Worldwide Buzz: Planetary-Scale Views on an Instant-Messaging Network

Jure Leskovec*
Machine Learning Department
Carnegie Mellon University
Pittsburgh, PA, USA
Eric Horvitz
Microsoft Research
Redmond, WA, USA

Microsoft Research Technical Report
MSR-TR-2006-186

June 2007

Abstract

We present a study of anonymized data capturing high-level communication activities within the Microsoft Instant Messenger network. We analyze properties of the communication network defined by user interactions and demographics, as reported and as derived from one month of data collected in June 2006. The compressed dataset occupies 4.5 terabytes, composed from 1 billion conversations per day (150 gigabytes) over one month of logging. The dataset contains more than 30 billion conversations among 240 million people. The network is the largest social network analyzed up to date. We focus on analyses of high-level characteristics and patterns that emerge from the collective dynamics of large numbers of people,rather than the actions and characteristics of individuals. Analyses center on numbers and durations of conversations; the content of communications was neither available nor pursued. From the data we construct a communication graph with 190 million nodes and 1.3 billion undirected edges. We find that the graph is well connected, with an effective diameter of 7.8, and is highly clustered, with a clustering coefficient decaying slowly with exponent -0.4. We also find strong influences of homophily in activities, where people with similar characteristics overall tend to communicate more with one another, with the exception of gender, where we find cross-gender conversations are both more frequent and of longer duration than conversations with the same gender.

*This work was performed while the first author was an intern at Microsoft Research.

Contents

1 Introduction

Large-scale web services provide opportunities to capture and analyze behavioral data. We discuss findings from analyses of anonymized data representing one month (June 2006) of high-level communication activities of people using the Microsoft Messenger instant messaging network. We did not have nor seek access to the content of messages. Rather we have examined information about the structure, properties, and relationships of a communication graph that represents how anonymous system users communicate with one another. We consider, the time of initiation, the duration, and the total number of messages exchanged within conversations. We also consider information about demographic properties reported by users upon registration, including gender, age group, language used, and location. In addition, location information was inferred via network address look up. The data occupies 4.5 terabytes in compressed form.

To our knowledge, the analyses in this paper represent the largest and most comprehensive study of interactions within an instant messaging system to date. Several studies on smaller datasets are related to this work. Avrahami and Hudson (Avrahami and Hudson, 2006) conducted a study of communication characteristics of 16 users. Similarly, Shi et al (Shi et al., 2007) also analyzed static buddy (contact) lists that were submitted by the users to a public website. The analyzed the static contact network of 140,000 people. In contrast with regard to the size of the dataset we analyze the static and dynamic communication and network characteristics of all 240 million users of Microsoft Instant Messenger who were active during the month of observation.

Over the month, 240 million distinct accounts logged in and generated 30 billion distinct conversations. We found that approximately 90 million distinct people logged into the Instant Messenger each day and that these users produced about 1 billion conversations per day, with around 7 billion exchanged messages per day. 190 million of the 240 million active accounts had at least one conversation. We found that about half of all conversations occurred between 2 people. The complementary half of conversations involved 3 or more people.

A recent report (IDC Market Analysis, 2005) estimated that around 12 billion instant messages are sent each day. Should this estimate be true, our work analyzes more than one half of all of the world's instant messaging communication during the observational period.

We created an undirected *communication graph* from the data where each user is represented by a node and an edge is placed between users if they exchanged at least one message during the month of observation. Thus, the graph represents accounts that were active during June 2006. The communication graph has 190 million nodes, representing users who participated in at least one conversation, and 1.3 billion undirected edges among active users, where an edge indicates that a pair of people communicated. Note this graph is different from a *buddy graph* where two people are connected if they are registered as "buddies", *i.e.,* they appear on each other people's contact lists. The buddy graph for the data contains 240 million nodes, and 9.1 billion edges, which means an average account has approximately 50 buddies on a contact list.

To summarize key findings about the graphical structure of the communication graph, we discovered that the network is well connected, with 99.9% of the nodes belonging to the largest connected component, and that the 90% effective diameter (Tauro et al., 2001) of the network is 7.8. We found that longer paths exist in the graph, with lengths up to 29. We also found that the network is well clustered, with a clustering coefficient (Watts

and Strogatz, 1998) that decays with exponent -0.37. Intuitively, clustering coefficient is a measure of network transitivity, *i.e.*, it measures the proportion of triangles ni the network. The observed decay is significantly lower than we had expected. For example, it has been observed that in real-networks clustering coefficient decays as k^{-1} with the degree k (Ravasz and Barabasi, 2003).

We also find strong patterns of homophily (McPherson et al., 2001), the tendency of like to associate with like, in communications. People have more conversations and converse for longer durations with people who are like themselves. We find the strongest homophily for the language used, following by conversants' geographic location, and then age. As a salient exception, we found that homophily does not hold for gender. For gender, people tend to converse more frequently and for longer durations with the opposite gender. We found that the number of conversations tends to decrease with geographical distance and age difference between the conversants.

2 Instant Messaging

The use of Instant Messaging, or IM, has been become widely adopted in private and corporate communication. Instant Messaging clients allow users fast "near-synchronous" communication, placing it between synchronous communication mediums, such as speech, and asynchronous communication mediums, such as email (Voida et al., 2002). Instant Messenger users are engaged in a communication by exchanging short text messages with one or more users from their list of contacts. Contact lists are also commonly referred to as a "buddy-list". Similarly, users on the buddy lists are referred to as "buddies". Instant Messaging allows users to easily send and receive short textual messages (instant messages) from their buddies. Recently users can also exchange files and are given various options to express their feelings through small animations and icons called "emoticons"[1]. As conversations and messages are usually very short, it has also been observed that users employ informal language, loose grammar, numerous abbreviations, and minimal punctuation (Nardi et al., 2000).

To construct the Microsoft Instant Messenger communication data from June 2006, we combined three different sources of data: (1) user demographic information (*e.g.*, age, gender, location); (2) time and user stamped events describing the presence of a particular user (login, logout, add/remove buddy); and (3) communication session logs, where, for all participants, the number of exchanged messages and the time participating in the session is recorded.

We use the terms *session* and *conversation* interchangeably. By a session, we mean an Instant Messaging interaction between 2 or more people. A two user session corresponds to a telephone call, and a multi user session corresponds to a conference telephone call. Note that for large sessions, people can come and go over time, so conversations can be long with many different people participating. We observed some very long sessions of more than 50 people even though the software places the limit with at most 20 people communicating at the same time.

[1] `http://en.wikipedia.org/wiki/Emoticon`

2.1 Data description

The data analyzed includes *presence data*, a log of user events when interacting with the system, *communication data* which captures high-level information about conversations, and *user* data which contains information given by the user. The data includes the following:

- **Presence:** login, logout, first ever login, add/remove/block a buddy, add unregistered buddy (invite new user), change of status (busy, away, be-right-back, idle,...). Each event is user and time stamped.

- **Communication:** For each user participating in the conversation/session the log contains a tuple: session id, user id, time joined the session, time left the session, number of messages sent, number of messages received.

- **User data:** For every user the following self reported information is stored: age, gender, location (country, ZIP), language, and IP address. We use the IP address number to decode the geographical locations/coordinates, which we then use to pinpoint users on the globe and calculate the distances.

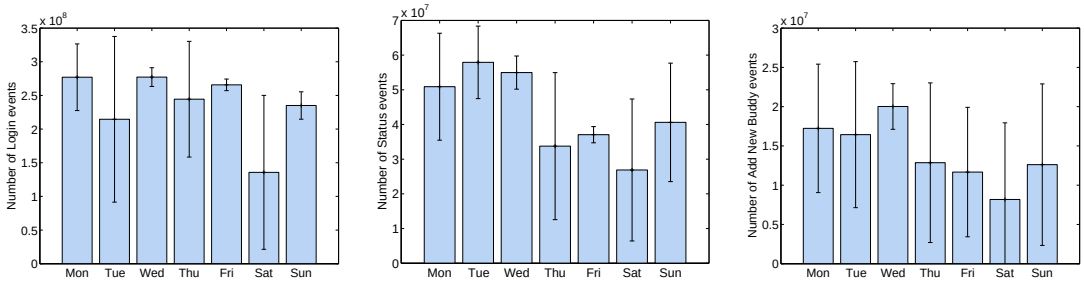
2.2 Privacy considerations

All our data was anonymized; we had no access to personally identifiable information. Also, we had no access to text of the messages exchanged or any other information that could be used to uniquely identify users. Given the size of the data and the scale of our experiments, we focused on analyzing high-level characteristics and patterns that emerge from the collective dynamics of 240 million people, rather than the actions and characteristics of individuals.

2.3 Data collection

We gathered the data for 30 days of June 2006. Each day of data collection yielded about 150 GB of compressed text logs. Copying over the network to a dedicated eight-processor server with 4 terabytes of disk space and 32 gigabytes of memory took 12 hours. Our log parsing system uses a pipeline of four threads that parse the data in parallel, collapse the session join/leave events into sets of conversations, and saves the data in a compact compressed binary format. This process compresses the data to 45 GB per day. Parsing and processing the data took additional 4 to 5 hours per day.

A special challenge was to account for missing and dropped events, and session id recycling across different instant IM servers in a server farm. For example, at user login, several events are triggered, and we collapsed all of them into a single "login" event. However, events get logged in different orders and sometimes they can also be dropped. We performed careful error-checking so as to ensure that errors would not propagate over time and corrupt the working queues. As part of this process, we closed a conversation 48 hours after the last leave session event. We also automatically closed sessions if only one user is left in the conversation. We also took care in saving the state of conversations that took place over the midnight so as to allow a fluid processing of conversations straddling days.



(a) Login (b) Status (c) Add buddy

Figure 1: Number of events per day of the week. We collected the data over a period of 5 weeks starting on May 29 2006.

3 Usage and population statistics

We shall now review several statistics about users and their communication activities. The statistics reflect the portion of the Microsoft Instant Messenger population that showed some activity during June 2006.

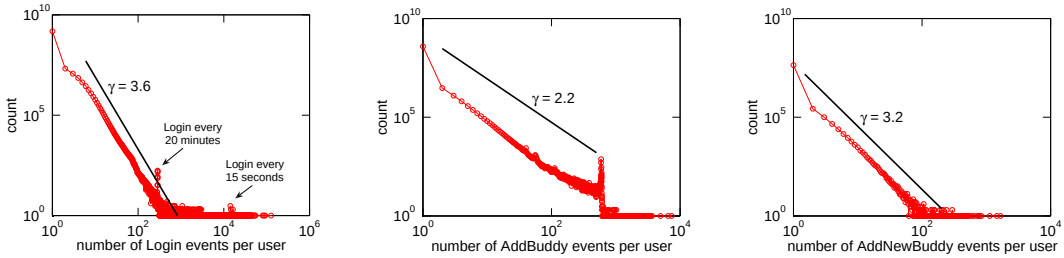
3.1 Per month and per day activity

Activity over the 30 days of June 2006 reveals that 242,720,596 users logged into the Microsoft Instant Messenger and 179,792,538 of these users were actively engaged in conversations during this time. During the one month observation period 17,510,905 new accounts were activated.

On a typical day, June 1 2006, there were 982,005,323 different sessions, *i.e.,* conversations between any number of people. About half of them (508,315,719) are two-user conversations, and the other half (471,837,591) are conversations between three or more users. We also see around 93 million users login with 64 million different users being engaged in conversations on that particular day.

Figure 1 shows the number of logins, status change and add buddy events by day of the week over a period of 5 weeks starting in June 2006. We count the number of particular events per day of the week, and we use the data from 5 weeks to compute the error bars. Figure 1(a) shows the average number of logins per day of the week over a 5 week period. Note that number of login events is larger than the number of distinct users logging in, since a user can login multiple times a day. Figure 1(b) plots the average number of status change evens per day of the week. Status events include a group of 8 events describing the current status of the users, *i.e.,* away, be right back, online, busy, idle, at lunch, and on the phone. Last, Figure 1(c) shows the average number of add buddy events per day of the week. Add buddy event is triggered every time user adds a new contact to their contact list.

Figure 2 shows the distribution of the number of events per user for the observation period of June 2006. Figure 2(a) shows the number of logins per user for the observation period. We observe the total of 250 million users login into the Messenger system in the observation period. Per day about 93 million different users login. The number of logins per user follows heavy tailed distribution with exponent 3.6. Also notice spikes in the distribution for users



(a) Login event (b) AddBuddy event (c) AddNewBuddy event

Figure 2: Distribution of the number of events per user over the observation period. (a) Number of logins per user. (b) Number of added buddies per user, *i.e.,* people added to the contact list. (c) Number of newly invited buddies (people that are not yet users of MSN Messenger) per user.

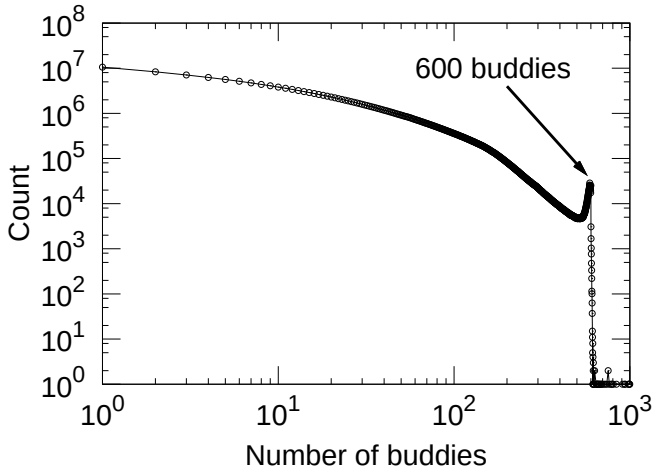


Figure 3: Distribution over the number of people on the buddy list (length of the contact list).

that login every 20 minutes and every 15 seconds. We speculate that the latter activities may reflect the activities of automated, robotic systems.

Figure 2(b) shows the number of added buddies per user in the observation period. Notice the spike at 600 added buddies. This means that many users quickly fill their contact lists with 600 buddies, which is the maximal length of the contact list set by the system. Figure 2(c) shows the number of "add new buddy" events per user for the observation period. A new buddy is a new user that does not yet have a Messenger account. This demonstrates that new users that were previously not part of the Microsoft Instant Messenger system are frequently invited. There are about 1.5 million new users invited every day. We plot all the data on logarithmic scales and power-law exponents are 2.2 for number of added buddies per user and 3.2 for number of newly added buddies.

Next, Figure 3) displays the number of buddies per user. This plot corresponds to degree distribution of a buddy-graph, a graph where there is an edge between a pair of people if

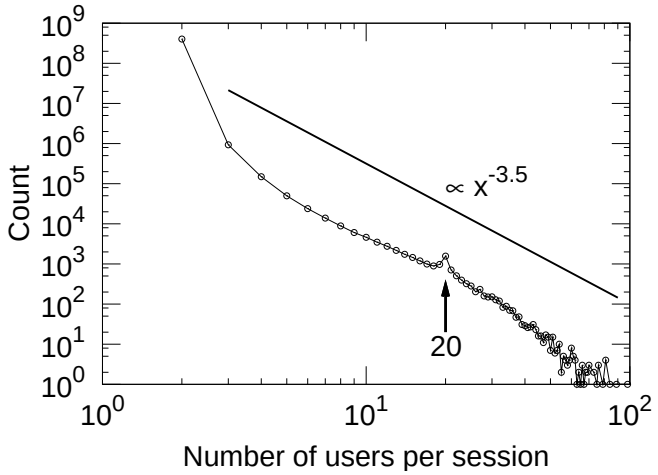


Figure 4: Distribution of the number of people participating in a conversation.

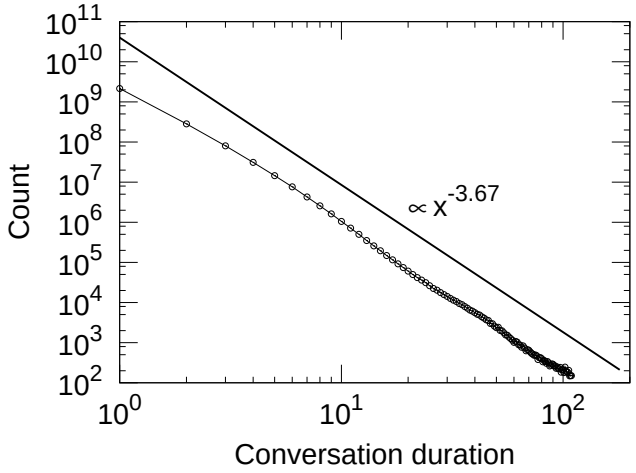


Figure 5: Distribution of the durations of conversations. Notice nice power-law in conversation duration distribution.

they are buddies, *i.e.,* they appear on each others contact lists. Notice this is different than the number of people that a person communicates to, as we observe the communication for just one month and people may not show communication with all of their buddies during this time. We found a total of 9.1 billion buddy edges in the graph with 49 buddies per user.

Also, notice a spike at 600 which is the limit on the maximal number of buddies set by the software client. The maximal number of buddies was increased from 150 to 300 in March 2005, which was then later raised to 600. However, there do not appear to be bumps at 150 and 300, only a peak at 600.

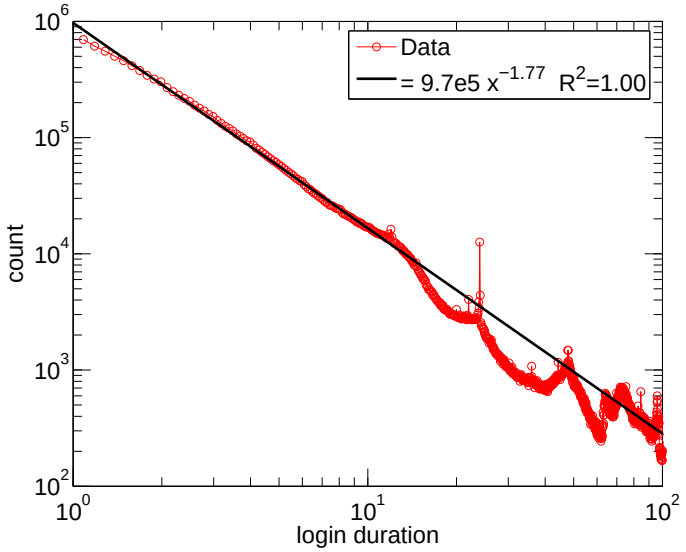


Figure 6: Distribution over the login duration of a user. For every user we determine times between logins to the system and plot the distribution.

Figure 4 shows the number of users per session. In Microsoft Messenger, more than two people can chat at once, as multiple people can be invited to the conversation. We observe a peak at 20 users which is the limit on the number of people who can join in a conversation and chat at the same time. Notice some sessions are larger as people come and go but there are never more than 20 people engaged simultaneously.

Figure 5 shows the conversation duration distribution, which appears to ne described by a power-law with exponent 3.6. On x-axis we plot the duration of a conversation and the y-axis is the number of the conversations with the duration of x time units.

Next we examine the distribution of the time durations that people are logged in and logged out of the system. Let (ti_j, to_j) denote a time ordered $(ti_j < to_j < ti_{j+1})$ sequence of login and logout times of a user, where ti_j is the time of j-th login, and to_j is the corresponding logout time. Then Figure 6 plots the distribution of $to_j - ti_j$ over all j over all users. Similarly, Figure 7 shows the distribution of logout duration, *i.e.* $ti_{j+1} - to_j$ over all j and over all users. The power-law fit shows exponents of 1.77 and 1.3 respectively. The data reveals that the distribution of login durations tend to be shorter and decay faster, while durations of the times that users are not logged in are longer and decay slower. We also notice periodic effects of login duration of 12, 24 and 48 hours. We observe similar periodicities for logout duration (Figure 7) of 24, 48, 72, ... hours.

3.2 Demographic characteristics of the users

Now we briefly review the demographic characteristics of the Microsoft Instant Messenger user population.

Figure 8(a) shows self-reported user age distribution. The distribution is skewed to the right and has a mode at age of 18. We note that the distribution has exponential tails. The

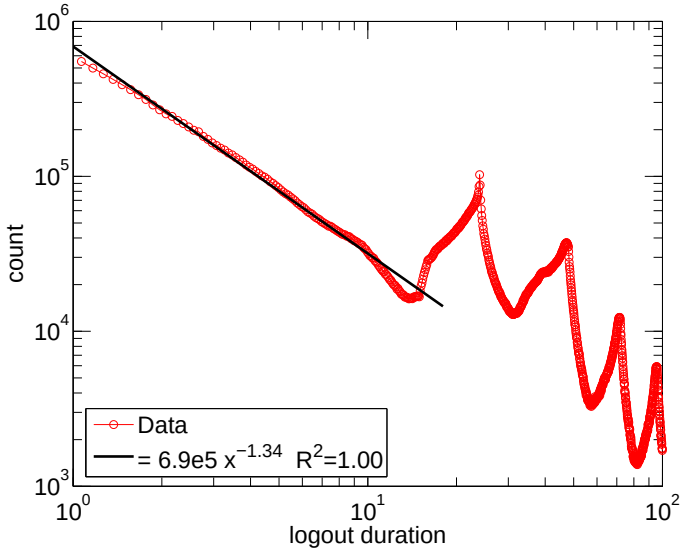


Figure 7: Durations of times when people are not logged into the system. For every user we determine times between the logout and login to the system and plot the distribution.

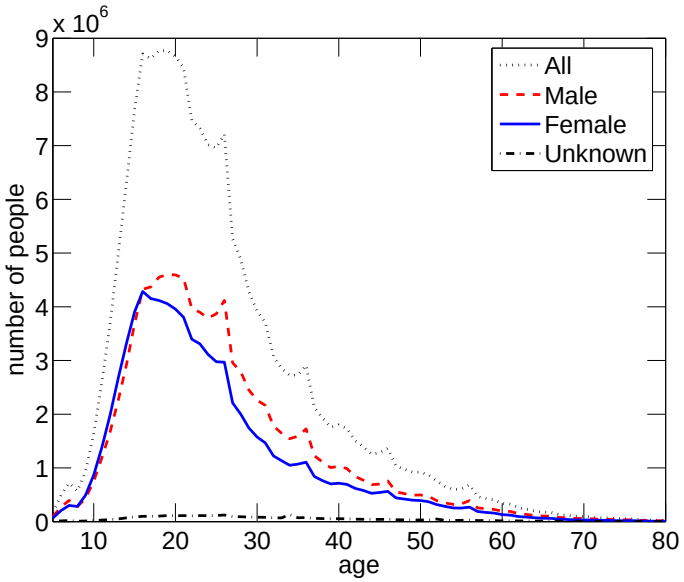
spike at age of zero represents people who did not submit their age to the system. These people account for 6.5% of the population, which is a surprisingly low fraction. We also notice that people not specifying their gender tend to misreport the age also. We have no way to tell whether someone misreported their age, but we suspect that the people with ages above 90 are largely false reports.

Interestingly, as Figure 8(a) shows for ages between 10 and 16, the absolute number of female users is greater than the number of male users. For ages above 17 the male users dominate the population (red dotted line are above the blue solid line). We also notice that majority of population with very high (> 100) or very low (≈ 0) age also does not reveal their gender.

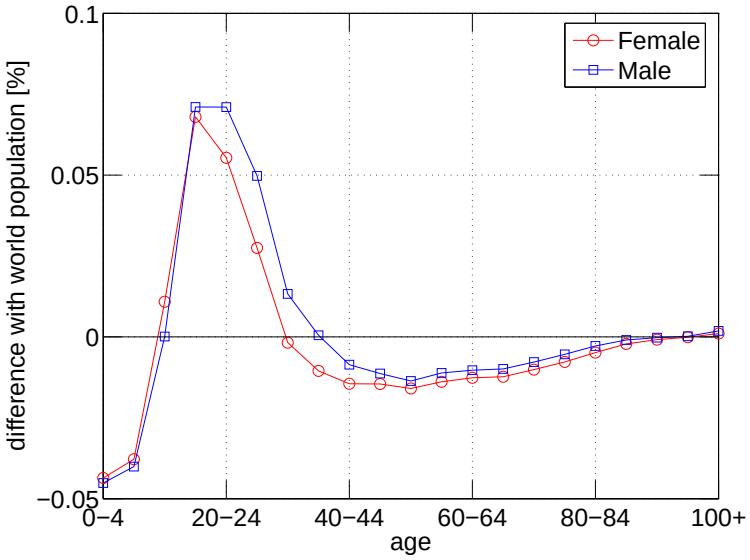
We can compare the age and gender statistics from the Messenger data with world population and distribution data. Figure 9 shows a comparison of the age distribution of MSN users with the world population distribution in 2005[2]. Notice that the active Messenger population of ages from 15 to 35 is heavily over represented when compared with the world population. Interestingly, in comparison to the world population females are over represented for the 10–14 age interval. For male users, we see a match with the world population for ages 10–14 and 35–39, and for women, a match for age group 30–34.

To emphasize the disparity between the populations, Figure 8(b) shows the relative differences in the fraction of Messenger and world population. Let m_a be the fraction of Messenger users of age a, and let w_a be the fraction of world population of age a. Figure 8(b) then for every age a plots $m_a - w_a$, the difference in fractions of population. We see that children (who presumably cannot type) and older parts of the population are under

[2]We obtained the world population data from the UN Population Division website.



(a) Age distribution



(b) Age difference

Figure 8: Distribution of self-reported ages of Messenger users and the difference of ages of Messenger population with the world population. (a) Age distribution for all users, females, males and unknown users. (b) Relative difference of Messenger population and the world population. Ages 15–30 are over-represented in the Messenger user population.

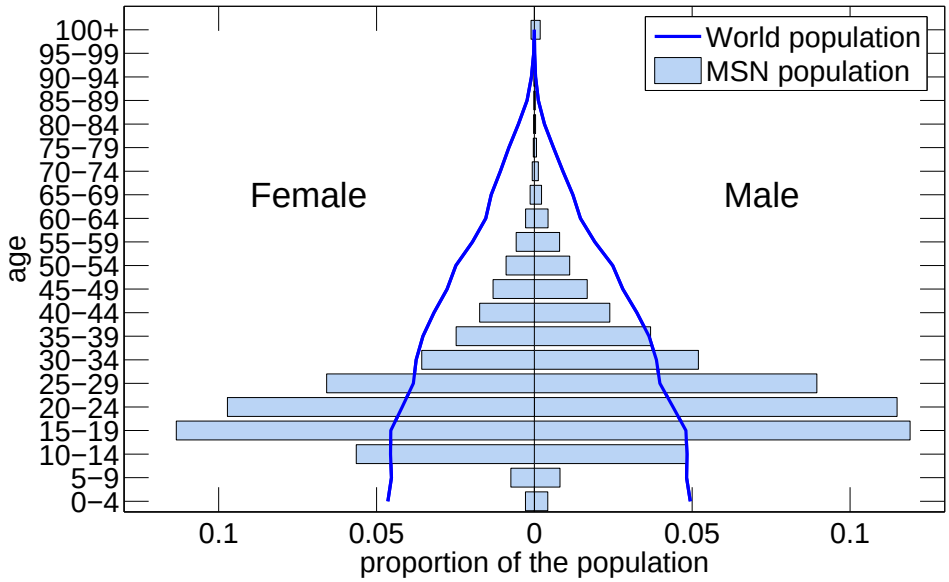


Figure 9: World and Messenger population age pyramid. Notice ages 15–30 are over represented in the Messenger user population.

represented and active population is over represented when compared with the population of the world.

4 Communication characteristics

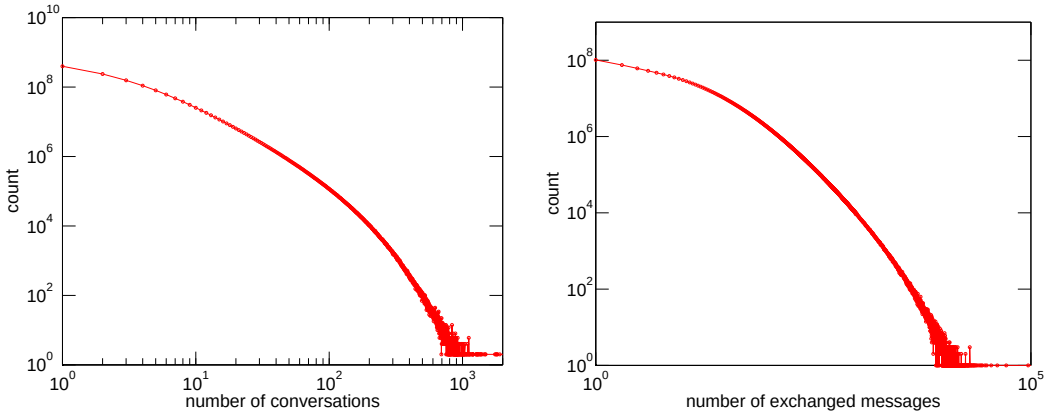
We now focus on an analysis of the large scale communication characteristics. We limit the analysis only to conversations with 2 exactly participants, which appear almost 3 orders of magnitude more often than conversations between 3 or more people.

Figure 10(a) shows the number of conversations per user for the observation period of June 2007. Similarly, Figure 10(b) shows the number of exchanged messages over all 2-user conversations. Both distributions seem to be heavy tailed but not power-law.

Similarly as in Figures 6 and 7, where we plotted the durations of users' login and logout times, we next plot the distribution of the durations of conversations and the periods between the successive conversations of a user.

Let user u have C conversations in the observation period. Then for every conversation i of user u we create a tuple $(ts_{u,i}, te_{u,i}, m_{u,i})$, where $ts_{u,i}$ denotes the start time of the conversation, $te_{u,i}$ is the end time of the conversation, and $m_{u,i}$ is the number of exchanged messages between the two users. We order the conversations by their start time ($ts_{u,i} < ts_{u,i+1}$). Then for every user u we can calculate the average conversation duration

$$\bar{d}(u) = \frac{1}{C} \sum_{i} te_{u,i} - ts_{u,i}$$



(a) Number of conversations (b) Exchanged messages

Figure 10: Conversation statistics: (a) Number of conversations of a user in a month; (b) Number of messages exchanged per conversation;

where the sum goes over all the u's conversations. Figure 11(a) plots the distribution of $\bar{d}(u)$ over all the users u, for every user, we calculate the average conversation length in the observation period and plot the distribution. Notice heavy tailed distribution with exponent -3.7. The distribution has a mode of the distribution at 4 minutes, which means that most likely conversations take about 4 minutes.

Similarly, Figure 11(b) shows between start times of consecutive conversations of a user. More precisely we plot the distribution of $ts_{u,i+1} - ts_{u,i}$, where $ts_{u,i+1}$ and $ts_{u,i}$ denote start times of two consecutive conversations of user u.

The power-law exponent in Figure 11(b) is 1.5 which was observed for many other types of human activity (Barabasi, 2005). Also notice the increased number of conversations exactly 24, 48, etc. hours apart. These may reflect robotic systems and automatic network probing.

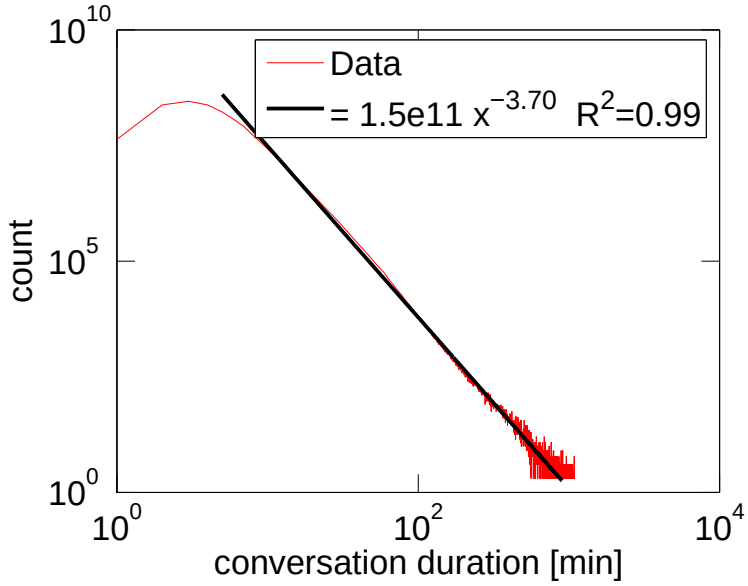
5 Communication by user demographics

Let us now examine the interplay of communication and user demographic attributes. We shall examine how geography, user location, age, and gender influence the observed communication patterns.

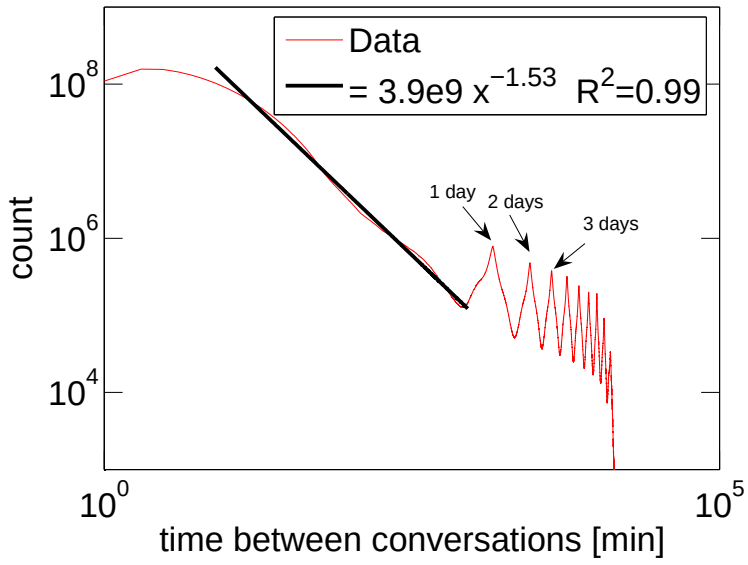
5.1 Communication by age

First, we examine how communication changes with the age of the participating users. We focus on the two-person conversations, and measure various characteristics of communication as a function of the reported ages of both users.

Figure 12 shows the communication characteristics by age. Row and column determine the ages of both parties participating on the conversation, and the color indicates the log of value at particular age–age cell. Because of potential misreporting at very low and high

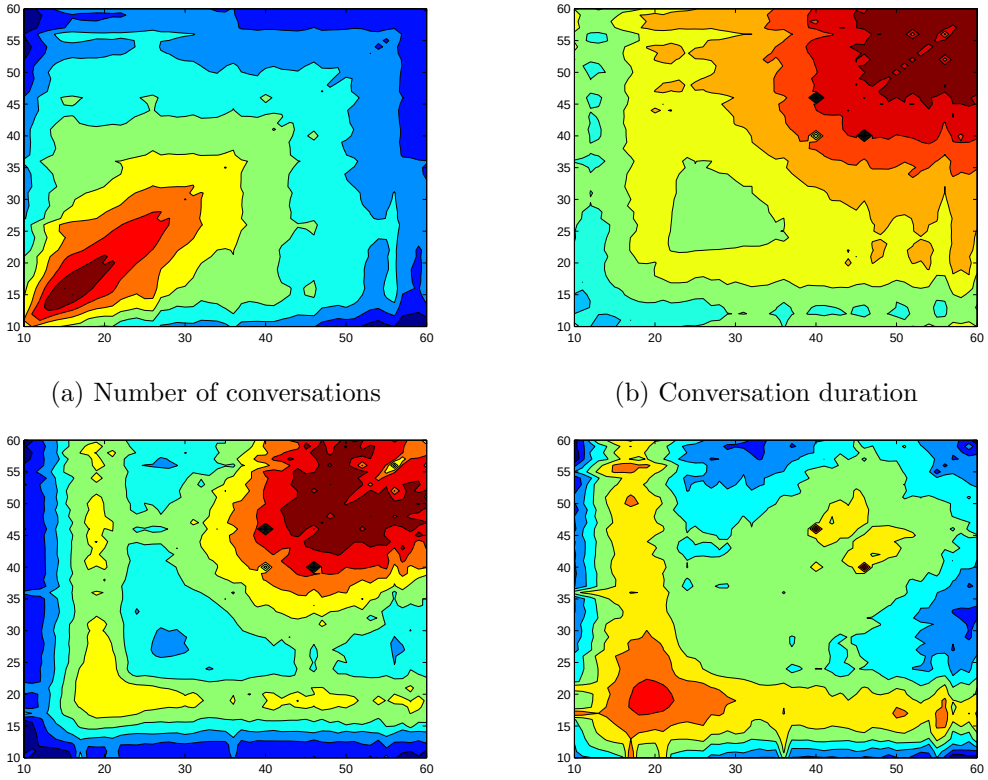


(a) Conversation duration



(b) Time between conversations

Figure 11: Temporal characteristics of conversations. (a) Average conversation duration per user; (b) Time between conversations of a user.



(a) Number of conversations

(b) Conversation duration

(c) Messages per conversation

(d) Messages per unit time

Figure 12: Communication characteristics of users by reported age. (a) Number of conversation between two people of different ages; (b) Average conversation duration; (c) Number of sent messages per conversation; (d) Number of sent messages per conversation per second. The color scale is defined in Figure 13.

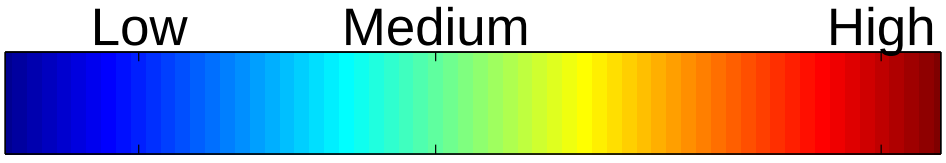


Figure 13: Color map that we use in several places: Cold blue colors correspond to low values, green to yellow colors to medium values, and hot red colors correspond to high values in the plots.

ages, we concentrate on the portion of the users with self-reported ages between 10 and 60 years.

Let a tuple (a_i, b_i, d_i, m_i) denote ith conversation in the entire dataset that occurred between users of ages a_i and b_i. The conversation had duration of m_i seconds during which m_i messages were exchanged. And let $C_{a,b} = \{(a_i, b_i, d_i, m_i) : a_i = a \wedge b_i = b\}$ denote a set of all conversations between users of ages a and b, respectively. Note that the notion of a conversation is symmetric, so also plots are symmetric.

Figure 12(a) shows the number of conversations between people of different ages. For every pair of user ages (a, b) the color indicates the size of set $C_{a,b}$, number of different conversations between users of ages a and b. Notice that the most conversations happen between people of ages 10 to 20. Also notice the diagonal trend which indicates that people tend to talk to people of similar age. This is especially true for age groups between 10 and 30. We will explore this observation in more detail in Section 6.

Figure 12(b) gives the similar type of plot but this time for every pair of user ages (a, b) we plot the average conversation duration: $\frac{1}{|C_{a,b}|} \sum_{i \in C_{a,b}} d_i$. Notice that older people tend to have longer conversations. Also notice a valley in the duration of conversation conversations for ages 25 – 35. We speculate that this dip may represent the shorter conversations associated with work-related communications. We also see that conversations between youngsters tend to be shorter.

We observe similar phenomena when plotting the average number of exchanged messages per conversation show in Figure 12(c). For every pair of ages (a, b) we plot $\frac{1}{|C_{a,b}|} \sum_{i \in C_{a,b}} m_i$. Again, older people exchange more messages, we observe a valley for ages 25 – 45 and a slight peak for ages 15 – 25.

Lastly, Figure 12(d) displays the number of exchanged messages per unit time. For every pair of user ages (a, b) we plot $\frac{1}{|C_{a,b}|} \sum_{i \in C_{a,b}} \frac{m_i}{d_i}$. Here, we see that younger people then to type/talk faster, while older people have a slower pace of exchanging messages.

5.2 Communication by age difference

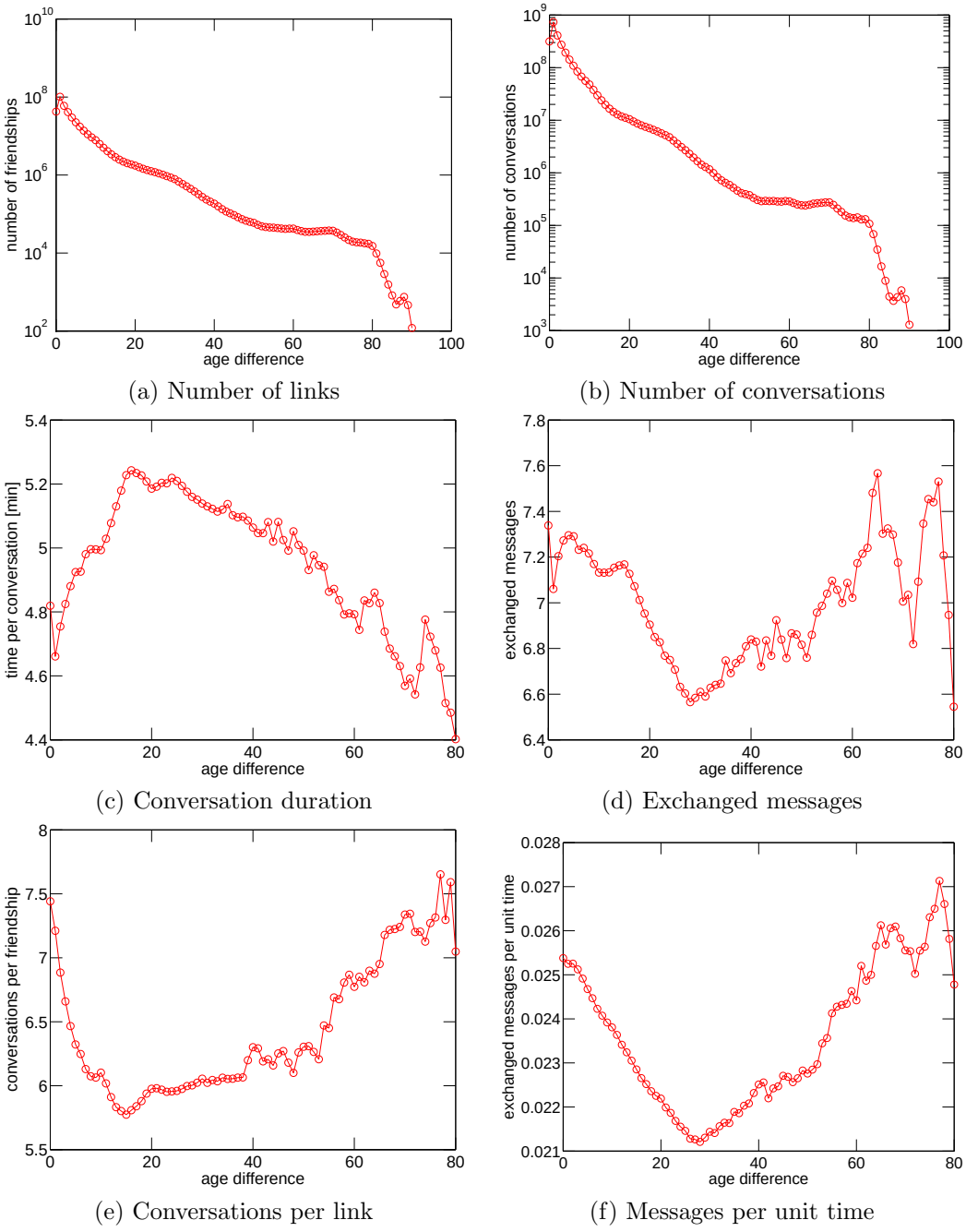
Next, we explore various communication patterns by the age difference between the users. We use the same data and the notation as in previous section. Let $C_a = \{(a_i, b_i, d_i, m_i) : |a_i - b_i| = a\}$ denote a set of conversations where the users were born a years apart.

Figure 14(a) plots the number of links in the social network with the endpoints having certain age difference on the log-linear scales. This plot shows the number of *unique* pairs of users of age difference a talking to each other. Figure 14(b) a "weighted" version of the above plot by also considering the number of conversations between users of age difference a, *i.e.,* it plots the number of conversations between users of different ages, $|C_a|$ vs. a.

Notice that links and conversations are strongly correlated with the age difference, which means that people tend to communicate more with the ones of the similar age.

Figure 14(c) shows the average conversation duration with the age difference between the users: $\frac{1}{|C_a|} \sum_{i \in C_a} d_i$. Interestingly, the mean conversation duration peaks at age difference of 20 years. This might be taken to correspond roughly to the gap between generations.

Figure 14(d) shows the average number of exchanged messages per conversation with the age difference between the users: $\frac{1}{|C_a|} \sum_{i \in C_a} m_i$. The plot follows the reverse pattern on conversation duration plot (Figure 14(c)). Notice the number of exchanged messages is the lowest for age difference of 25 years and the highest for people of the same ages and where the age difference is about 60 to 80 years. It is not clear whether this last peak can



(a) Number of links (b) Number of conversations

(c) Conversation duration (d) Exchanged messages

(e) Conversations per link (f) Messages per unit time

Figure 14: Communication characteristics with age difference between the users. (a) Number of links (pairs communicating) with the age difference. (b) Number of conversations. (c) Average conversation duration with the age difference. (d) Average number of exchanged messages per conversation as a function of the age difference between the users. (e) Number of conversations per link in the observation period with the age difference. (f) Number of exchanged messages per unit time as a function of age difference between the users.

be attributed to people that did not report their age truthfully.

Last, Figure 14(e) plots the number of conversations per link, *i.e.,* we show how the number of conversations between a pair of people that communicate varies with users' age difference. We plot $\frac{1}{q_a}\sum_{i \in C_a} m_i$, where q_a denotes the number of unique pairs of people of age difference a that communicate – they exchanged at least one message in the observation period. Notice that most used links are between people of same ages, again the curve reaches the bottom at around 20 year age difference.

And Figure 14(f) shows the number of exchanged messages per unit time. For every age difference a we calculate $\frac{1}{|C_a|}\sum_{i \in C_a} \frac{m_i}{d_i}$. The conversation intensity measured in the number of exchanged messages per unit time decays linearly with age difference and reaches the bottom at the age difference of 30 years. We speculate that the high communication per link for large differences may be explained by patterns in which people misreport their age.

Homophily (*i.e.*, love of the same) among people is the tendency of individuals to associate and bond with people who are similar. The plots reveal that there is strong homophily in the communication network. People tend to communicate more with people of similar reported age. This is especially evident for the number of buddies and conversations between people of the same ages. We also observe that the links between similar age people tend to be used more often with shorter and more intense (more exchanged messages) communication. The intensity of communication decays linearly with the difference in age.

5.3 Communication by gender

A different aspect of user demographics is the self-reported gender of communicating users. We again hone the focus down to consider the 15 billion two-user conversations that are in our dataset and present the results on gender–gender communication characteristics in Table 1.

Again, let $C_{g,h} = \{(g_i, h_i, d_i, m_i) : g_i = g \wedge h_i = h\}$ denote a set of conversations where the two participating users were of genders g and h Note g takes 3 possible values: female, male and unknown (not reported).

Table 1(a) plots $|C_{g,h}|$ for every combination of genders g and h. Table shows that 50% of conversations are between female and a male, while around 40% of the conversations occur between the users of same gender (20% for male, and 20% for female). Also we note a very small number of conversations between people who did not reveal their gender.

Table 1(b) shows the average conversation length in seconds broken down by the gender of the two participating users: $\frac{1}{|C_{g,h}|}\sum_{i \in C_{g,h}} d_i$. We find that male–male conversations tend to be shortest. Such conversations last approximately 4 minutes. Female–female conversations last 4.5 minutes on the average. The longest conversations are female–male conversations. These take more then 5 minutes on the average. Not only that female–male conversations are longer, but as can be seen from table 1(c), they also exchange 7.6 messages per conversation on the average as opposed to 6.6 and 5.9 for female–female and male–male, respectively. More precisely, table 1(c) plots $\frac{1}{|C_{g,h}|}\sum_{i \in C_{g,h}} m_i$.

Last, Table 1(d) shows the intensity of communication, *i.e.,* the average number of exchanged messages per minute of conversation: $\frac{1}{|C_{g,h}|}\sum_{i \in C_{g,h}} \frac{m_i}{d_i}$. Notice male–female conversations occur at higher rate of 1.5 messages per minute than mixed gender conversations where the rate is 1.43 messages per minute.

Interestingly, by looking at the number of edges in the network, where an edge indicates that two people exchanged at least 1 message in June 2006, we see that there are 300 million

	Unknown	Female	Male
Unknown	1.3	3.6	3.7
Female		21.3	49.9
Male			20.2

(a) Proportion of conversations

	Unknown	Female	Male
Unknown	277	301	277
Female		275	304
Male			252

(b) Conversation duration (seconds)

	Unknown	Female	Male
Unknown	5.7	7.1	6.7
Female		6.6	7.6
Male			5.9

(c) Exchanged messages per conversation

	Unknown	Female	Male
Unknown	1.25	1.42	1.38
Female		1.43	1.50
Male			1.42

(d) Exchanged messages per minute

Table 1: Cross gender communication. Data is based on 15.5 billion 2-person conversations from June 2006. (a) Percentage of conversations between users of different self-reported gender. (b) Average conversation length in seconds. (c) Number of exchanged messages per conversation. (d) Number of exchanged messages per minute of conversation.

edges between males, and 255 million between females. However there are 640 million mixed gender edges.

5.4 World geography and communication

We shall now describe geographical findings. We examine in particular how geography and distance influence communication patterns.

Figure 15 shows the geographical locations of Messenger users. The location of the user was obtained via reverse IP lookup. We simply plot all latitude/longitude positions of where users login from. The color of the circle corresponds to the *log* of the number of logins from the particular location. Notice how North America, Europe, and Japan are very dense, which means there are many users from those regions using Messenger. Also notice how, for the rest of the world, the population comes mostly from coastal regions. The maps were built solely from the location data. However, viewers can clearly recognize the shapes of the continents, *e.g.*, South America, Africa, Australia.

Figure 16 displays the same data, but now overlays it on top of a map of the world. The color of each dot corresponds to the number of users at a particular location. To denote locations with more than a million users we use circles, where the size of the circle is proportional to the log number users at the geographical location. These locations are large cities like New York, London, Paris, Madrid, San Francisco, Tokyo, Houston, Chicago and Boston.

Next, we use the United Nations gridded world population data. This data provides us the number of people estimated to be living in each cell. Given this data and the data from Figure 15, we calculate the number of users per capita, as displayed in Figure 17. In this case, the central region of the United States, Canada, Scandinavia, Ireland, Crete, Australia, New Zealand, and South Korea stand out as areas with the highest numbers of Messenger users per capita.

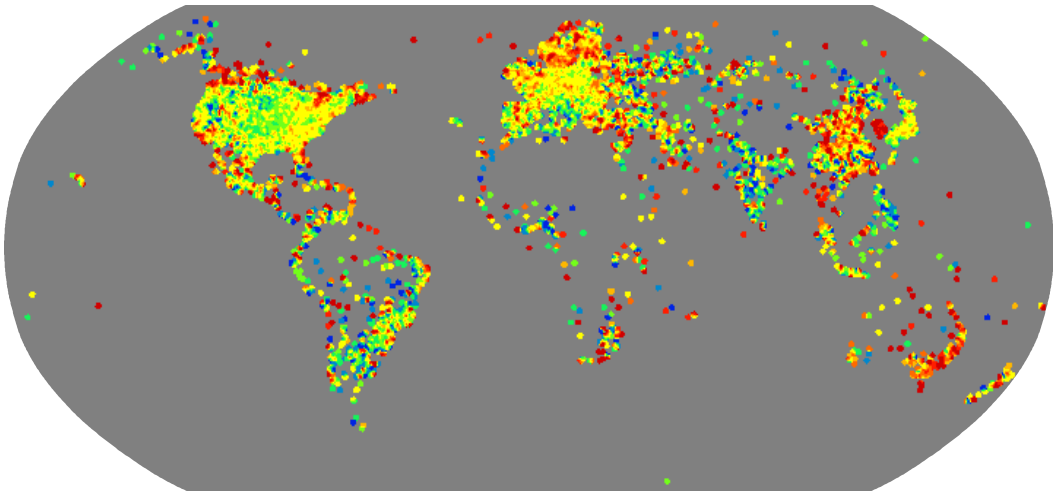


Figure 15: Geographic characteristics of communication. Each point is a location of a user login obtained from a mapping of IP address to geographical coordinates. The color of the dot corresponds to the number of logins from that particular location. Color scale is defined in Figure 13.

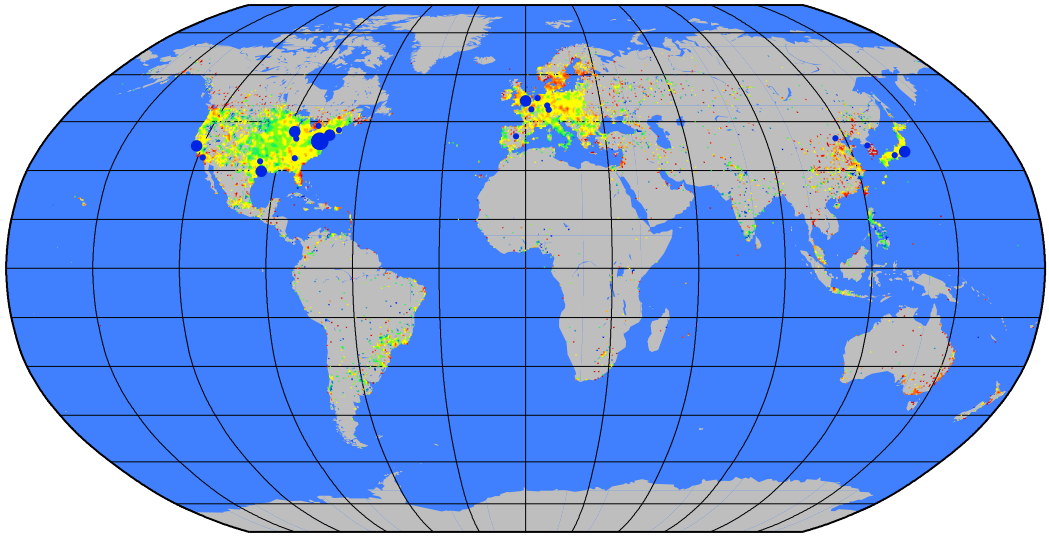


Figure 16: Number of users at particular geographic location superimposed on the map of the world. Color represents the number of users, and blue circles represent locations with more than 1 million users. Color scale is defined in Figure 13.

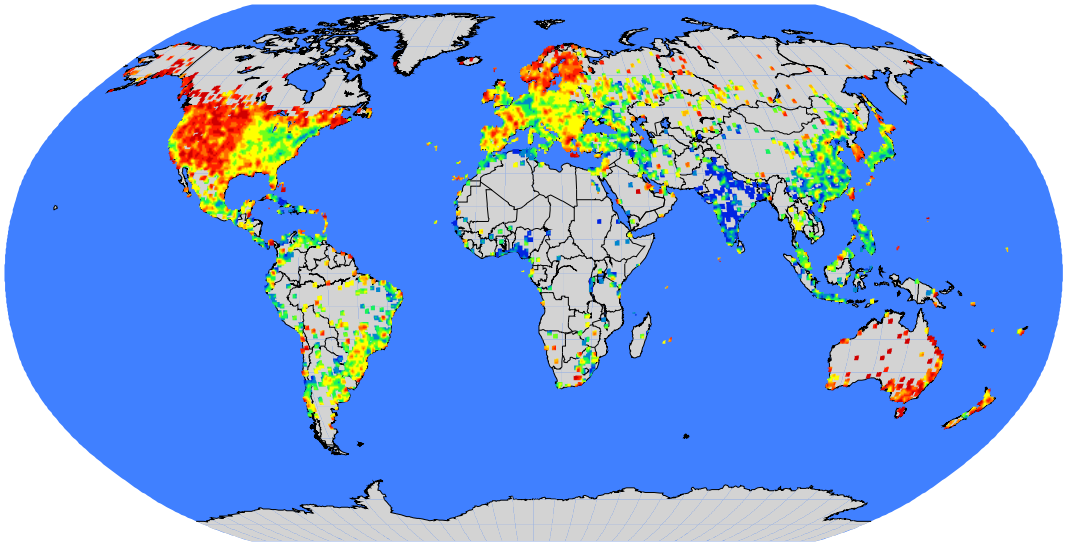


Figure 17: Number of Messenger users per capita. Color intensity corresponds the number of users per capita in the cell of the grid. Color scale is defined in Figure 13.

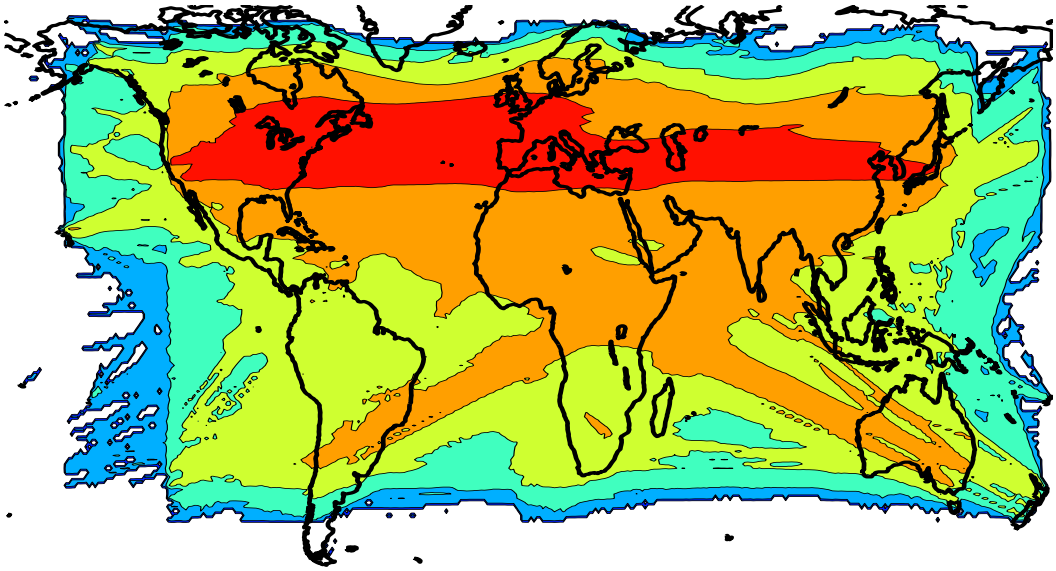


Figure 18: Communication heat map. For every conversation between two points on the Earth, we increase the counts of the cells on the line connecting them. Thus, every point on the map contains the number of conversations that cross it. The intensity of the color is proportional to the number of conversations. Color scale is defined in Figure 13.

Country	Fraction of population
Iceland	0.35
Spain	0.28
Netherlands	0.27
Canada	0.26
Sweden	0.25
Norway	0.25
Bahamas, The	0.24
Netherlands Antilles	0.24
Belgium	0.23
France	0.18
United Kingdom	0.17
Brazil	0.08
United States	0.08

Table 2: Top 10 countries with most the largest number of Messenger users. Fraction of country's population actively using Messenger.

Last, Figure 18 shows a communication heat map. We imagine the world map on a fine grid, where each cell of the grid contains the count of how many conversations passed through that point. We created the plot in the following manner: For all 2 user conversations, we geo-locate the participants, and then increased the count of all cells on the straight line between the locations of the two participants. In Figure 18, the color indicates the number of conversations crossing each point, showing the main directions of communication. For example, we see how Australia and New Zealand have their communication directed towards Europe and United States. Similar observations hold for Japan. Interestingly we see that Brazilian communications are weighted toward Europe and Asia. We can also explore and seek to visualize cross-Atlantic and cross-US communication.

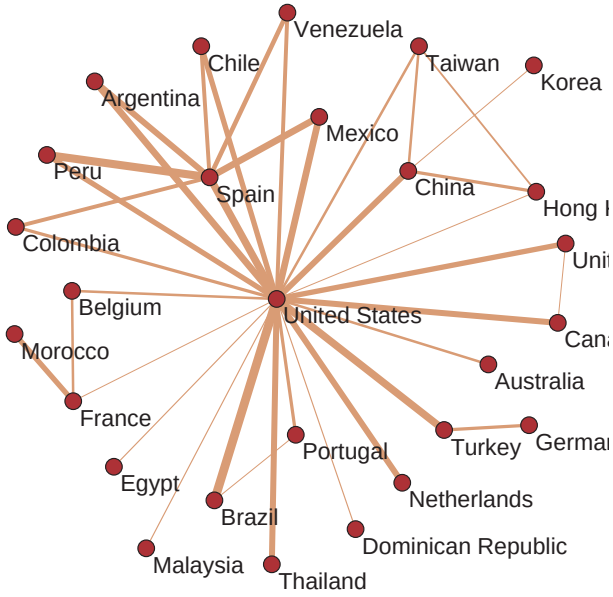
5.5 Communication between countries

Next, we briefly look at the communication characteristics between the countries.

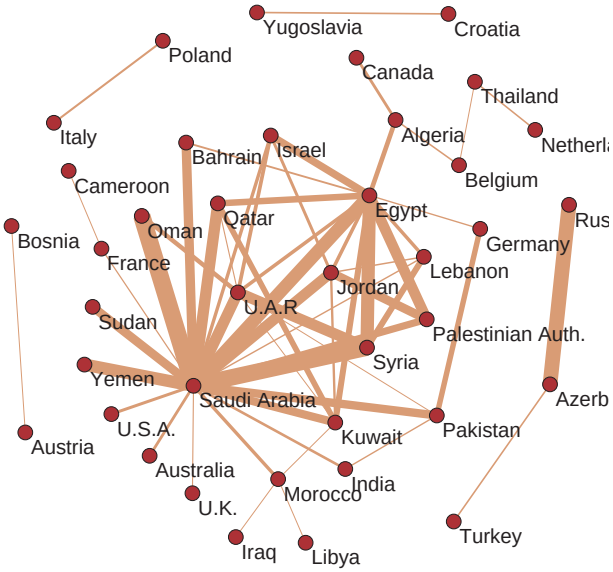
Figure 19(a) shows the top countries by the number of conversations. We examined all pairs of countries with more than 10 million conversations per month. The width of edge is proportional to the number of conversations between the countries. Notice that United States and Spain appear to serve as hubs and that edges are mainly between historically connected countries: Spain is connected with the Spanish speaking countries in South America, Germany has a link to Turkey, Portugal to Brazil, China to Korea, and Taiwan to Hong Kong.

Figure 19(b) gives a similar plot where we take country pairs by the average length of the conversation. Again, the width of the edge is proportional to the mean conversation length between the countries. The core of the network appears to be Arabic countries, *e.g.*, Saudi Arabia, Egypt, United Arab Emirates, Jordan, Syria, etc.

Table 2 shows the top 10 countries with the highest fraction of population using Messenger. These are mainly northern European countries and Canada. Countries with most of the users (US, Brazil) tend to have smaller fraction of population using Messenger.



(a) Number of conversations



(b) Conversation length

Figure 19: Communication among countries. (a) Conversations among countries with at least 10 million conversations in June 2006. Edge width corresponds to the log number conversations between the countries. (b) Countries by average length of the conversation. Edge width is proportional to the average length of the conversations between the countries.

Country	Conversations per user per day
Afghanistan	4.37
Netherlands Antilles	3.79
Jamaica	2.63
Cyprus	2.33
Hong Kong	2.27
Tunisia	2.25
Serbia	2.15
Dominican Republic	2.06
Bulgaria	2.07

Table 3: Top 10 countries by the number of conversations per user per day.

Country	Messages per user per day	Minutes talking per user per day
Afghanistan	32.00	20.91
Netherlands Antilles	24.12	17.43
Serbia	22.41	12.01
Bosnia and Herzegovina	22.40	11.41
Macedonia	19.52	10.46
Cyprus	19.33	12.37
Tunisia	19.17	13.54
Bulgaria	18.94	11.38
Croatia	17.78	10.05

Table 4: Top 10 countries by the number of messages and minutes talking per user per day.

Table 3 shows the top 10 countries by the number of conversations per user per day. Here the countries are very diverse with Afghanistan topping the list. The Netherlands Antilles appears on top 10 list for both the fraction of the population using Messenger and the number of conversations per user.

Last, Table 4 shows the top 10 countries by the number of messages and minutes talking per user per day. We note that the list of the countries is similar to those in Table 3. Afghanistan still tops the list but now most of the talkative counties come from Eastern Europe (Serbia, Bosnia, Bulgaria, Croatia).

5.6 Communication and geographical distance

We now examine how communications changes as the distance between people increases. Our hypothesis was that the number of conversations will decrease with the geographical distance between the users as they might be doing less coordination with one another on a daily basis. We expected that conversations between people who are farther apart would be somewhat longer as there might be a stronger need to catch up, given the low frequency of the conversations. We found that the conversation length does not increase with the distance between the users.

We group the conversations by the distance l between the users that communicate. Let

$$C_l = \{(u_i, v_i, d_i, m_i) : \text{geo-distance}(x_i, y_i) = l\}$$

denote a set of all conversations that occurred between users at geographical distance of l kilometers. Now, we examine how communication differs with distance l between the users.

Figure 20 shows the results of the communications and distance study. Figure 20(a) plots the number of unique pairs of users that communicate and are at geographical distance l. Figure 20(b) shows the weighted version of the plot where, each link is weighted by the number of conversations between the users, *i.e.,* we plot the total number of conversations $|C_l|$ at distance l. We see that the number of links rapidly decreases with distance. This may suggest that users use messenger mainly for short communication in their current environment. Notice that number of conversations also decreases with distance, however we observe a peak at distance of around 500 kilometers. Also notice a big drop in communication at distance of 5,000 kilometers which is roughly 3,500 miles, *i.e.,* the width of the Atlantic ocean. Also notice a peak at distance 3,600 kilometers (2,200 miles) which is the distance between east and west coast of the United States. The next peak occurs at 6,400 kilometers (4,000 miles), which can be attributed to the communication across the Atlantic ocean. The number of conversations again peaks 9,100 kilometers (6,000 miles, the distance between Japan and USA or Europe; West coast and Europe; Brazil and USA), and we observe the last peak at 11,000 kilometers (7,000 miles, Japan to East coast).

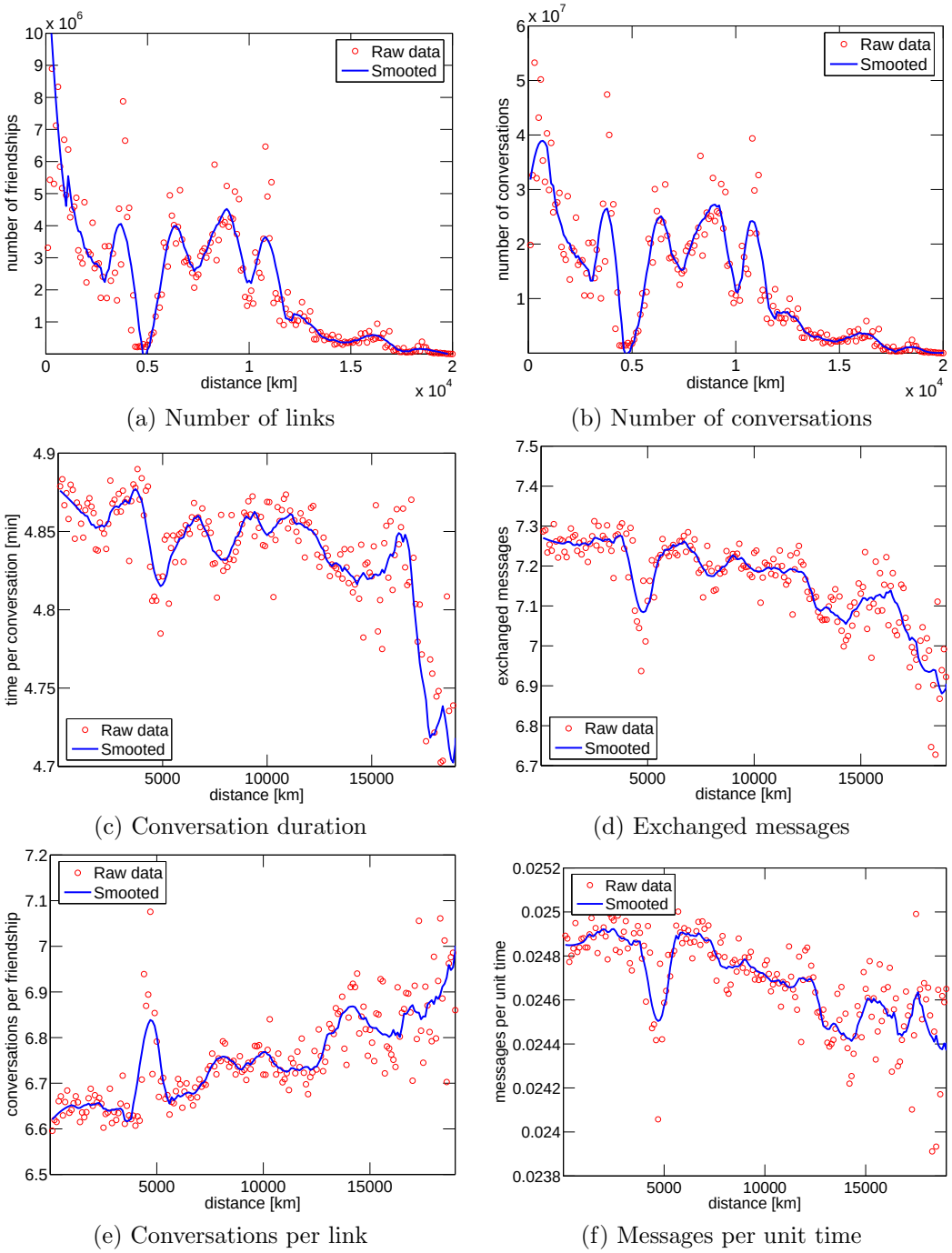
Figures 20(c) and 20(d) show conversation duration and the number of exchanged messages with distance. We plot $\frac{1}{|C_l|}\sum_{i \in C_l} d_i$ and $\frac{1}{|C_l|}\sum_{i \in C_l} m_i$, respectively. Observe that the number of exchanged messages and the conversation length do not increase with the distance. The conversation duration seems to decrease with the distance, while number of exchanged messages remains constant and then slowly decreases.

Last, Figure 20(e) shows the communication per link versus the distance. We plot $\frac{|C_l|}{q_l}$, where q_l denotes the number of unique pairs of users that communicate and are at distance l. The plot shows that longer links, *i.e.,* connections between people who are farther apart, are more often used than short links. This means that people who are farther apart tend to use Messenger more frequently. Figure 20(f) shows the number of exchanged messages per unit time with the distance: $\frac{1}{|C_l|}\sum_{i \in C_l} \frac{m_i}{d_i}$. Again, the communication tends to be slower as the distance increases, and we observe a dip at distance of 5000 kilometers.

In summary, we observe that the number of links and conversations decreases with distance. Similar observations hold for the conversation duration, the number of exchanged messages per conversation, and number of exchanged messages per unit time. On the other hand, the number of times a link is used tends to increase with the distance between the users. This suggests that people who are father apart tend to converse more often. We also notice that 5000 kilometers (3500 miles) seems to be a distance where communication exhibits drop offs. We speculate that these represent communications between the east and west coasts of the of the United States.

6 Homophily of communication

To measure the level at which people tend to communicate with similar people, we performed several experiments. First, we consider all 1.3 billion pairs of people who exchanged at least one message in June 2006, and calculate the correlation coefficient of various user attributes.



(a) Number of links (b) Number of conversations

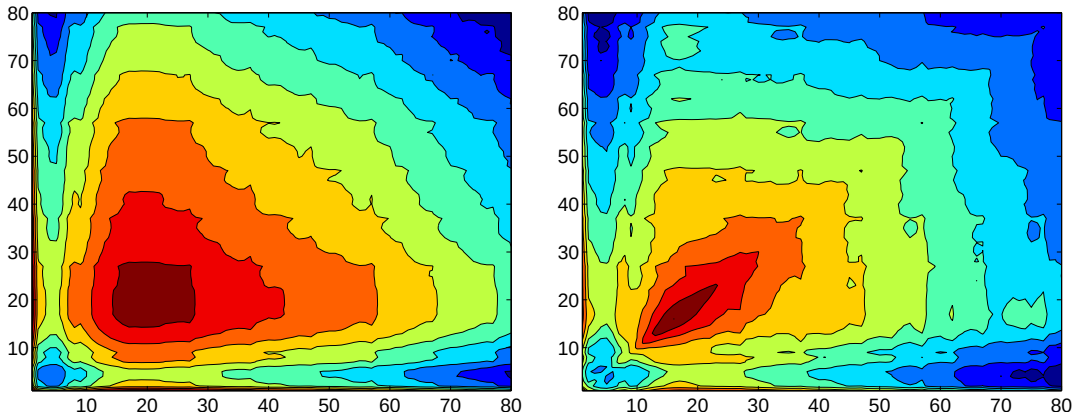
(c) Conversation duration (d) Exchanged messages

(e) Conversations per link (f) Messages per unit time

Figure 20: Communication with the distance. (a) Number of links (pairs of people that communicate) with the distance. (b) Number of conversations between people at particular distance. (c) Average conversation duration. (d) Number of exchanged messages per conversation. (e) Number of conversations per link (per pair of communicating users). (f) Number of exchanged messages per unit time.

Attribute	Random	Communicate
Age	-0.0001	0.297
Gender	0.0001	-0.032
ZIP	-0.0003	0.557
County	0.0005	0.704
Language	-0.0001	0.694

Table 5: Correlation coefficients for random pairs of people and pairs of people who communicate. We compare the degree of homophily of random pairs of users with pairs of users that communicate.



(a) Random (b) Communicate

Figure 21: Number of pairs of people of different ages. We plot ages of two people and color corresponds to the number of such pairs. (a) Ages of randomly selected pairs of people; we note there is little correlation. (b) Ages of people who communicate with one another, *i.e.,* ages of people at the endpoints of links in the communication network. The high correlation is captured by the diagonal trend.

We contrast this statistic with the correlation coefficient where we choose users via a process of uniform random sampling across 1.3 billion users.

We also consider two measures of similarity—the correlation coefficient and the probability that users have the same attribute value, *e.g.,* that users come from the same countries.

Table 5 compares correlation coefficient of various user attributes when pairs of users are chosen uniformly at random with pairs of users that communicate. As attributes are not correlated for random pairs of people, they are highly correlated for users who communicate. Also, notice that gender communication is negatively correlated—people tend to communicate more with people of a different gender.

Figure 21 further illustrates the results of table 5. We plot the number of pairs of people of particular age. Figure 21(a) shows the distribution over the randomly sampled pairs, *i.e.,* the Messenger user base created by sampling from 1.3 billion random user pairs, and plot the distribution over reported ages. As most of the population comes from the age group 10-30, the distribution of random pairs of people reaches the mode at those ages but there is no correlation. Figure 21(b) shows the distribution of ages over the pairs of people that

Attribute	Random	Communicate
Age	0.030	0.162
Gender	0.434	0.426
ZIP	0.001	0.23
County	0.046	0.734
Language	0.030	0.798

Table 6: Probability of observing a user with the same value of the attribute. We compare the homophily seen in random pairs of users with pairs of users who communicate.

communicate. Notice the significantly higher correlation (as represented by the diagonal trend on the plot), where people tend to communicate more with others at a similar age. Notice that the distribution again reaches the mode at the age group 10–30.

A different method to measure association is to measure the probability that a pair of users will show an exact match in values of an attribute, *i.e.,* whether the two users come from the same country, speak the same language, etc. Table 6 shows the results for the probability of users sharing the same attribute value. We make similar observations as in Table 5. People that communicate are more likely to share common characteristics – age, location, language, with the exception of gender. Also notice that the most common attribute of people who communicate is the language.

We also note that the amount of communication decreases with user dissimilarity. This can be see from Figure 20 where we notice that the amount of communication between a pair of people decreases with distance. Similar observations can also be made for age difference between the users (Figure 14).

7 The communication network

So far we have examined communication patterns as framed by pairwise communications. We now create a communication network from the data where nodes represent users and there is an edge connecting a pair of users if they exchanged at least one message during our observation period. Using this network, we can examine the typical *social distances* between people, *i.e.,* the number of associations that separate a random pair. This analysis seeks to understand how many people can be reached within certain numbers of hops among people who communicate. Also, we test the transitivity of the network, *i.e.,* the degree at which pairs with a common friend tend to be connected.

We took only two-user conversations from June 2006 and we build a graph, where each node corresponds to person and there is an undirected edges between the two nodes if the two users were engaged in an active conversation during the observation time. The graph contains 179,792,538 nodes, and 1,342,246,427 edges, that were extracted based on 15,010,572,090 two-user conversations.

Note that we do not have complete buddy lists but we infer the buddy network from the communication patterns in June 2006. This means that for every person, we have a node and there is an undirected edge between two nodes if they had at least one conversation (one message exchanged) in June 2006.

Figures 22–26 show the properties of the structure of the communication network. The degree distribution shown in Figure 22 is heavy tailed but does not follow a power-law

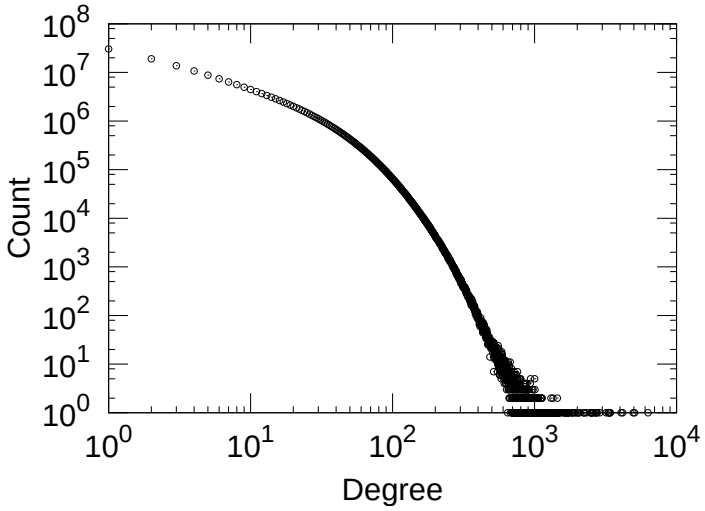


Figure 22: Degree distribution (number of people a person exchanged at least one message in the observation period).

distribution.

Figure 23 shows the distribution of the distances in the network. We plot the log number of pairs of nodes that are reachable in less than h hops (steps). We see that the diameter of the network is 7.8. We used the ANF algorithm by Palmer et al (Palmer et al., 2002) that uses approximate hashing to calculate the hop plot efficiently. Figure 24 gives a similar plot where we show the log number of nodes reachable at particular distance (non-cumulative version of Figure 23). To approximate the distribution of the distances we randomly sampled 1000 nodes and for each node then calculated shortest paths to all other nodes. Notice that the distribution reaches the mode at 7 hops, 90-th percentile at 7.8, and the average at 5.5. This means that a random pair of nodes is less than 6 (5.5) hops apart on the average. Also, notice that long paths, *i.e.,* nodes that are far apart, also exist. The longest shortest path we were able to find had length 29.

Figure 25 graphs the clustering coefficient, which is defined as the average number of triangles around a nodes of degree k (Watts and Strogatz, 1998). Previous results on measuring web graph, and theoretical analyses show that clustering coefficient decays as k^{-1} (exponent -1) with node degree k (Ravasz and Barabasi, 2003). Interestingly, the clustering coefficient decays very slowly with exponent -0.37 with the degree of a node. This suggest that clustering in the messenger network is much higher than expected and people with common friends also tend to be connected.

Figure 26 displays the distribution of the connected components in the network. Notice that there is a giant component comprising of 99.9% of the nodes in the network, while the rest of the components are small, and the distribution follows a power-law.

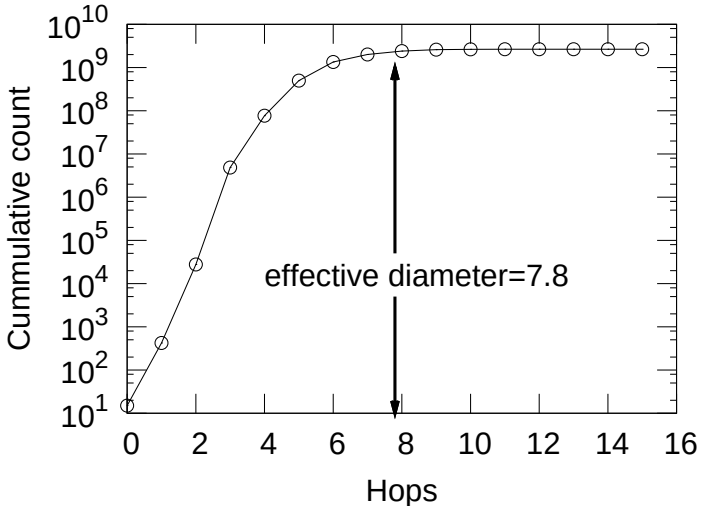


Figure 23: Hop plot. The number of pairs of nodes reachable at x or less hops. Notice the 90% effective diameter is 7.8, which means that in 8 hops more than 90% of pairs of nodes can be reached.

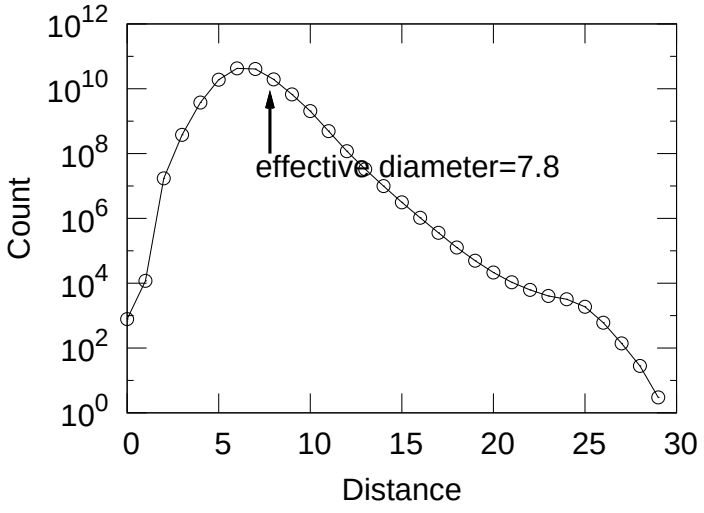


Figure 24: Distribution over the shortest path lengths. Figure shows a non-cumulative version of Figure **??**. Average shortest path has length 5.5, the distribution reaches the mode at 7 hops, and 90% effective diameter is 7.8. This means that majority of network is reachable in few hops, however long paths of around 30 hops exist.

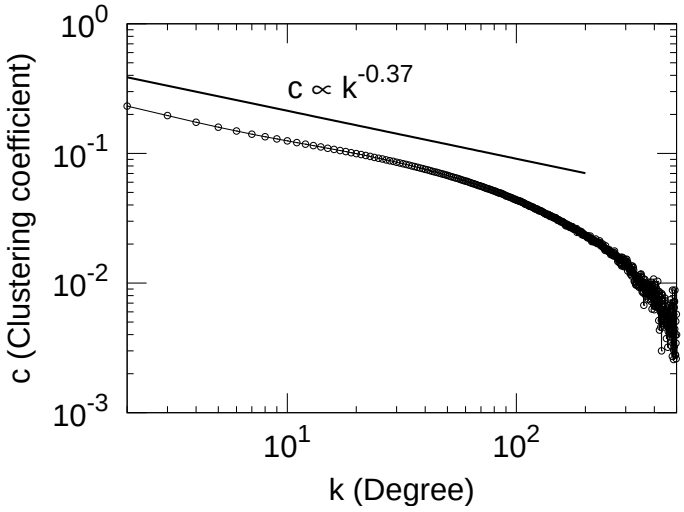


Figure 25: Clustering coefficient. We sampled 150 million nodes; The exponent is very high and not -1 as expected and predicted by the Theory, suggesting that the clustering in the messenger network is much higher than expected and people with common friends also tend to be connected.

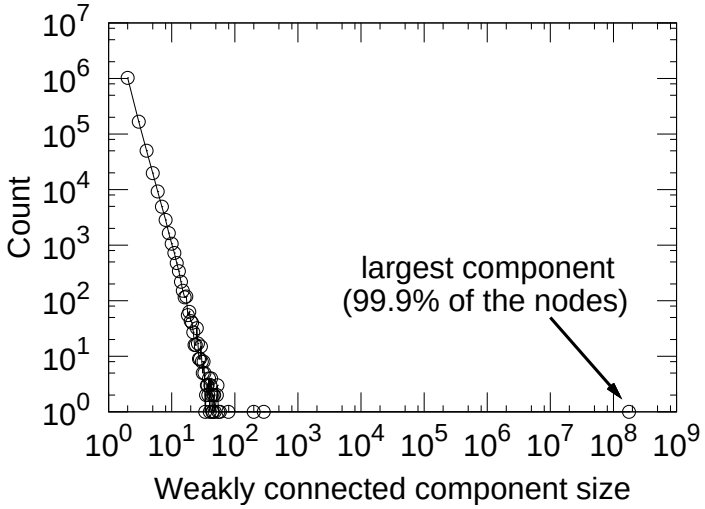


Figure 26: Distribution of connected components. 99.9% of the nodes belong to largest connected component.

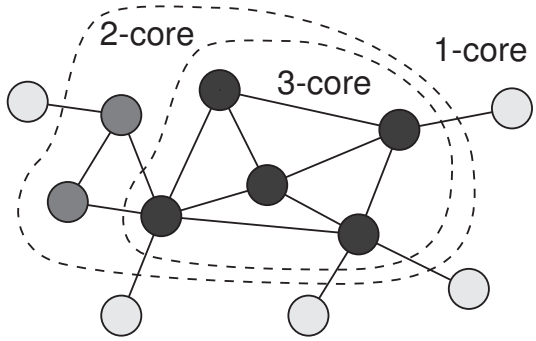


Figure 27: k-core decomposition of a small graph. Nodes contained in each closed line belong to a given k-core. Inside each k-core all nodes have degree larger than k (after removing all nodes with degree less than k.

7.1 Network cores

Next, we further study connectivity of the communication network by examining the k-cores (Batagelj and Zaversnik, 2002). The concept of k-core is a generalization of the giant connected component. The k-core of a network is a set of vertices K, where each vertex in K has at least k edges to other vertices in K. Figure 27 gives an example of a network where all the nodes belong to 1-core, and all the nodes inside the inner circle belong to a 3-core. The distribution of the sizes of k-cores gives us an idea of how quickly does the network shrink as we move towards the core of it.

The k-core of a graph may be obtained by the following algorithm: delete from the network all vertices of degree less than k. This process will decrease degrees of some non-deleted vertices, so more vertices will have degree less than k. We keep pruning vertices until all remaining vertices have degree at least k. We call the remaining vertices a k-core. The algorithm to compute k-core is very efficient as it runs in linear time.

We carried out the computation of the k-cores on the communication network. Figure 28 plots the number of nodes in a core of order k. Notice that up to value of $k \approx 20$ the core sizes are remarkably stable – the number of nodes in the core drops for only an order of magnitude. After $k > 20$ the core size rapidly drops. The central part of the communication network is composed of 79 nodes where each of them has more than 68 edges inside the set. The structure of Messenger communication network is quite different from the graph of internet (autonomous systems) where it has been observed (Alvarez-Hamelin et al., 2005) that the size of a k-core decays as a power-law with k. Here we see that the core sizes remains very stable up to a degree ≈ 20, and only then start to rapidly degrease. This means that the nodes with degrees of less than 20 are on the fringe of the network, and the core starts to rapidly decrease as nodes of degree 20 or more are deleted.

7.2 Strength of the ties

It has been observed by Albert et al (Albert et al., 2000) that many real-world networks are very robust to node-level attacks. Authors showed that networks like the World Wide Web, Internet, and a number of social networks display a high degree of robustness to random

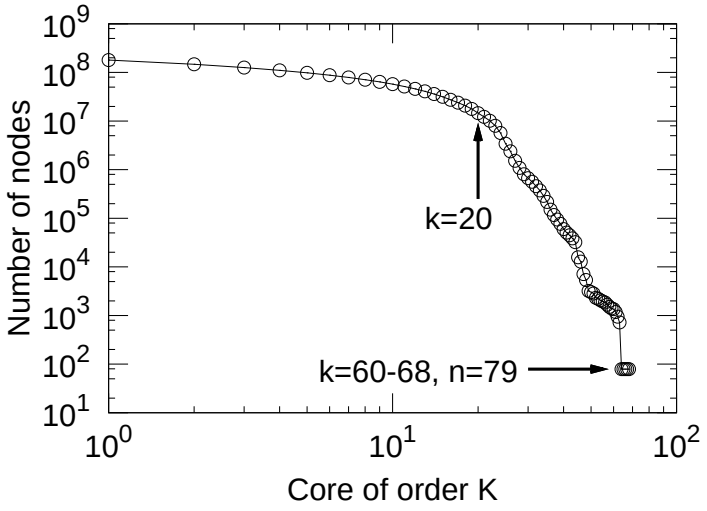


Figure 28: Distribution of the sizes of cores of order k.

node removals, *i.e.,* one has to remove lots of uniformly at random chosen nodes to make the network disconnected. On the contrary, targeted attacks are very effective. Removing a few high degree nodes has a drastic impart on the connectivity of the network.

Here, we perform a similar experiment where we study how does the network fall apart if "strong", *i.e.,* heavily used, edges are removed from the network. We consider several different definitions of "heavily used", e.g., number of exchanged messages between a pair of users, number of conversations, etc. We measure what kind of edges are most important for network connectivity.

In the context of a small instant messaging buddy network, a similar experiment was performed by Shi et al (Shi et al., 2007). Authors used the number of common friends of at then ends of an edge as a measure of the tie strength. Since the number of edges here is too large (1.3 billion) to remove edges one by one, we use the following approach. We order the nodes by the decreasing value of some criteria that measures how engaged is the user (*e.g.*, number of messages sent by the user in the observation period). Given this ordering of the nodes we start deleting nodes from users that are heavily engaged towards the low engagement users. We repeatedly delete nodes and edges attached to them. As we keep deleting nodes we observe the evolution of the properties of the communication network.

We consider the following criteria that measure how engaged is the user in the communication:

- Avg. sent: average number of sent messages per user's conversation
- Avg. time: average duration of user's conversations
- Links: number of links of a user (node degree), *i.e.* number of different people he or she exchanged messages with
- Conversations: total number of conversations of a user in the observation period

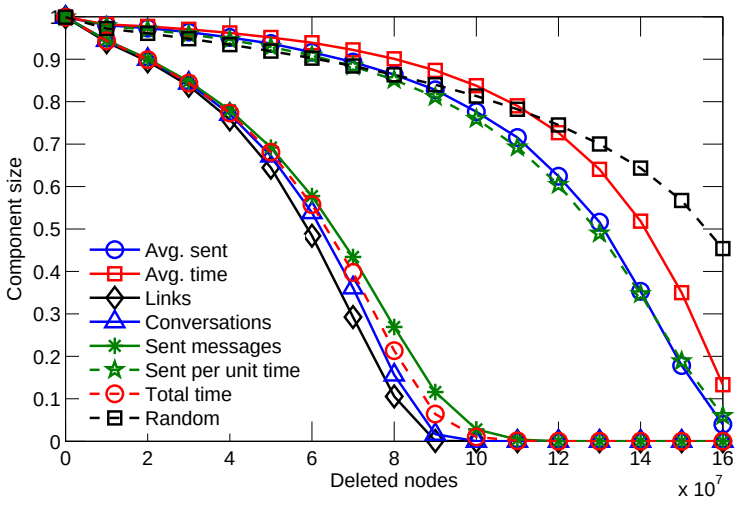


Figure 29: We order nodes in the decreasing order by some criteria and keep deleting nodes. We record the relative size of the largest connected component with the number of nodes removed.

- Sent messages: total number of sent messages by a user in the observation period
- Sent per unit time: number of sent messages per unit time of a conversation
- Total time: total conversation time of a user in the observation period

At each step of the experiment we removed 10 million nodes starting with highly engaged users. We then measured the relative size of the largest connected component, *i.e.* given the network at particular step we find the fraction of the nodes belonging to the largest connected component of the network.

Figure 29 plots the evolution of the fraction of nodes in the largest connected component with the number of deleted nodes. For each of 7 different measures of how engaged is the node we plot a separate curve. For comparison, we also consider random deletion (ordering) of the nodes.

First, notice two different groups of behavior: using number of links, number of conversations, total conversation time or number of sent messages very quickly decreases the size of the largest component. On the other hand, average time per conversation, average number of sent messages and number of sent messages per unit time very slowly decrease the component size.

Not surprisingly the component size decreases most rapidly when the nodes are deleted in the decreasing number of links they have, *i.e.,* number of different people they communicate to. Random ordering of the nodes shrinks the component the slowest. After removing 160 million out of 180 million nodes the largest component still contain about the half of the node. Surprisingly, when deleting up to 100 million nodes the average time per conversation criteria shrinks the component even slower than random node deletion. Next, we further investigate this phenomena.

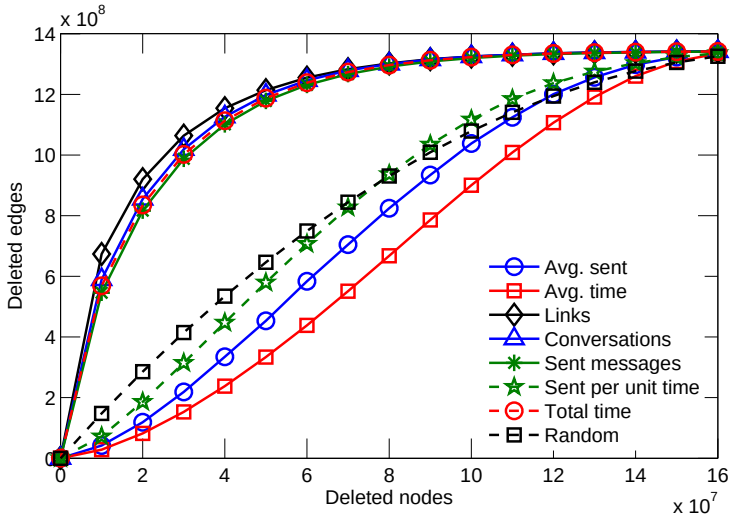


Figure 30: We order nodes in the decreasing order by some criteria and keep deleting nodes. We record the number of edges that got deleted when removing the chosen number of nodes.

Figure 30 plots the number of removed edges from the network as the nodes are being deleted. Similarly as in Figure 29 deleting nodes by the inverse number of edges, removes edges the fastest. As in Figure 30 the same group of node ordering criteria (number of conversations, total conversation time or number of sent messages) removes edges from the networks as fast as the number of links criteria.

However, we notice the difference for the other group. We see that random node removal removes about the linear time of the edges. Surprisingly, deleting nodes by average time per conversation, average numbers of sent messages, or numbers of sent messages per unit time, less edges get deleted. This shows that users with long conversations and many messages per conversation tend to have smaller degrees, even though that in Figure 29 we saw that these users are more effective in breaking the network connectivity than random users, *i.e.*, random node deletion. Figure 30 also shows that average number of per conversation removes edges in the slowest manner. Intuitively, this makes sense:If every user spends the same amount of time talking, then people with short conversations will talk to many people in a given amount of time, while users with long conversations with only be able to interact with a few users within the same time budget.

8 Conclusion

We described the creation and analysis of an anonymized dataset representing the communication patterns of people using a large, popular instant-messaging system. We reviewed the creation of the dataset, capturing anonymized, high-level communication activities and demographics in June 2006. The dataset contains more than 30 billion conversations among 240 million people. We described the creation and analysis of a communication graph from the data containing 190 million nodes and 1.3 billion edges. We discovered that the graph is

well connected and highly clustered. We reviewed the influence of multiple factors on communication frequency and duration. We found strong influences of homophily in activities, where people with similar characteristics tend to communicate more, with the exception of gender, where we found that cross-gender conversations are both more frequent and of longer duration than conversations with users of the same reported gender. In other directions of research with the dataset, we have pursued the use of machine learning and inference to learn predictive models that can forecast such properties as communication frequencies and durations of conversations among people as a function of the structural and demographic attributes of the communicators. Our future directions for research include gaining an understanding of the dynamics of the structure of the communication network via a study of the evolution of the network over time.

Acknowledgments

We thank Dan Liebling for help with generated world map plots, and Dimitris Achlioptas and Susan Dumais for helpful suggestions.

References

Albert, R., Jeong, H., and Barabasi, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406:378.

Alvarez-Hamelin, J. I., Dall'Asta, L., Barrat, A., and Vespignani, A. (2005). Analysis and visualization of large scale networks using the k-core decomposition. In *ECCS '05: European Conference on Complex Systems*.

Avrahami, D. and Hudson, S. E. (2006). Communication characteristics of instant messaging: effects and predictions of interpersonal relationships. In *CSCW '06: Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 505–514, New York, NY, USA. ACM Press.

Barabasi, A.-L. (2005). The origin of bursts and heavy tails in human dynamics. *Nature*, 435:207.

Batagelj, V. and Zaversnik, M. (2002). Generalized cores. *ArXiv*, (cs.DS/0202039).

IDC Market Analysis (2005). *Worldwide Enterprise Instant Messaging Applications 2005–2009 Forecast and 2004 Vendor Shares: Clearing the Decks for Substantial Growth*.

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

Nardi, B. A., Whittaker, S., and Bradner, E. (2000). Interaction and outeraction: instant messaging in action. In *CSCW '00: Proceedings of the 2000 ACM conference on Computer supported cooperative work*, pages 79–88.

Palmer, C. R., Gibbons, P. B., and Faloutsos, C. (2002). Anf: a fast and scalable tool for data mining in massive graphs. In *KDD '02: Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 81–90.

Ravasz, E. and Barabasi, A.-L. (2003). Hierarchical organization in complex networks. *Physical Review E*, 67(2):026112.

Shi, X., Adamic, L. A., and Strauss, M. J. (2007). Networks of strong ties. *Physica A Statistical Mechanics and its Applications*, 378:33–47.

Tauro, S. L., Palmer, C., Siganos, G., and Faloutsos, M. (2001). A simple conceptual model for the internet topology. In *GLOBECOM '01: Global Telecommunications Conference*, volume 3, pages 1667 – 1671.

Voida, A., Newstetter, W. C., and Mynatt, E. D. (2002). When conventions collide: the tensions of instant messaging attributed. In *CHI '02: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 187–194.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393:440–442.