

Comparative Statistical Analysis of Large-Scale Calling and SMS Network

Jian Li, Wenjun Wang, Pengfei Jiao^(✉), and Haodong Lyu

Tianjin University, Tianjin 300072, China
pjiao@tju.edu.cn

Abstract. Mobile phone call and SMS are one the most popular communication means in modern society. The interactions between individuals result in a complex community structure that embody the social evolution. The real time call and SMS records of 36 million mobile phone users provide us with a valuable proxy to understand the change of communication behaviors embedded in social networks. Mobile phone users call each other and send SMS forming two paralleled directed social networks. We perform a detailed analysis on these two weighted networks and their derivative networks by examining their degree, weight, strength distribution, clustering coefficients and topological overlapa, as well as the correlations among these quantities. We focus on comparing the statistical properties of these networks and try to discover and interpret the discrepancy between calling and SMS networks. The finings shows that these networks have many structural features in common and exhibit idiosyncratic features when compared with each other. These findings offer insight into the pattern differences between the two large networks.

Keywords: Mobile phone network · Human dynamic · Complex network

1 Introduction

Mobile phones are wildly used in modern society. People mainly used phone to make calls and send SMS before the birth of smartphone. With the boom of mobile Internet, instant message applications such as WeChat change people's communication patterns. It was released in October 2015 by the Ministry of Industry and Information Technology of China that there were 1.302 billion mobile phone users and a quarter of them were 4G network users. The massive data brought by billions of users provide us with opportunities to understand human's communication patterns and the evolution of social networks.

Gonzalez et al. studied mobile phone records of more than 6 million anonymous mobile phone users over a period of 6 months in an European country and found that the density function followed a shifted power law with an exponential cutoff. Though the individual's communication patterns might not be homogeneous, the communication pattern of human beings are largely predictable at the aggregate level [1].

Individual communication patterns are too frequently changing to be explored. However, the records of mobile phone provide us with opportunity to study human communication dynamics via complex networks. We know phone call and SMS are two widely used communication methods in our daily life, and the duration analysis on two consecutive calls [2,3] and short message correspondences has achieved significant gains that the distribution of inter-communication durations have a fat tail and human interactions exhibit non-Poissonian characteristics. However, the communication networks' evolution and comparative analysis between calling network and SMS network have long been ignored.

In this paper we present a comparative analysis on calling and SMS networks constructed from a data set of 36 million anonymous users' mobile phone records over one month. We investigate these weighted, large scale, peer to peer social interaction networks, with emphasis on the eight derivative networks of calling and SMS networks. The eight networks include directed network, mutual network, statistically validated directed network, statistically validated mutual network and their corresponding giant components. We carry out a systematic analysis of basic and advanced network characteristics, and our work mainly focus on the qualitative analysis and the comparison between the calling and SMS networks in the idiosyncratic characteristics.

2 Data and Network Construction

The networks are constructed from the detailed mobile phone records of more than 36 million mobile phone users. We split the data set into call set and SMS set. In each part we treat all users in our data set as nodes and a directed edge is drawn from a call maker to a call receiver. Both data sets have an original network, a mutual network and corresponding statistically validated network. The statistically validated network is performed by comparing the number of connection between each pair of originator and receiver with a null hypothesis of random matching between the originator and receiver. The method is a variant of the concept proposed in Ref. [4] which evolves to different versions in different systems [5-7]. The statistically validated method is attained by considering the number of calls (SMS) originated by the originator and the number of call (SMS) received by the receiver. Follow that, the compatibility is checked between the number of connections between them and the null hypothesis that these calls an SMS are attained by setting receivers randomly. The test allows us to assign a p -value to every tested originator and receiver pair and the p -values are then compared with the 1% of a statistical threshold. If the p -values are less than the threshold, we assume the connection has a social original. More detail information can be found in Ref. [4]

It is generally held that the frequency of SMS usage is much higher than calling, but our statistical results of mobile phone records show a contrary phenomenon. Due to the rapid development of instant messaging applications and SMS has been on a decline reported by the MIIT, we believe the result achieved

Table 1. Sizes of the calling and SMS networks and their giant components. N_{node} and N_{edge} are respectively the number of nodes and edges of a calling network. N_{Comp} , $N_{GC,node}$ and $N_{GC,edge}$ are respectively the number of components, the number of nodes and edges of the giant component of network.

CALL	N_{node}	N_{edge}	N_{Comp}	$N_{GC,node}$	$N_{GC,edge}$
DCN	35,637,916	166,315,663	402,940	34,567,395	165,614,958
SVDCN	18,514,998	77,367,187	418,599	17,407,810	46,083,991
MCN	5,302,334	25,937,081	164,756	4,917,763	51,401,150
SVMCN	4,190,455	8,983,869	157,462	3,822,224	17,520,235
SMS	N_{node}	N_{edge}	N_{Comp}	$N_{GC,node}$	$N_{GC,edge}$
DMN	16,261,236	104,211,414	211,992	15,752,994	52,010,649
SVDMN	9,820,402	35,026,653	254,541	9,179,469	17,199,919
MMN	322,117	403,552	9,110	11,889	25,099
SVMMN	138,247	77,443	63,893	60	124

in our study are reasonable. All the detail information about the network are summarized in Table 1.

3 Network Characteristics

3.1 Degree and Degree-Degree Distribution

As a basic network characteristic, degree distributions of networks are shown in Fig. 1. Since DCN and DMN and their statistical validated networks are directed, we respectively show their in-degree and out-degree distributions in Fig. 1(a) and (b). We find that all the eight probability distributions are well fitted by an exponentially truncated power law [10]: $P(k) = ak^{-\gamma_k} e^{-k/k_c}$

The distributions are all skewed with a fat tail, which indicates that most users communicate with only a few individuals. There exists noticeable differences between in-degree and out-degree distribution in panels (a) and (b). The out-degree distributions have a fatter tail and narrower distribution range than in-degree distributions. This indicates that most users' initiative usage frequency of mobile phone is lower than passive usage frequency.

In general, communication interaction via SMS between individuals should be more frequently than interactions through phone calls, because message reply is more convenience than dialing phone number and SMS is easily formed dialogic conversation. In panels (c) and (d), the degree distributions of MMN and SVMMN are quite sparse and the distribution range are also very narrow. It appears plausible that the mutual network is dominated by trusted interaction, that's to say people only reply to whom they know well. So we can safely infer that users mainly use phone call for interaction rather than SMS. An interesting phenomenon in (a) and (c) is that the MCN's right-end tail is much narrower

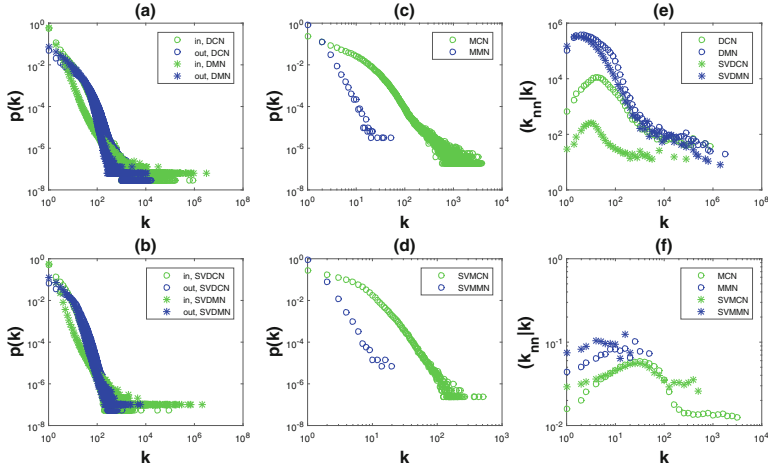


Fig. 1. Degree and Degree-degree distribution. (a) Distributions of in-degree and out-degree of DCN and DMN. (b) Distributions of in-degree and out-degree of SVDCN and SVDMM. (c) Degree distributions of MCN and MMN. (d) Degree distributions of SVMCN and SVMMN. (e) Average nearest neighbor degree $\langle k_{nn}|k \rangle$ as a function of degree k for DCN, DMN, SVDCN and SVDMM. (f) Average nearest neighbor degree $\langle k_{nn}|k \rangle$ as a function of degree k for MCN, MMN, SVMCN and SVMMN.

than DCN’s, because the calling criterion for the construction of MCN is capable of filtering out most unusual calls associated with sale calling and hot lines.

Social networks are expected to be assortative: people with many friends tend to connect with other people who also have many friends. Nodes in the networks are not independent if the degrees of adjacent nodes have degree-degree correlation. In a practical way, we define the average nearest neighbor degree of a node $k_{nn,i} = (1/k_i) \sum_{j \in N(v_i)} k_j$, where N_{v_i} denotes the neighbor nodes of node v_i and edge direction is ignored when dealing with directed networks. By averaging this over all nodes in the network of a given degree k , we can calculate the average degree of nearest neighbors with degree k denoted by $\langle k_{nn}|k \rangle$, which corresponds to $\sum_{k'} k' P(k'|k)$ [11]. It’s said that the network exhibits assortative mixing if $\langle k_{nn}|k \rangle$ monotonously increasing and disassortative mixing if it monotonously decreasing as a function of k [12].

We present the results of degree-degree distributions in Fig. 1. In order to clearly demonstrate the degree correlation between adjacent nodes, we bin the experimental results. In panel (e) and (f), all distributions exhibit a similar feature that the curves keep increasing before reaching the peak and then decrease gradually. This interesting phenomenon shows that the network is assortative mixing if the k value is below a certain threshold and disassortative mixing if the k value is the threshold.

In panel (e), we find in SMS’s directed networks that nodes with k less than 10 only interact with large degree nodes. However, we can’t find similar feature

in SMS’s mutual networks. This can partly be explained by the fact that nodes with a few edges often only connect with the nodes with large degree which usually corresponds to short messages robots or short message advertisers.

3.2 Edge Weight Distribution

For calling and SMS networks, we define the edge weight w_{ij}^N as the number of calls or messages occurred between user i and user j and the edge direction in the networks is ignored. Here we focus on the distribution comparison between SMS networks and calling networks. Figure 2 shows the distribution of the main networks from the two data sets. From the Fig. 2 we can see that all the distribution curves of SMS have a shallower slope and a wider distribution range. This indicates that the number of interactions between users by SMS are higher than phone call. It can be partly explained by the convenience of SMS and its relatively cheaper price. In Fig. 2(a) and (b), the distributions of DCN and SVDCN exhibit an obvious Kink at $w^N \approx 80$. It’s not clear why directed calling networks have kicks but mutual networks and SMS networks do not. We can use a bi-power-law distribution to fit the data: $p(w) \sim w^{-\alpha_1}, 1 < w < 80; p(w) \sim w^{-\alpha_2}, w > 80$

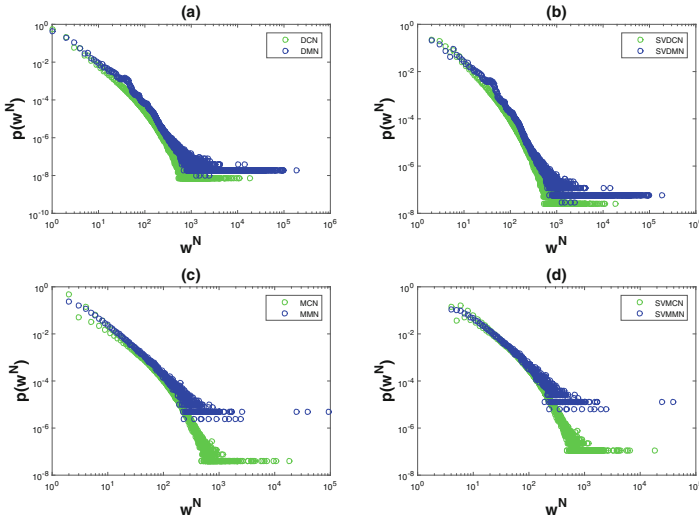


Fig. 2. Edge weight distribution. (a) Distribution of number-based edge weight w^N for DCN and DMN. (b) Distribution of number-based edge weight w^N for SVDCN and SVMN. (c) Distribution of number-based edge weight w^N for MCN and MMN. (d) Distribution of number-based edge weight w^N for SVMCN and SVMMN.

3.3 Node Strength Distribution

For each node in the networks, we define the node strength S_i^N based on the number of calls or messages, and $s_i^N = \sum_{j \in N_i} w_{ij}^N$ as the total number of calls

or messages user made where w_i^N is the calls or SMS between i and j . As the edges in directed networks have directions, we will further distinguish the in-degree and out-degree node strength in our experiments. We present the node strength distributions in Fig. 3, and the distributions for all networks in our experiments can be fitted by an exponentially truncated power-law function $p(s) \sim -\gamma_s \exp(-s/s_c)$. In panel (a) and (b) we find that there is no difference in in-degree and out-degree distributions among the four directed networks. But there is a remarkable difference between the distributions of the networks and the corresponding statistical validated networks. This seems to indicate that the statistical validated method is able to filter out the users who rarely contact with others.

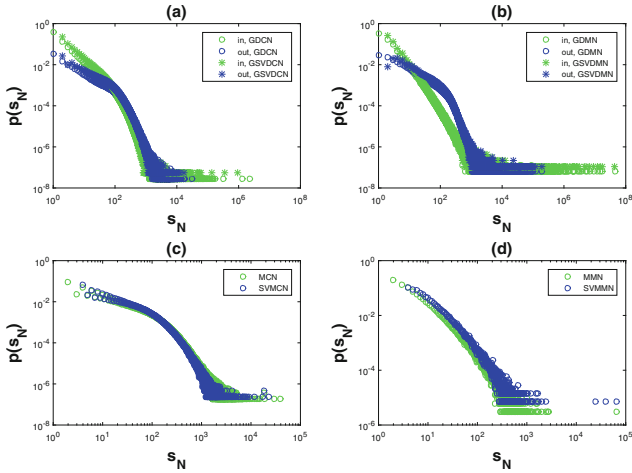


Fig. 3. Node strength distribution. (a) Distribution of number based node strength for GDCN and GSVDCN. (b) Distribution of number based node strength for GDMN and GSDVMN. (c) Distribution of number based node strength for MCN and SVMCN. (d) Distribution of number based node strength for MMN and SVMMN.

3.4 Clustering Coefficient and Topological Overlap

In order to quantify the local cohesiveness around the nodes, we import the concept of clustering coefficient in our study, and use it as an important method to metric the calling and SMS networks. The clustering coefficient of node i is defined as $C_i = 2t_i/[k_i(k_i - 1)]$, where t_i denotes the number of triangles around node i [8] and the edge direction is ignored in directed networks. Panels (a) and (b) in Fig. 4 show the dependence of $\langle C|k \rangle$ on k for the eight networks. We find that the curves in the panels are fitted by the function $\langle C|k \rangle \sim k^{-1}$ which is commonly found in many empirical networks [9]. Every user has a quite small clustering coefficient value in the networks and high-degree users have a relativity low clustering coefficient. This indicates that triangle relationships are

quite frequent in local structure of the networks and users with few contacts may only connect with people in the same social circle. And this can partly be explained by the fact that people have been accustomed to using mobile phone for communication in their real world social networks.

In panel (a) we find that even though the clustering coefficient variation tendency of calling networks and SMS networks are similar, there still exist some noteworthy discriminations between the four curves. The curves of DCN and DMN are above their corresponding statistical validated networks. This observation reflects the fact that statistical validated approach, while minimizing the presence of links not related to an underlying social relationship, may also remove some edges with real social relationships. The calling networks' clustering coefficient curves are also above the SMS networks' and this reflects the fact that user's phone call friend circle is bigger than SMS's. Phone call play a more important role than SMS in people's daily social contact.

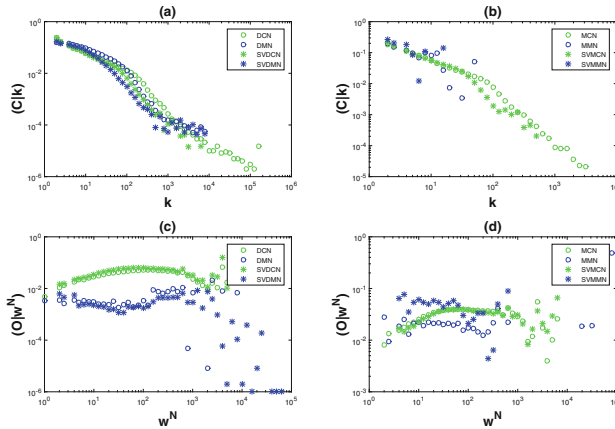


Fig. 4. Clustering coefficient and Topological overlap. (a) Average clustering coefficient $\langle C|k \rangle$ as a function of k for DCN, DMN and their SV networks. (b) Average clustering coefficient $\langle C|k \rangle$ as a function of k for MCN, MMN and their SV networks. (c) Average overlap $\langle O|w^N \rangle$ as a function of number-based edge weight w^N for DCN, DMN and their SV networks. (d) Average overlap $\langle O|w^N \rangle$ as a function of number-based edge weight w^N for MCN, MMN and their SV networks.

Finally we study the properties of links and their neighbor nodes. We quantify the topological overlap of the neighborhood of two connected nodes i and j by the relative overlap of their common neighbors. It's defined as

$$O_{ij} = \frac{n_{ij}}{(k_i - 1) + (k_j - 1) - n_{ij}} \quad (1)$$

where k_i and k_j are the degrees of the two nodes and n_{ij} is the number of the common neighbors of node i and j . When we calculate the overlap for the

directed networks the edge directions is ignored. We present the result in Fig. 4 (a) and (b) and find that the two curves for DMN and SVDCN are quite similar, while the curve for the DCN is lower than SVDCN. This contrasts with the phenomenon in Ref. [7]. This indicates a significant fraction of abnormal nodes have been removed by the statistical validated method and the nodes in the network contact more closely. In addition all curves of calling exhibit an increasing trend and seem to decrease after $w^N \approx 100$, but the curves of SMS fluctuate more furiously. This remarkable difference reflects the SMS communication network is quite different from the phone call network.

4 Discussion

Due to the dramatic development of science and technology we have the opportunity to study large-scale complex networks and allow us perceive our social from a totally new perspective. In this paper we constructed networks from mobile phone records of 36 million users and used the cumulative number of calls and SMS as a measure of the strength of the social tie. We construct the social communication network by taking users as nodes and the connections between users as edges. We can study human communication patterns and human dynamics at the societal level.

From the original networks, we construct directed networks, mutual networks and their statistical validated networks. Since statistical validation is used to remove the edges that have no social origin, we take these statistical validated networks as control to be compared with their original networks. In order to understand the networks, we investigate the networks from the distributions of the degree, the edge weight, the node strength and etc. By comparative analysis of these network attributes, we found that these networks share many common properties and also exhibit idiosyncratic characteristics in both qualitative and quantitative ways.

Our work mainly focus on the comparison between calling networks and SMS networks. As we know that traditional mobile communication means have been challenged by mobile Internet. We all experience the powerful influence but we can't quantity the influence brought by mobile Internet. So through comparison between calling networks and SMS networks, we can find the discrepancy between calling and SMS and quantify the influence of burgeoning mobile Internet.

Our work find that usage of SMS is performance poor than calling no matter from quality of user or number of user. Users prefer to use phone call rather than SMS for mutual social communications. The communities constructed by SMS are sparse and unstable. And these phenomenon might be used to formation explanation and deletion of social ties indicate the presence of different elementary mechanisms governing social dynamics under different culture and social norm.

The result of our work can be used as a basis for social communication pattern evolution models of our society. In particular the basic mechanisms of the remarkable discrepancy between SMS and calling need a detailed investigation.

The lessons learnt from our work are not limited to understanding human social networks, but may be used in other domains as well. Finally, we believe our work provides a meaningful method for the study on the evolution of social communication patterns.

Acknowledgement. This work was supported by the major research plan of the National Natural Science Foundation (91224009, 51438009), Technology Commission (13ZCZDZX01099), and the Ocean Public Welfare Scientific under Grant No. 201305033

References

1. Yan, X.Y., Han, X.P., Wang, B.H., et al.: Diversity of individual mobility patterns and emergence of aggregated scaling laws. *Sci. Rep.* **3**(9), 454–454 (2013)
2. Gonzalez, M.C., Hidalgo, C.A., Barabasi, A.L.: Understanding individual human mobility patterns. *Nature* **453**, 779–782 (2008)
3. Candia, J., Gonzalez, M.C., Wang, P., Schoenharl, T., et al.: Uncovering individual and collective human dynamics from mobile phone records. *J. Phys. A Math. Theor.* **41**(22), 1441–1446 (2007)
4. Scebran, M., Palladini, A., Maggio, S., et al.: Statistically validated networks in bipartite complex systems. *Plos One* **6**(3), e17994 (2011)
5. Hatzopoulos, V., Iori, G., Mantegna, R.N.: Quantifying preferential trading in the e-MID interbank market. *SSRN Electron. J.* 15(4), 693–710(18) (2013)
6. Tumminello, M., Lillo, F., Piilo, J., et al.: Identification of clusters of investors from their real trading activity in a financial market. *New J. Phys.* (2011)
7. Li, M.X., Palchykov, V., Jiang, Z.Q., et al.: Statistically validated mobile communication networks: evolution of motifs in European and Chinese data. *New J. Phys.* (2014)
8. Watts, D.J., Strogatz, S.H.: Collective dynamics of small-world networks. *Nature* **393**, 440–442 (1998)
9. Szab, G., Alava, M., Kertsz, J.: Clustering in complex networks. In: Ben-Naim, E., Frauenfelder, H., Toroczkai, Z. (eds.) *Complex Networks*. LNP, vol. 650, pp. 139–162. Springer, Heidelberg (2004)
10. Clauset, A., Newman, M.E.J.: Power-law distributions in empirical data. *SIAM Rev.* **51**(4), 661–703 (2007)
11. Bril', A.I., Kabashnikov, V.P., Popov, V.M.: Dynamical and correlation properties of the Internet. *Phys. Rev. Lett.* **87**(25), 527–537 (2001)
12. Newman, M.E.J.: Assortative mixing in networks. *Phys. Rev. Lett.* **89**, 208701 (2002)