 3, but the average month contains several thousand users in the snowball sample (average of 41,617 users per month).

——— Figure 3 About Here ———

Representation for Individual Users

The top two graphs in Figure 3 plot network size in and beyond the snowball sample. Taking all registered Weibo users as potential contacts, the top graph in Figure 3 shows average outdegree (number of others followed), indegree (number of others following posts by the user), and mutuals (number of others following, and followed by, the user). The second graph in Figure 3 shows the same scores when connections are limited to other users captured by the snowball sampling. The all-Weibo networks are of course larger. The average sample user follows 224.5 other sample users, but also another 267.2 Weibo users outside the sample (hollow-dot lines in Figure 3). Similarly, the average sample user has 54.6 mutual connections within the sample, and another 136.0 outside the sample (solid-dot lines in Figure 3). Indegree is most underestimated by the snowball data: the average user has 144.2 followers within the sample, and another 4,650.2 in the broader population of Weibo users (triangle-dot lines in Figure 3).

Regardless of mean differences, the sample users with larger networks within the snowball sample have larger networks in the broader Weibo population. For outdegree, indegree, and mutuals respectively, the correlation is .85, .78, and .84 between scores

within the sample versus scores within Weibo more generally. To make a good guess about a user's number of connections in all of Weibo, the user's connections in sample should be multiplied by 1.54 for outdegree, 2.01 for mutuals, and 31.12 for in degree (these are regression coefficients predicting all-Weibo degree from in-sample degree, excluding sample celebrities and organizations; 1.50, 1.87, and 54.06 respectively if celebrities and organizations are included).

Representation for Kinds of Users

Relative network size around kinds of users are also similar in and beyond the snowball sample. Table 1 shows mean degree, messaging, and changes beyond (left) and within (right) the sample for kinds of users we distinguish with the user profile data (the user distinctions in Table 1 are discussed in a moment). For example, sample common users in the first row of Table 1 cite an average of 218.8 other sample users to follow. They cite an average of 432.0 other users across Weibo more generally. The relative mean numbers of citations are correlated down corresponding columns for the 15 kinds of users distinguished down the rows of Table 1. With respect to number of followers, for example, mean indegree from within the snowball sample is correlated .99 with mean indegree from all Weibo users. Kinds of users who have numerous followers within the snowball sample also have numerous followers within Weibo more generally. With respect to number of people followed, mean outdegree within the snowball sample is correlated .78 with mean outdegree across all Weibo users. And mean number of mutual connections in sample is correlated .88 with mean number of mutuals across Weibo.

——— Table 1 About Here ———

Representation in Making and Breaking Connections

Third, patterns of making and breaking connections over time are similar in and beyond the snowball sample. Consider Figure 4, which shows monotonic associations between user time in Weibo and the cumulative number of messages a user posts. For everyday people (common and star users plotted in Figure 3), celebrities, and organizations, Figure 4 shows user cumulative number of messages increasing continuously with user time in Weibo. In fact, cumulative number of messages is a good indicator of how long

a user has been using Weibo (98%, 95%, and 86% of the variance in message volume by persons, celebrities, and organizations respectively is predicted from user months in Weibo and months squared).

——— Figure 4 About Here ———

In contrast to the monotonic association in Figure 4, network statistics in Figure 3 show three distinct periods of user experience. We use dashed vertical lines in Figure 3 to distinguish the three periods. The specific times distinguishing the periods could be shifted a little earlier or a little later. Our interest is not in the boundaries between periods so much as the network activity characteristic within each period.

To the left in the graphs, "late, inexperienced" users have the most volatile networks. They rapidly add connections and are most likely to disconnect during their first year using Weibo. This first-year learning period is evident in both graphs at the top of Figure 3. It is also evident in the graph at the bottom of Figure 3, which shows the probability that a user will change a relationship, and the probability that a contact will be unfollowed.[3] Both probabilities are high during the initial six months of using Weibo, then go into a long, continuous decline as users become more experienced in making connections.

We use "late" and "inexperienced" to label these users, however, we cannot say whether the users exhibit the behavior they do because they have less experience with

[3]Probabilities here measure a user's tendency to change connections. Probability of change is the number of changes made by a user during the observation period divided by the number of people the user followed during the period. Probability of disconnect is the number of unfollow actions a user took during the observation period divided by the number of people the user followed during the period. Adding and deleting contacts need not be a one-time event during the period. Most relations that were changed were changed once – either added during the observation period and kept throughout the period, or deleted and left deleted throughout the period. Of 18 million relations changed during the observation period, 87.8% were only changed once. Another two million were changed twice; either added then deleted, or deleted then added back. The extreme is three relationships changed 22 times. All three were from male star users to celebrity users (see Table 1 below for types of users). For example, one of the three relations was added during the first day of the observation period, then deleted on day two, then added again on day three, then deleted on day four, then added on day five, then deleted two days later, then added back the next day, then deleted two days later, and so on. The additions and deletions summed to 22 changes in the user's relationship with the celebrity.

Weibo, or because they are a kind of person who adopted Weibo late. The volatility of their networks in Figure 3 could be a symptom of inexperience, or their late adoption. Evidence of their late adoption is given in Figure 5. Weibo adoptions followed the usual S-shaped curve describing the spread of a new idea through a population (Rogers, 2003). The horizontal axis in Figure 5 is reversed from Figures 3 and 4 to put early adopters to the left of the graph. In the first month that Weibo was available — the extreme left in Figure 5 (36 months before the observation period), an initial 7,991 users in the snowball sample began using the service. From then on, the line of hollow dots in Figure 5 shows the cumulative number of users increasing with a few new adopters each month, then spreading quickly month-after-month through a bandwagon period, then slowing as most of the sample users had already adopted. Late, inexperienced adopters are to the right in the graph, past the vertical dashed line marked B. By the time they began using Weibo, the rapid spread of Weibo was established and there were relatively few new users each month.

The late adopters are not defined as “late” because they began to use Weibo two or more years after it became available. Absolute time is not the essential criteria for their distinction. What makes them late adopters is the fact that they adopted after the bandwagon rush to Weibo was over (flat hollow-dot line to the right of dashed-line B relative to the preceding steep line preceding dashed-line B).

——— Figure 5 and Figure 6 About Here ———

The emphasis on social time rather than physical time is a significant point because it means that distinguishing late from early adopters in the snowball data is not dependent on when the snowball data were gathered. The point is illustrated by the two graphs in Figure 6. In Figure 6A, we extrapolate Weibo’s diffusion for another twelve months past the September, 2012 observation period during which the snowball data were gathered. In the last four months of the observation period, the snowball sampling captured an average of 5,643 users each month who had opened their Weibo account just that month. The dashed line in Figure 6A shows how the new users would have expanded the snowball population by another sixty-seven thousand users if the data had been gathered a year later, in September 2013. But all of the additional new users

would have been “late” users — the bandwagon rush to Weibo ended about two years after the service had been available. In other words, the socially-defined “late” users defined by the snowball data are “late” adopters regardless of when the snowball data were gathered.

That statement is true if diffusion continues at a constant rate after the September 2012 observation period. We do not know what the snowball sampling would have captured, but the available data on Weibo’s regional diffusion provides reassurance. Weibo’s diffusion would not be consistent, for example, if the service diffused rapidly in Beijing, then began diffusing in Shanghai a couple months later. If the diffusion curve in Figure 5 were an aggregation of distinct regional diffusion curves, then a late adopter in one region need not be a late adopter in another region. Figure 6B shows that Weibo’s diffusion curves were very similar for the five regions distinguished (discussed below) in that the rush of new users to Weibo had slowed to a late-adopter trickle in all five.

At the other end of the diffusion process, there are similarly “early, experienced” users. Early, experienced users adopted Weibo in the first six months after it was available. They have large, growing networks in which users are unlikely to disconnect from their contacts. The top two graphs in Figure 3 both show increasing in and out degree for experienced users to the right in the graphs, and the bottom graph shows the lowest probability of discontinued relationships. Again, we cannot say whether these experienced users exhibit the behavior they do because of their longer time in Weibo, or because they are the kind of users who first joined Weibo. The ambiguity is more obvious for organizations. Experienced organizations to the right in Figure 4 post a large number of messages, which could result from their longer experience with Weibo, or a tendency for public-relations conscious, message-posting organizations to join Weibo as soon as the service was available.

Regardless, the diffusion curve in Figure 5 is again useful here. We define early, experienced users to be those who began using Weibo during the first six months that it was available (that is, sample users who had 30 months or more of experience before the observation period). But the growing, rare-decay networks characteristic of early, experienced users can be seen in Figure 3 for users who entered seven, eight, nine,

even ten months after Weibo was launched. We selected six months as the boundary time based on the results in Figure 5. The solid dots at the bottom of the graph show the growth rate of new users for each month during Weibo's diffusion. Growth rate is measured this month by the number of new users who adopted next month minus the number who had already adopted last month. The fastest growth for Weibo was between 14 and 18 months after the service was released. The solid-dot line in Figure 5 shows that new sample users were entering at a rate of almost a hundred thousand each month.[4] New users entered more slowing in the months before and after the peak between 14 and 18 months after Weibo was released. To distinguish early adopters, we looked for the early month in which growth matched the growth distinguishing late adopters. Late adopters are distinguished by the 12-month experience marker in Figure 3 and the 24-month late-entry marker in Figure 5. Follow the late-adopter dashed line B in Figure 5 down to the solid-dot line describing monthly growth. At 24 months, new users were entering at a rate of 90,802 every two months. The corresponding growth early in Weibo's diffusion is at six months, when growth is 86,581 new users every two months. Again, there is no substantive importance to the six months in physical time. What makes the early adopters distinct is the fact that they began using Weibo before it was clear that the platform would become a national standard, before the bandwagon rush of new users to the platform. The six months of experienced users to the right in Figure 3 show the growth and rare decay characteristic of early, experienced users. The first six months of Weibo diffusion in Figure 5 show the tentative pre-bandwagon adoptions of early adopters. The extrapolated graph in Figure 6A shows that the same people would be early adopters if the observation period were twelve months later than the one used by Zhang et al. (2015), assuming a constant rate of diffusion.

---

[4]Specifically, new users were entering at a rate of 90,441 per month. Growth through month 15 is the cumulative number of sample users who began using Weibo in month 16, minus the number who had begun by month 14 — which is the 190,581 for month 15 on the solid-dot line in Figure 5. Averaging the corresponding growth for months 14 through 18 yields an average 2-month growth of 180,881, or 90,441 per month. Growth is slower in the months immediately before and after the peak during months 14 through 18 (148,576 is the 2-month growth through month 13 and 146,100 is the growth through month 19).

Between the early and the late adopters are the bandwagon users — users in the middle of Figures 3, 4, 5, and 6, who began using Weibo when new users were rushing at peak rates to the platform.

SAMPLING PARAMETERS

We are guided in our search for sampling parameters by the two often-observed correlates of strong relations in social networks: closure and homophily. With respect to closure, strong relations are more likely to occur, and less likely to decay between people embedded with numerous mutual friends in a closed network (Burt, 2005: Chps. 3-4, for review). With respect to homophily, strong relations are more likely to occur, and less likely to decay, between people who share attributes relevant to a relationship (Mcpherson, Smith-Lovin, and Cook, 2001, for review).

Four Kinds of Users

Four kinds of users distinguished by the Weibo platform are listed at the top of Table 1. On average, the largest networks — especially with respect to having numerous followers — are around organization and celebrity users. An "Organization" is an account verified by Weibo to be the organization purported to be reached through the account. The majority of organization users are businesses (63% of the sample organization users), but the category also includes schools, universities, government agencies, media companies, websites, clubs, and other kinds of represented groups (77,926 organization users in the sample). The most prominent organization user in the sample data is Weibo customer service, which had over 41 million users following its posted messages during the sample observation period, and agents of the organization followed messages posted by 81 other users (almost all organizations).[5] A "Celebrity" is an account verified by Weibo to represent the celebrity purported to be reached through the account. Weibo management determines who is a celebrity, and persuaded many

[5]Of the 81 users followed by Weibo customer service, 70 are in the snowball sample. All but one of the 70 was another organization user.

prominent Chinese celebrities to become Weibo users. Celebrity users include movie stars, singers, business and religious figures, athletes, scholars, artists, and prominent government officials (162,784 celebrity users in the snowball sample). The most prominent celebrity user in the sample data is actress Chen Yao, who had 29,382,466 users following her during the sample observation period. Ms. Yao was registered as following messages posted by 690 other users (most of whom were other celebrities).[6]

Then there are everyday people sorted by their level of activity on the Weibo platform. Users earn points each time they sign into Weibo and send original messages. Three-star users have more than a thousand points, two-star users have 200-999 points, and a one-star user has up to 199 points (502,150 star users in the sample). Everyone else is a "common user" (912,818 sample users). Common users tend to be more prominent than star users in terms of having more followers, but star users have more volatile networks in the sense of being more likely to add new contacts and delete current contacts.

Early Versus Late Adopters

As explained with Figures 5 and 6, the people and organizations who adopted Weibo in the first six months it was available can be distinguished as early, experienced users. Table 1 shows that early users have larger networks, and relations in their networks are at lower risk of decay than relations in the networks of late adopters. Users who began more than 24 months after the platform was available can be distinguished as late, inexperienced users. Table 1 and Figure 3 show that, relative to the average Weibo user, late users have smaller, but faster growing, networks combined with the highest probability of decay.

In addition to the differences already described, early and late users can be distinguished by the kinds of contacts they follow. Figure 7 shows the composition of user networks by when a user created a Weibo account. To the left in the graph, users

[6]Of the 690 users Chen Yao was registered as following, 666 are in the snowball sample. Most are other celebrity users (77.6%), with some non-celebrity people (14.7%), and a few organizations (5.7%). The remaining 13% are users not found when user profiles were scraped.

who began using Weibo late in its diffusion typically follow common users, decreasing attention to celebrities, and increasing attention to star users. To the right in the graph, early Weibo users typically follow celebrities, giving the least attention to common and star users. The difference in preferred contacts seems to be a selection bias more than a correlate of experience. Figure 8 shows that the users who began using Weibo early tend — during the observation period three years later — to follow other users who began early.[7] In other words, the preference for celebrities among early, experienced users is a preference for celebrities who were also early users. Although numerous celebrity users adopt Weibo as the platform becomes popular, early adopters prefer to follow celebrities who were also early users (celebrities followed by early adopters began using Weibo five months ahead of celebrities followed by late adopters, respectively 22.1 months with Weibo on average versus 27.5 months). At the other end of the diffusion process, users who adopted late tend to follow other users who adopted late. In other words, the preference that late, inexperienced users have for common users is for other common users new to Weibo.

——— Figure 7 and Figure 8 About Here ———

Five Geographic Areas

Users provide their city and province when they register with Weibo. Not surprisingly, the largest concentrations of sample users are in the three Chinese population concentrations: 320,549 sample users live to the north in Beijing or the surrounding Hebei and Tianjin provinces, 341,194 sample users live to the south in Guangdong or the adjacent Hong Kong province, and 308,064 sample users live to the east in Shanghai or the adjacent Jiangsu and Zhejiang provinces. The next two concentrations are substantially smaller: 56,242 sample users live in Sichuan province to the southwest of China, and 55,636 live in Fujian province, sandwiched between the Guangdong and

[7]For each user, we averaged the months since each of the user's close contacts had created their Weibo account. A close contact is a person or organization that the user follows and was followed by (bold lines in Figures 1 and 2, "mutuals" in Figure 3). Figure 8 shows how the distribution of contact time with Weibo increases as the user's time with Weibo increases.

Shanghai areas. We focus on the three population concentrations, distinguishing also Weibo users who live outside China. Table 2 shows the percentage of each kind of user in each geographic area. We focus on broad differences between geographic areas since routine statistical inference is not useful here, given the several million observations, and those observations interdependent from the snowball sampling design.

——— Table 2 About Here ———

The Beijing area is characterized by larger networks (Table 1) around a disproportionate number of celebrity users and early adopters of Weibo (Table 2). Beijing has the smallest percentages of common and star users. The regional diffusion curves in Figure 6B show that Weibo diffused much more quickly in Beijing than elsewhere. The early participation of Beijing celebrities presumably built momentum for Weibo's rapid diffusion across China.

The Guangdong area is characterized by a disproportionate number of star users and a lack of early adopters (Table 2). Users here maintain average size networks, prone to additions and deleted connections (Table 1). Weibo became established in Guangdong during the bandwagon period of Weibo's diffusion. Table 2 shows 84.3% of Guangdong sample users adopted during the bandwagon period and Figure 6B shows that Weibo's diffusion began later in Guangdong than elsewhere. The lack of early adopters in Guangdong, and the disproportionate number of changes in Guangdong connections, offers another explanation for the volatility in the networks of late adopters; perhaps it is the late entry of change-prone Guangdong users that is responsible for the volatility displayed to the left of the graphs in Figure 3.

The Shanghai area is characterized by larger networks (Table 1), but without Beijing's concentration of celebrities (Table 2). Common and star users in Shanghai just have larger networks than the average user outside Beijing. For example, the average common user in Shanghai has 7,783.0 followers versus 5,005.5 for the average user elsewhere outside Beijing.

Users networks overseas are characterized by numerous followers and changing connections (Table 1). Common users, disproportionately female, are particularly

typical overseas (Table 2). In other words, Weibo seems to be a popular platform for everyday Chinese ex-pats staying in touch with other ex-pats. The inference is supported by the concentration of connections among overseas users. The snowball data show 3,005,552 connections between overseas users (observed frequency, f). If connections occur independent of geography, there should be 1,443,994 connections between overseas users (expected frequency, ef, is the probability of a connection from an overseas user [.04239088], times the probability of a connection to an overseas user [.08992549], times the total number of connections observed with known location, 378,800,059). The ratio of observed frequency over the frequency expected under independence is a measure of homophily — the extent to which people of a kind prefer connections with other people of the same kind. A ratio of 1.0 means that the observed number of connections is consistent with connections being independent of whatever categories are used to distinguish kinds of connected people. Ratios over 1.0 indicate homophily, a preference for connections with people like oneself. Ratios of observed connections over number expected under independence are central in Blau's models of social structure, generating visual displays of differential attachment described as "Blau space" (Blau, 1977; Blau and Schwartz, 1984, for initial work; McPherson and Ranger-Moore, 1991, on Blau space). The above number of observed connections between overseas users is more than twice the number expected if connections were independent of geography (f/ef is 2.08), revealing a preference among overseas users for other overseas users.

——— Table 3 About Here ———

Table 3 displays homophily ratios within and across geographic areas. The ratios in the first row are all larger than 1.0, showing that users in every area prefer connections with other users in the same area. The tendency is most pronounced for overseas users, but Guangdong users are close behind in their high preference for other Guangdong users. Common and star users have no special preference for other such users, but celebrities prefer to follow other celebrities and organizations prefer to follow other organizations. There is no tendency for women to prefer following women, or men to prefer following men. Extending the adoption-date homophily displayed in

Figure 8, early, experienced users prefer to follow other early, experienced users. Stronger still is the preference that late, inexperienced users have for other late, inexperienced users. The strongest homophily preference in Table 4 is the tendency for late Beijing users to follow the posts of other late Beijing users.

Fourteen Geographic Locations

There is differentiation within the five geographic areas that could be useful in sampling. Before aggregating users into the five geographic areas in Tables 1, 2, and 3, we analyzed connections within and across 14 geographic locations in Table 4 and Figure 9. Figure 9 is a "Blau space" in which connections are more likely between people in locations adjacent in the space. The space is a NetDraw scaling of the ratios in Table 4. Connections are symmetric (we were interested in relative strength of connections between locations before aggregating them into broader areas). Results for connections between common and star users are in the lower-diagonal cells (and used to generate Figure 9). To look for consistency with Weibo users more generally, the upper-diagonal cells contain results on connections involving celebrity and organization users.

——— Table 4 and Figure 9 About Here ———

Locations aggregated together are adjacent in the table. For example, the Beijing area in the earlier tables combine the first three rows and columns in Table 4. The Guangdong area best illustrates what we were looking for in the aggregation. Users in Guangdong province are combined with users in two cities surrounded by the province, Hong Kong and Macao. The diagonal cell in the fourth row shows Guangdong user preference for other users in Guangdong (1.29 ratio for all users), Hong Kong and Macao users prefer connections with other users in Hong Kong and Macao (2.06 ratio for all users), and outside their own location, users in both locations prefer connections with users from the other location over users in any of the other 12 locations (1.42 ratio for all users, the largest off-diagonal in the fourth and fifth rows and columns). Given the frequent connections between users in the two locations, the two appear close to one another in the northeast corner of Figure 9. Similarly, the three locations combined

as a Beijing area are close together in the northwest corner of Figure 9, and the three locations combined as a Shanghai area are together in the southeast corner of Figure 9.

We take three points from the results. First, there is substantial evidence of location homophily. Users connect more often with other users in their area, and connect still more with users in the immediate location. The ratios in Table 4 average 2.25 in the 28 diagonal cells, drop to a lower 1.03 average in the 28 off-diagonal cells between locations aggregated into the five broad geographic areas used in Tables 1, 2, and 3, and drop still further to a 0.94 average in the 154 off-diagonal cells between the five areas.[8] It is not surprising to see location homophily, but its obvious existence in the Weibo data means there is a substantive research design choice between sampling users from broad areas or from more narrow geographic locations within the broad areas.

Second, putting aside connections within areas, connections between areas are close to random. Off the diagonal in Table 4, most ratios of observed to expected frequencies are within a decimal place of complete independence from geography. Of 182 off-diagonal cells, 112 are between .9 and 1.1 (61.5%). In short, connections with a user outside one's own location are pretty well predicted by how many connections the outside user has, regardless of the outsider's geographic location.

But the prediction would be incorrect in underestimating connections between certain locations — represented by heavy lines in Figure 9. The thin lines in Figure 9 indicate where connections are close to random between locations. No line between

---

[8]The preference for local contacts continues past the 14 locations in Table 4, down to the district level within cities. We tabulated observed and expected frequencies of connections within and across districts in the municipal cities. Some districts we put aside because they had so few Weibo users that the expected frequency was less than one, which meant that one or two connections in the district generated extreme ratios of observed to expected. The average district for which we computed homophily ratios had hundreds of connections. In Beijing, the average homophily ratio is 6.48 for 13 districts. In Tiangjin, the average is 13.54 for 18 districts. In Shanghai, 8.85 for 17 districts. In Chongqing, 6.12 for 14 districts. We also computed district homophily in Guangdong province; 21.91 average homophily ratio for 18 districts (perhaps so much higher because named subdivisions in the province surrounding Hong Kong are likely discrete villages or towns). These are all very high and based on thousands of connections, showing strong preference for connections with others in one's immediate area.

two locations indicates a lack of connections between users in the two locations. Heavy lines indicate more connections than would be expected if connections occurred independent of geographic location. The heaviest lines indicate the five strongest connections in the sociogram. Our third point from the location results is that there is noteworthy geographic variation within the five broad geographic areas.

Begin with the three largest population areas. Each area holds a corner of the space in Figure 9. The closer together the locations combined in an area, the more similar their user connections with users in other locations. To the northeast in Figure 9, the Guangdong area most serves as a portal to the world beyond the China mainland, connecting with users outside and inside the mainland. To the northeast are Hong Kong, Macao, Taiwan, and users in the other overseas locations. Guangdong province is the core location for volume and connections in and outside the mainland. Hong Kong, Macao, and other overseas locations are distinct for their relative disconnection from the mainland. Taiwan is especially different for its extensive connections into the mainland. To the southeast in Figure 9, Shanghai city is the core location in the Shanghai area for connections to Bejing and overseas. Zhejiang province is especially different for its relative isolation from other locations. To the northwest in Figure 9, Beijing city is the core location in the Bejing area for connections into China and its lack of connections to outsiders, direct or indirect through Guangdong. Adjacent Tianjin city is especially different for its relative lack of connections into China and its strong connections with Taiwan.

None of the large population centers are in the center of Figure 9. The scaling algorithm puts in the center of the space the network elements most connected to other elements. We expected to see Beijing in the center of the space. Beijing has a concentration of early Weibo adopters and a connection drawn at random from the Weibo data is most likely to involve a user in Beijing (right-most column in Table 4). Regardless, none of the three large population areas are in the center of the Figure 9 space. Each stands structurally distinct in a corner of the space. What they have in common is disproportionate connections to users in provincial China. Provincial China stands in the center of Figure 9. The “Provincial China” category contains all locations

in China that are relatively non-urban. We divided "Other China" — used as a broad, residual category in the previous tables — into four more-narrow categories in Table 4 and Figure 9. The first distinction was Chongqing city. This is the fourth of four municipal cities administered directly by the national government. In contrast to the other three municipal cities (Beijing, Tianjin, and Shanghai), which are historically central commercial and government locations on the east coast, Chongqing is a new major industrial center in the center of China that was promoted in 1997 to the rank of municipal city. The city's social isolation is apparent in the Weibo data. It is the location in which users most prefer to connect with other users in the same location (homophily ratios of 7.46 and 6.46 in Table 4), and the city's only strong connection in Figure 9 is to locations in surrounding provincial China. The great many other areas outside Chongqing and the three large population centers were categorized using a typology provided by a large Chinese real estate organization to guide people looking for a place to live. The company typology sorts areas by GDP, population, wealth, investment, retail sales, household savings, education infrastructure, housing costs, and number of retailers.[9] We distinguish substantial cities outside the Beijing, Shanghai, and Guangdong population centers (18,234 median number of users in the snowball data, e.g., Xi'an with 20,791 users, in Shaanxi province), from smaller cities (4,727 median users, e.g., Urumqi with 5,900 users, out west in Xinjiang province), from places (448 median users, e.g., Jixi with 458 users in the snowball data, on the northeast boarder above Korea in Heilongjiang province). Places like Jixi are cities — Jixi has two million residents — but such places and smaller constitute the "Provincial China" category in Table 4 and Figure 9. Like the three large population centers, users in the substantial other cities and smaller cities are differentiated on the periphery of Figure 9. They have in common connections to Bejing city, Taiwan, and provincial China. Only provincial China stands in the center of the space, widely connected to users in all the greater and

[9]See the company website (http://henan.china.com.cn/finance/2015/0624/519547.shtml).

lesser population centers. Weibo clearly stands out as a communication network between China's urban centers and the great mass of people outside those centers.

——— Table 5 About Here ———

Table 5 lists location averages for users and their networks on sampling criteria discussed above as distinguishing geographic areas. For example, Beijing is characterized by early adoption of Weibo, few star users, many celebrity users, and a low probability of relations being disconnected. The first three rows of Table 5 show that disconnection is least likely in all three locations in the Beijing area (.02 probability) and the highest frequency of early adopters occur in the three locations, but Beijing city is where early adaptors are most likely (19%). Late adoption is least likely in Beijing city, but more likely in the surrounding area than in any of the other 12 locations. Celebrity users are most likely in Beijing city, but as unlikely in the surrounding area as they are absent in Guangdong province. In short, Beijing city is the place to same users for the characteristics of the Beijing area. Similarly, Guangdong province is the place to sample Guangdong users for the lack of early adopters, high presence of star users, lack of celebrity users and network instability characteristic of the Guangdong area.

Gender

User-defined gender seems to be a minor consideration relative to the above user differences. Users register as male or female. On average, male users follow more (528.1 average versus 413.5 for females) and have more followers (11,082.3 average versus 8,835.9 for females). However, the raw averages included organization users. All users have to register as male or female, even organizations. The sample users are 44.1% men, 51.1% women, and 4.7% organizations (of which 54.4% register as men). Since organizations have larger networks than everyday people, we exclude organization users from the gender averages in Table 1. The gender differences in the table are smaller than averages when organizations are included, but males still follow more often and are more often followed (and the gender difference remains if celebrities, who are 62.6% male and have larger networks on average, are excluded from the table). Table 1 shows little gender difference in the probabilities of changing, or discontinuing, a connection. Tables 2 and 4 show that users in the geographic areas

containing Weibo concentrations are almost equally men and women, as are the contacts users follow.

Network Clustering

The Weibo network in Figure 1 corresponds to a traditional image of social networks as a source of interpersonal influence and dynamics (e.g., Festinger, Schachter and Back, 1950; Coleman, Katz, and Menzel, 1957; see Burt, 2010:329-365, for historical review): The user is embedded in a cluster of densely-interconnected contacts. Relationships would be expected to persist, and new ones to develop to further close the network. The user would be expected to conform to group norms of opinion and behavior, with the opinion or behavior of the user's contacts averaged to predict the user's own opinion or behavior.

The image of a dense social fabric around the user does not apply so well to the Weibo network in Figure 2, which displayed the network around a person who adopted early. The Figure 2 network is twice as large as the network in Figure 1, contains twice as many disconnected contacts, and three distinct social clusters of contacts (where clusters are distinguished by stronger connections within a cluster than across clusters). Figure 2 is the social environment of a network broker. Relatively weak connections between the three social clusters define structural holes across which the user can broker information between the clusters (again, characteristic of opinion leaders, Burt, 1999; 2005:84-86). If opinion and behavior differ between the social clusters — as they usually do, providing the user's brokerage opportunities — it is not clear how average contact opinion or behavior would predict the user's own. Prediction would depend on agreement across clusters, or the user's relative attachment to each cluster. It is not clear how the average connection between contacts would predict the probability of a relationship persisting over time. The average would underestimate closure around relations to contacts within clusters, and overestimate closure around relations to contacts outside, or on the periphery of clusters. In short, contact influence on a user, and durable connection with the user, can be expected to depend on where the contact is located in a multi-cluster network.

The implication is that clustering is a parameter to hold constant in analyzing social media networks, either by sampling or by measuring it as a control in statistical analysis. We do not know the extent to which Figure 1 versus Figure 2 characterizes Weibo networks. As discussed in the introduction, network analyses for the two users in Figures 1 and 2 would be impractical to repeat for the 1.8 million users in 414 million connections captured in the snowball data, let alone the Weibo population more generally. However, we suspect there is considerable clustering in the networks because we know that the Weibo networks are much larger than the close, personal networks traditionally studied for social influence, and sample results show two or three clusters in most Weibo networks.

We use modularity to measure clustering. The modularity of a user's network can be used to hold constant clustering differences between users, or distinguish users with multi-cluster networks. For example, Figure 10 displays modularity and clustering in the networks displayed earlier in Figures 1 and 2. The closed network in Figure 1 shows little clustering. The network in Figure 2 disaggregates into three distinct clusters. Modularity was proposed as a scalable method for distinguishing groups within large networks (Newman and Girvan, 2004; Newman, 2006, 2010:224). The method is related to the factor analyses used in early network analysis to distinguish cliques within a larger network (e.g., Coleman and MacRae, 1960; Bonacich, 1972), and its application resembles the bifurcating correlations used to distinguish clusters of structurally equivalent network elements (Breiger, Boorman, and Arabie, 1975). Modularity has the virtues of these early methods, but in addition provides an attractive measure of how much clustering is in a network.

——— Figure 10 About Here ———

The results in Figure 10 were obtained in five steps. (1) Remove isolates and pendants (respectively, contacts who have no reciprocal connections in the user's network, or a reciprocal connection with just one other contact). The white dots in Figure 10 are excluded isolates and pendants. It became clear after analyzing several Weibo networks that clusters are more apparent when the noise of isolates and pendants is removed. Also, it helped to focus on mutual connections, ignoring

asymmetric connections. Asymmetric connections (only one contact in a pair follows the other) are so prevalent that clustering is obscured – as illustrated in the sociogram at the bottom of Figure 10, in which thin lines connect contacts across the three clusters, but mutual connections reveal the three clusters. (2) Transform connections to be deviations from random chance. The probability of observing a random connection from contact j to contact k within a user's network is the probability of j following someone in the network, times the probability of k being followed by someone in the network, times the number of connections observed in the network. Let b_{jk} equal the observed connection from j to k in the network minus the connection expected by random chance. Score b_{jk} is positive when j is connected to k, negative in the absence of a connection. (3) For a cluster of contacts, extract the first eigenvector from the matrix of b_{jk} among the contacts to partition the cluster into two sub-clusters, one composed of contacts with negative eigenvector scores and the other composed of contacts with non-negative scores. (4) Compute Newman's modularity coefficient, Q, to see how well the current partition assigns connections inside clusters (positive b_{jk}) and disconnections (negative b_{jk}) between clusters: $Q = \sum_{jk} (b_{jk}w_{jk}) / M$), where w_{jk} is 1 if contacts j and k are in same cluster (else zero), M is the sum of all relations among ego's contacts, $M = \sum_{jk} z_{jk}$, and summation is across all (N^2) relations among ego's N contacts, including self relations. Modularity is zero if all contacts are assigned to the same cluster, or if connections are randomly distributed within and between clusters. The measure is a positive fraction to the extent that connections occur within clusters. It is a negative fraction to the extent that connections occur between clusters. (5) Repeat steps three and four, sequentially dividing clusters into pairs of subclusters.

For example, the early-adopter network at the bottom of Figure 10 began with 118 directly or indirectly connected contacts separated from 80 isolates and pendants (white dots in Figure 10). The eigenvector for the 118 contacts partitioned the contacts into 29 as one cluster (squares in Figure 10) and 89 in the other (triangles). The .367 modularity score for the 2-cluster partition shows that many connections are inside one or the other cluster. We selected the cluster with the greatest internal differentiation to disaggregate on the next iteration (differentiation measured by the density of negative

b_{jk} within a cluster). Of the b_{jk} among the contacts indicated by squares in Figure 10, 78% are negative. Of the b_{jk} among the triangles, 94% are negative. The density of negative b_{jk} among the triangles is higher because there are many disconnections between contacts in the cluster of down-triangles at the top of the sociogram in Figure 10 and contacts in the cluster of up-triangles in the middle of the sociogram. The eigenvector for the b_{jk} among the 89 triangle contacts distinguishes the two triangle clusters in the sociogram, and modularity increases to .552, showing a better fit to the observed connections. A fourth iteration distinguishes a minor subcluster of three contacts. Distinguishing the minor subcluster in the 4-cluster partition creates several connections between clusters, so modularity decreases slightly to .546, and we return to the 3-cluster partition.

——— Figure 11 About Here ———

While we do not have clustering results on the 1.8 million users in the snowball data, we do have results on a stratified random sample of 2,000 networks around early and late-adopter star users in China's four municipal cities and the population concentration around Hong Kong — Guangdong province. Figure 11 plots the sample networks by size and modularity. To better represent typical users, extremely small and large networks are excluded from the sample (networks between the 10^{th} and 90^{th} percentiles of network size in each area — note the truncation in Figure 11 at networks below size 31 and above size 591).

Networks are distributed well above zero modularity. Median modularity is .34 and 75% of networks have a modularity of .27 or higher. There is a statistically significant tendency for modularity to be higher in larger networks, but the association is slight (.15 correlation between modularity and network size, $t = 6.86$, $P < .001$). Using modularity scores to classify users by the number of clusters in their networks, the sample users break down as follows: two networks contain four clusters, 21% contain three clusters (e.g., the network in Figure 2 and the bottom of Figure 10), 65% contain two clusters, and 14% can be treated as a single cluster (either because contacts are connected

around a core cluster as illustrated by the network in Figure 1 and the top of Figure 10, or because there is no clustering in the network).[10]

——— Table 6 About Here ———

The multi-cluster issue does not go away if analysis is limited to a user's strongest connections. It might seem reasonable to assume that clustering can be ignored by focusing on a user's strongest connections, which are presumably with contacts at the "core" of the user's network. There is some truth to the assumption, but not much. Contacts followed by the 2,000 sample users in Figure 11 are displayed in Table 6 by the cluster to which a contact is assigned (row) and strength of user connection with the contact (column). We distinguish two levels of strong connection: mutual (user follows contact and contact follows user), and Simmel. Following Krackhardt (1999; Tortoriello and Krackhardt, 2010), a Simmel connection is embedded in shared mutual connections, which is associated with trust (Burt, 2005: Chps. 3-4, for review), and facilitates the flow of complex information across clusters (Centola and Macy, 2007; Tortoriello and Krackhardt, 2010). A Simmel connection in the Weibo data occurs when a user and contact linked by a mutual connection also have mutual connections with one or more of the same other users.

The tabulation in Table 6 shows that strong connections are not concentrated in a single "core" cluster in each user's network. Unreinforced mutual connections tend to be with isolated contacts — close contacts apart, or on the periphery, of a cluster in the user's network. Reinforced mutual connections, that is, Simmel connections, are most likely in the first cluster, but then equally likely into subsequent clusters. Late adopters are more likely than early adopters to have Simmel connections with their contacts,

---

[10]Classification requires a decision rule, and different rules can be appropriate for different analytical goals, but for the purposes here, we use the following sequential rule: Run four modularity iterations for a user's network so every network begins partitioned into four clusters with modularity defined by the four-cluster partition (Q4). If the three-cluster partition, Q3, is larger than Q4, or Q4 provides less than a 10% improvement in fit over Q3, set modularity for the network to Q3 and code the network as having three clusters. If Q2 is larger than Q3 or Q3 provides less than a 10% improvement over Q2, set modularity to Q2 and code the network as having two clusters. For modularity below .2, code the network as having one cluster.

illustrating the more-closed networks often observed around late adopters (and 20% of late-adopter networks contain a single cluster versus 8% of early-adopter networks).

We infer from the sample results in Figure 11 and Table 6 that there is probably substantial clustering within the networks around individual Weibo users, typically two or three distinct clusters each of which contains contacts strongly connected with the user. Therefore, users are likely exposed to conflicting social pressures from multiple social clusters so modularity is likely to be an important variable to hold constant in an analysis of the networks.

CONCLUSION

This note was prepared as foundation for further analyses of Zhang et al.'s (2015) snowball data on Weibo networks, but two points stand out as likely to be significant for network analyses of social media more generally: time and clustering. The most consequential sampling parameter is social time, the distinctions between early, bandwagon, and late adopters. User networks are different in size, growth, decay, and composition by the period during Weibo's diffusion when a user opened a Weibo account, and there are well-known substantive differences between early and late adopters. Fortunately, the Weibo data provide a robust distinction between early and late adopters (Figure 6). Second, the networks around individual users are often differentiated into distinct clusters (Figure 10). Sample data imply there are two or three clusters in most Weibo networks, and there are strong connections into each of the clusters (Figure 11 and Table 6). Clustering has implications for estimating interpersonal influence and network dynamics, so modularity scores could be useful to hold constant clustering differences between users. Also significant are kind of user (common, star, celebrity, organization), and geographic location (Beijing, Guangdong, Shanghai, other China, and overseas, though differentiation within the aggregate locations could be used to sample from more narrow areas within each location, Table 4 and Figure 9). At minimum, models of making or breaking connections in Weibo should be tested for consistency across the sampling parameters.

On the other hand, rich variation on certain topics is more likely to be found by focusing on certain parameters. For example, the richest information on network dynamics will be found among users in their first year using Weibo, especially star users, especially in Guangdong province. These are the late, inexperienced users most involved in expanding and editing their networks. In contrast, the richest information on steady growth will be found among the users who were early adopters of Weibo, especially celebrities, especially in Beijing city. These are the users with large, growing networks least subject to decay.

Certain user attributes can be ignored or used to identify kinds of users to be put aside for separate study. Gender need not be a sampling parameter because men and women occur in relatively constant proportion across the adopter, kind, and location categories of users (Tables 1, 2, 3). The organization users might be put aside for study in their own right. In one sense, organizations can be ignored in sampling because they are a relatively consistent proportion of the contacts cited by early and late users (Figure 7), and organizations are a low proportion of users in most locations, especially overseas (Table 2). More significantly, organization users are likely to be represented by employees or agents, so their decisions to follow other users involve processes different from the ones by which individual people decide whose posted messages to follow. Similarly, celebrity users could be put aside in that the most popular of celebrities have press agents who handle messaging, however, many of the people designated celebrities by Weibo are just prominent people who likely handle their own messaging. Regardless, any analysis that includes celebrity or organization users should test for robust effects across the two kinds of users. Overseas users are a final category to sample with caution. They stand apart from the usual user in that they are more likely to change contacts than are users within China (Table 1), they are disproportionately common users and the only category of user that stands out for its number of females (Table 2), and they are the category most likely to be focused on others in their own location category (Table 3, suggesting that expats use Weibo to stay in touch).

REFERENCES

Blau, Peter M. (1977) “A macrosociological theory of social structure.” American Journal of Sociology 83: 26-54.

Blau, Peter M. and Joseph E. Schwartz. (1984) Crosscutting Social Circles: Testing a Macrostructural Theory of Intergroup Relations. New York: Academic Press.

Bonacich, Phillip. (1972) “Factoring and weighting approaches to status scores and clique identification.” Journal of Mathematical Sociology 2: 113-120.

Borgatti, Stephen P. (2002) NetDraw Network Visualization. Harvard, MA: Analytic Technologies.

Breiger, Ronald L., Scott A. Boorman, and Phipps Arabie. (1975) “An algorithm for clustering relational data with applications to social network analysis and comparison with multidimensional scaling.” Journal of Mathematical Psychology 12: 328-383.

Burt, Ronald S. (1999) “The social capital of opinion leaders.” Annals 566: 37-54.

Burt, Ronald S. (2005) Brokerage and Closure: An Introduction to Social Capital. New York: Oxford University Press.

Burt, Ronald S. (2010) Neighbor Networks: Competitive Advantage Local and Personal. New York: Oxford University Press.

Burt, Ronald S., Martin Kilduff, and Stefano Tasselli. (2013) “Social network analysis: foundations and frontiers on advantage.” Annual Review of Psychology 64:527-547.

Centola, Damon and Michael Macy. (2007) “Complex contagions and the weakness of long ties.” American Journal of Sociology 113: 702-734.

Coleman, James S., Elihu Katz, and Herbert Menzel. (1957) “The diffusion of an innovation among physicians.” Sociometry 20:253-270.

Coleman, James S. and Duncan MacRae, Jr. (1960) “Electronically processing of sociometric data for groups up to 1,000 in size.” American Sociological Review 25:722-727.

Festinger, Leon, Stanley Schachter, and Kurt Back. (1950) Social Pressures in Informal Groups. Stanford, CA: Stanford University Press.

Krackhardt, David. (1999) “The ties that torture: Simmelian tie analysis in organizations.” Research in the Sociology of Organizations 16:183-210.

Newman, Mark E. J. (2006) “Modularity and community structure in networks.” Proceedings of the National Academy of Sciences 103: 8577-8582.

Newman, Mark E. J. (2010) Networks: An Introduction. New York: Oxford University Press.

Newman, Mark E. J. and M. Girvan. (2004) “Finding and evaluating community structure in networks.” Physical Review E 69: 026113.

McPherson, Miller, Lynn Smith-Lovin, and Karen Cook. (2001) “Birds of a feather: homophily in social networks.” Annual Review of Sociology 27: 415-444.

McPherson, Miller and James R. Ranger-Moore. (1991) “Evolution on a dancing landscape: organizations and networks in dynamic Blau space.” Social Forces 70:19-42.

Rogers, Everett M. (2003) Diffusion of Innovations. New York: Free Press.

Tortoriello, Marco and David Krackhardt. (2010) “Activating cross-boundary knowledge: the role of Simmelian ties in the generation of innovations.” Academy of Management Journal 53: 167-181.

Zhang, Jing, Jie Tang, Juanzi Li, Yang Liu, and Chunxiao Xing. (2015) "Who Influenced You? Predicting Retweet via Social Influence Locality." ACM Transactions on Knowledge Discovery from Data 9, 3, Article 25 (April).

Table 1. Weibo Networks Beyond and Within the Sample

	Means with All Weibo Users as Contacts				Means with Sample Users as Contacts				
	Outdegree	Indegree	Mutuals	Messages	Outdegree	Indegree	Mutuals	P(change)	P(unfollow)
Type of User									
Common User	432.0	6155.6	161.1	1095.1	218.8	155.3	51.4	.06	.02
Star User	487.5	2320.4	250.1	2194.5	234.8	124.1	66.1	.09	.04
Celebrity	548.9	41036.2	323.9	2032.9	323.9	853.2	151.6	.05	.02
Organization	573.9	37225.1	298.5	1840.9	265.6	853.0	103.2	.06	.03
Gender									
Female	408.2	7838.5	174.3	1581.6	206.6	180.2	51.0	.07	.03
Male	523.9	9340.7	243.1	1495.5	267.4	260.6	84.3	.06	.03
Weibo Entry									
Early, Experienced	521.2	20781.1	232.7	2568.3	321.2	554.3	103.2	.06	.02
Bandwagon User	452.3	8213.1	205.8	1506.1	226.8	212.7	64.3	.06	.03
Late, Inexperienced	521.5	11001.1	222.7	850.0	216.9	183.5	60.2	.08	.04
Location									
Beijing	513.8	18505.7	233.1	1751.6	288.3	426.6	94.7	.05	.02
Guangdong	433.7	8025.9	213.9	1375.7	189.7	206.3	54.3	.08	.04
Shanghai	494.8	9508.1	220.3	1821.0	251.7	234.3	71.1	.06	.03
Other China	458.0	6266.4	196.7	1407.3	233.0	156.7	62.9	.06	.03
Overseas	390.1	18776.0	173.5	1634.9	193.7	422.2	50.2	.08	.04

NOTE: N = 1,655,678 (excludes 131,765 sample users whose profiles were not found, and for location another 1,304 whose location is unknown). User type is defined by Weibo. Star users post many messages and check messages frequently. Gender and location are user-defined. Organizations register as male or female, so gender means exclude organizations. Categories of Weibo entry are defined by months in Weibo (explained in text). Guangdong includes Hong Kong and Macao. Beijing Tianjin and Hebei province. Shanghai includes the adjacent areas of Jiangsu and Zhejiang provinces. Overseas includes Taiwan.

Table 2. Kinds of Users by Location

	Beijing	Guangdong	Shanghai	Other China	Overseas	All Locations
Type of User						
Common User	50.9	53.0	53.1	58.7	61.4	55.1
Star User	23.3	37.1	32.1	29.7	27.4	30.3
Celebrity	20.1	5.9	9.0	7.3	8.2	9.8
Organization	5.7	4.0	5.8	4.3	3.1	4.7
Gender						
Female	51.2	52.0	55.5	53.3	62.7	53.3
Weibo Entry						
Early, Experienced	17.9	5.3	10.7	11.2	8.7	11.0
Bandwagon User	73.9	84.3	79.5	77.2	82.3	78.7
Late, Inexperienced	8.2	10.4	9.9	11.6	9.0	10.3

NOTE: N = 1,654,374 (excludes 133,069 sample users whose profiles were not found or whose location is unknown). Rows give percent of column users in row category. Types of users are defined by Weibo. Star users post many messages and check messages frequently. Gender and location are user-defined. Organizations register as male or female, so organizations are excluded from the gender averages. Categories of Weibo entry are defined by months in Weibo (see text). Guangdong includes Hong Kong and Macao. Beijing includes the adjacent areas of Tianjin and Hebei province. Shanghai includes the adjacent areas of Jiangsu and Zhejiang provinces. Overseas includes Taiwan.

Table 3. User Homophily by Location

	Beijing	Guangdong	Shanghai	Other China	Overseas
Location	1.47	1.96	1.74	1.39	2.08
Type of User					
Common User	1.15	1.04	1.08	1.06	1.04
Star User	1.12	1.03	1.02	1.08	1.07
Celebrity	1.37	1.76	1.47	1.47	1.54
Organization	1.49	1.51	1.42	1.70	1.43
Gender					
Female	1.10	1.11	1.12	1.12	1.10
Male	1.06	1.08	1.09	1.07	1.13
Weibo Entry					
Early, Experienced	1.39	1.87	1.52	1.55	1.60
Bandwagon User	1.05	1.03	1.03	1.03	1.03
Late, Inexperienced	2.80	2.45	2.52	2.37	2.37

NOTE: N = 1,654,374 (excludes 133,069 sample users whose location is unknown). Homophily is the tendency for users to connect with users just like themselves (e.g., women connecting with other women). The cell measure is frequency of self-citation divided by the frequency expected if connections were independent of user category within the location (1.00 indicates independence, larger numbers indicate homophily; see text). Types of users are defined by Weibo. Star users post many messages and check messages frequently. Gender and location are user-defined. Organizations register as male or female, so gender means exclude organizations. Categories of Weibo entry are defined by months in Weibo (see text). Guangdong includes Hong Kong and Macao. Beijing includes the adjacent areas of Hebei and Tianjin. Shanghai includes the adjacent areas of Jiangsu and Zhejiang. Overseas includes Taiwan.

Table 4. Weibo Connections between Locations

Beijing City	0.98 1.05	1.18	1.15	0.93	0.73	1.03	1.01	0.98	1.00	1.07	1.06	1.1	0.67	0.96	161,041,520
Tianjin City	1.11	6.40 4.60	0.92	0.83	0.96	0.86	0.79	0.78	0.74	0.77	0.73	0.90	1.23	0.98	8,333,975
Hebei Province	1.07	1.00	4.40 3.34	0.89	0.87	0.83	0.89	0.85	0.8	0.84	0.84	1.01	1.06	0.82	8,909,530
Guangdong Province	0.92	0.91	0.92	1.29 1.13	1.42	0.90	0.91	0.93	0.88	0.89	0.88	1.01	1.01	1.09	92,723,880
Hong Kong and Macao	0.89	0.98	0.87	1.17	2.06 1.26	0.92	0.92	0.93	0.98	0.96	0.98	0.99	0.78	1.11	18,140,580
Shanghai City	1.00	0.91	0.88	0.93	0.97	1.43 1.32	0.99	1.01	0.84	0.87	0.83	0.92	1.02	1.06	66,121,692
Jiangsu Province	0.97	0.90	0.96	0.95	0.93	1.00	2.28 1.73	0.93	0.82	0.83	0.82	0.97	1.15	0.94	27,395,120
Zhejiang Province	0.95	0.85	0.9	0.96	0.95	1.00	0.97	2.28 1.86	0.81	0.83	0.81	0.94	1.15	0.96	32,249,030
Chongqing City	0.94	0.82	0.85	0.91	0.9	0.86	0.88	0.85	7.46 6.46	0.89	0.8	0.95	1.15	0.95	8,823,377
Other Cities	1.02	0.85	0.92	0.95	0.98	0.92	0.91	0.90	0.93	1.42 1.21	0.86	1.06	1.21	0.95	55,661,550
Smaller Cities	1.00	0.84	0.93	0.95	0.98	0.89	0.91	0.87	0.86	0.94	1.51 1.26	1.13	1.29	0.92	46,508,040
Provincial China	1.07	0.99	1.07	1.04	0.99	0.99	1.03	0.99	1.00	1.10	1.15	0.78 0.74	1.14	0.99	86,884,000
Taiwan	0.88	1.06	0.95	1.00	1.01	0.99	0.99	0.97	0.98	1.07	1.08	1.04	1.05 1.53	1.15	14,274,480
Other Overseas	0.97	0.99	0.79	1.08	1.05	1.05	0.94	0.97	0.93	0.98	0.95	0.99	1.00	1.07 1.02	33,678,720

NOTE: These are ratios of observed connections over number expected if connections were independent of geography. Counts are symmetric (connections from row to column are combined with connections from column to row). Number of observed connections involving row users are given in the column to the far right (excluding 44,547,846 connections with users whose profiles were not found when user profiles were scraped, so their location is unknown). Lower diagonal is common and star users only (Figure 9). Upper diagonal includes connections with celebrities and organizations.

Table 5. Average User Characteristics within Locations

	Kinds of Users				Kinds of User Networks					
Location	% Early	% Late	% Star	% Celebrity	Outdegree	Indegree	Mutuals	P(unfollow)	P(same location)	P(overseas)
Beijing City	18.5*	7.7*	22.6*	22.3*	291.6*	474.3	98.9*	0.02*	0.53*	0.08*
Tianjin City	13.9	10.3	30.6	8.7	278.5	156.6	69.5	0.02*	0.13	0.10
Hebei Province	14.5	12.5*	25.3	5.5*	261.0	140.7	70.6	0.02*	0.10*	0.09
Guangdong Province	5.3	10.5	38.6*	4.9	187.9*	173.6	53.4	0.04*	0.35	0.11
Hong Kong and Macao	5.8	9.0	15.3	19.4	220.3	681.6	67.0	0.03	0.27	0.13
Shanghai City	10.8	8.5	32.6	10.8	266.9	293.3	76.5	0.03	0.28	0.11
Jiangsu Province	11.3	10.6	29.6	7.5	250.3	174.5	68.4	0.03	0.16	0.11
Zhejiang Province	9.8	11.3	33.2	7.5	228.8	190.8	64.8	0.03	0.21	0.11
Chongqing City	10.5	10.5	34.4	7.1	218.0	157.5	60.6	0.03	0.18	0.11
Other Cities	11.1	10.8	32.6	9.0	238.2	163.1	65.7	0.03	0.19	0.11
Smaller Cities	11.0	11.7	34.0	8.4	235.0	138.8*	63.9	0.03	0.16	0.11
Provincial China	11.3	12.3	24.9	5.6	230.3	162.7	60.8	0.03	0.16	0.11
Taiwan	5.1*	9.9	11.1	22.5	195.0	1384.5*	59.8	0.02*	0.27	0.34*
Other Overseas	9.2	8.8	29.5	6.4	193.6	301.3	49.0*	0.04	0.17	0.22

NOTE: : N = 1,655,678 (excludes 131,765 users whose profiles were not found, plus another 1,304 whose location is unknown). Asterisks mark the highest and lowest averages in each column.

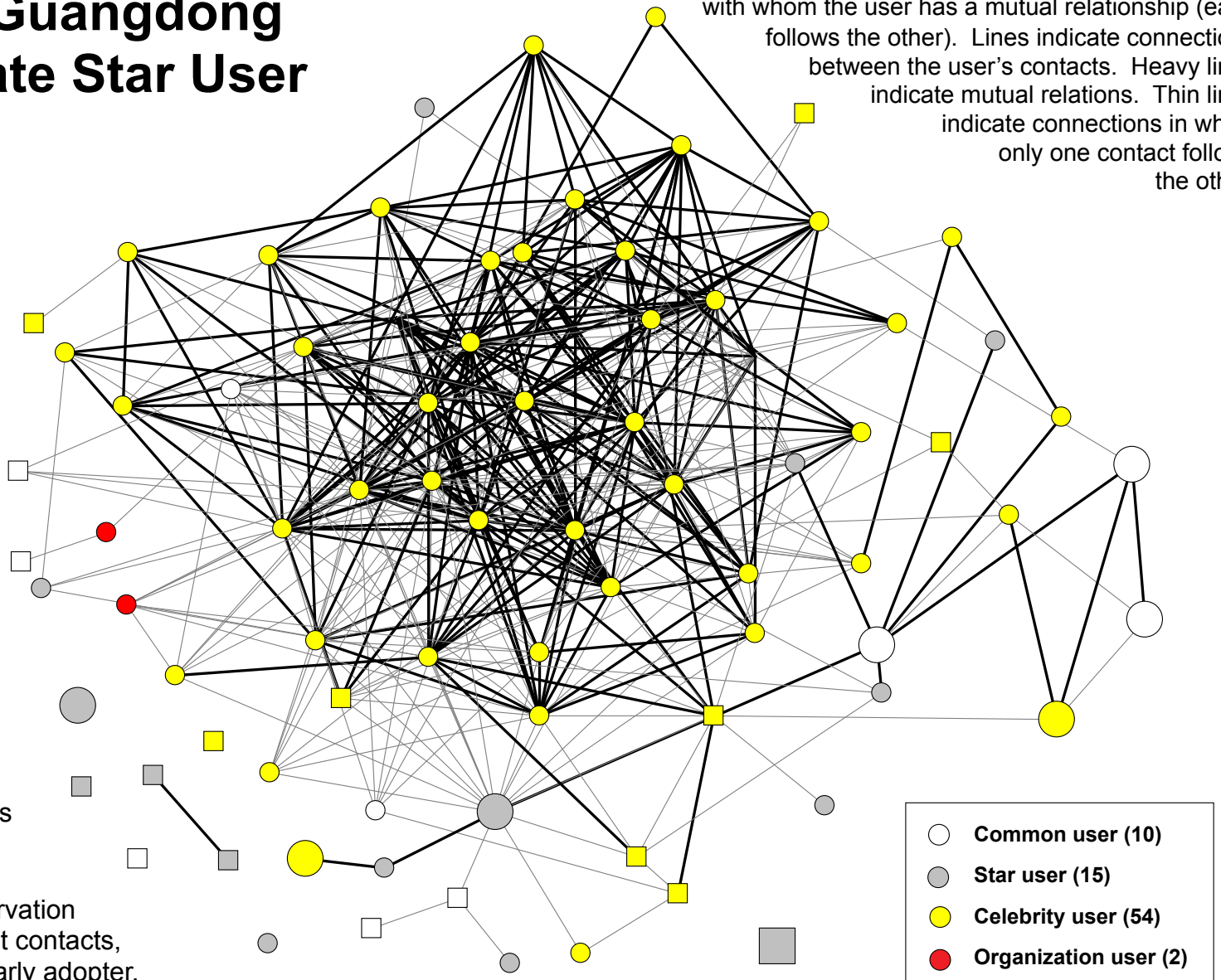
Table 6. Users Have Close Connections with Contacts in All Clusters

Cluster Containing the Contact	Contacts of Early Adopters			Contacts of Late Adopters		
	% Mutual	% Simmel	N	% Mutual	% Simmel	N
Isolate	16.4	9.8	85,510	21.7	12.6	89,136
First Cluster	0.5	36.8	63,177	0.4	52.7	39,322
Second Cluster	0.7	24.3	68,133	0.6	42.7	34,475
Third or Fourth Cluster	0.6	23.3	8,539	0.6	46.2	5,504
All Contacts	6.6	22.2	225,359	11.7	29.2	168,437

NOTE: These are the contacts selected by a random stratified sample of 1000 early adopters and a corresponding sample of 1000 late adopters. Rows sort by the cluster to which contact was assigned as discussed in the text (third and fourth clusters are combined because only two users had four clusters in their networks). "Isolate" contains contacts who have a strong connection with none, or only one, of a user's other contacts. Columns give the percent of row contacts who have a mutual connection with the user (user follows contact and contact follows user), and the percent of who are "Simmel" ties (mutual with contact, both of whom have mutual connection with one or more other contacts in the user's network).

Figure 1. Guangdong Median Late Star User

Symbols are people the user followed during the observation period (7 changed during the period). Circles are contacts who also live in Guangdong. Squares are contacts who live elsewhere. Larger symbols are contacts with whom the user has a mutual relationship (each follows the other). Lines indicate connections between the user's contacts. Heavy lines indicate mutual relations. Thin lines indicate connections in which only one contact follows the other.



(# 764,197)

This is a male 9.90 months in Weibo posting 1,916 messages and following 81 others during the observation period (73.0 nonredundant contacts, 10% mutual contacts, 1 early adopter, 79% in Guangdong, 67% celebrities, 13.6% network density among contacts (50% mutual), 4.5 constraint from contacts, 1,774.1 betweenness).

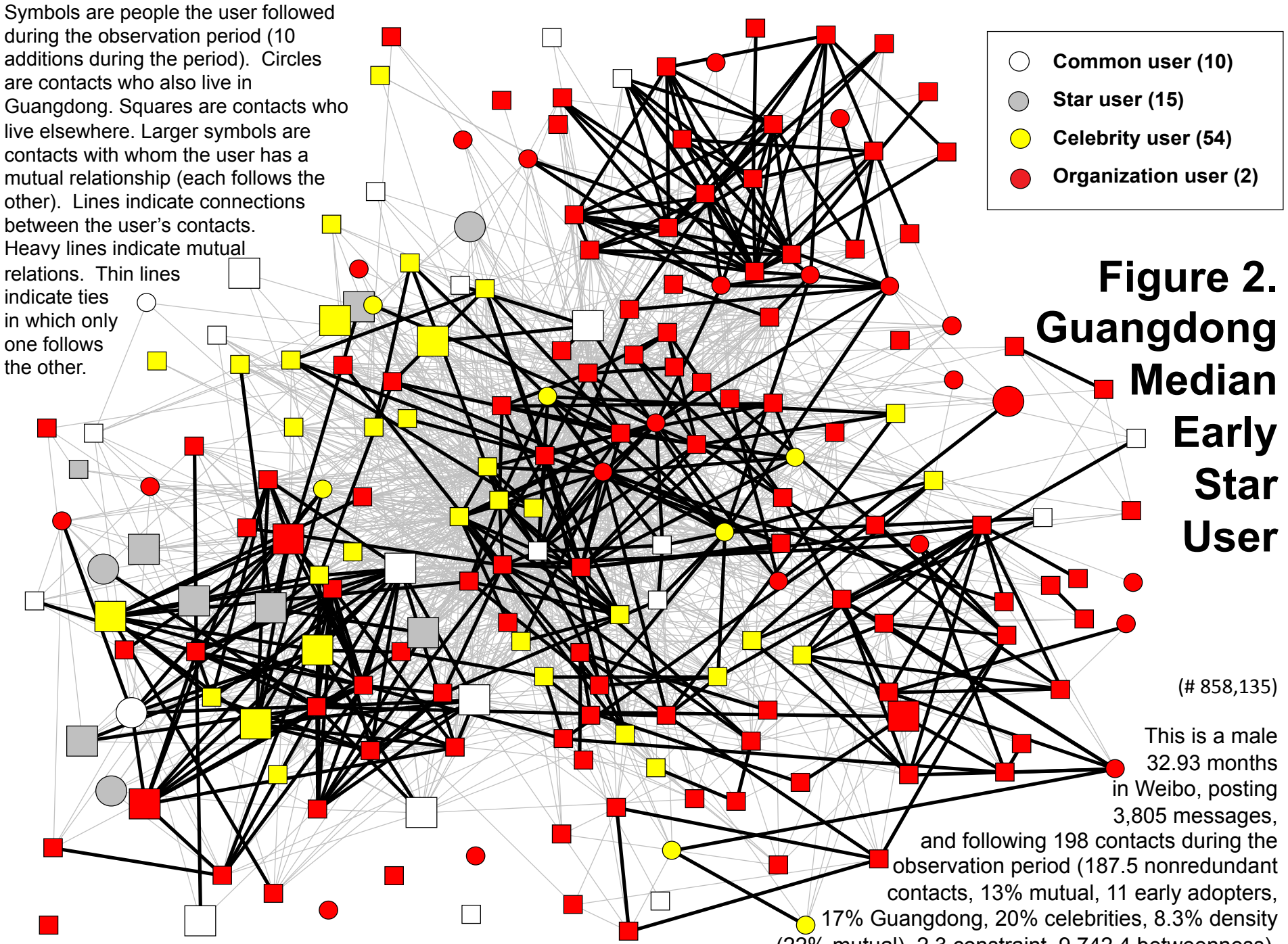

Symbols are people the user followed during the observation period (10 additions during the period). Circles are contacts who also live in Guangdong. Squares are contacts who live elsewhere. Larger symbols are contacts with whom the user has a mutual relationship (each follows the other). Lines indicate connections between the user's contacts. Heavy lines indicate mutual relations. Thin lines indicate ties in which only one follows the other.
Common user (10)
Star user (15)
Celebrity user (54)
Organization user (2)
Figure 2. Guangdong Median Early Star User
(# 858,135)
This is a male 32.93 months in Weibo, posting 3,805 messages, and following 198 contacts during the observation period (187.5 nonredundant contacts, 13% mutual, 11 early adopters, 17% Guangdong, 20% celebrities, 8.3% density (22% mutual), 2.3 constraint, 9,742.4 betweenness).

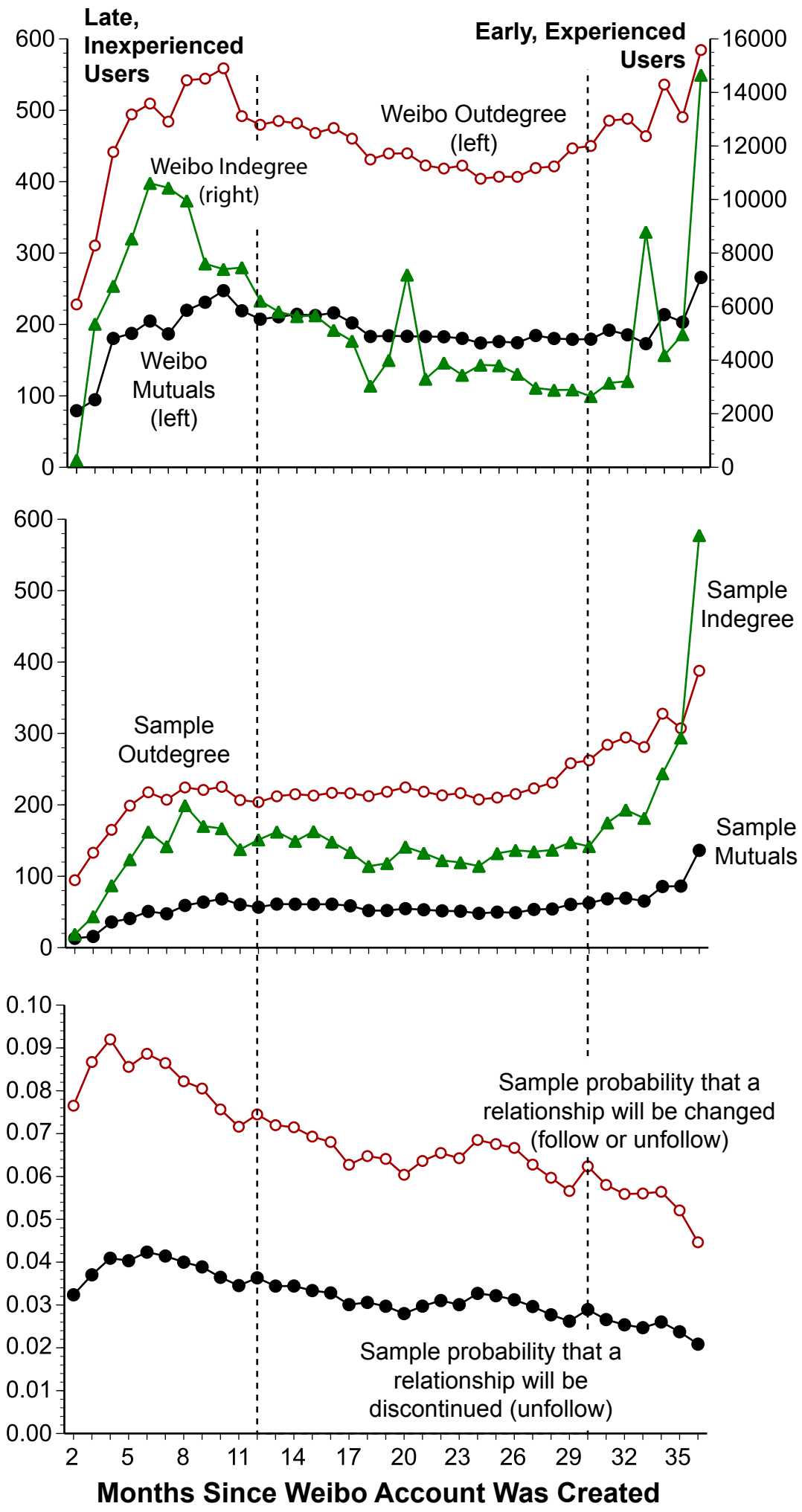
Figure 3

Network Degree and Change by Time since Weibo Account Was Created

Averages are within whole months, based on 1,414,968 common and star users (no celebrities or organizations).

First graph plots user connections with any contact in Weibo.

Second and third graphs plot user connections with any contact in the snowball sample.



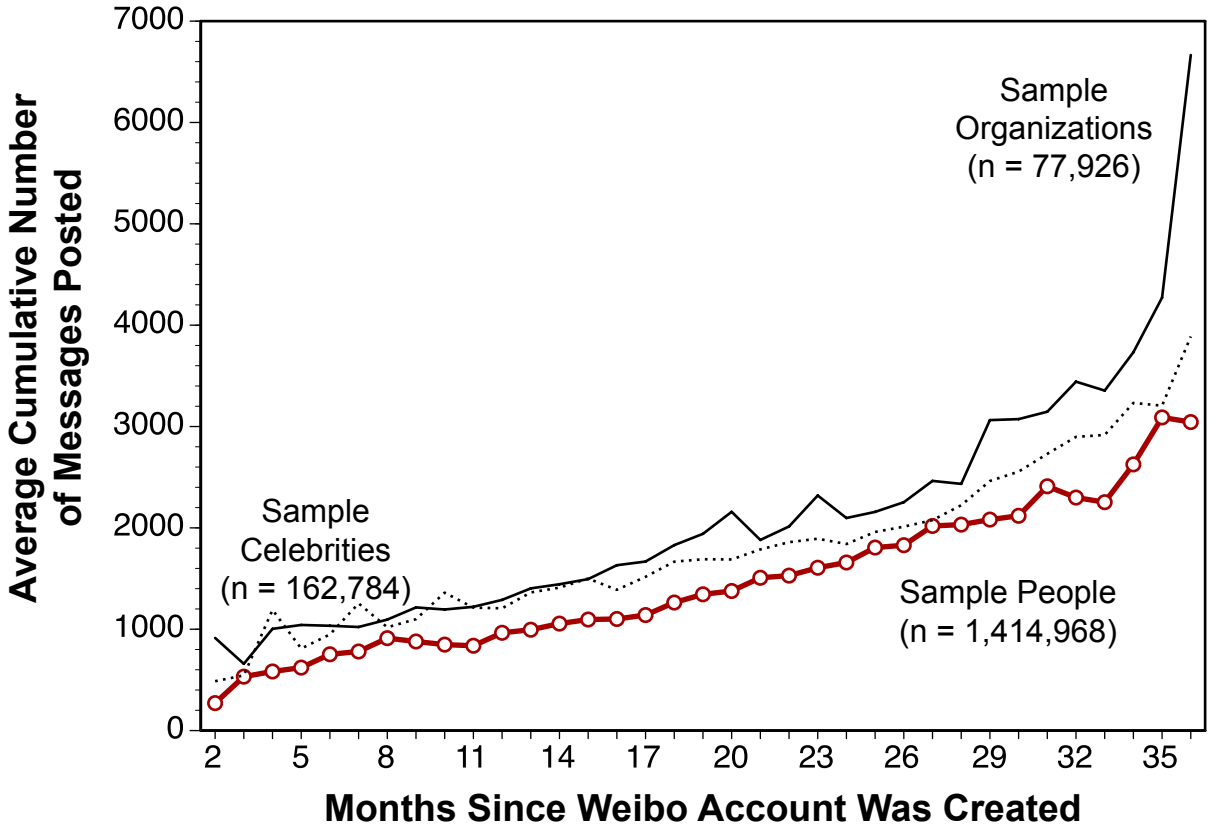


Figure 4. Cumulative Messages Posted

Figure 5. Early versus Late Users

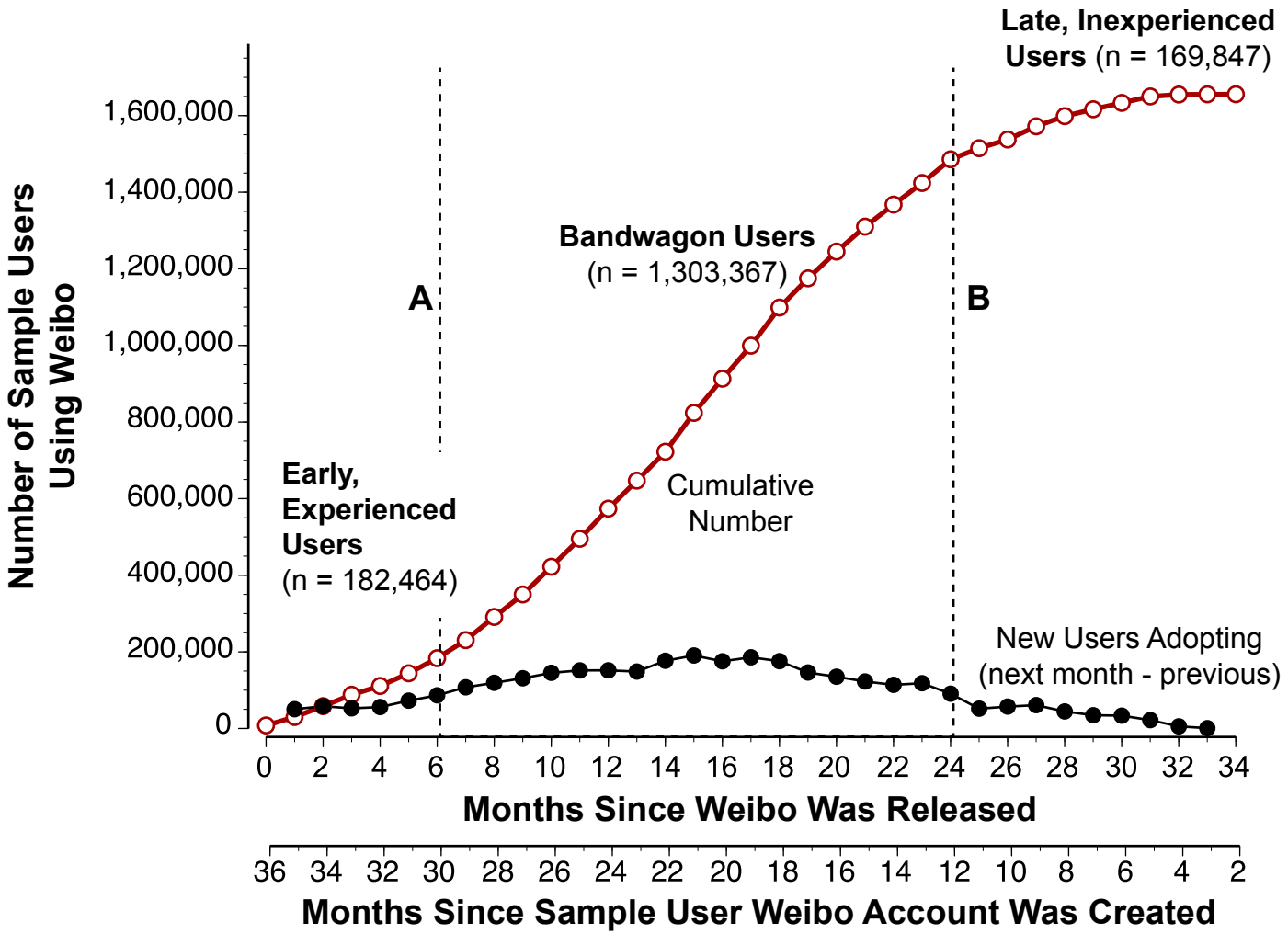


Figure 6. Snowball Distinction between Early and Late Users Is Robust

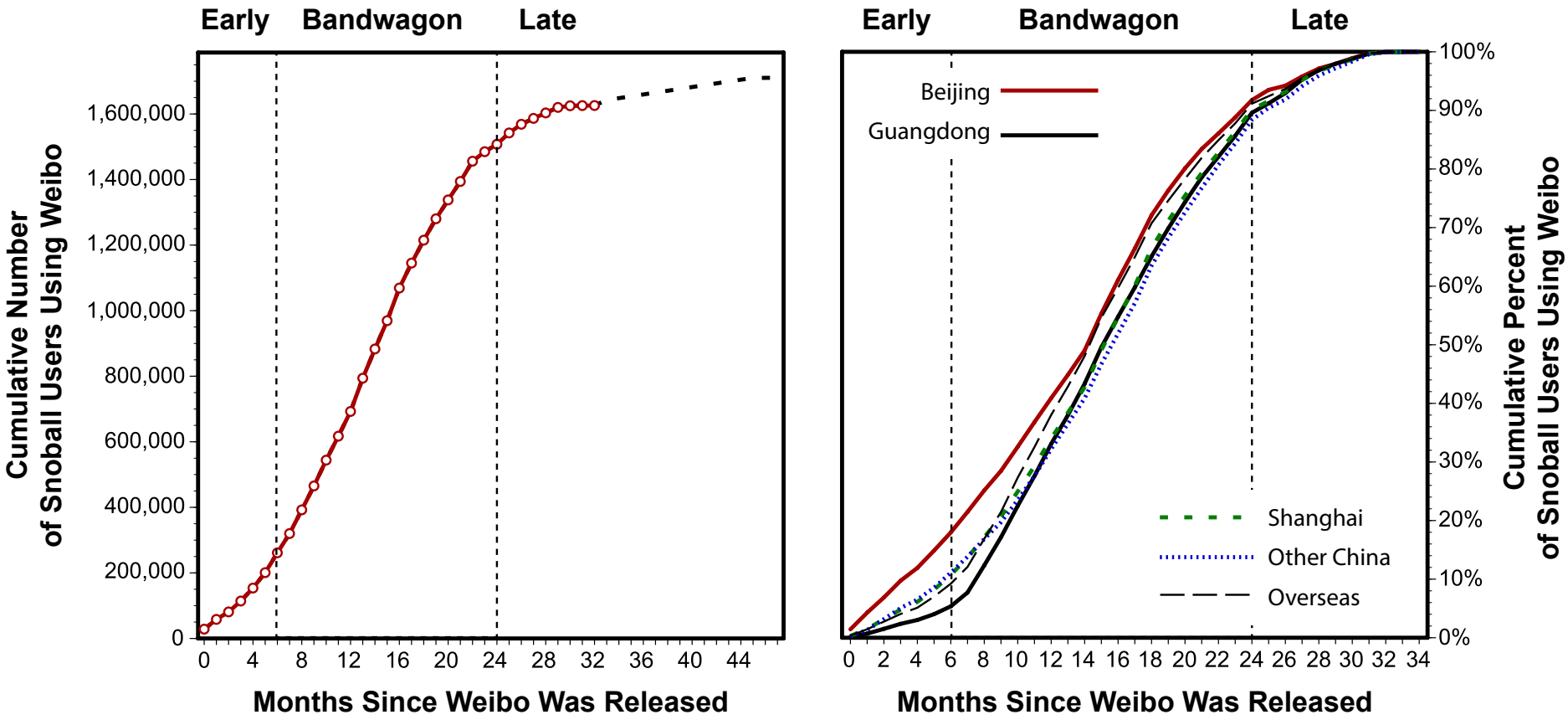


Figure 7. Network Composition for Early versus Late Users

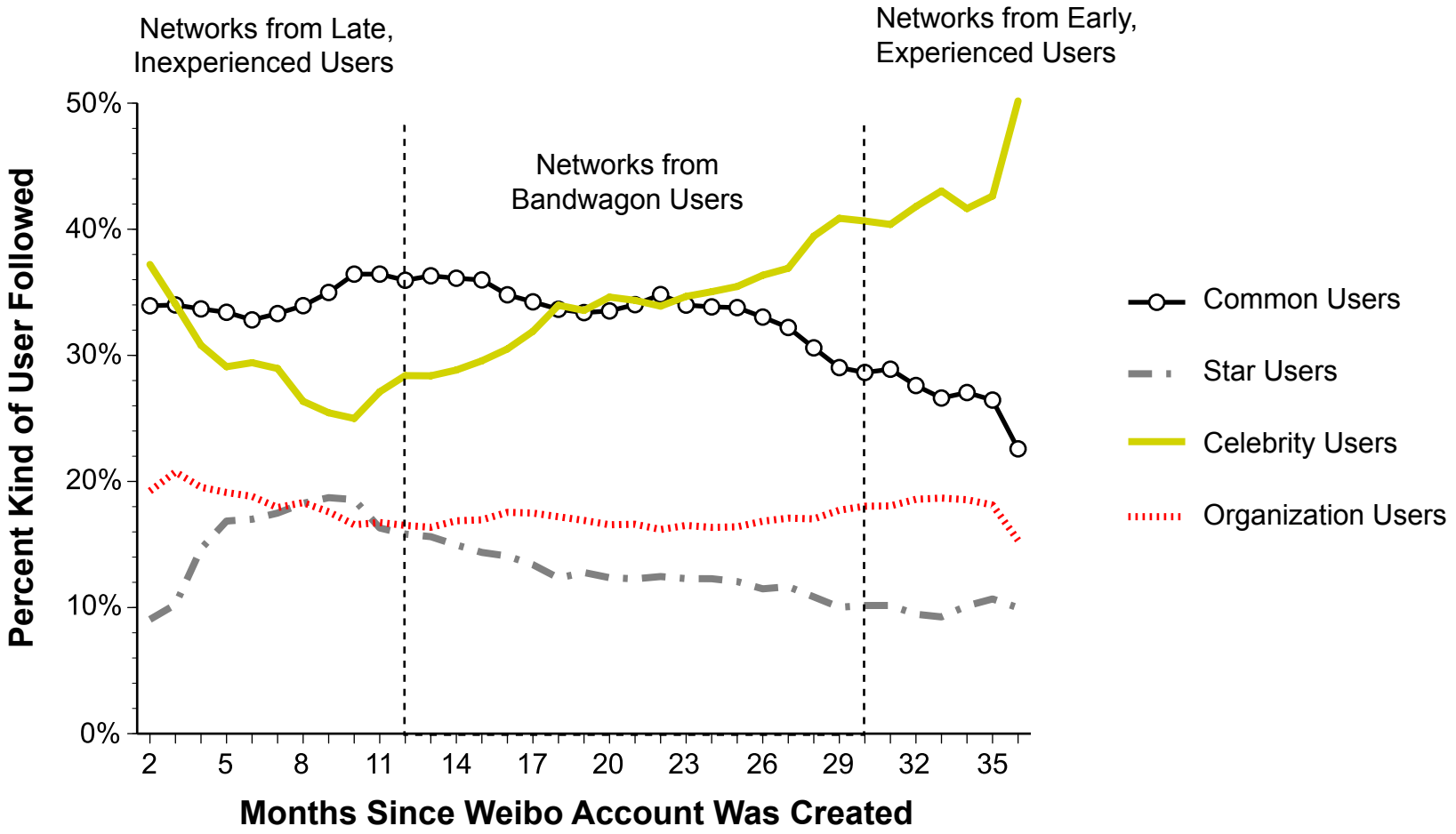


Figure 8. Early Follows Early, Late Follows Late

(User months in Weibo are on the horizontal. Average contact months are on the vertical. Axes cross at their mean values. Boxes extend from 25th to 75th percentile of when contacts began using Weibo, around median in the center. Light boxes extent to 10th percentile on either end of contact adoption dates.)

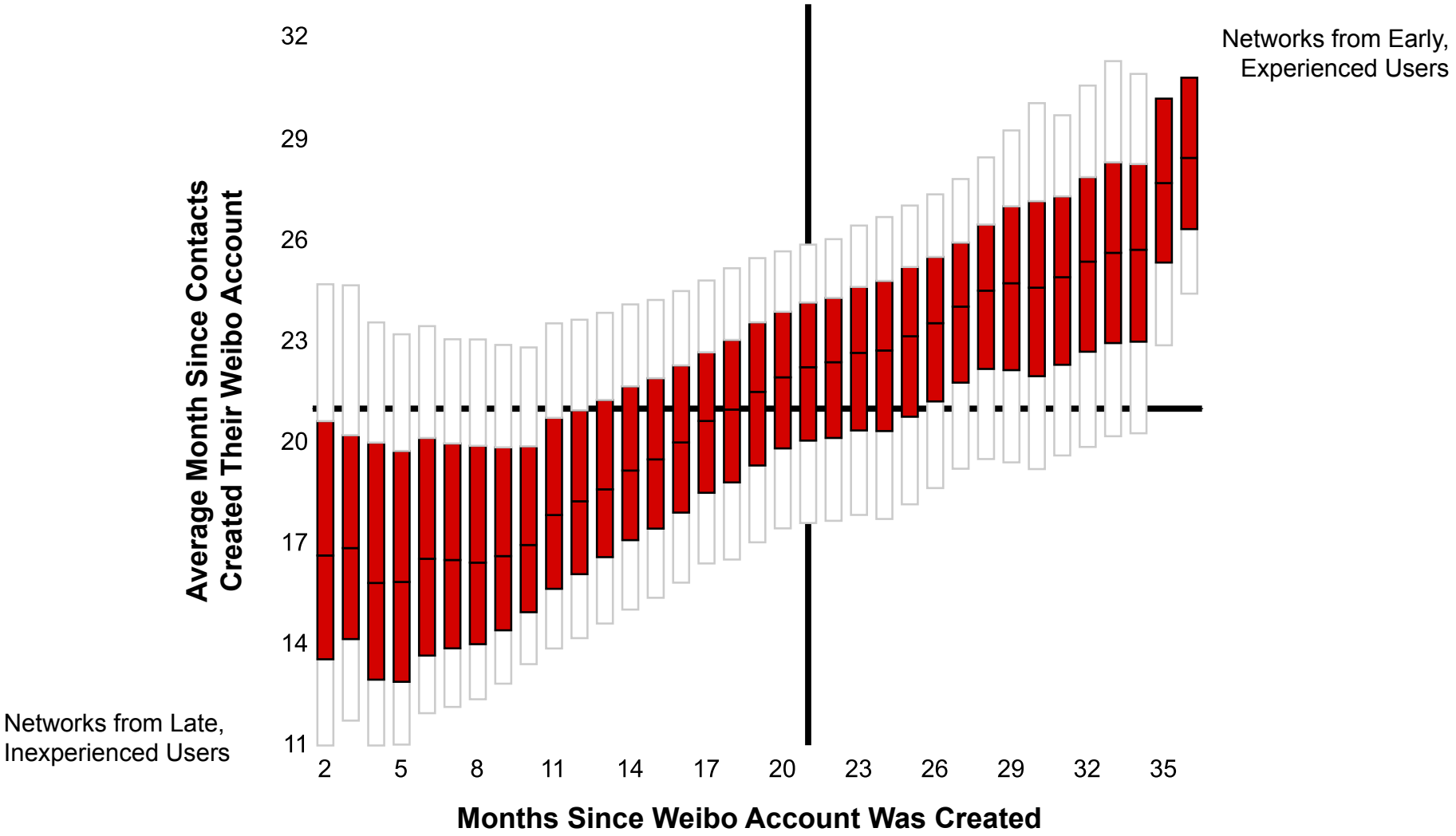
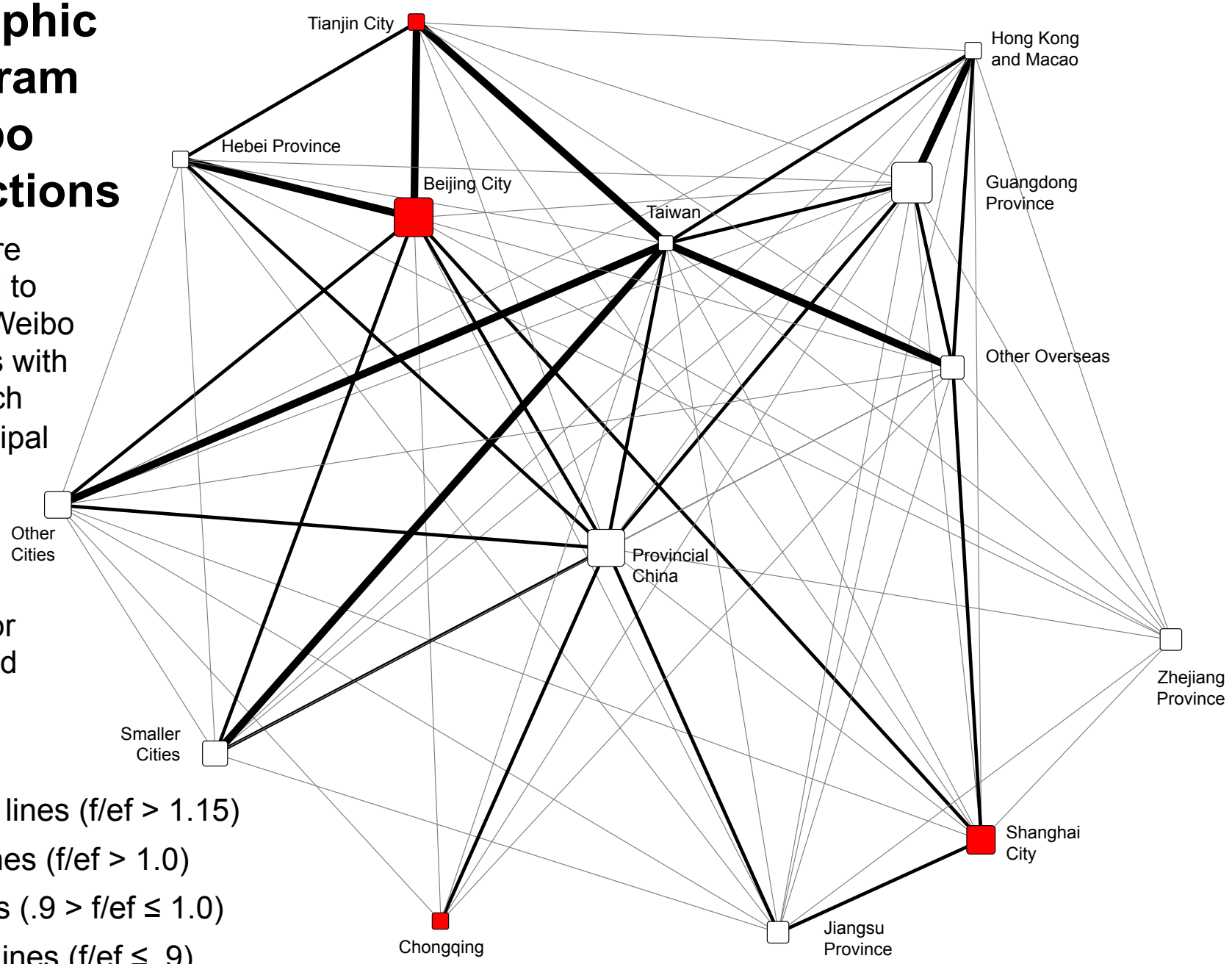


Figure 9. Geographic Sociogram of Weibo Connections

(Symbols are proportional to number of Weibo connections with users in each area. Municipal areas are red. Line strength is given in Table 4 for common and star users.)



7 heaviest lines (f/ef > 1.15)

17 heavy lines (f/ef > 1.0)

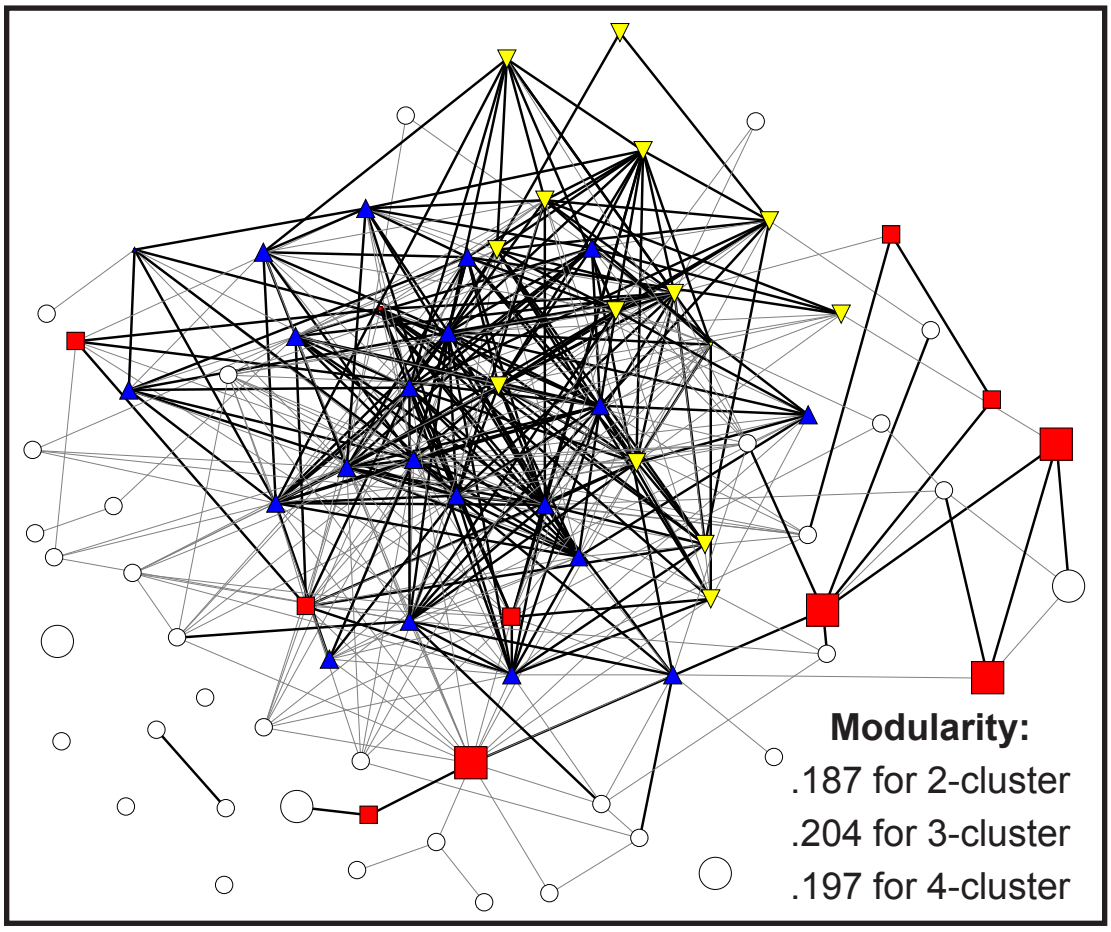
48 light lines (.9 > f/ef ≤ 1.0)

19 missing lines (f/ef ≤ .9)

Figure 10

Network Clustering

These networks are displayed in Figure 1 (above) and Figure 2 (below), but here distinguish four categories of contacts: isolates and pendants (white circles) versus three groups revealed by an eigenvector decomposition of relations among the nonwhite triangles and squares.



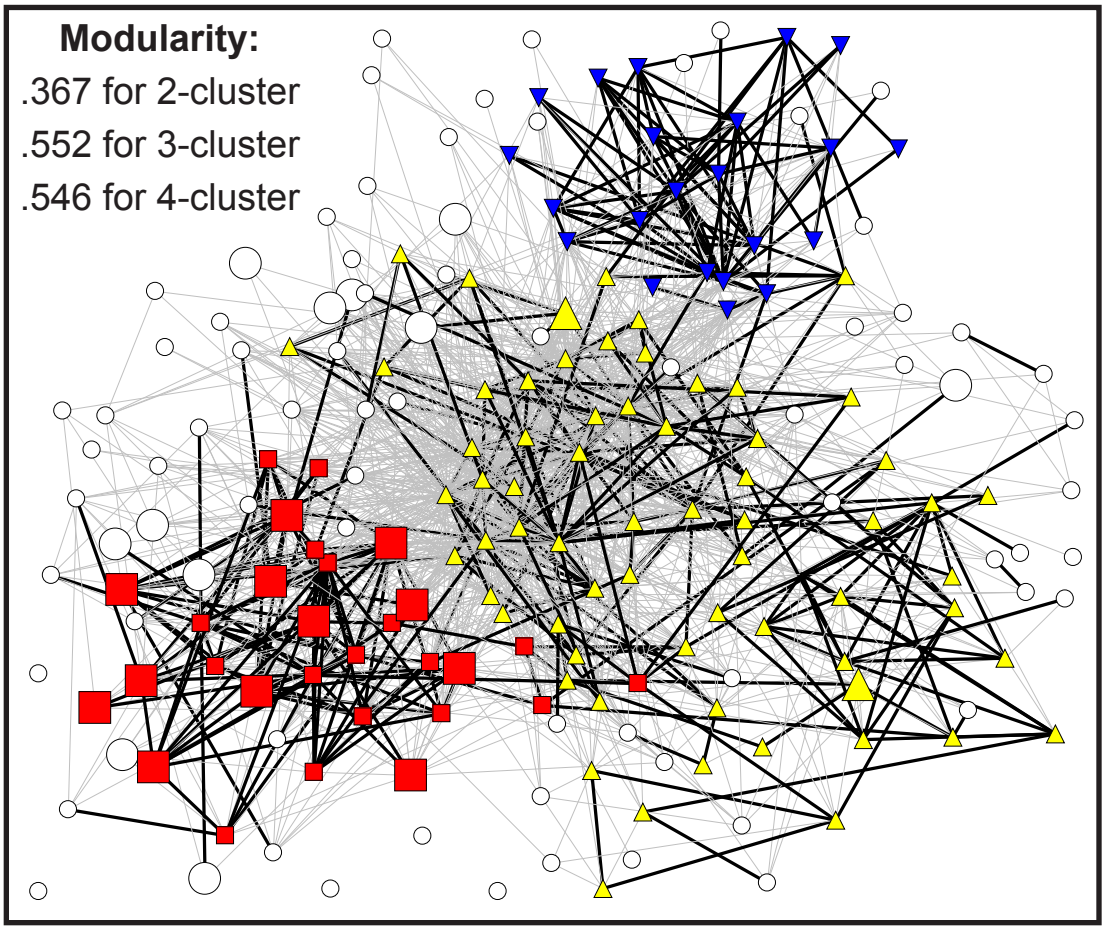


Figure 11. Sample Distribution of Size and Modularity

(stratified random sample of 1,000 early adopters and 1,000 late adopters)