

Heterogeneous patterns in human communication

Zhechao Zhou

2008 NSF/REU Program

Physics Department, University of Notre Dame

Advisor: Professor Zoltán Toroczkai

Abstract

My work contributes to an ongoing social network study based on a statistical analysis of communication patterns among millions of individuals as inferred from cell-phone records. I have studied the corresponding communication network on individual level (network nodes), the level of dyads (edges in the network) and on the triad level (triangles in the network). On the dyad level I explored various communication tie properties such as type distribution, reciprocity, asymmetry and tie strength, while at the triad level I investigated the distribution of the 16 triad types following J. Moody's [1] classification. Since the size of the cell-phone network data is very large, the number of the triangles is on the order of $\mathcal{O}(10^{19})$. The matrix computation techniques for the triad census introduced by Moody cannot be implemented due to limitations in computational memory. I have introduced an iterative procedure, however, which avoids this problem and performs an exhaustive analysis of the triad types. My work indicates some strong patterns in large-scale human communications, with very good statistical properties, which might lead to a systematic understanding of large-scale human communication patterns and their laws. For example, the distributions of the number of people a person communicates to and of the number of calls made by a person follow clean Pareto distributions to a cut-off degree. I give a possible explanation based on multiplicative stochastic processes. My study on the dyad and triad levels also indicates that the network is loosely connected globally, while within local communities communication ties are fairly strong, information spreading well among local groups.

1 Introduction

In the past, social scientific research on human behavior and social networks has been rather limited due to lack of high-quality large-scale data and quantitative methods to analyze such data. Although individual human behavior is rather complex, one expects that there might be quantitative measures characteristic to large *collectives* of humans that are reproducible and thus subject to measurement, bringing the possibility of applying the classical scientific method (very successful in physics, chemistry, biology, etc.) more

extensively into social sciences. We hope to be able to develop an understanding of large-scale social systems and to find general laws that govern aggregate properties of the more complex human behavior in a similar way that statistical physics is able to describe the collective properties of many-component complex systems such as gases, solids, fluids, etc.

In general social system data obtained through surveys is usually small, non-longitudinal (independent of time) and biased by the existence of the observer and the methods in the experiment. In order to be able to successfully apply statistical physics methods to analyze human behavior, one needs large amounts of pure data without observational intervention. The social network in this study is generated from real-world cell-phone communication logs from among more than 6.7 million people. This data has four key traits that will transform the way researchers analyze and understand social networks and human behavior : 1) quality of statistics (the data comes from millions of users), 2) purely observational (void of any bias induced by obtrusive measurement), 3) complete network data (not just information on the ego-networks of a sample of people) 4) longitudinal (spanning several years).

My work contributes to an ongoing social network study at Center of Complex Networks, University of Notre Dame, based on a statistical analysis of communication patterns among millions of individuals as inferred from cell-phone usage pattern logs. I have studied the corresponding communication network on individual level (network nodes), the level of dyads (edges in the network) and on the triad level (triangles in the network). The first part of this paper is devoted to the structure of the original data and node census, the second and third to weighted dyadic and triadic distributions of the network.

2 Data processing

The data is clustered from a cell phone network of about 6.7 million users during 4/16-4/30/2006, a 16 days period. The data comes in the format shown in Table 1.

Table 1: This is from the original data file

call_from	call_type	call_to	des_code	num_calls	tot_duration (s)
Daus+8...	5	qDGKtI...	8	12	0
awe7Hg...	2	gdRY4M...	470	4	963
rhrDhX...	2	U1KZYf...	8	9	1433
dseJxO...	5	6uMo12...	8	1	0

The communication record between two users is clustered separately into voice calls and text messages. Call type “2” indicates voice calls. Call type “5” indicates text messages with a zero total duration. There are also a few entries that have call_type other than “2” and “5”, indicating other types of communication such as fax transmission, emergency calls, etc. These counts are relatively small compared to “2” and “5” types, and their affect on the census is negligible. Only 24,441,476 entries, out of 99,421,911 from the original data are valid to the study after the original data is filtered by the destination code which indicates the caller and the callee of that entry are both in the cell phone network and are subject to our study. The filtered hashed data is further mapped into integer indices starting from 0, generating a mapping log `mapping.csv` and the data file `data.txt` as shown in Table 2. I use Unix command `wc -l -w data.txt` to check the mapped file. It has 24,441,476 lines and 122,082,380 words. The file has 5 words in each line, $5 \times 24,441,476 = 122,082,380$. The mapped data is not sorted. I used Unix’s `sort` command to sort by the first column, caller index, then by the second, callee index, then by the third, call type and the fourth, number of calls.

Table 2: From the file `data.txt`:

caller	callee	call type	nr. of calls	tot durations (s)
26016	26015	2	1	908
26028	1325723	5	4	0

```
sort -n -k 1,1 -k 2,2 -k 3,3 -k 4,4 -o data_s.txt data.txt
```

The output file is `data_s.txt`. By observation the largest user index in `data_s.txt` is 6743619. There are self-called entries that the caller and the callee's index are identical. The self-calling entries should not be counted in the later dyadic and triadic census. The actual users under the study is 6,733,758 as counted after all the self-calling entries and unwanted call-type entries have been filtered out.

The *degree* of a user (node) is the number of lines that are incident with it, or the number of nodes adjacent to it. In the digraph of cell phone network, a node can be either *adjacent to*, or *adjacent from* another node, depending on the direction of the arc. The code `degree_census.cpp` [A.0.1] is to derive the distribution of out-degrees, how many users each users have initiated communication to in the network. Running this code:

```
$ ulimit -s 65530
$ ./out_degree data_s.txt out_degree.txt 6743619 50000
50000 is the estimated largest out-degree.
```

output:

```
out_degree.txt
number of active callers = 4836455
```

[A.0.1] can be used to generate the distribution of in-degree if `c1` and `c2` are switched in `if (c1!=c2) insert(&head[c1],c2)`. Running the new code:

```
$ ulimit -s 65530
$ ./in_degree data_s.txt in_degree.txt 6743619 50000
```

output:

```
in_degree.txt
number of active callees = 6200129
```

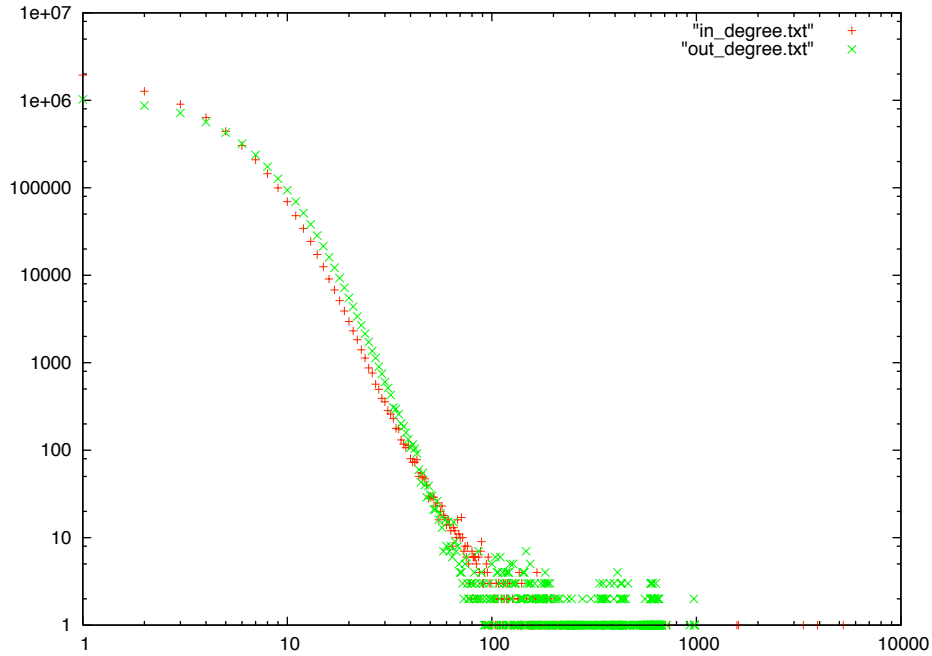


Figure 1: In and out degree of the users

Plotting `in_degree.txt` and `out_degree.txt` in log scale gets the histogram [1]:

```
$ gnuplot
set logsc
plot "in_degree.txt", "out_degree.txt"
```

The distribution for the degree has a fat-tail slowly decreasing to zero with some out-degree noise near the end of the tail. So the network is a scale-free network [3] meaning that the distribution of the degree follows the power-law, $P(d) \propto d^{-\gamma}$ up to a cut-off degree. In our case, for the degree distribution exponents are $\gamma = \gamma_{in} = \gamma_{out} = \dots$. Such distributions are also called in statistics Pareto distributions.

The code `tot_out.cpp` [A.0.2] is to derive the distribution of the *total number of outgoing calls* for a user. Running this code:

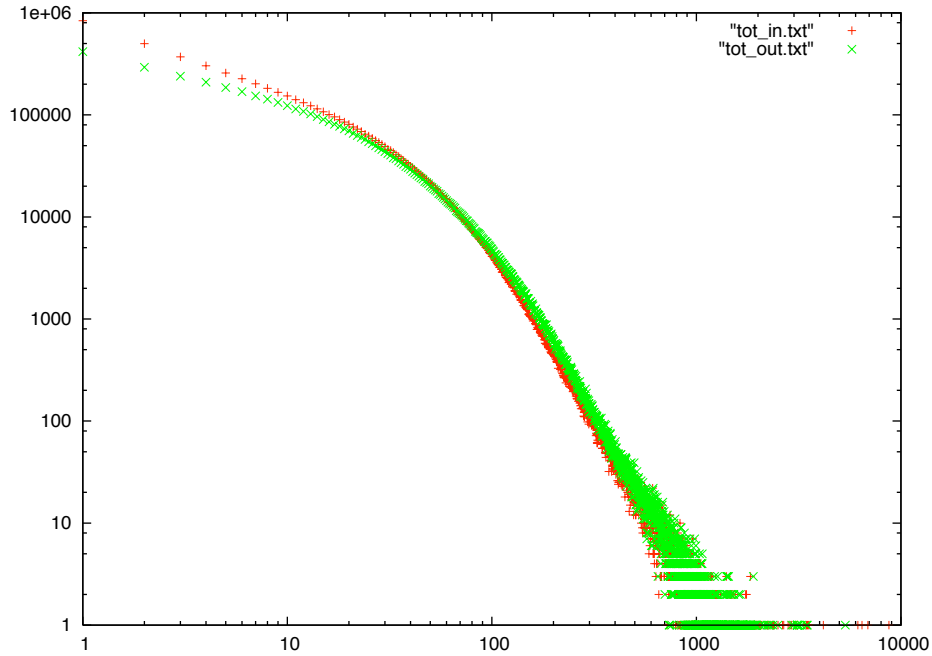


Figure 2: total number of incoming and outgoing communication of the users

```
$ ulimit -s 65530
$ ./tot_out data_s.txt tot_out.txt 6743619 500000
500000 is the estimated largest out calls.
```

output:

```
tot_out.txt
```

Similarly, the census on the distribution of total incoming calls of each users `tot_in.txt` can be done with a little revision of [A.0.2]. Both outcomes are plotted in Figure 2.

3 Multiplicative process as a possible explanation

The deviation of the distributions above (in and out degrees and call numbers) from the pure power-law (Pareto) form for smaller values of the x -axis

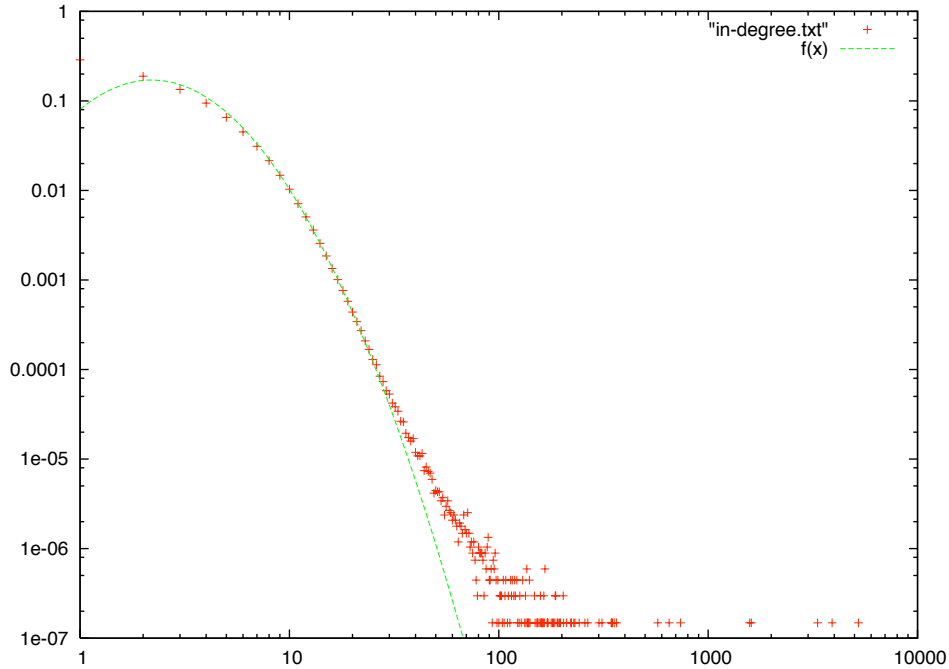


Figure 3: in-degree distribution fit by $f(x) = e^{-2.52073+1.92244x-1.22243x^2}$

variable suggests that instead of a linear fit $Y = A - BX$ (in log-log scale) it would be better to use a quadratic simple polynomial $Y = A - BX - CX^2$ (which is just the next level approximation to the curve, see Figure [3] [4]). However, this means in normal scale x and y ($X = \ln x$ and $Y = \ln y$) that the distribution is lognormal.

Assume that you have a quantity X which evolves in successive steps i and the corresponding values are X_i . You start with X_0 . The multiplicative process is given by:

$$X_i = (1 + \delta_i)X_{i-1} \quad (1)$$

where δ_i are iid (independent, identically distributed) random variables. For example, X_i could be the amount of cracks in the material at time step i , or the size of a snow ball rolling down the hill. Obviously, the amount of new cracks developing in the next step, or the amount of new snow picked up by

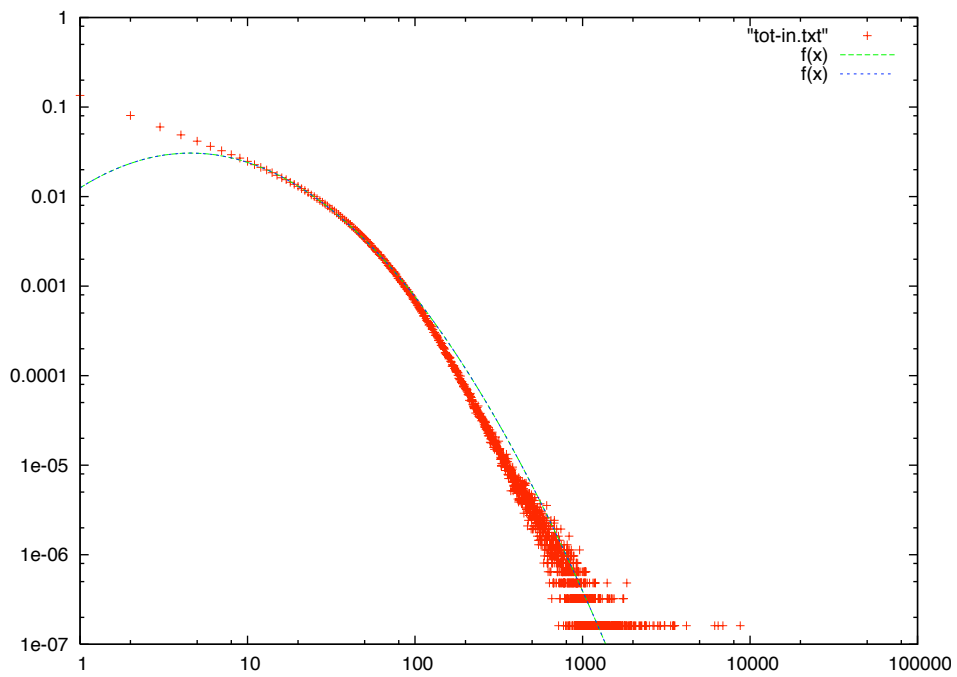


Figure 4: total in distribution fit by $f(x) = e^{-4.37921+1.17609x-0.387552x^2}$

the ball depends on the existing number of cracks or the existing size of the snow-ball (hence proportionality to X_{i-1}), modified by a random amount (δ_i) which depends on the medium (where the cracks are actually propagating, or what kind of snow is there, how sticky it is, where the ball is rolling at that moment). Thus, from the above equation:

$$X_n = \left[\prod_{i=1}^n (1 + \delta_i) \right] X_0 \quad (2)$$

Now if you take the log of the above:

$$\ln X_n = \sum_{i=1}^n \ln(1 + \delta_i) + \ln X_0 \quad (3)$$

Since the δ_i are iid, then also the $\ln(1 + \delta_i)$ will be iid. So, as long as the distribution of the $\ln(1 + \delta_i)$ random variable has a finite second moment, by the central limit theorem, the distribution of (3) will be described by a Gaussian, and thus the distribution of X_n by a lognormal. It does not matter what is the distribution of the δ_i as long as n is large enough. If $\ln(1 + \delta_i)$ does not have a finite second order moment, then a central limit theorem still applies, however, that's when you get Lévy distributions (Gnedenko and Kolmogorov).

Equation (1) describes avalanche-like processes, and for this reason lognormal distributions are considered as a signature of self-organized criticality (SOC). It was used successfully to model failure propagations, crack propagations, chemical reactions, ion migration, etc.

Next we describe how this might apply to human communications. Assume that an acquaintance j of a given person i introduces δ_{ij} new acquaintances in the next time interval (a day, for example). When viewed over a large population, these δ_{ij} numbers can be thought of as random variables. If

$N_i(t)$ is the number of acquaintances of i at step t then after the next step:

$$N_i(t+1) = N_i(t) + \delta_{ij}N_i(t) = [1 + \delta_{ij}] N_i(t) \quad (4)$$

(we assume that N_i is the number of acquaintances of i who gave her their phone number and thus she can call them at any time). This evolution equation is identical to (1), and thus it captures the log-normal behavior as measured. Certainly, (4) is expected to fail, since after a while a saturation effect has to step in, and people may lose connection to old acquaintances as time goes, however, the scaling (Pareto) behavior seems to be reproduced by it. Further longitudinal analysis is needed to connect the properties of δ_{ij} with that of the dyads from the data and seeing if indeed these properties are consistent with the exponent measured and thus if this simple theory presented here is supported by it.

4 Dyadic census

A dyad is a subgraph consisting of an unordered pair of nodes and the possible ties between the nodes. In our case, any two users in the network can form only three dichotomous dyads, mutual where the two users have mutual communication, asymmetric where the two users have single directional communication, and null where they are not related at all. In my study, communication dyads between two users contain information on how frequently the pair communicates (number of calls), and how durable the edge is (total duration). Code `dyad.cpp` [A.0.3] counts the number of mutual dyads and asymmetric dyads in the network.

```
input:    $ ulimit -s 65530
          ./dyad data_s.txt 6743619
```

```
output:  the number of mutual dyads is 5446810
          the number of asymmetric dyads is 9645463
```

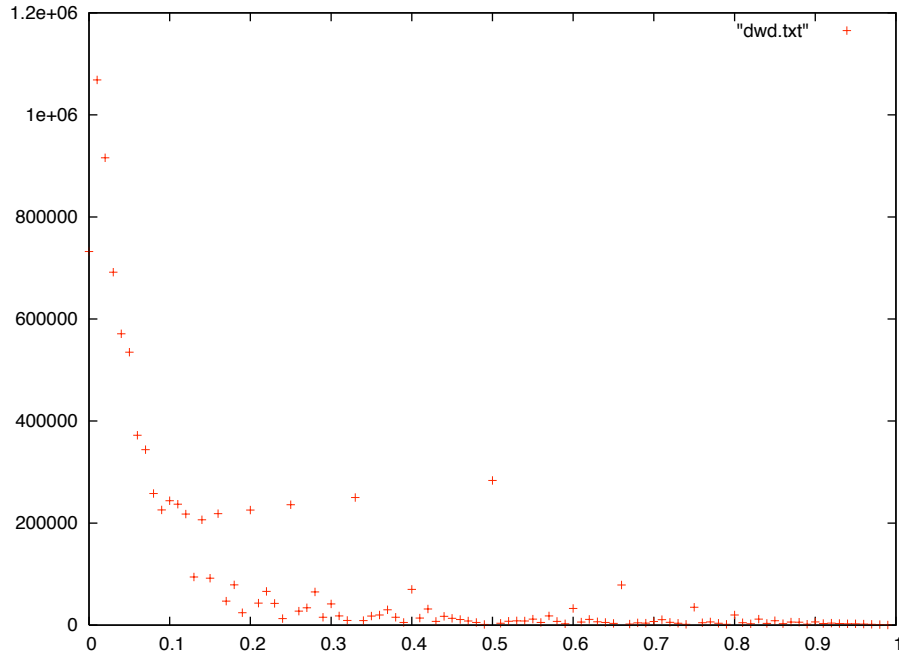


Figure 5: u_{AB} distribution

By simple calculation, the number of null dyads in the network is about 2.267×10^{13} . Code `dwd.cpp` [A.0.4] is an example on weighed dyad census. We define

$$w_{AB} = \frac{n_{AB}}{n_A} \quad (5)$$

where n_{AB} is the total number of times that A calls B. n_A is the total number of calls that A makes to all the callees including B. Code `dwd.cpp` [A.0.4] measures $u_{AB} = |w_{AB} - w_{BA}|$, the relative skewness between A and B for all the mutual and asymmetric dyads in the network.

```
input:          $ ulimit -s 65530
                ./dwd data.s.txt 6743619
```

```
output: dwd.txt
```

From Figure 5, u_{AB} of most dyads are distributed near 0, which means that

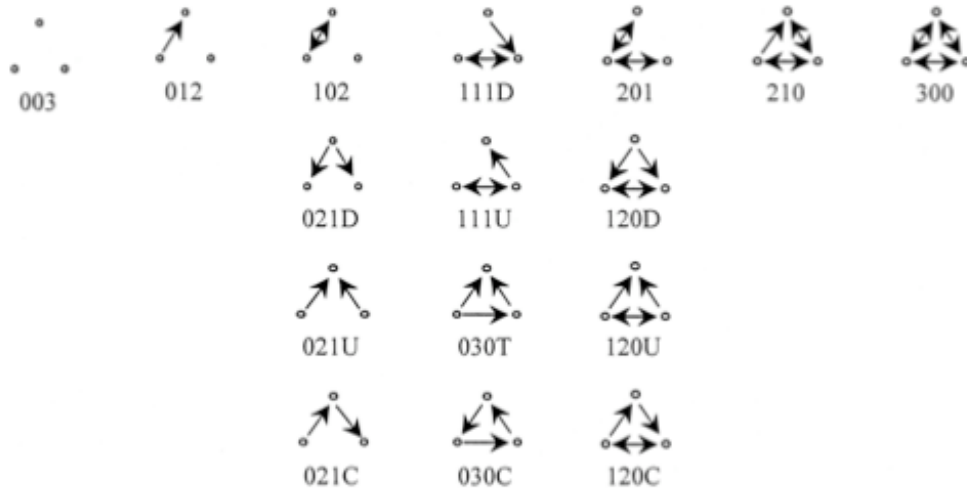


Figure 6: 16 triads

the two users are almost the same devoted in the communication.

5 Triadic census

A *triad* is a subgraph consisting of three nodes and the possible lines among them. Social scientists classify triads into 16 different types regardless of node combinations [6]. J. Moody provides a triad census method by matrix operation. However, matrix method doesn't work well on large scale network due to the limit of computer memory space and time. The codes `triad_1.cpp` and `triad_2.cpp` yields the census of 15 triad types in the network. [A.0.5] [A.0.6]

```

input:  ulimit -s 65530
        ./triad_1 data_s.txt 6743619

output: T201: 13586387  $\approx 13.6M$ 
        T210: 1542087  $\approx 1.5M$ 
        T300: 1014963  $\approx 1.0M$ 
        T021D: 47874450  $\approx 47.9M$ 
        T111U: 25874930  $\approx 25.9M$ 
        T120D: 340590  $\approx 0.34M$ 
        T021C: 12862894  $\approx 12.9M$ 
        T030T: 607288  $\approx 0.61M$ 
        T030C: 23400  $\approx 0.023M$ 
        T120C: 350722  $\approx 0.35M$ 
        T111D: 14456103  $\approx 14.5M$ 
        T012: 64950052609794  $\approx 65T$ 
        T102: 36677428963655  $\approx 36.7T$ 

```

and

```

input:  ./triad_2 data_s.txt 6743619

output: T021U: 45747952  $\approx 45.7M$ 
        T120U: 1027216  $\approx 1M$ 

```

T003 is the most common triad in the network for about 5.09×10^{19} .

By observation, the most common triads besides T003, T012, T102 types are T111D, T201, T021D, T111U, T021U and T021C triads, the triads that contain one null dyad, i.e. two nodes in the triad has no communication with each other. It is true that cell phone users don't actually know most of the other users in the same network.

A *transitive configuration* is an important structure existing in a triad that $i \rightarrow j, j \rightarrow k$, and $i \rightarrow k$. For example, T300 triad has 6 transitive triples, T210 have three, T120U have two and T030C triad has no transitive triples. To our surprise, among the rest 7 types of triads where any two of the three nodes have at least single-directional communication, T300, T210, T120 triads of more than 2 transitive triples are more common than those less transi-

tive triads (T030T, T030C, T120C), and T030C of no transitive triples is the most rare triad in the network for only 23,400. This suggests that the global network is loosely connected while within local communities communication ties are fairly strong, information spread well among those users in the same community.

6 Conclusion

I have studied a social network of 6.7 million people on individual level (network nodes), the level of dyads (edges in the network) and on the triad level (triangles in the network). On the dyad level I explored various communication tie properties such as type distribution, reciprocity, asymmetry and tie strength, while at the triad level I investigated the distribution of the 16 triad types following the Moody classification. I have introduced a novel iterative procedure, which allows for the study of the statistics of these triangles. and performed an exhaustive analysis of the triad types. My work indicates that distributions of the number of people a person communicates to and of the number of calls made by a person follow clean Pareto distributions to a cut-off degree. I gave a possible explanation based on multiplicative stochastic processes. My dyad and triad census shows weak global network property but strong local ties in large-scale human behavior. These results might pave the way to a systematic understanding of large-scale human communication patterns and their laws.

7 Acknowledgements

I would like to thank Prof. Toroczkai for his instruction, generosity and free coffee, Prof. Hachen for providing the data and computing facility, Prof. Garg for organizing this REU program that I have a wonderful experience with, Science, Mathematics & Computing Division of Bard College for partial fi-

nancial support.

References

- [1] J. Moody, *Matrix methods for calculating the triad census*, Social Networks **20** (1998), 291-299
- [2] S. Wasserman, and K. Faust, *Social Network Analysis: Methods and Applications*, Cambridge Univ. Press, Cambridge, 1994.
- [3] G. Caldarelli, *Scale-Free Networks*, Oxford Univ. Press, Oxford, 2007.

A Codes

A.0.1 `out_degree.cpp`

A.0.2 `tot_out.cpp`

A.0.3 `dyad.cpp`

A.0.4 `dwd.cpp`

A.0.5 `triad_1.cpp`

A.0.6 `triad_2.cpp`

Due to the limit of pages, please contact Zhechao Zhou at zzhou4@nd.edu or Prof. Zoltan Torozckai at toro@nd.edu for specific codes.