IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 2 November 2015; revised 11 September 2016; accepted 5 November 2016.
Date of publication 9 November 2016; date of current version 5 June 2019.

Digital Object Identifier 10.1109/TETC.2016.2627034

An Analysis of Social Networks Based on Tera-Scale Telecommunication Datasets

HIDAYET AKSU, (Member, IEEE), IBRAHIM KORPEOGLU, (Senior Member, IEEE), AND ÖZGÜR ULUSOY, (Member, IEEE)

The authors are with the Department of Computer Engineering, Bilkent University, Ankara 06800, Turkey
CORRESPONDING AUTHOR: H. AKSU (hidayet.aksu@gmail.com)

ABSTRACT With the popularization of mobile phone usage, telecommunication networks have turned into a socially binding medium. Considering the traces of human communication held inside these networks, telecommunication networks are now able to provide a proxy for human social networks. To study degree characteristics and structural properties in large-scale social networks, we gathered a tera-scale dataset of call detail records that contains $\approx 5 \times 10^7$ nodes and $\approx 3.6 \times 10^{10}$ links for three GSM (mobile) networks, as well as $\approx 1.4 \times 10^7$ nodes and $\approx 1.9 \times 10^9$ links for one PSTN (fixed-line) network. In this paper, we first empirically evaluate some statistical models against the degree distribution of the country's call graph and determine that a Pareto log-normal distribution provides the best fit, despite claims in the literature that power-law distribution is the best model. We then question how network operator, size, density, and location affect degree distribution to understand the parameters governing it in social networks. Our empirical analysis indicates that changes in density, operator and location do not show a particular correlation with degree distribution; however, the average degree of social networks is proportional to the logarithm of network size. We also report on the structural properties of the communication network. These novel results are useful for managing and planning communication networks.

INDEX TERMS Social networks, degree analysis, call graph, empirical analysis, tera-scale dataset

I. INTRODUCTION

Human communication behavior is the root of the usage pattern in physical and virtual communication networks, including telecommunication (telco) networks and online social networks. While fixed-line phones and shared computers in homes and offices reflect family or colleague behavior; mobile phones and portable computers better reflect individual usage behavior. Technological developments in the last two decades have resulted in two significant trends in human behavior: 1) going frequently online and 2) owning personal mobile computing and communication devices. Thus, the end-user behavior of communication networks has changed from group behavior to individual behavior.

Human communication behavior is highly related to underlying social network relationships. Mobile phone communication patterns provide strong insights into human social relationships [28]. For instance, person A calls person B usually because of a social relationship, e.g., B is a friend of A or B does business with A. The more social interactions dominate communication networks and online media, the more user behavior on those networks is dominated by human social relationships and networks. Hence, managing and planning today's communication networks require a deep understanding of user behavior on those networks and their social structures.

Early studies on social networks were limited by manual data collection and considered at most hundreds of individuals [39]. Later, social network analysis (SNA) became an interesting topic for many other sectors and research fields, including recommender systems [24], [31]; marketing [7]; intelligence analysis [35]; network structure [16]; modeling epidemics spreading [44]; clustering and community detection [6], [9], [15], [17], [18], [23] and complex systems [19]. Massive use of electronic devices and online communication leaves traces of human interaction and relationships, such as phone call records, e-mail records, etc. Using these traces, collective human behavior and social interactions can be understood on a large scale, which was previously impossible [40]. Recently telecommunication datasets with location

information have been used to conduct research on human behavioral patterns [8], [13], [21], [22], [25], [41], [42], mobile network behavior [43] and inferring hierarchies [38].

Social network analysis tries to understand the characteristics a social network exhibits. The first and most-cited characteristic among others is *degree distribution* of nodes constituting a social network. A bulk of studies in the literature on this topic reports that power-law best fits with certain parameters [1], [10], [30]. Other studies, however, propose different statistical fit models [4], [34], [36].

Since current studies are limited by the used datasets from which their proposals are derived/obtained, it is necessary to explore the influence of dataset specific parameters on discovered social network characteristics. This observation motivates us to conduct research on degree distribution on larger scales to discover the parameters governing degree distribution in social networks. Among many current research issues to be investigated, we prefer this less studied problem which requires a complete dataset.

Therefore, we explore how

- network operator,
- network size,
- population density and
- geographic location

affect degree distribution in social networks.

To investigate these issues, we perform degree analysis on different social networks derived from the telecommunication network call data of a country's[1] different mobile (GSM[2]) and fixed-line (PSTN[3]) telco operators. We obtain degree distribution results for these networks to understand how well existing distribution models fit reality.

In this study, our scope is limited to empirically revealing the parameters that govern degree distribution, and comparing a limited number of structural properties with other studies.

Our paper contributes to the field in the following ways:

- We first construct a countrywide call graph utilizing a full call detail record (CDR) set of all mobile and fixed-line telco network operators. This comprehensive dataset allows us to analyze a social network without wondering about possible bias from single-operator, size, location or density-limited datasets.
- We question the root cause of different conclusions in the literature about degree distribution in social networks, suggesting that they might be related to utilized datasets' density, location, size and source operator.
- We perform controlled empirical analyses for various densities, sizes, locations, and operators, and form conclusions on density-degree, location-degree, size-degree and operator-degree distribution relations.
- We analyze call graph for structural properties and compare it with other social graphs.

The paper proceeds as follows: In Section III, we describe the dataset used in this study and highlight its unique features. In Section IV, we discuss the statistical modeling of degree distribution in social networks and report the results of our empirical analysis. We also provide an analysis and interpretation for each of the following factors, any or all of which may affect social network characteristics: network operator, network size, network density and network location. Then we provide structural properties of the communication network in Section V. Finally, in Section VI, we present our conclusions.

II. RELATED WORK

Aiello *et al.* [1] study the statistics of phone call graphs for long-distance fixed-lines and report that in-degree distribution is fitted by power-law distribution with exponent $\gamma = 2.1$. In [30], Onnela *et al.* work on mobile phone data containing $N = 4.6 \times 10^6$ nodes and $L = 7.0 \times 10^6$ links and report a power-law distribution fit with exponent $\gamma = 8.4$. They describe the dataset as "all mobile phone call records of calls among ≈ 20 percent of the entire population of the country", which implies that they used a sub-network of a country's operator network. Dasgupta *et al.* [10] present another study on mobile phone data, with a reciprocal call graph containing $N = 2.1 \times 10^6$ nodes and $L = 9.3 \times 10^6$ directed edges. That dataset belongs to one of the world's largest mobile operators. The authors report that degree distribution is fitted well by power-law distribution with exponent $\gamma = 2.91$. Another study by Nanavati *et al.* [29] reports similar results. On the other hand, Bi *et al.* [4] propose the discrete Gaussian exponential (DGX) and report that it provides a very good fit with many datasets, including telco data. Moreover, Seshadri *et al.* [36], using mobile phone data from an anonymous operator in the US, study modeling degree characteristics and report that degree distribution significantly deviates from what would be expected by power-law and log-normal distributions. Their findings suggest that double Pareto log-normal distribution (DPLN) provides better fits for degree distribution. In [34], Sala *et al.* analyze Facebook's social network data and report that Pareto log-normal (PLN) distributions are much better predictors of degree distributions in real graphs than power-law distributions are.

III. DATASET

Obtaining necessary and sufficient data is one of the most difficult steps in social network analysis. Until the current pervasive use of mobile phones, the lack of large-scale data has limited our knowledge regarding human relationships and social networks. Now, however, the situation has changed. Call detail records are records of communication traces stored by operators primarily for billing purposes. Mobile phone companies can collect CDRs for all subscriber calls going through their networks, and this CDR database is

[1]Data was provided on the condition of anonymization, including country anonymity.
[2]Global System for Mobile Communications (GSM) is a digital cellular network standard used by mobile phones.
[3]Public switched telephone network (PSTN) stands for the circuit switched telephone network and in this paper all PSTN data is originated from fixed-line telephone networks.

the most exhaustive dataset to date on human mobility and social interactions. For billing purposes, GSM networks record the base station each mobile phone call is made from, and this data thus holds the details of individual user movements. Having almost 100 percent penetration of mobile phones, the GSM network can now function as the most comprehensive proxy of a large-scale social network available today [37].

Lack of large and comprehensive data was one of the main reasons for doubts behind social network claims like Milgram's six degrees of separation (his small-world experiment) [27]. Now, however, one can (with permission) access anonymized CDRs from all network carriers providing service in a country. Particularly, European Union Data Retention Directive 2006/24/EC requires "the retention of data generated or processed in connection with the provision of publicly available electronic communications services or of public communications networks" [14] and each country has its own specific application of this requirement. In this study case, application of this direction is managed by a government agency which stores and processes the data of all network operators in its data-center. Upon our request to access the data for academic research purpose, we are granted access to anonymized data with a non-disclosure agreement and a data access agreement which limits study to be done on their own premises, i.e., no data movement, and limits access time to a specific duration. Thus, we can extract information about social interactions and construct a social network of the whole country from data provided by all mobile and fixed-line operators. This situation has the following advantages over previous studies:

- To the best of our knowledge, the dataset we use is much larger than the largest dataset containing trajectories and social interactions analyzed to date [37].
- Our data represents all country communication interaction, which is free from bias for a particular operator, size, location or density.
- The data contains spatial positions so we can also analyze the effect of location on social networks.

We are aware of the following limitations of our dataset:

- It covers calls of a one-month period and therefore some infrequent links might be missing.
- It comprises data from only voice and SMS communications. People might be using many other communication channels including e-mails, instant messaging tools, smartphone apps, etc.

Consequently, our dataset does not contain whole social network but a projection of it. It also contains many non-social entities.

The dataset used in this study covers all GSM (three networks) and PSTN (one network) CDRs for a whole country between 1 January 2010 and 31 January 2010.[4] Data is anonymized and used solely for this research. The structure of the data is presented in Table 1. The dataset contains $N \approx 5 \times 10^7$ nodes and $L \approx 3.6 \times 10^{10}$ links for the GSM networks, and $N \approx 1.4 \times 10^7$ nodes and $L \approx 1.9 \times 10^9$ links for the PSTN network. In this dataset, GSM penetration was approximately 82 percent while PSTN penetration was 23 percent in 2010. We compute penetration as the ratio of phone users to the total population of 10+ years olds.[5] Assuming single subscription per user, 82 percent mobile penetration covers 70 percent of the total population. In this study, we also refer to this dataset as the social network analysis database.

TABLE 1. Structure of data used in this work.

Field name	Value description
source	source party of communication: calling party
destination	destination party of communication: called party
operator	network operator ID
communication type	voice, SMS services, etc.
date time	time of communication in seconds resolution
duration	duration of communication in seconds resolution
cell ID	location of communication in connected base-station
	location resolution

IV. ANALYSIS

For a sound and complete understanding of degree distribution in a large-scale social network, we investigate the effects of the following factors: 1) network operator to which the dataset belongs; 2) size of the community network; 3) population density; 4) location of the community live. For each factor, we perform an analysis to determine how it affects degree distribution.

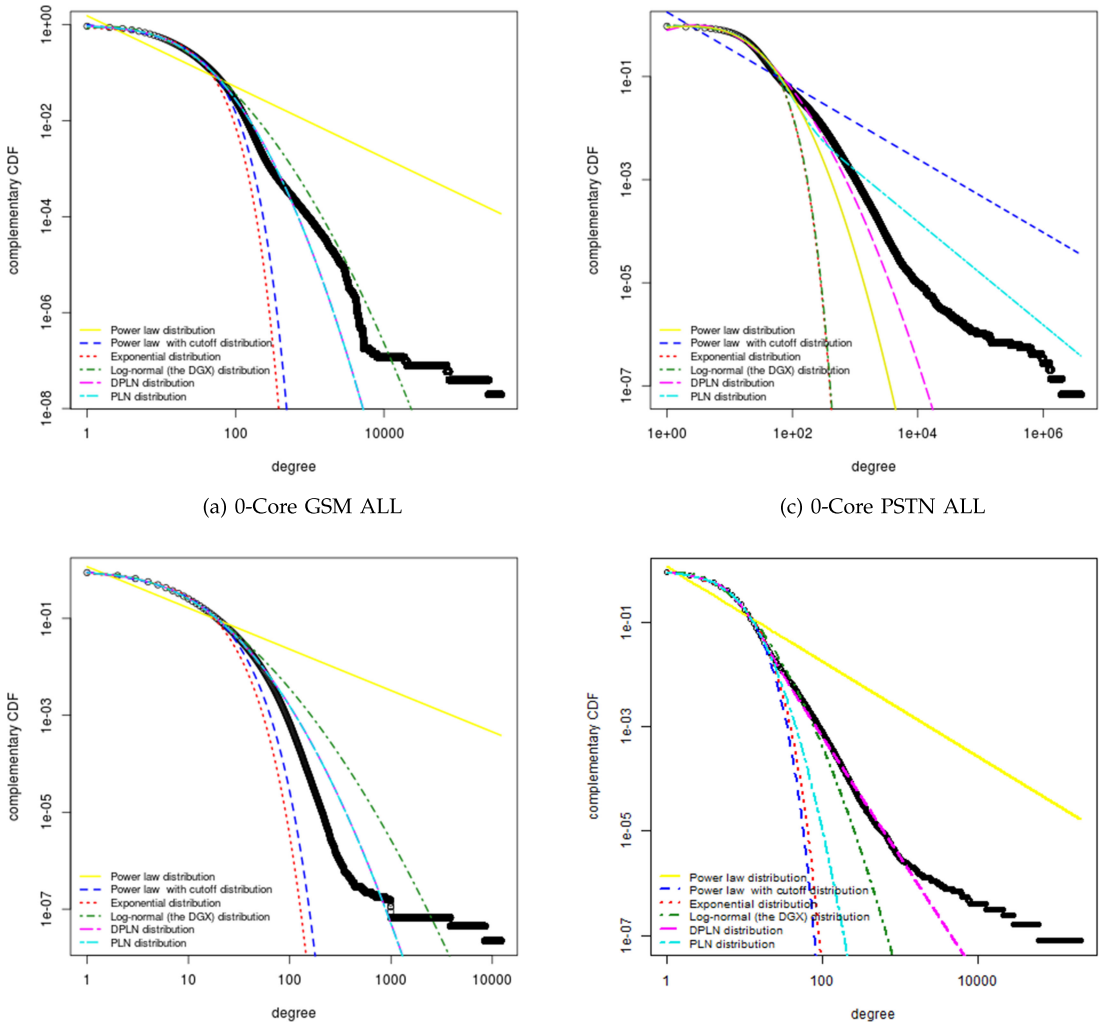
A. SOCIAL NETWORK MODELING

A call graph is a projection of a social graph and reflects some properties of it (i.e., a call graph is considered to reveal citizens' social interactions). Our dataset consists of call traces from the one PSTN and the three GSM operators in the country. Hence, we separately construct call graphs of the whole country for the three GSM operators and one PSTN operator. We also construct a call graph of the whole country for all GSM networks. Then we try to analyze degree distribution characteristics.

We first compute the degree distribution of the call graph with no filtering. We call such a network *0-Core network*. Then we filter out automated one-way calls which may not imply a work-, family-, leisure- or service-based relationship [30]. To eliminate the automated calls, we use our so-called *1-Core network* (reciprocal network) to also characterize degree distribution. if A has called B then 0-Core network has an edge. However, each pair of nodes (A, B) in the 1-Core network has an edge if and only if A has called B and B has called A at least once in the observation duration. Please note that this filtering eliminates only non-social entities which make one-way calls. Still, there may be many

[4]Unfortunately, we cannot make this dataset available due to a non-disclosure agreement signed.

[5]The country population of 10+ years olds was 61 M in 2010.



(a) 0-Core GSM ALL

(c) 0-Core PSTN ALL

(b) 1-Core GSM ALL

(d) 1-Core PSTN ALL

FIGURE 1. Network degree distributions and model fits for (a) 0-Core GSM ALL network (b) 1-Core GSM All network (c) 0-Core PSTN ALL network (d) 1-Core PSTN All network. Qualitative visual analysis suggest that PNL and DPLN distributions provides tightest fit while power-law distribution deviates most. See Table 2 for p-value based quantitative results.

non-social entities in the dataset like customer support lines and business lines.

When we plot the degree distributions (i.e., degree versus frequency of appearance of that degree in the call graph) on linear x-y scales, all distributions resemble an L shape (the curve quickly declines and most of the x-axis is close-to-zero valued). Visually, it is hard to interpret behavior from these plots. If we plot the degree distributions in log-log scales, however, the plots are easier to follow. Thus, we use log-log plots in this study. Degree distributions in Figure 1 are heavy tailed until a certain degree; then it takes an out-of-pattern fat-tail like shape. This means that the probability of having very high degree nodes is higher than what you would expect under a model fitting low-degree nodes. In Figure 1(a) we see a slope change around degree 5,000 where $1/10^6$ of the nodes are covered. We can see a similar situation in parts 1 (b), 1(c), and 1(d). Nodes with large degrees present a particular behavior, which we think is caused by non-social entities (e.g., business-related phone numbers, customer support lines, etc.). Comparing 0-C GSM, 1-C GSM, 0-C PSTN and 1-C PSTN graphs, we see that out-of-pattern vertex ratio is higher in the PSTN network than the GSM network. Also in both PSTN and GSM networks, 1-C networks show lower out-of-pattern vertex ratio compared to 0-C networks. This observation supports that out-of-pattern vertices are business phones or automated agents since 1-C networks cover less number of such non-social entities. Moreover, the horizontal nature of the tails on 0-C networks can be explained by the fact that automated agents may call fixed numbers of people in a 30 day period.

The literature related to degree distribution in call graphs and social networks includes various works on power-law distributions, power-law with cutoff distributions, log-normal distributions, exponential distributions, DPLN distributions and PLN distributions. All these distributions are possible candidates to statistically model degree distribution

in a complex network with an L-shape-like degree-frequency distribution.

For each constructed social network (call graph) in our dataset, we try to fit all candidate distributions and compute their goodness of fit. For each hypothesized distribution, we modeled datasets with the distribution and then solved least-squares estimates of the distribution parameters of the nonlinear model using Gauss-Newton algorithm [5]. We used the R language [32] for statistical computations and graphics. All analysis code including our fitness function implementation is available online.[6]

Figure 1 shows GSM 0-Core, GSM 1-Core, PSTN 0-Core and PSTN 1-Core network fit results. In GSM 0-Core and 1-Core networks, power-law distribution provides the worst fit, while DPLN and PLN provide the best fit. When we look at each operator network shown in Figure 2, DPLN and PLN continue to be the best-fitting models.

It is clear that none of the curves fit the tail of the network particularly for $degree \gtrsim 150$. The tail of network for such large degrees, i.e., $degree \gtrsim 150$, represents less than one percent of nodes. Dunbar's study [12], [20] on the maximum number of individuals with whom any person can maintain stable social relationships suggests that number lies between 100 and 230; it is usually assumed to be 150. Considering Dunbar's study, the tail of the network most probably represents non-human complex nodes. Since in this study our scope is social networks with human subjects, curve fitting to the network body is sufficient to model the social network.

We also evaluate the fit success of these distribution models numerically. Table 2 summarizes the residual sum of squares (RSS)-based fit success values for each network-distribution pair. The best fits are shown in bold in the table. To compute model fit success (p-value), we first compute normalized distance where distance is the residual sum-of-squares, then subtract it from 1. Thus we get a p-value which measures how tight the model fits the real dataset. A large p-value indicates better fit to the empirical data.

The fit success results in Table 2 put forward two distributions: DPLN and PLN. The former provides the best fit for three social networks (0C PSTN, 1C PSTN, and 1C GSM C), while the latter provides the best fit for four social networks (0C GSM A, 0C GSM ALL, 1C GSM Aand 1C GSM B). Both distributions provide equally good fits for three social networks (1C GSM ALL, 0C GSM B0C GSM C). There is no significant difference in their fit success; PLN is only slightly better than DPLN. In fact, DPLN and PLN do not lead to significantly better fits than the other models except power-law distribution. It is only a marginal improvement and should not be accepted as a generalized improvement. *power-law with cutoff, log-normal, exponential, PLN* and *DPLN* are possible representative distributions. Nevertheless, considering its lower number of parameters than DPLN and its slightly better fits than other distributions, we choose PLN distribution as the representative distribution for our social network datasets. Hereafter, when we need to model a network, we will use PLN.

[6]see www.cs.bilkent.edu.tr/∼haksu/callgraph/

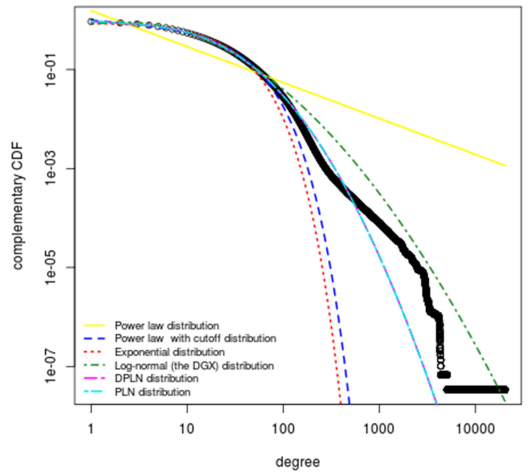
1) WORKING WITH LARGE DATASETS

We encountered some limitations while working with large datasets. Initially, we started with a commercial relational database management system (RDBMS) on high-end hardware with ∼ 45 terabyte disk, 24 CPU cores, and 96 GB memory. Extract, transform, and load processes take three days and require careful performance tuning. Using this RDBMS solution, we are able to compute and export the degree distributions used in Sections IV-B, IV-C, IV-D, and IV-E. 8 GB memory is sufficient for R programs to compute our fitting models, statistics, and plots. On the other hand, relational databases perform poorly on graph traversal operations, i.e., multiple self-joins of large edges table become computationally infeasible. In order to be able to compute traversal-based network properties (e.g., clustering-coefficients) we setup a Hadoop/HBase cluster and loaded our dataset into HBase tables. We then implemented network analysis algorithms for graphs stored in HBase (see [2] for used platform details). Hadoop/HBase cluster solution enables us to compute the network properties reported in this study.

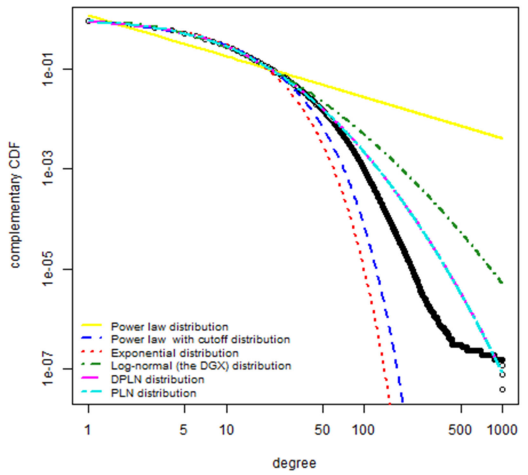
B. NETWORK OPERATOR

By comparing the degree distribution characteristics of social networks derived from different operator data, we try to answer the question of whether characteristics are dependent on network operators or not. Doing so will clarify if investigating one operator's social network of users is sufficient for social network analysis.

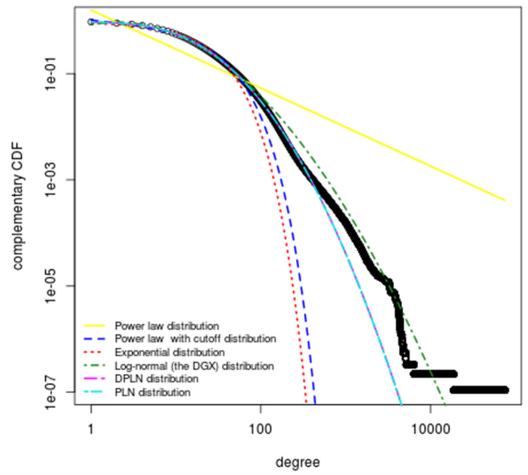
To analyze the effect of the network operator, we again use the social networks constructed in Section IV-A, i.e., three GSM operators' social networks, one PSTN operator's social network and the GSM operators' joint social network. Figure 3 illustrates and compares degree distribution in the GSM and PSTN networks. The former displays a higher density for lower degrees, while the latter displays a higher density for degrees larger than 122. We think that the high density for higher degrees in the PSTN network might be because fixed-line phones are used as household items rather than personal belongings, and are shared by many members in the house. Thus, PSTN node degrees can be considered as the sum of social degrees of multiple individuals. Figure 4 shows the degree distributions of the various GSM operator networks. We can see that there is no significant difference between degree distributions of the three GSM operators' networks and the joint network derived from the three operators. We also apply the Kruskal-Wallis Test to compare the degree distribution of complex communication networks breakdown by network-operator. As the result of this test, the p-value turns out to be greater than the 0.05 significance level (p-value=0.84). Hence, we conclude that the degree distributions of the analyzed social networks at network-operator breakdown are statistically identical.



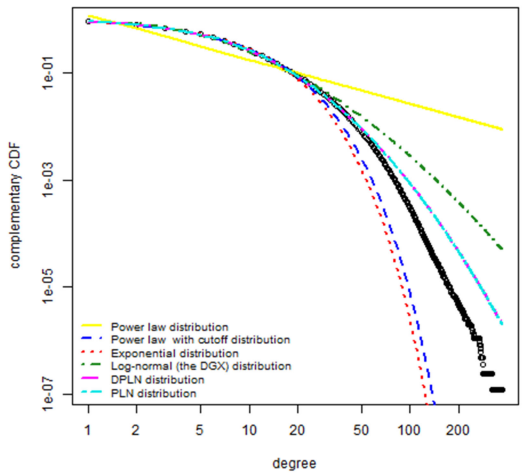
(a) 0-Core GSM A



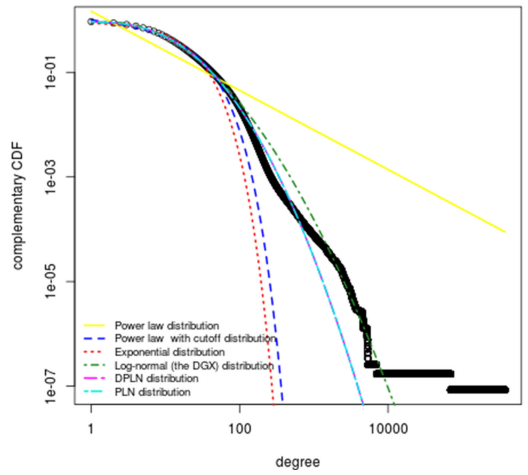
(b) 1-Core GSM A



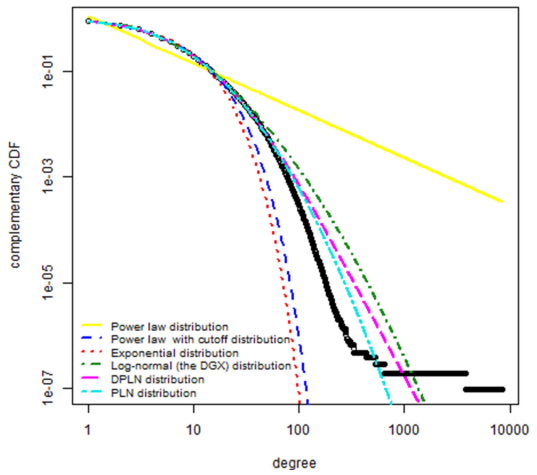
(c) 0-Core GSM B



(d) 1-Core GSM B



(e) 0-Core GSM C



(f) 1-Core GSM C

FIGURE 2. Model fits for 0-Core and 1-Core variations of GSM A, GSM B and GSM C networks are illustrated. In all networks DPLN and PLN models perform better than the rest of models. See Table 2 for p-value based quantitative results.

TABLE 2. Numerical distribution fit success results for various networks.

Network \ Distribution	Power-law	Power-law with cutoff	Exponential	Log-normal (DGX)	DPLN	PLN
1-Core GSM ALL	0.8597156	0.9980274	0.9983446	0.9954544	**0.9999636**	**0.9999639**
1-Core GSM B	0.8579531	0.9985913	0.9976061	0.9978552	0.9999707	**0.9999709**
1-Core GSM A	0.8579372	0.9981947	0.997876	0.9950699	0.9999429	**0.9999432**
1-Core GSM C	0.8799332	0.9977323	0.9991961	0.9961851	**0.9999637**	0.9999612
1-Core PSTN ALL	0.8473295	0.9991812	0.9955966	0.9976018	**0.9999069**	0.9996437
0-Core GSM ALL	0.7714906	0.9966974	0.9953066	0.991538	0.999826	**0.9998263**
0-Core GSM B	0.7733198	0.994963	0.9966673	0.9902132	**0.9999488**	**0.9999488**
0-Core GSM A	0.7642553	0.997863	0.9933416	0.993648	0.9997411	**0.9997416**
0-Core GSM C	0.7957198	0.9938651	0.997852	0.9879222	**0.9997517**	**0.9997517**
0-Core PSTN ALL	0.7228171	0.986819	0.9904483	0.9867846	**0.9969739**	0.9946071

C. NETWORK SIZE

To analyze the effect of network size on degree distribution, we start with a network around one base station and then expand it by including neighbor base station networks, just like snowball sampling. Thus, we construct social networks of different sizes for a city.[7] Then for each social network of a different size, we compute and plot the corresponding degree distribution, resulting in a chart of network size versus degree distribution parameters.

To obtain networks of various sizes, we use the SNA database, which contains the cell IDs and geographic coordinates of the GSM base stations. We divide a dense urban part of city X into 1,000 sub-parts, each of which hosts an equal number of base stations. Since each base station can serve a certain number of cell phones, we safely assume that an equal number of base stations will serve an equal number of cell phones (users). Using Google Maps, we determine the coordinates of the urban part of city X. The dataset lists around 17,000 base stations in this region, so each sub-part hosts 17 base stations. Starting from the center of the city, we draw rings around the nearest 17 base stations and label the rings from 1 to 1,000. Thus, in each iteration, we draw a new ring around the nearest 17 base stations that are not yet covered by a ring as shown in Figure 5.

Having 1,000 rings determined, we start to filter the calls in these rings so that we have networks with an increasing number of nodes inside. We define a *circle* as a ring containing all other rings with a label lower than its label. More precisely, $ring_N$ is the set of nodes R_m, where $m \leq N$. In this manner, 1,000 circles $(circle_1, \ldots, circle_{1,000})$ are defined. By filtering the calls established in each circle, we come up with 1,000 networks that differ only in size (i.e., density, location, etc., are not considered).

To determine whether there is any effect of size on degree distribution we plot the pdf of degree versus network size. Since there are 1,000 networks with increasing size, in order to make the plot easier to interpret we create a color list with a gradient of 1,000 green-blue-red colors. As illustrated in Figure 6, for increasing network size, the degree distribution curves in a specific direction: the pdf for low degrees decreases while the pdf for high degrees increases. We also apply the Kruskal-Wallis Test to compare the degree distribution of complex communication networks breakdown by network-size. As the result of this test, the p-value turns out to be less than the 0.05 significance level (p-value=5.122e-5). Hence, we conclude that the degree distributions of the analyzed social networks at network-size breakdown are statistically nonidentical.

To further investigate the effect of network size, we fit the PLN distribution to all 1,000 networks with increasing size. Then we analyze each PLN distribution model parameter against the change in size. The PLN distribution has the following pdf function:

$$pdf_{PLN}(x) = \beta x^{\beta-1} e^{(-\beta\nu+\frac{\beta^2\tau^2}{2})}\left(1-\Phi\left(\frac{log(x)-\nu+\beta\tau^2}{\tau}\right)\right)$$

and $E[X] = \nu - \frac{1}{\beta}$.

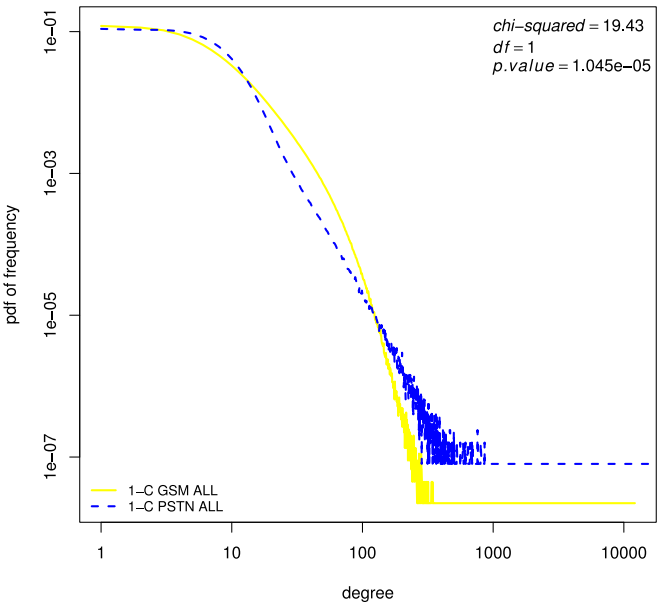


FIGURE 3. 1-Core GSM and PSTN network operators' degree pdf distribution. Test shows that GSM and PSTN are not identical distribution at 0.05 significance.

[7]As part of anonymization, we refer to the chosen city as city X.

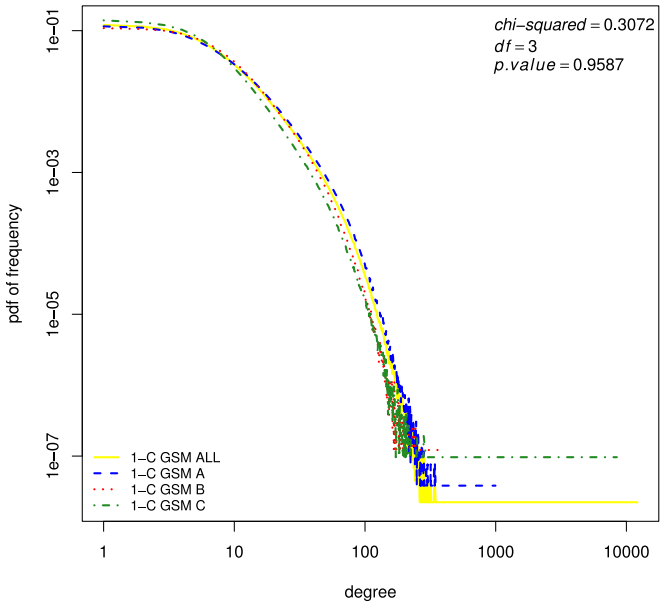


FIGURE 4. Degree distributions for different network operators are compared. Degree distributions are statistically identical for different network operators.

Figures 7 and 8 show the β and ν parameters behavior of the PLN distribution as a function of network size respectively. Figures indicate that $\beta \sim log(size)$ and $\nu \sim log(size)$. Thus, when we try to fit $\beta = a * log(size) + b$ and $\nu = a * log(size) + b$ to the results separately, we get tight fits as illustrated by blue dashed lines. Since $E[X] = \nu - \frac{1}{\beta}$, considering the $\nu \sim log(size)$ and $\beta \sim log(size)$ observations together, we conclude that the average degree of observed networks is proportional to the logarithm of the network size.

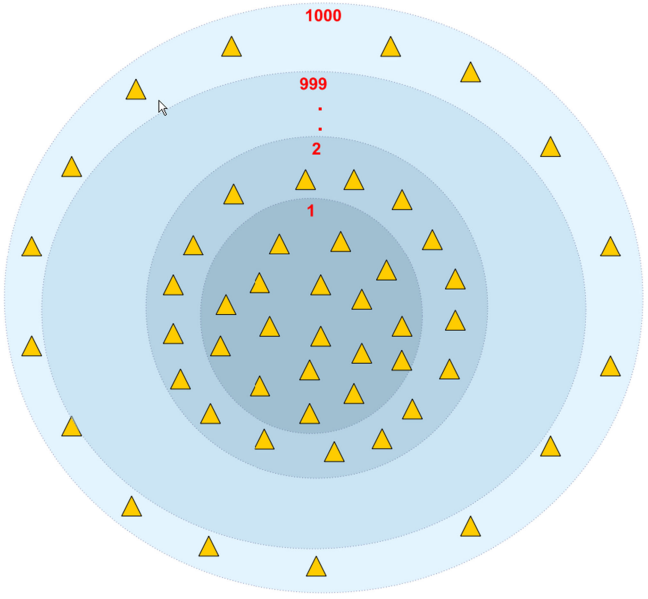


FIGURE 5. 1,000 rings around base stations. Each ring is drawn to cover the nearest 17 base stations that are not yet covered by a ring.

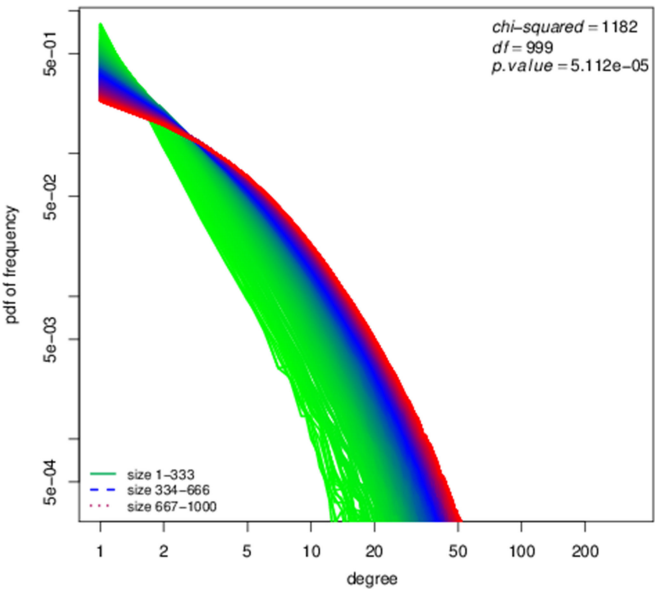


FIGURE 6. Degree distribution for increasing network size. Size unit is 17 base station, e.g., 100 means network size is 1,700 base stations. Degree distribution for 1,000 samples are plotted with gradient colors in green-blue-red range to visually follow network size versus distribution shape change. Statistical test reject the hypothesis claiming that degree distributions for varied sized networks are identical.

Following green-blue-red transition in Figure 6 size versus degree distribution, we see that the distribution function shape changes from a line into a curve while the size of network increases. This empirical result does not follow power-law generating evolution models discussed in [11]. We know that our dataset is composed of both social and non-social (complex) entities. Considering the evolution of complex networks study, we think that while complex network entities follow preferential attachment, social entities do not, due to the natural upper-bound on a node degree. Therefore, small-size samples might result in overestimating the density of popular nodes where this natural upper bound is not hit. For instance,

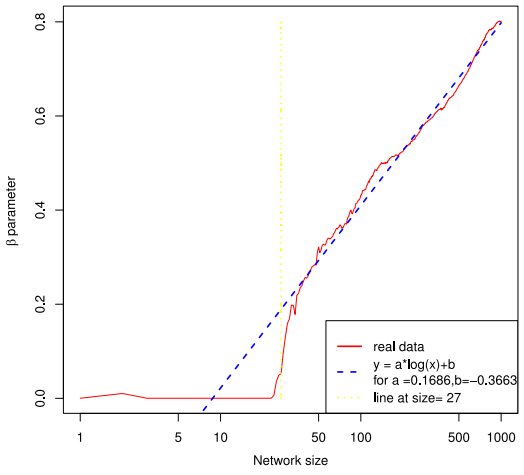


FIGURE 7. PLN β parameter versus network size in linear-log scale.

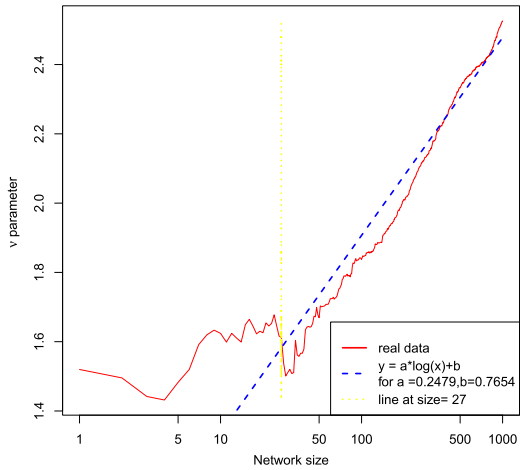


FIGURE 8. PLN ν parameter versus network size in linear-log scale.

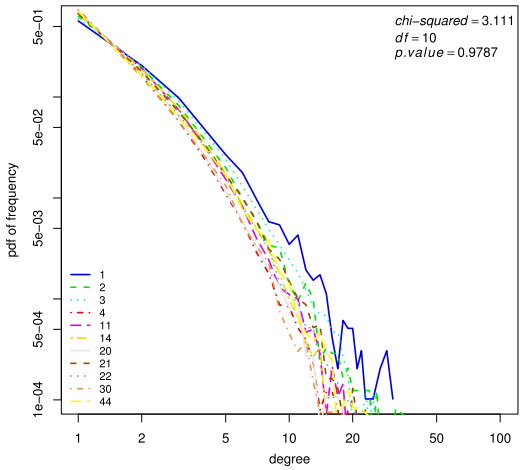


FIGURE 9. Network degree pdf versus network density plots. Kruskal-Wallis rank sum test results.

the average number of received calls (in-degree) is less than 2 in the telephone call graph sample analyzed in [11]. Thus, power-law fit for in-degree, in this case, may not remain valid for a larger sample. In fact, the study reports that it was impossible to fit out-degree by any power-law dependence.

D. POPULATION DENSITY

Here we aim to understand the effect of population density (number of users in a geographic region) on degree distribution in social networks. We would like to see whether, for example, a denser region has a denser social network. For this analysis, we again use the SNA database with GSM base station cell IDs and geographic coordinates. We draw a rectangle that incorporates the dense urban area and neighboring sparse rural areas. We divide the rectangle into 10 parts with an equal number of base stations. The entire rectangle covers nearly 450 base stations, therefore, starting from the city center, each of 45 base station cells is grouped as a ring. Then, by filtering the calls made in each ring, we get 10 social networks. For each ring, density is computed as the number of base stations per kilometer square.[8]

Figure 9 shows the degree distributions for social networks of different densities. These distributions have no specific behavior regarding increasing network density. All distributions are close to each other and they cross many times. The highest-density line (dashed blue line) falls in the middle of all the density lines. Rural areas, where the number of base stations per kilometer square is lower, show slightly higher degree density. This might be the result of outdoor based work culture in which communication is more dominated by mobile phone usage compared to the urban office based work culture where communication is achieved via Internet-based tools as well.

We also apply the Kruskal-Wallis Test to compare the degree distribution of complex communication networks breakdown by network-density. As the result of this test, the p-value turns out to be greater than the 0.05 significance level (p-value=0.98). Hence, we conclude that the degree distributions of the analyzed social networks at network-density breakdown are statistically identical.

E. GEOGRAPHIC LOCATION

Next, we aim to understand the impact of geographic location on degree distribution characteristics. We investigate how degree distribution in social networks changes when the networks are physically located in different places. For this analysis, we need social networks for which geographic locations are different while network size, density, etc., are as close as possible. To derive such networks, we sort all cities in the country by the number of base stations they have, and then we look for a consecutive sub-list in which cities are located as far apart as possible while their number of base stations are not different more than ten percent. As illustrated in Figure 10, we choose 10 such cities, each having $1,000 \pm 100$ base stations. We filter the calls made in each city and then construct 10 social networks.

Figure 11 shows degree distributions of the social networks of the selected cities. The anonymized list of cities north to south is: E, Z, G, T, B, Y, A, I, M, R; and west to east is: E, T, M, I, A, B, Z, Y, G, R. As can be observed from the figure, degree distribution curves are very close to each other and there is no specific curve behavior following city locations.

We also apply the Kruskal-Wallis Test to compare the degree distribution of complex communication networks breakdown by network-location. As the result of this test, the p-value turns out to be greater than the 0.05 significance level

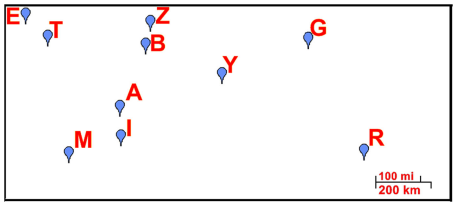


FIGURE 10. Locations of chosen cities in the country.

[8]Because base stations are located with a density proportional to population density, we consider base station density to be a measure of population density.

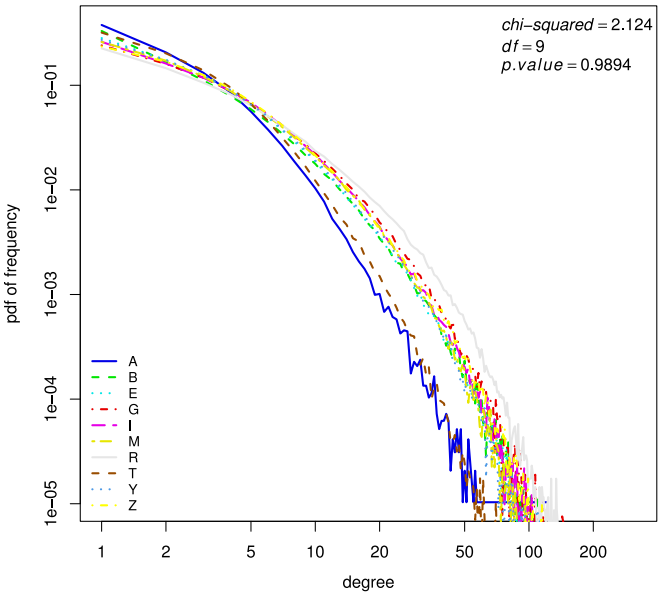


FIGURE 11. Network degree pdf versus network location.

(p-value=0.99). Hence, we conclude that the degree distributions of the analyzed social networks at network-location breakdown are statistically identical.

V. STRUCTURAL PROPERTIES OF THE COMMUNICATION NETWORK

So far we have examined the effects of certain parameters on degree distribution. We now construct a general communication network from the dataset and analyze it for structural properties. *Clustering coefficient* is defined as the fraction of triangles around a node. This measure says how well a node's neighbors are connected. Social networks are known to have large clustering coefficients. Figure 12 displays the clustering coefficient values as a function of the degree of a node for GSM and PSTN networks. The clustering coefficient decays slowly with exponent -0.37 ($c\ d^{-0.57}$) with the degree of a node till degree d (150), and then scatters around. Results on web graphs and theoretical analysis on hierarchical networks report decays with exponent -1 [33], while results on

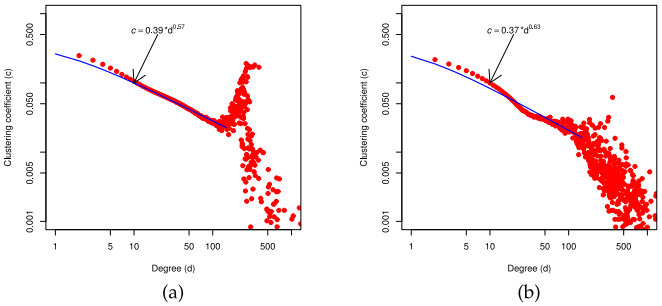


FIGURE 12. Average clustering coefficient distribution versus node degree for (a) 1-Core GSM and (b) 1-Core PSTN networks. Clustering coefficients decay with node degree with exponents (a) -0.57 and (b)-0.63, respectively. Variance increases after $d \sim 150$ where non-social entities appear more.

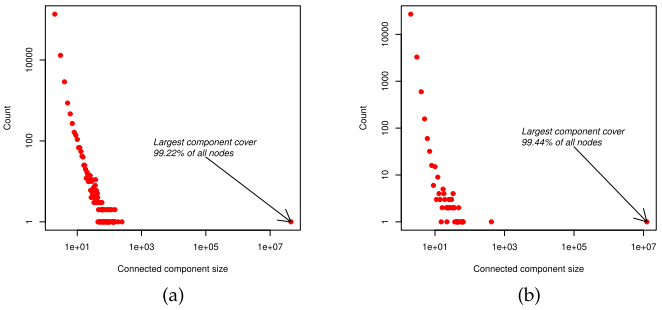


FIGURE 13. Distribution of connected components in (a) GSM (b) PSTN networks. Over 99 percent of the nodes belong to the largest connected component. Many small components exist against a few large components.

Messenger network report decays with exponent -0.37 [26]. Comparatively, our results suggest that clustering in phone call graphs is much higher than the theoretical expectation and web graph results, however, it is lower compared to the clustering in Messenger communication graph. In other words, phone users with common friends tend to be connected more probably than the theoretical expectation and connected less probably than Messenger users with common friends. Scattering after a certain degree d (150) implies that neighbors with high degree nodes know each other less, thus such nodes are non-social entities like customer support lines.

Figure 13 displays size distribution of connected components in networks. Over 99 percent of the nodes belong to the largest connected component, and the remaining small components show a power-law like distribution. This high connected component indicates that vast majority of users have communication with society and society is well connected. In other words, most of the users are reachable from the community. When the connectivity threshold is made higher, the size of the largest connected component is dropped as displayed at Figure 14(a).

We further study *community structure* in the networks by computing k-core decomposition of the graph. k-core

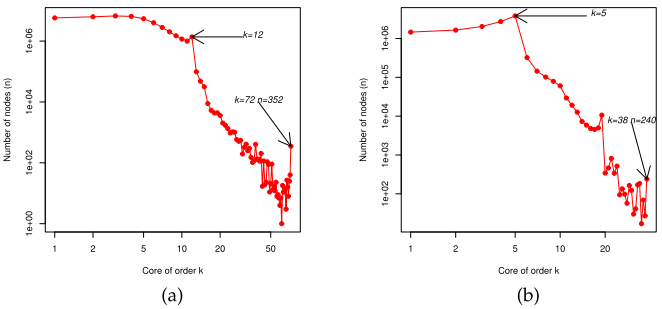


FIGURE 14. Size distribution of k-cores in (a) GSM (b) PSTN networks. The densest region in GSM network is composed of 352 nodes where each node has more than 72 edges inside the set, while the densest region in PSTN network is composed of 236 nodes where each node has more than 38 edges inside the set. The decay in k-core sizes is stable up to a cutoff value $k_{pstn\_cutoff} \approx 5$ in PSTN and $k_{gsm\_cutoff} \approx 12$ in GSM, and then the k-core size drops rapidly which means that the nodes with degrees less than the cutoff value are on the fringe of the network.

decomposition is a subgraph density measure and it identifies dense regions in the graph.[9] Figure 14 displays the distribution of k-core sizes for (a) GSM and (b) PSTN networks. The decay in k-core sizes is stable up to a cutoff value ($k_{pstn\_cutoff} \approx 5$ in PSTN and $k_{gsm\_cutoff} \approx 12$ in GSM), then the k-core size drops rapidly which means that the nodes with degrees less than the cutoff value are on the fringe of the network. This structure is similar to the Messenger communication network with $k_{msn\_cutoff} \approx 20$ [26], while it is quite different from the Internet graph in which k-core size decays as a power-law with k [3]. The densest region in GSM network is composed of 352 nodes where each of the nodes has more than 72 edges inside the set.

VI. CONCLUSION AND FUTURE WORK

In this study, we attempt to empirically test degree distribution versus different dataset scenarios to understand the parameters governing degree distribution in social networks. We observe that degree distribution in social networks does not show a significant correlation with population density, user telco operator, and user geographic location; however, population size directly affects the average degree of the social network. Therefore, in social network studies it is important to keep social network size as a parameter while interpreting degree distribution. It also seems acceptable to study a social network without considering its location, density and referred telco operator. For instance, a researcher could gather data from an urban part or a rural part of a country, or may choose a specific city or telco operator. However, any change in the size of the studied network would result in a considerable change in degree distribution characteristics and overall network topology. Hence, social network studies must indicate the size of the studied network and consider different sizes to come up with a sound and complete conclusion. As a future work, multivariate regression / mixed-effects modeling can be used which will eliminate possible effects of the heuristics that are used to fix parameters in this study. Considering the size of the dataset and lack of distributed multivariate regression algorithm for Hadoop cluster, we did not attempt to use multivariate regression at this study.

ACKNOWLEDGMENTS

We thank The Scientific and Technological Research Council of Turkey (TUBITAK) for supporting this work in part with project 113E274. We are grateful to Rana Nelson for proofreading and suggestions. In addition, we would like to thank Mahmut Kutlukaya for his expert contributions on statistical tests. We also deeply thank anonymous reviewers for their insightful comments and suggestions.

REFERENCES

[1] W. Aiello, F. Chung, and L. Lu, "A random graph model for massive graphs," in *Proc. 32nd Annu. ACM Symp. Theory Comput.*, 2000, pp. 171–180.

[2] H. Aksu, M. Canim, Y. Chang, I. Korpeoglu, and O. Ulusoy, "Distributed k-core view materialization and maintenance for large dynamic graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 10, pp. 2439–2452, Oct. 2014.

[3] J. I. Alvarez-Hamelin, L. Dall'Asta, A. Barrat, and A. Vespignani, "Analysis and visualization of large scale networks using the k-core decomposition," in *Proc. Eur. Conf. Complex Syst.*, 2005.

[4] Z. Bi, C. Faloutsos, and F. Korn, "The "DGX" distribution for mining massive, skewed data," in *Proc. 7th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2001, pp. 17–26.

[5] Å. Björck, *Numerical Methods for Least Squares Problems*. Philadelphia, PA, USA: SIAM, 1996.

[6] A. Buscarino, M. Frasca, L. Fortuna, and A. S. Fiore, "A new model for growing social networks," *IEEE Syst. J.*, vol. 6, no. 3, pp. 531–538, Sep. 2012.

[7] J. Carrasco, D. Fain, K. Lang, and L. Zhukov, "Clustering of bipartite advertiser-keyword graph," presented at the *Int. Conf. Data Mining*, Melbourne, FL, USA, Nov. 2003.

[8] S. Chiappetta, C. Mazzariello, R. Presta, and S. Romano, "An anomaly-based approach to the analysis of the social behavior of VoIP users," *Comput. Netw.*, vol. 57, no. 6, pp. 1545–1559, 2013.

[9] A. Clauset, M. E. J. Newman, and C. Moore, "Finding community structure in very large networks," *Phys. Rev. E*, vol. 70, no. 6, Dec. 2004, Art. no. 066111.

[10] K. Dasgupta, *et al.* "Social ties and their relevance to churn in mobile telecom networks," in *Proc. 11th Int. Conf. Extending Database Technol.: Advances Database Technology*, 2008, pp. 668–677.

[11] S. N. Dorogovtsev and J. F. Mendes, "Evolution of networks," *Advances Phys.*, vol. 51, no. 4, pp. 1079–1187, 2002.

[12] R. I. M. Dunbar, "Neocortex size as a constraint on group size in primates," *J. Human Evol.*, vol. 22, no. 6, pp. 469–493, Jun. 1992.

[13] N. Eagle, A. S. Pentland, and D. Lazer, "Inferring friendship network structure by using mobile phone data," *Proc. Nat. Academy Sci. USA*, vol. 106, no. 36, pp. 15274–15278, 2009.

[14] European Union, "Directive 2006/24/EC of the european parliament and of the council," 2006. [Online]. Available: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:105:0054:0063:EN:PDF, Accessed on: Jun. 20, 2016.

[15] S. Fortunato and M. Barthélemy, "Resolution limit in community detection," *Proc. Nat. Academy Sci. USA*, vol. 104, no. 1, pp. 36–41, Jan. 2007.

[16] D. A. Gianetto and B. Heydari, "Catalysts of cooperation in system of systems: The role of diversity and network structure," *IEEE Syst. J.*, vol. 9, no. 1, pp. 303–311, Mar. 2015.

[17] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proc. Nat. Academy Sci. USA*, vol. 99, no. 12, pp. 7821–7826, Jun. 2002.

[18] R. Guimerà, M. Sales-Pardo, and L. A. N. Amaral, "Modularity from fluctuations in random graphs and complex networks, " *Phys. Rev. E: Statistical Nonlinear Soft Matter Phys.*, vol. 70, no. 2, 2004, Art. no. 025101.

[19] M. Haghnevis and R. G. Askin, "A modeling framework for engineered complex adaptive systems," *IEEE Syst. J.*, vol. 6, no. 3, pp. 520–530, Sep. 2012.

[20] R. A. Hill and R. I. M. Dunbar, "Social network size in humans," *Human Nature*, vol. 14, no. 1, pp. 53–72, Mar. 2003.

[21] S. Hoteit, S. Secci, S. Sobolevsky, C. Ratti, and G. Pujolle, "Estimating human trajectories and hotspots through mobile phone data," *Comput. Netw.*, vol. 64, pp. 296–307, 2014.

[22] D. Jiang, Y. Wang, C. Yao, and Y. Han, "An effective dynamic spectrum access algorithm for multi-hop cognitive wireless networks," *Comput. Netw.*, vol. 84, pp. 1–16, 2015.

[23] B. Karrer, E. Levina, and M. E. J. Newman, "Robustness of community structure in networks," *Phys. Rev. E*, vol. 77, no. 4, Sep. 2007, Art. no. 046119.

[24] P. Kazienko, K. Musial, and T. Kajdanowicz, "Multidimensional social network in the social recommender system," *IEEE Trans. Syst. Man Cybern.—Part A*, vol. 41, no. 4, pp. 746–759, Jul. 2011.

[25] A. Le Menach, *et al.*, "Travel risk, malaria importation and malaria transmission in Zanzibar," *Sci. Rep.*, vol. 1, 2011, Art. no. 93.

[26] J. Leskovec and E. Horvitz, "Planetary-scale views on a large instant-messaging network," in *Proc. 17th Int. Conf. World Wide Web*, 2008, pp. 915–924.

[27] S. Milgram, "The small world problem," *Psychology Today*, vol. 2, pp. 60–67, 1967.

[28] J.-K. Min and S.-B. Cho, "Mobile human network management and recommendation by probabilistic social mining," *IEEE Trans. Syst. Man Cybern. Part B: Cybern.*, vol. 41, no. 3, pp. 761–771, Jun. 2011.

[9]The k-core of a graph is a subgraph K, where each vertex in K has at least k edges to other vertices in K.

[29] A. Nanavati, *et al*., "Analyzing the structure and evolution of massive telecom graphs," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 5, pp. 703–718, May 2008.
[30] J. P. Onnela, *et al*., "Structure and tie strengths in mobile communication networks," *Proc. Nat. Academy Sci. USA*, vol. 104, no. 18, pp. 7332–7336, May 2007.
[31] J. Palau, M. Montaner, B. López, and J. L. D. L. Rosa, "Collaboration analysis in recommender systems using social networks" in *Proc. 8th Int. Workshop Cooperative Inf. Agents*, 2004, pp. 137–151.
[32] R Development Core Team, *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Found. Statistical Comput., 2010.
[33] E. Ravasz and A. L. Barabási, "Hierarchical organization in complex networks," *Phys. Rev. E*, vol. 67, no. 2, Feb. 2003, Art. no. 026112.
[34] A. Sala, S. Gaito, G. P. Rossi, H. Zheng, and B. Y. Zhao, "Revisiting degree distribution models for social graph analysis," *CoRR*, 2011.
[35] J. Schroeder, J. Xu, and H. Chen, "CrimeLink explorer: Using domain knowledge to facilitate automated crime association analysis," in *Proc. 1st NSF/NIJ Conf. Intell. Secur. Informat.*, 2003, pp. 168–180.
[36] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskove, "Mobile call graphs: Beyond power-law and lognormal distributions," in *Proc. 14th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2008, pp. 596–604.
[37] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. L. Barabasi, "Human mobility, social ties, and link prediction," in *Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2011, pp. 1100–1108.
[38] Y. Wang, M. Iliofotou, M. Faloutsos, and B. Wu, "Analyzing communication interaction networks (CINs) in enterprises and inferring hierarchies," *Comput. Netw.*, vol. 57, no. 10, pp. 2147–2158, 2013.
[39] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge, U.K.: Cambridge Univ. Press, Jan. 1995.
[40] D. J. Watts, "A twenty-first century science," *Nature*, vol. 445, no. 7127, Jan. 2007, Art. no. 489.
[41] A. Wesolowski, *et al*., "Quantifying the impact of human mobility on malaria," *Science*, vol. 338, no. 6104, pp. 267–270, 2012.
[42] J. Yang, Y. Qiao, X. Zhang, H. He, F. Liu, and G. Cheng, "Characterizing user behavior in mobile internet," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 1, pp. 95–106, Mar. 2015.
[43] S. Zhang, D. Yin, Y. Zhang, and W. Zhou, "Computing on base station behavior using erlang measurement and call detail record," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, no. 3, pp. 444–453, Sep. 2015.
[44] Z. Zhang, H. Wang, C. Wang, and H. Fang, "Modeling epidemics spreading on social contact networks," *IEEE Trans. Emerg. Topics Comput.*, vol. 3, pp. 3, pp. 410–419, Sep. 2015.

HIDAYET AKSU received the BS, MS, and PhD degrees from the Department of Computer Engineering, Bilkent University, in 2005, 2008, and 2014, respectively. He is currently a postdoctoral associate in the Department of Electrical & Computer Engineering, Florida International University. Before that, he worked as an adjunct faculty in the Computer Engineering Department, Bilkent University. He conducted research as visiting scholar with IBM Thomas J. Watson Research Center, Yorktown Heights, New York, in 2012-2013. He also worked with Scientific and Technological Research Council of Turkey (TUBITAK). His research interests include security for cyber-physical systems, Internet of Things, security for critical infrastructure networks, IoT security, security analytics, social networks, big data analytics, distributed computing, wireless networks, wireless ad hoc and sensor networks, localization, and p2p networks. He is a member of the IEEE.

IBRAHIM KORPEOGLU received the BS degree in computer engineering from Bilkent University, in 1994. He received the MS and PhD degrees in computer science from the University of Maryland, College Park, in 1996 and 2000, respectively. He joined Bilkent University in 2002, and he is an associate professor in the Department of Computer Engineering. Before that, he worked in several research and development companies in USA including Ericsson, IBM Thomas J. Watson Research Center, Bell Laboratories, and Bell Communications Research (Bellcore). He received Bilkent University Distinguished Teaching Award in 2006 and IBM Faculty Award in 2009. He is a member of the ACM and a senior member of the IEEE.

ÖZGÜR ULUSOY received the PhD degree in computer science from the University of Illinois at Urbana-Champaign. He is currently a professor in the Computer Engineering Department, Bilkent University, Ankara, Turkey. His current research interests include web databases and web information retrieval, multimedia database systems, social network analysis, and peer-to-peer systems. He has published more than 120 articles in archived journals and conference proceedings. He is a member of the IEEE and the ACM.