Reconstruction of a Social Network Graph from Incomplete Call Detail Records

Mariusz Kamola
Research and Academic Komputer Network (NASK)
ul. Wąwozowa 18, 02-796 Warsaw, Poland
Mariusz.Kamola@nask.pl

Bartłomiej Cezary Piech, Mariusz Kamola
Institute of Control and Computation Engineering
Warsaw University of Technology
ul. Nowowiejska 15/19, 00-665 Warsaw, Poland
B.Piech.1@stud.elka.pw.edu.pl

Abstract—**Real-life call detail data (CDR) are used to build a graph of a social network of telecommunication operator customers. Affiliation network is used in graph construction since CDR data are partially kept anonymous. A number of the resulting network properties are examined to prove the correctness of the graph construction algorithm. Cliques in the network and network dynamics are analyzed; suggestions are given about possible utilization of the obtained information in the operation of a telecommunication operator.**

Keywords-social network, affiliation graph, bipartite graph, parallel processing, CDR, cliques

I. INTRODUCTION

Social network analysis is a valuable tool for knowledge extraction from raw and massive data. In the telecommunication industry the fundamental set of data is call detail records, CDR, describing call attempts and successful calls made by telecommunication operator customers.

The standard data mining techniques serve inferring phenomena of interest form the unstructured data. They are: daily activity profiles, call duration distributions, phone usage deviations preceding customer churn – and other kinds of statistics useful for sales strategy making. Constructing a graph of interpersonal connections from the same data is another step further of the knowledge extraction process. It allows behavior analysis of a user in the context of her network of relations with the – direct and indirect – neighbors. Data mining performed on such graphs gives results focused on individual customers, but is still capable of producing interesting statistical network parameters. Personalized results help addressing customers with customized sales offers or churn prevention actions, while the new statistical parameters give insight into coherence of customers community. There are numerous opinions [1] that social network analysis (SNA) is one of the top technologies of interest for the industry and the suite of off-the-shelf commercial SNA tools is already impressive.

In the classical scenario CDRs contain plain phone numbers of both the caller and the callee. In such case the social graph construction consists in simply using phone numbers as vertices and calls as edges, calls frequency being the weight of each edge. Certainly, the data must prior be cleaned and filtered. However, in the case described in this paper the CDR data have been anonymized. Those data properties are described in Sec. II. Next, procedure of social graph construction from the data is presented in Sec. III. Properties of the obtained graph and their comparison to other typical kinds of social networks are given in Sec. IV. The paper concludes in Sec. V where possibilities of the result application and future work are outlined.

II. CALL DATA

The data subject to analysis contain call details from a wired telephony operator, from October to December several years ago. Call records have been deprived of the calling number, and the number being called has been scrambled in an unknown, uniform way, thus preserving uniqueness of numbers being called. The intent of the operator was to preserve privacy of its customers, while making it possible to performed data mining outside the company. Social network analysis was not intended initially, but social network reconstruction from the scrambled data was performed under the consent.

The major research question posed was if it is possible to reconstruct classical social network graph, provided that a CDR contains the following fields:

- customer ID, scrambled,
- number being called, scrambled,
- call date, time and duration,
- call flags (successful, busy, timed out on waiting),
- "business customer" flag.

It is evident that customer IDs and the numbers being called belong to two different spaces, and there is no mapping available between a customer ID and a set of phone lines (and numbers) that he or she possesses. Therefore, no simple approach (as the one mentioned in Sec. I) is possible.

The number of CDRs available was 133 million; all they were moved from plain files into relational database for more efficient processing. They cover both individual and commercial customer activities; however, these two market segments differ so much that they cannot be treated uniformly. Commercial customers generate and receive an order of magnitude bigger traffic, and do not fit some assumptions in Sec. III on the nature of interpersonal relations. That is why it was decided that commercial customers will be excluded from analysis. The first exclusion criterion is the "business customer" flag explicitly set; but it was noticed that some of the remaining "individual" customers exhibit business-like characteristics, for instance, a number of calls made that would be physically impossible

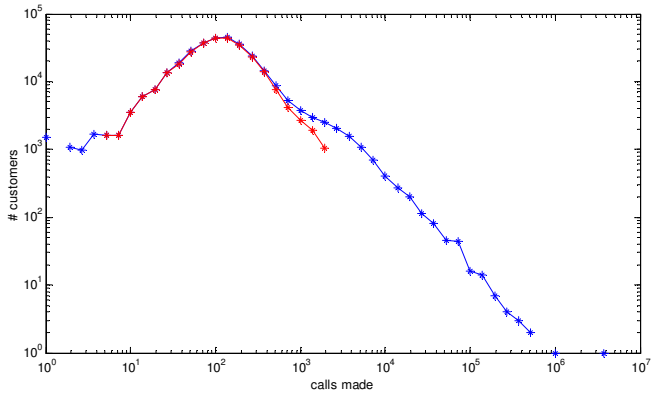

Figure 1. Histogram of calls made count by all customers (blue) vs. the calls made by the selected individual customers

for an individual to generate (e.g. the order of tens of thousands). Therefore, an extra exclusion criteria being the maximum number of calls made in the quarterly period being less that 2,400 (i.e. 24 daily, on average), was implemented. On the other hand, some data cleaning was needed in order to exclude invalid or very passive customers not providing any useful information. It was assumed that customers with less than 6 calls in the considered period will be removed.

To verify correctness of these criteria, a histogram of number of calls was prepared, as in Figure 1. The histogram of number of calls made by the filtered customers covers almost all calls in its domain, and resembles a reasonable Gaussian distribution, indicating that most customers make 2 calls a day on average. It is noticeable that the number of filtered inactive or over-active customers amounts to 6.7 percent of the total of 316,000 customers.

III. Reconstruction Method

In absence of direct call details between the customers, the social network graph must be reconstructed using different approach. We followed an old observation by Simmel [2] that dyads and triads are the basic building blocks of a society. Extending this approach, we can build an affiliation network, using customers as ordinary nodes and called numbers as affiliation objects. This leads to creation of a bipartite graph. Next, we can assume that customers affiliated at commonly called numbers are also connected.

Such approach has several drawbacks, though. First, it may lead to creation of very dense, complete subgraphs of individuals calling the same number. Hopefully, the customers exhibiting inhumanly intense behavior have already been excluded, but the upper qualification limit of 2400 calls per quarter applies only to calling side; therefore there can exist numbers being called more frequently. Indeed, the most called number received as many as 176,000 calls from nearly 5,100 different customers. It can hardly be expected that all of those customers maintain personal relationships.

Therefore, the graph must be reconstructed using only the numbers most frequently dialed by customers. Let us denote the list of r most often dialed numbers by customer c, in descending order, by $D_c=(n_{c,1}, n_{c,2},\ldots, n_{c,r})$. The reconstructed social network graph is then defined by sets of vertices and edges, $\{V,E\}$, where $V=\{c_1, c_2,\ldots\}$ denotes all qualified customers and

$$E=\left\{\left(c_x,c_y\right): c_x,c_y \in V \wedge D_{c_x} \cap D_{c_y} \neq \varnothing \right\}$$

The second problem is the choice of value for r, the ranking length or algorithm sensitivity. Keeping it small, e.g. r=2 goes in line with the observation about triads but prevents creation of bigger clusters in the graph. With r growing, the graph density grows unnecessarily, unimportant and accidental affiliations are treated with equal care as the important ones, and graph processing algorithms need more time to run. A series of attempts was made by the authors to reveal the most typical length of numbers frequently called, and apply it as r. The idea was inspired by Zipf's law [8] and confirmed in [3] where it was shown that calls distribution is power-like wrt. the rank of the numbers called. In our case, a quick insight into the shapes of calls distribution revealed that they may definitely differ from Zipf's law, and frequently taking a reverse "S" shape. A number of parametric models (power, arcus tangent and hyperbolic tangent) have been tried to model the actual call density distributions for every user. However, neither of them has lead to any kind of segmentation, i.e. to any typical cut-off values for r.

The third and not the least problem encountered were resource requirements of algorithms for graph reconstruction. The number of graph vertices was n=295,000, computational complexity was $O(n^2r^2)$ and memory requirements depended on the final graph density, difficult to assess. Similarly, algorithms for graph analysis (calculating diameter, betweenness, cliques etc.) have their own, substantial complexities. Since reconstruction algorithm and graph analysis algorithms implementation language was Java, a dedicated parallel machine, Azul, was harnessed for computations. Azul [4] is a specialized parallel Java coprocessor – in the configuration used by the authors 96 cores and 60 GB RAM were available. Jung [5] was used as the library for graph analysis, with necessary improvements for parallel processing, developed by the authors.

IV. Social Network Properties

Verification of correctness of the reconstruction algorithm proposed in Sec. III is difficult because no testing (i.e. unanonymized) data are available. This is why we decided to calculate a number of statistical parameters of the obtained network in order to confront them with typical social networks. First, we expect power-law node degree distribution to hold for the reconstructed network. Figure 2 shows probability $P(k)$ of node with degree k to be found in the reconstruction network, for r=5. It has a shape commonly observed in other social networks.

Second, we calculate a number of statistical properties of the reconstructed graph:

- mean node degree $\bar{k}$,

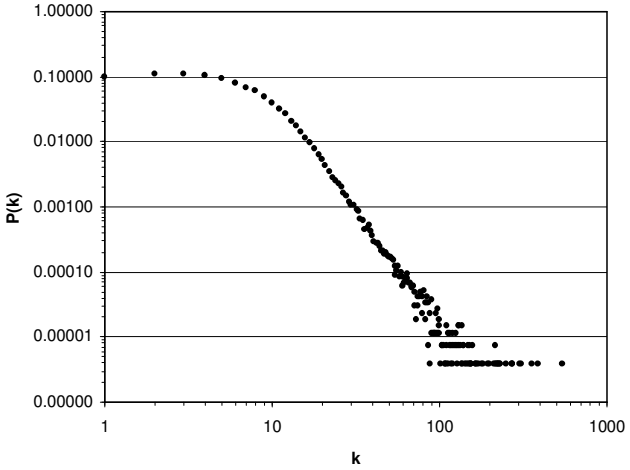

Figure 2. Power-law node degree scaling in the reconstructed network

- clustering coefficient c, as the ratio of the number of triads in the graph, tripled, to the number of paths in the graph where path lengths equals 2,
- number of graph vertices N,
- estimated power coefficient α, determining the power law $P(k) = Ck^{-\alpha}$, with C being some constant.

These properties, calculated for two reconstructed graphs for r=4 and r=5, are compared in Table 1 to properties of other typical social graphs. The graph density, the mean clustering coefficient and the graph size grow with r but the exponent α value does not change much. Similarly, the distribution of node degree (not shown) remains similar to the one in Figure 2. Consequently, one may hope that reconstructing social graph by analysis of commonly called numbers is stable w.r.t. the choice of cutoff parameter, r. Also, the reconstructed graph parameters fall quite well in ranges laid by other network examples, listed down in Table 1.

TABLE 1. GRAPH STATISTICAL PROPERTIES FOR CHANGING ALGORITHM SENSITIVITY r, VS. OTHER NETWORKS PROPERTIES

The reconstructed networks				
r	$\bar{k}$	c	N	α
4	5.34	0.21	263,000	2.7
5	7.02	0.24	278,000	2.75
Other networks				
Interconnection topology in IP network	5.98	0.18-0.3	10,700	2.5
Actors and their contacts	113.00	0.79	450,000	2.3
Words appearing together in written texts	70.00	0.77	461,000	2.7

Structured data mining from social graphs is already used for sorting out important users, using some centrality measure, and utilizing their influential position in product promotion or churn management. Figure 3 presents the distribution of nodes betweenness [9] for the reconstructed network. (From now on we will refer to the graph reconstructed with r=5.) We can spot that power law holds also in this case, and we can easily pick most influential nodes to be addressed effectively with viral marketing or social campaigns, or with churn prevention schemes. Another meaningful measure is the mean distance between graph nodes: in our case this is 6.83, which nicely confirms Milgram's observation [6] and is a useful hint for viral marketers. Noticeably, the graph dimension is only 19.

The network presented in Table 1, for r=5 has exactly 277,795 nodes, out of which 277,141 constitutes the largest connected component. The remaining 320 edges form triads and dyads. This should be considered as good result too, although, in general, all customers in a real telecommunication network should stay connected [10].

Let us now take a look at communities forming within the reconstructed network. By definition, a k-clique is a complete subgraph formed with k nodes, while a community contains cliques that share at least k-1 nodes. An exemplary graph of communities formed by 8-cliques is shown in Figure 4. With k decreasing, the number of detected communities grows; also the processing time is growing. Analyzing communities of various sizes may be a valuable help for construction of tariffs taking into account real users' social habits.

Another useful information can be observed from the graph dynamics. The input data cover last quarter of a year, and therefore it is possible to perform reconstruction only for the data from first two months, using December CDRs (including Xmas season) to discover attaching users' preferences. There were 12,000 new customers in December, but only 28% of those have connected to any of the existing communities, preferring 3 to 4-person groups. It is also important to observe that newcomers do not create their own communities but prefer the existing ones.

Figure 5 shows linking preferences for newly connected nodes. Probability of being linked to an existing node is linearly proportional to that node degree k – this is evident as far as k<20, where power law node degree scaling does not hold yet (cf. Figure 2). Such preferential linking is still in force for $k \geq 20$; it shows by the slope being less steep than in Figure 2, just due to preferential linking. Therefore, it has been shown that our network dynamics follows the well

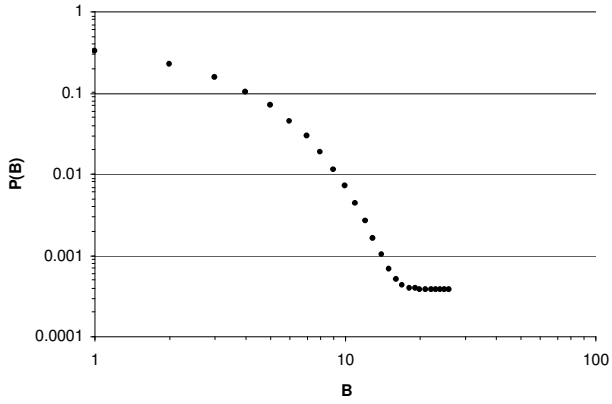

Figure 3. Power-law betweenness centrality scaling in the reconstructed network

known Barabási-Albert preferential linking mechanism [7]. This, once again, proves correctness of the reconstruction algorithm and is a valuable hint for telecommunication marketing, too.

V. CONCLUSION

By showing a number of similarities to existent networks and by examining network dynamics, we tried in Sec. IV to prove that the reconstruction algorithm proposed in Sec. III is correct, stable and useful in daily telecommunications practice. Therefore, one can quite reliably use affiliation networks to model direct relationships between telephone users.

Additionally, by observing amount of data possessed and processed, it can be hoped that operating on affiliation network can easily be done using modest computing resources. The presented approach can be applied where no direct users relationship data are available, e.g. in inferring relationships from publicly available traces of users activity in the Internet (forum posts, records of events at clubs and organizations etc.). Specifically, it can help in telecommunication tariffs construction, churn reduction and prediction, and in viral marketing.

Further work, being now in progress, concerns using edge weights and directed graphs to model interactions more precisely: new metrics (based on call frequency and rank of callees) and asymmetric relationships are under concern.

REFERENCES

[1] Gartner Identifies the Top 10 Strategic Technologies for 2011, http://www.gartner.com/it/page.jsp?id=1454221, Accessed 2011.06.30

[2] G. Simmel, The Sociology of Georg Simmel. The Free Press, New York, 1908/1950

[3] D. Lam, D.C. Cox and J. Widom, "Teletraffic Modeling for Personal Communications Services"IEEE Communications Magazine, Feb. 1997, pp. 79-87

[4] Azul Systems, http://www.azulsystems.com, Accessed 2011.06.30

[5] J. O'Madadhain, D. Fisher, P. Smyth, S. White, Y.-B. Boey, "Analysis and Visualization of Network Data using JUNG", unpublished, http://jung.sourceforge.net/doc/JUNG_journal.pdf, Accessed 2011.06.30

[6] S. Milgram, "The Small World Problem", Psychology Today, vol. 1, pp. 60-67, 1967

[7] A.-L. Barabási and R. Albert, "Emergence of Scaling in Random Networks", Science, vol. 286, pp. 509-512, 1999

[8] G.K. Zipf, Human Behavior and the Principle of Least Effort, Addison-Wesley, 1949

[9] K.-I. Goh, E. Oh, H. Jeong, B. Kahng and D. Kim, "Classification of Scale-Free Networks", Proceedings of the National Academy of Sciences of the United States, vol. 99, 2002

[10] A.A. Nanavati et al., "Analyzing the Structure and Evolution of Massive Telecom Graphs", IEEE Transactions on Knowledge and Data Engineering, vol. 20, pp.703-718, 2008

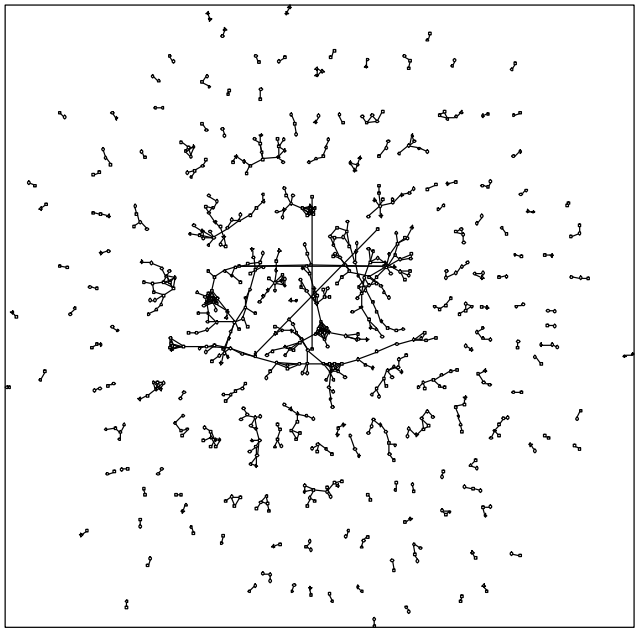

Figure 4. Exemplary community graph formed by 8-node cliques

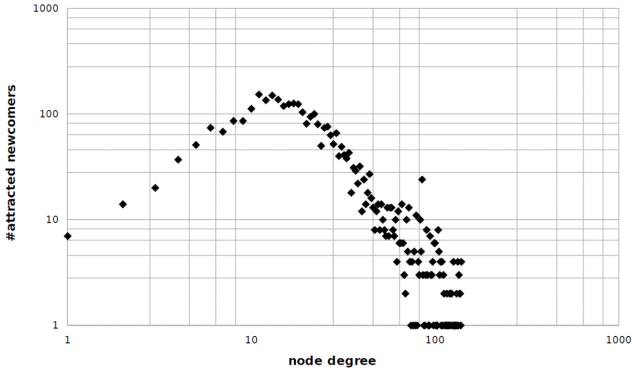

Figure 5. Linking preferences histogram w.r.t. node degree being chosen