

Exploration of Structure and Dynamics of Large Phone and SMS Networks

Leman Akoglu
Carnegie Mellon University
Computer Science Department
lakoglu@cs.cmu.edu

Bhavana Dalvi
Carnegie Mellon University
Language Technologies Institute
bbd@cs.cmu.edu

ABSTRACT

The effect of the network structure on the dynamics of social and communication networks has been of interest in recent years. It has been observed that network properties such as neighborhood overlap, clustering coefficient, etc. influence the tie strengths and link persistence between individuals. In this paper we study the communication records (both phonecall and SMS) of 2 million anonymized customers of a large mobile phone network over a period of 6 months. We analyze different network parameters in these call/SMS networks and the correlations between them. We also formulate a learning problem to determine whether existing links between users will persist in the future. Experimental results show that our method performs better than existing rule based techniques. We also detect change points in the collective behavior of users using eigenvalue analysis of various behavioral features extracted over time. Our analysis shows that there exists correlation between detected behavior change-points and social events and festivals during those points in time.

1. INTRODUCTION

Social network analysis has always been of great interest to economists, physicists and social scientists. After the emergence of telecommunications, phones have become the central source of communication and an integral part of our lives. As a result, the analysis of phone networks has indeed become very important and attracted a lot of attention. Initially the studies were carried out using questionnaire data. In recent years analysis of large scale networks has also been explored. For example, Onnela et. al. [6, 7] have done such a large-scale network analysis of one to one human communications using phone data. In particular, they observed the coupling between the tie strengths and the local network structure of users. They also analyzed the information diffusion through strong ties versus weak ties. Hidalgo et. al. [3] analyzed different attributes from local network structure and found their correlation with tie persistence. They used

rule based techniques to predict whether existing ties would persist in the future.

Nanavati et. al. [5] study the structure and the global shape of four geographically disparate mobile call graphs and propose the Treasure-Hunt model to fit their observations. [10] study the formation of social communities in temporal telecommunications records. Finally, Eagle et. al. [1, 2] study massive amounts of mobile phone records and infer social structure and behavior of users by their mobile phone interactions.

There has also been work on the analysis of time-varying networks. For example, Papadimitriou et. al. [8] introduced a technique for pattern discovery in streaming graphs which can incrementally and efficiently capture correlations and discover trends and anomalies. Sun et. al. [9] proposed GraphScope to dynamically detect communities and spot discontinuities in time-evolving streaming large graphs.

In our work, we study the anonymous phone call records of millions of users collected in a large city over a period of 6 months. These records include both their calling and SMS text messaging communications. Our phone call and SMS graphs constructed from this data include up to 2 million users with 50 million interactions. Given these large time-varying communication networks of millions of users, the following important questions come up:

- What is the level of reciprocity in human communications? Do mutual and non-mutual networks differ significantly? What can we say about the number of times i calls j , given j calls i n times?
- Is a user's number of contacts (degree) related to the number of contacts of his/her contacts (neighbors' degrees)?
- What is the relation between the topology of the communication networks and the tie strengths between individuals?
- How can we predict which ties will continue existing in the future?
- Can we detect the points in time when the user calling/texting behaviors change? Can we characterize the users mainly causing this change?

In this paper we answer the above questions. Our work is mainly divided into three major parts:

1. **Structure analysis:** We study the structural properties of our phonecall/SMS graphs such as link reciprocities, tie strengths and topological neighborhood overlaps of users.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

KDD '10, July 25–28, 2010, Washington, DC, USA.

Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$5.00.

2. **Tie persistence prediction:** We extract node/tie attributes and develop a learning model for the prediction of tie persistence over time and show that our model yields high prediction accuracy and outperforms the earlier rule based methods.
3. **Event detection:** We develop a change-point detection method using the time-varying attributes of users and show that our method can be used to find important days in which the collective calling and texting patterns of users change.

Next, we give a more detailed description of the data we used in this work. Then, we provide the methodology behind our study for each part above in Sections 3.1, 3.2 and 3.3, respectively. We show the experimental results and major conclusions in corresponding sections 4.1, 4.2 and 4.3 for each study. Finally, we conclude the paper in Section 5.

2. DATA DESCRIPTION

Our data consists of anonymous mobile communication records of millions of users over a time period of six months (December 1, 2007 through May 31, 2008). The dataset contains both phone-call and SMS interactions.

From the whole six months' of activity, we build two graphs in which nodes represent users and directed edges represent (phone-call and SMS) interactions between these users. We call the who-calls-whom graph as the MCG (for Mobile Call Graph) and the who-texts-whom graph as the MTG (Mobile Text Graph).

By construction, our graphs are weighted and directed. Here, we consider two types of weights on the arcs e_{ij} : (1) total number of phone-calls w_N , similarly total number of SMSs w_{SMS} ; and (2) total duration of phone-calls w_D from node i to j (only for MCG). Since our graphs are directed, we study the reciprocity of edges between user pairs. An edge e_{ij} from i to j is called reciprocated if there also exists an edge e_{ji} from j to i . The MTG and MCG graphs that contain *only* the reciprocated edges are called *mutual*.

To give a sense of the scale of the data we studied, we show the number of customers (with at least one contact), the number of (un)directed interactions, the total number of phone-calls/SMSs, $w_{N,SMS}$ and the total duration of phone-calls w_D , for both the mutual and non-mutual MCG and MTG in Table 1.

Notice that the MTG shrinks considerably when only reciprocated edges are considered, whereas the MCG remains almost intact.

		MTG	MCG
nonmutual	Number of nodes	1,87M	1,87M
	Number of directed edges	8,70M	49,50M
	Number of undirected edges	7,70M	28,57M
	Total number of SMS/calls	119,50M	483,70M
	Total duration of calls	N/A	5,49x10 ¹⁰
mutual	Number of nodes	0,53M	1,75M
	Number of reciprocated edges	1,99M	41,84M
	Total number of SMS/calls	91,80M	468,70M
	Total duration of calls	N/A	5,31x10 ¹⁰

Table 1: Data size statistics for (top) non-mutual and (bottom) mutual networks MTG and MCG.

3. METHODOLOGY

3.1 Network Characteristics

We start our study by analyzing our MTG and MCG graphs in terms of several network measures. Specifically, the questions that we answer in this work can be listed as follows:

1. How do mutual and non-mutual networks differ?
2. Given that a user i calls/texts user j n times, what can we say about the reciprocity, that is how many times j calls/texts user i ?
3. Is there a correlation between a node's degree and its neighbors' degrees?
4. How does the total duration or the number of phonecalls and SMSs grow by the number of contacts a user has?
5. Does strength of a tie between i and j depend on the overlap between their neighborhoods?

We study and answer these motivating questions in detail in Section 4.1.

3.2 Analysis of Tie Persistence

Our major goal here is to figure out the effect of several tie and node attributes defined below on the tie persistence between users over time. We use all or subset of these features to learn a logistic regression model and predict tie persistence in our data. We also evaluate and compare other methods to our method in terms of prediction accuracy.

We use the following definitions of tie persistence and user perseverance as defined in [3].

DEFINITION 1. Tie persistence: *It is the stability of ties across time as number of panels in which a link is observed, over the total number of panels. That is, $P_{ij} = \sum_t A_{ij}(t)/m$, where P_{ij} is the persistence of tie e_{ij} and A_{ij} is 1 if users i and j communicated in time panel t and 0 otherwise.*

DEFINITION 2. User perseverance: *Perseverance of a user is defined as the average of the persistences of all his/her ties. That is, $P_i = 1/K_i * \sum_j P_{ij}$, where P_i is the perseverance of user i , K_i is i 's degree (number of neighbors), and P_{ij} is the persistence of tie e_{ij} as defined above.*

Based on the analysis of several network measures we consider the effect of following attributes on link persistence. We divide them into two types: Tie attributes and Node attributes.

- Tie Attributes

- Reciprocity (R): R is a Boolean attribute which denotes whether the tie between i and j is reciprocated during a given time period. That is, R is 1 if both edges e_{ij} and e_{ji} exist, and 0 otherwise.
- Topological Overlap (TO): $TO(i, j) = \sqrt{\frac{O_{i,j}^2}{K_i * K_j}}$, where $O_{i,j}$ is the number of common neighbors of node i and node j , and K_i denotes the degree of node i .

- Node Attributes

- Degree (K): K_i denotes the number of neighbors of a given node i .

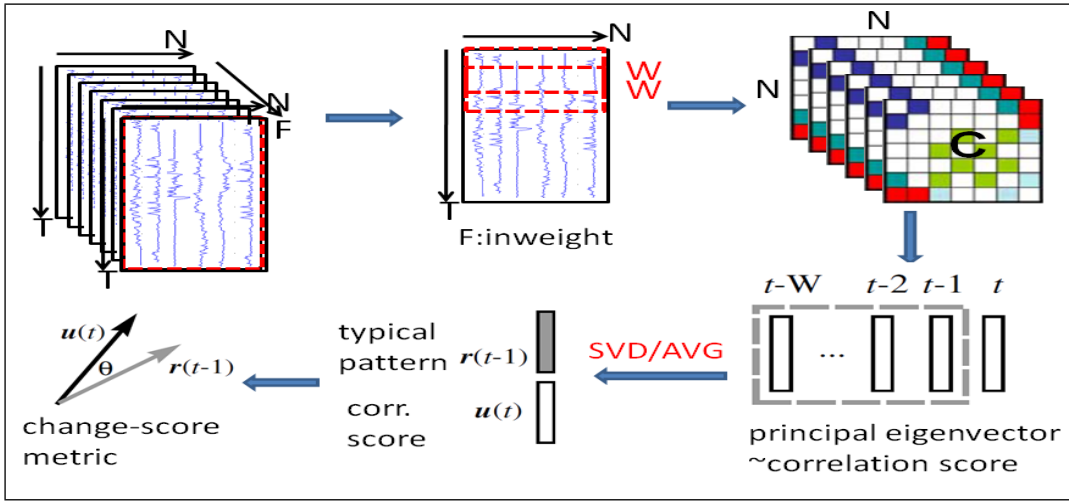


Figure 1: Flow of the change-point detection procedure.

- Cluster Coefficient (C): $C_i = \frac{2\delta_i}{K_i(K_i-1)}$, where δ_i is the number of local triangles that node i is involved in.
- User Reciprocity (r): r_i the fraction of ties of a given user i that are reciprocal.

The main questions we focus on in this section are the following:

1. Which link and node features are important in predicting link persistence?
2. How are they correlated to each other?
3. Which prediction method works best?

We answer these questions in detail in Section 4.2.

3.3 Behavior Change-Point Detection in Time

Anomaly detection has been studied widely in many settings from anomalous point detection on clouds of multi-dimensional points to spatio-temporal anomalous pattern detection with applications to network intrusion detection, medical insurance claim fraud, credit card fraud, electronic auction fraud and many others, with much less focus on anomaly detection in graph data.

In this section, we study the behavior of users in the MTG, in which nodes represent the users and edges represent the SMS interactions between these users. The data consists of six months of activity and is therefore time-evolving. Also, the edges are weighted, with weights denoting the total number of SMSs sent/received between individual pairs, for example (See Section 2 for data description). In such a setting of dynamic series of MTG graphs, the main questions we answer are the following: (1) *What points in time does the collective behavior of the nodes change?*, and (2) *Can we characterize which nodes cause most of that change?*

3.3.1 Feature Extraction from Nodes

In order to find patterns that nodes of a graph follow, we characterize the nodes with several features, so that each node becomes a multi-dimensional point. In particular, each node is summarized by a set of features extracted from its egonet (egonet of a node includes the node itself, its neighbors, and all the interactions between these nodes). The

7 features considered in this work are as follows: *indegree*, *outdegree*, *inweight*, *outweight*, *number of neighbors*, *number of reciprocal edges*, and *number of triangles in egonet*.

3.3.2 Change-Point Detection

The flow of the method used in this work to find change-points in the behavior of nodes is illustrated in Figure 1. This method is similar to Ide and Kashima [4], but differs in the construction of the “dependency” matrix C . We explain our methodology in more detail next.

Here, the data we have looks like the 3-D $T \times F \times N$ tensor on the top left corner of Figure 1, where $T=183$ days, $N \sim 2M$ nodes, and $F=7$ features. Each entry $e_{t f i}$ in this tensor denotes the value of feature f for node i at time t . To start, we take one slice of this tensor for a particular feature F_i , say *inweight*, which is a $T \times N$ matrix. Next, we define a sliding window of size W (days) over the time-series of values of nodes for that particular feature, i.e., we consider the $W \times N$ matrices. In each particular window, each node has a vector \mathbf{v} of length W of time-series of values for that particular feature. Then, for each pair of nodes i and j , we compute the correlation between their time-series vectors \mathbf{v}_i and \mathbf{v}_j for a given time window using Pearson’s ρ as,

$$\rho_{\mathbf{v}_i, \mathbf{v}_j} = \frac{\text{cov}(\mathbf{v}_i, \mathbf{v}_j)}{\sigma_{\mathbf{v}_i} \sigma_{\mathbf{v}_j}} = \frac{E[(\mathbf{v}_i - \mu_{\mathbf{v}_i})(\mathbf{v}_j - \mu_{\mathbf{v}_j})]}{\sigma_{\mathbf{v}_i} \sigma_{\mathbf{v}_j}}.$$

Then, for each window W we construct a correlation matrix C , where $C(i, j) = \rho(\mathbf{v}_i, \mathbf{v}_j)$ over that window W . For each C , we slide the window down one day and do the same for the next W days. As a result, we end up constructing 177 C matrices (top-right in Figure 1).

By the Perron-Frobenius theorem, the largest (principal) eigenvector u of each C matrix is positive. The value for each node in the eigenvector can be thought as the “activity” of that node; that is, the more correlated a node is to the majority of the nodes, the higher its “activity” value will be. Here, we call each such eigenvector as the “eigenbehavior” of the nodes.

After finding all the eigenvectors for all 177 C matrices, the change-point in the “eigenbehavior” of nodes is found as follows: For the eigenvector computed at time say t denoted by $u(t)$, we compute an “average” typical “eigenbehavior”

denoted by $r(t-1)$ from the last W eigenvectors back in time (See bottom-right in Figure 1). Next, the “eigenbehavior” at time t is compared to the “typical eigenbehavior” by taking the dot-product of those two unit vectors, $r(t-1) \cdot u(t)$. The change metric we use becomes $Z = (1 - r \cdot u)$. Here, if the new “eigenbehavior” $u(t)$ is perpendicular to the typical pattern $r(t-1)$, their dot-product gives a value of 0 or $Z=1$, whereas if $u(t)$ is the same as $r(t-1)$, then their dot-product gives a value of 1, or $Z=0$. Therefore, Z changes between 0 and 1 and a higher value of Z indicates a change point and is flagged accordingly. We present our experimental results in Section 4.3.

4. EMPIRICAL STUDY

4.1 Analysis of Network Characteristics

4.1.1 Mutuality in MTG and MCG

We notice that the MTG shrinks considerably when only mutual edges are considered, whereas MCG remains almost intact in Table 1 in Section 2.1. In particular, while there are around 2 million nodes in MTG, when we remove non-reciprocal edges the graph size drops to around half a million nodes. This shows that the vast majority of SMS interactions occur non-mutually suggesting for spam messages.

We depict the Negative Cumulative Density Function for the degree distribution in both mutual and non-mutual graphs in Figure 2. Notice that the probability there exists a node with degree $k \geq 1$ is only 0.3 for mutual MTG. This indicates that around 70% of the nodes become singletons with no incoming/outgoing links when non-reciprocal edges are removed.

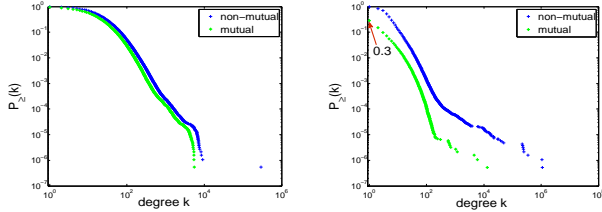


Figure 2: NCDF (Negative Cumulative Density Function) of degrees in (left) MCG and (right) MTG for mutual and non-mutual graphs, shown with green stars and blue pluses, respectively. Notice the MTG shrinks considerably when only reciprocated edges are considered.

OBSERVATION 1. *There is far less mutuality in the SMS network MTG, suggesting for spam messengers.*

4.1.2 Reciprocity in mutual MTG and MCG

Given a user i calls/texts user j n times, what can we say about how many times j calls/texts user i ?

In Figure 3 we show the reciprocal edge weights on a given pair of reciprocal edges e_{ij} and e_{ji} between i and j for all such pairs. (We denote the smaller weight as n_{ST} and the larger weight as n_{TS} , (S for Silent and T for Talkative) so that the points lie above the diagonal).

Each blue dot in Figure 3(a) represents a reciprocal edge pair and plots n_{TS} versus n_{ST} , where the weights denote the

total duration of phonecalls w_D between pairs (aggregated in 10 mins). The pairs close to the diagonal have a more balanced amount of reciprocity, whereas the pairs farther from the diagonal have an uneven communication.

Due to over-plotting, the density of points is missing in Figure 3(a). Therefore, we show the same plot using heatmaps in Figure 3(b), in which dark red regions have a higher concentration of points. Also, Figure 3(c) and Figure 3(d) show the reciprocity for total count of phonecalls w_N in MCG and SMSs w_{SMS} in MTG, respectively. In all three cases we notice the same pattern: the majority of the points reside closer to the origin and along the diagonal. This shows that users usually have a small and balanced amount of reciprocity in both their SMS and phonecall interactions. On the other hand, there also exists pairs where in return to a single SMS/call, over a thousand SMSs/calls have been made. These pairs can be easily flagged as suspicious, however, it is not the scope of our work since we do not have any ground truth in this dataset.

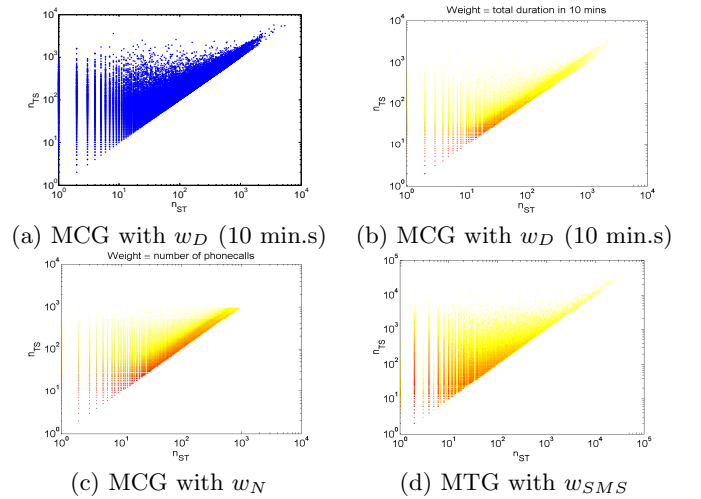


Figure 3: Weights n_{TS} versus n_{ST} on the reciprocal edge pairs in MCG and MTG. In (b), (c) and (d) weights denote the total duration w_D , total number w_N of phonecalls and total number w_{SMS} of SMSs, respectively. Figure best viewed in color.

OBSERVATION 2. *The weights on mutual edge pairs in MCG and MTG are mostly even and small.*

4.1.3 Assortative mixing of degrees

A network is said to show assortative mixing if the nodes that have high degree tend to be connected to other nodes with high degree. Here we study the mixing patterns in MCG and MTG by showing the average degree of neighbors k_{nn} versus its degree k for all the nodes. For a given node i , $k_{nn,i} = 1/k_i \sum_{j \in \mathcal{N}(i)} k_j$ denotes its average neighbor degree where $\mathcal{N}(i)$ is the set of direct neighbors of i . One can also weight the neighbors j of i by the amount of weight w_{ij} of the link between them, that is, $k_{nn,i}^w = 1/s_i \sum_{j \in \mathcal{N}(i)} w_{ij} k_j$ ($s_i = \sum_{j \in \mathcal{N}(i)} w_{ij}$).

We show both $k_{nn,i}$ and $k_{nn,i}^w$ for MTG in Figure 4 (a) and (b), respectively, where weights denote the count of SMSs w_{SMS} and each dot represent a node in the graph. Figure 4 (c) shows the average among the nodes for a given degree.

We notice that the nodes up to degree 20 exhibit disassortative mixing. That is, the higher the degree gets, the lower degree nodes they get linked to on average. On the other hand, nodes with increasing degree k larger than 20 tend to link to nodes with higher degree, pointing to assortative mixing.

We also notice two similar groups of nodes marked as **A** and **B** in Figure 4. These nodes are large hubs with very large degrees that are linked to low degree nodes. In addition, they overlap in Figure 4 (c), which shows that the weights on these links are all equal, and probably small.

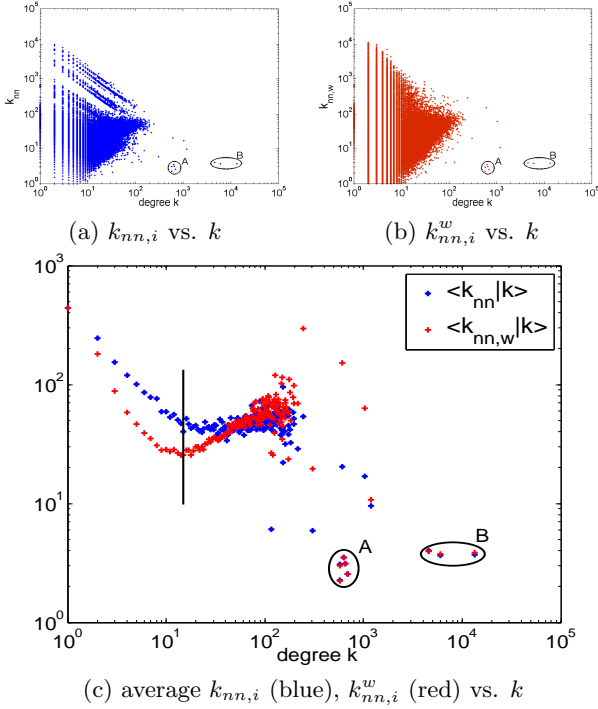


Figure 4: Average neighbor degree (a) $k_{nn,i}$ and (b) $k_{nn,i}^w$ versus degree k for all nodes in MTG. (c) shows the averages among nodes with a given degree. Notice the disassortativity for nodes up to degree 20 (low degree nodes are linked to hubs). Nodes with degree $k \geq 20$ show assortative mixing (they tend to connect to high degree nodes).

The results are similar for MCG: In Figure 5 we observe the assortative mixing of nodes when weights w_N and w_D are considered. This shows that high degree nodes tend to connect to other high degree nodes with stronger ties.

OBSERVATION 3. Degree of a node and average degree of its neighbors have an assortative mixing for nodes of degree $k \gtrsim 20$, i.e. high degree nodes tend to connect to other high degree nodes.

4.1.4 Node Strength w.r.t. Node Degree

Here, we want to understand how the strength s (total weight, $\sum_{j \in \mathcal{N}(i)} w_{ij}$) grows with increasing degree k among nodes in the MCG and MTG. In other words, we study how the amount of time a user spends on the phone is affected by the number of his/her contacts.

In Figure 6 we depict the (from top to bottom) total number of SMSs s_{SMS} , total number of phonecalls s_N , and the

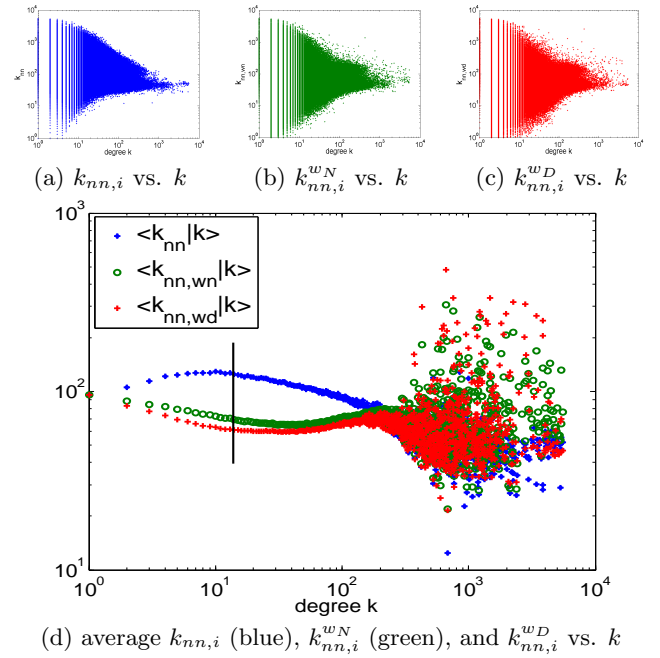


Figure 5: Average neighbor degree (a) $k_{nn,i}$ and (b,c) $k_{nn,i}^w$ versus degree k for all nodes in MCG. (d) shows the averages among nodes with a given degree.

total duration s_D of phonecalls versus the number of contacts (degree k) for each node in MTG and MCG, respectively. In the figures on the left, data is logarithmically binned (vertical dotted lines) and an LS line (red) is fit to the median values (blue circles) obtained for each bin. Similarly, in the right figures an LS line is fit to the median values among all the nodes for each given degree (all figures are in log-log scales). We observe that the fitting lines all have slope greater than 1, which point to a power-law. This indicates that the more number of contacts one has, super-linearly more time s/he spends on the phone.

Note that the blue dots in the right figures which we did not consider in our fitting (also shown with black squares in left figures) correspond to users with extreme behavior. That is, for example in Figure 6(c) they are the customers who spend up to 10 minutes with their more than 150 (a.k.a. the Dunbar's number) contacts each (the gray line depicts $x=y$). Similarly, those points in Figure 6(b) are the ones who exchange only 1 phonecall mutually with their contacts (the dots lie above the $x=y$ line as the mutual MCG is considered and the minimum edge weight w_N in that network is 2).

OBSERVATION 4. Total node strength (number of SMSs, duration of calls) grows super-linearly (power-law) by increasing degree (number of contacts).

4.1.5 Tie Strength w.r.t. Neighborhood Overlap

In this section, we study whether there is a correlation between the strength w_{ij} of the tie e_{ij} between nodes i and j and their neighborhood overlaps. Neighborhood overlap O_{ij} is taken to be the Jaccard coefficient of their common neighbors. That is, $O_{ij} = \frac{|\mathcal{N}(i) \cap \mathcal{N}(j)|}{|\mathcal{N}(i) \cup \mathcal{N}(j)|}$, where $\mathcal{N}(i)$ denotes the neighbor set of i .

Figure 6(left) depicts the average neighborhood overlap O_{ij} among the edges with each given w_{ij} for (from top to

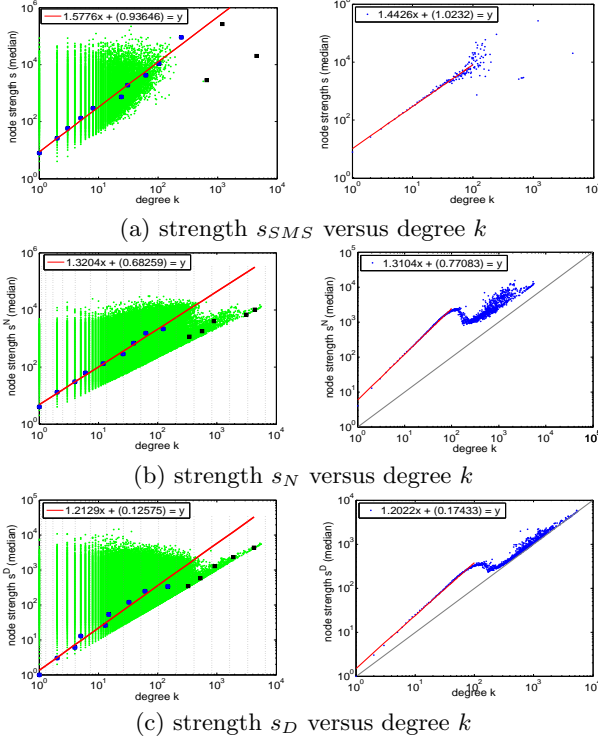


Figure 6: Node strength s (total weight) versus degree k (number of contacts) for (from top to bottom) number of SMSs s_{SMS} in MTG, number s_N , and duration s_D of phonecalls in MCG. Notice the super-linear growth in strength by increasing degree (the more number of contacts one has, even longer time s /he will spend on the phone).

bottom) MTG with weights w_{SMS} , MCG with weights w_N , and MCG with weights w_D (as the number of edges with larger weights is smaller, the averages gets noisy towards higher w_{ij}). Notice that in all three graphs, there exists a positive correlation between the two: tie strength w_{ij} gets larger with increasing neighborhood overlap O_{ij} on average, and vice versa.

To make this observation clearer, we sort all the edges e_{ij} in each graph in increasing order by weight w_{ij} . Then, we take the top α , $0 \leq \alpha \leq 1$, proportion of edges from that list and compute the average neighborhood overlap O_{ij} for the endpoints i and j of those edges. Figure 6(right) shows the average O_{ij} for a given α fraction of the least weighted edges. Again, we observe that the overlap gets higher when more weighted edges get included.

OBSERVATION 5. *Tie strength increases by increasing neighborhood overlap on average.*

4.2 Analysis of Tie Persistence

4.2.1 Tie strengths and persistence in MCG

For the purpose of studying persistence of ties, we took a random sample of nodes preserving the local network structure. The sample has around 5K users and 14.6K links between them. We divided the data into 6 panels of 15 days each spanning over a period of 3 months. We first give the list of notation used throughout text in Table 2.

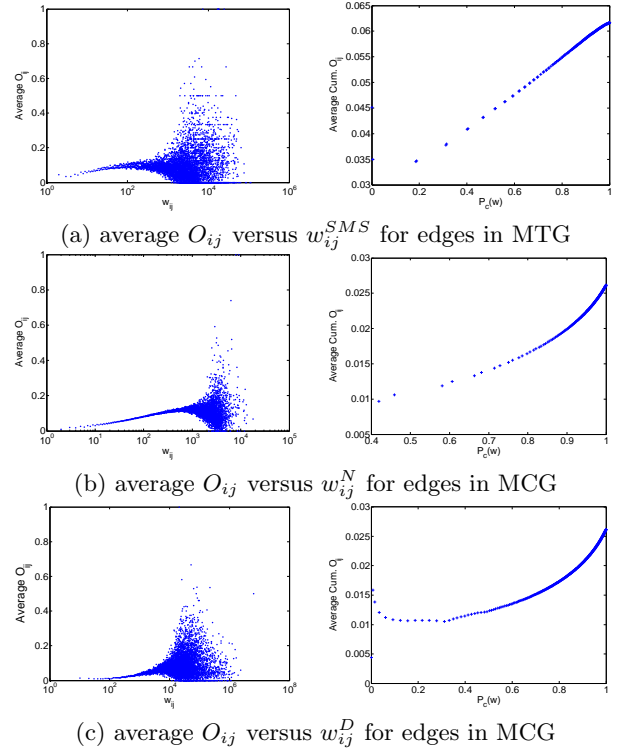


Figure 7: (left) Neighborhood overlap O_{ij} versus tie strength w_{ij} for all pairs (i,j) averaged among all links with a given w_{ij} . (right) Average cumulative neighborhood overlap O_{ij} versus the proportion of the links considered in the list of edge weights w_{ij} sorted in increasing order of weight. The more larger weighted edges are included, the larger the neighborhood overlap gets. Tie strengths represent edge weights as (from top to bottom) number of SMSs w_{SMS} in MTG, number w_N and total duration w_D of phonecalls in MCG.

Symbol	Definition
C_i	Cluster coefficient of user i
K_i	Degree of user i
r	User reciprocity
UP	User persistence
ΔC	Difference in cluster coefficients of two users
ΔK	Difference in degrees of two users
Δr	Difference in reciprocities of two users
R	Reciprocity of a tie
TO	Topological overlap between two users
TP	Tie persistence

Table 2: List of notations used in text.

In Figure 8, we show the distribution of the links in panel 1 w.r.t. their persistence over the 6 panels. Here, we observe that the tie persistence distribution is *bi-modal*, which indicates that the links are either always active (ties persist in all 6 panels) or rarely active (ties persist in only panel 1 and then disappear).

We also studied the influence of tie and node features (as described in Section 3.2) which exploit local network structure of nodes on the tie persistence. Pearson correlation

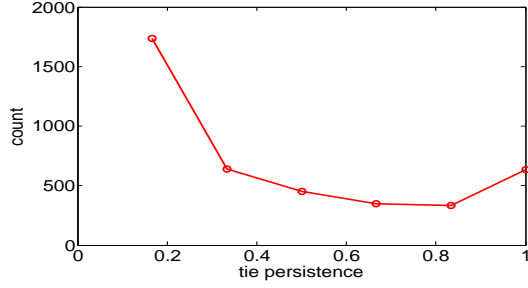


Figure 8: Count vs. Tie persistence

coefficient is widely used to measure dependence between two attributes. The results of our correlation analysis are shown in Table 3. We note that the correlation analysis shows the reciprocity R and topological overlap TO are the most correlated features to tie persistence TP . Also ΔC is negatively correlated to TP . Cluster coefficient indicates how tightly connected neighborhood of a user are. Large difference in cluster coefficients of two users is an indicator that users belong to different neighborhoods and therefore the tie may not persist (hence the negative correlation). Finally, we observe that ΔK and Δr are weakly correlated to TP .

	ΔC	ΔK	Δr	R	TO	TP
ΔC	1.0000	0.2184	0.1274	-	-	-
ΔK		1.0000	-	0.1226	0.0692	0.0019
Δr			1.0000	0.1271	0.4087	0.0735
R				1.0000	0.0835	0.0428
TO					1.0000	0.5064
TP						1.0000

Table 3: Pearson correlation coefficient for tie attributes and tie persistence

We conducted a similar analysis for user perseverance. As can be seen from Table 4, cluster coefficient C and reciprocity r of the users are highly correlated with user perseverance UP . This seems reasonable since high clustering coefficient indicates the user belongs to a tight neighborhood, which means s/he will be in touch with his neighbors and hence will have high perseverance. Also, degree k is weakly correlated with UP , since high degree may indicate some strong ties and many weak ties, resulting in weaker UP .

	C	K	r	UP
C	1.0000	0.0675	0.2740	0.2594
K		1.0000	0.0679	0.0695
r			1.0000	0.3853
UP				1.0000

Table 4: Pearson correlation coefficient for node attributes and user perseverance

4.2.2 Predicting Tie Persistence

Next, we formulate a learning problem on tie persistence. Given the links in panel 1 along with tie and node attributes

(using panel 1 data), we predict whether any link existing in panel 1 will persist in panel 2, 3, 4 etc. Hidalgo et. al. [3] use rule based techniques for this problem. The rule based method predicts that all links which are reciprocal and have topological overlap greater than some threshold will persist in the future. We, on the other hand, learn a logistic regression function to predict whether a tie will persist in future panels. We then compare the performance of the methods using F1 measure computed with 10-fold cross-validation (F1 measure is the harmonic mean of the precision and the recall, hence a better way to evaluate model performance).

In Figure 9, we show the prediction accuracy versus time in number of days for variants of logistic regression using different subsets of features. As we can see LR with *node attributes only* performs very poorly while LR with *tie attributes only* gives better than 0.7 F1 score. We also observe that using tie *as well as* node attributes gives the best performance on average. This indicates that though node attributes are weakly correlated to tie persistence, they together with tie attributes result in better prediction.

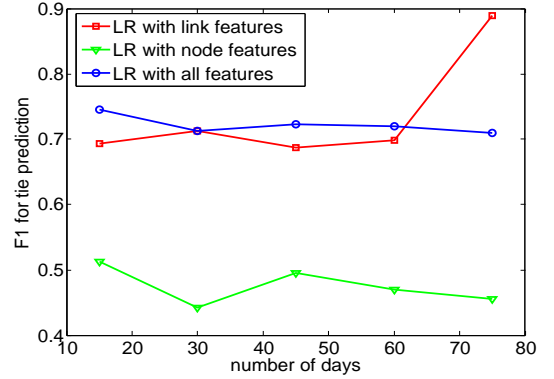


Figure 9: F1 score for predicting tie persistence using different sets of features with Logistic regression

We also compared our method's performance with the rule based method proposed in [3]. For the rule based method, we chose the best rule and threshold that gives the highest accuracy. As can be seen from Figure 10, our LR method with all attributes always outperforms the rule based method.

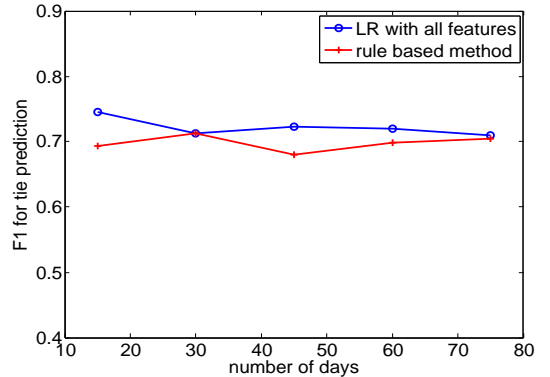


Figure 10: F1 score on tie persistence prediction with Logistic Regression vs. rule based method

4.3 Analysis of Change-Point Detection

4.3.1 Detected Change-Points

After computing the deviation scores Z as was explained in Section 3.4, we use a simple heuristic to flag the high Z scores. Rather than using a threshold value, we simply compute the difference between two consecutive Z scores and rank the time points according to $|Z(t) - Z(t-1)|$. Figure 11 shows the top 10 time ticks for which the difference score is the highest. Here, feature F is taken to be the “inweight”. Experiments with other features such as “number of reciprocal edges” and “outdegree” also flag similar time points which we will discuss later in this section.

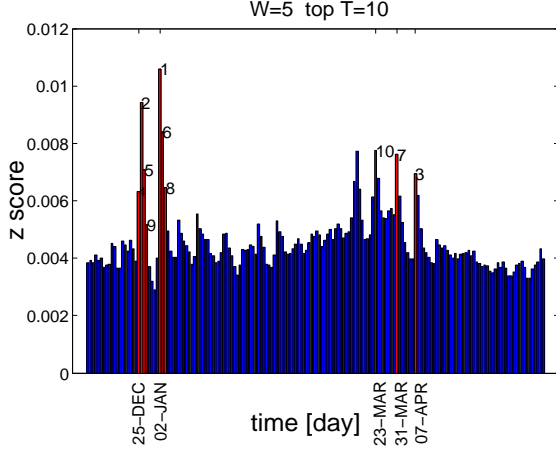


Figure 11: Top 10 time points with highest Z -scores flagged by our method (red bars) for feature F :inweight. Numbers on the bars indicate the rank of each day by the Z -score.

In Figure 11, we observe that the top 2 time periods correspond to the weeks of Christmas and New Year (Dec 26, Jan 2). This shows that even though our data comes from India and mostly people are not Christian, they would be “celebrating” the Christmas. The reason that Jan 2nd rather than Jan 1st is flagged is it shows that it is a change-point in which things went back to normal.

Another surprising finding is with the 3rd time tick which is Apr 7th. Similar to Jan 2nd, this is also a time-point where things turned back to normal. The actual interesting day here is indeed Apr 6th: <http://www.infoplease.com/ipa/A0777465.html> lists Apr 6th as the “Hind New Year” (our data is in 2008). These results suggest that our method is effective in finding points in time for which the collective behavior of the nodes deviate from the recent past.

As a sanity check, we ran our method on other features such as *number of reciprocal edges* and *outdegree*. Figure 12 shows that our method flags almost the same time points including Dec 26th and April 6th also with these features. Moreover, the difference/spike in the Z score is even clearer with these methods. This is intuitive in the sense that even though the “inweight” (number of SMSs received) is expected to increase on days such as Christmas and New Year, the number of reciprocated interactions are expected to increase more (people tend to reply to celebration messages on such days).

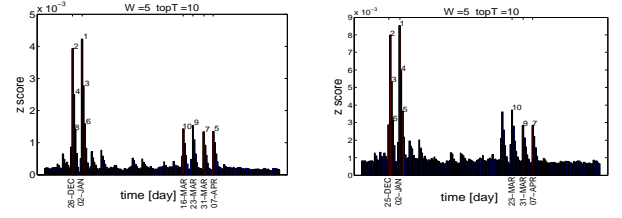


Figure 12: Top 10 time points with highest Z -scores flagged by our method (red bars) for (left) F :number of reciprocal edges and (right) F :outdegree. Notice the flagged time points are similar to those using F :inweight.

4.3.2 Detecting the nodes most effective in change

Here the question is for a given change-point detected in the previous section, can we go back and detect which node(s) contributed to the change the most?

Figure 13 shows the scatter plot of the values of the eigen-scores $u(t)$ versus the typical pattern $r(t-1)$ scores for all the nodes on December 26th. Here, we observe that most of the values lie on the diagonal, which shows that a majority of the nodes did not change much on their typical behavior. On the other hand, some points that are far off-diagonal (marked with red stars) contribute to the Z score the most.

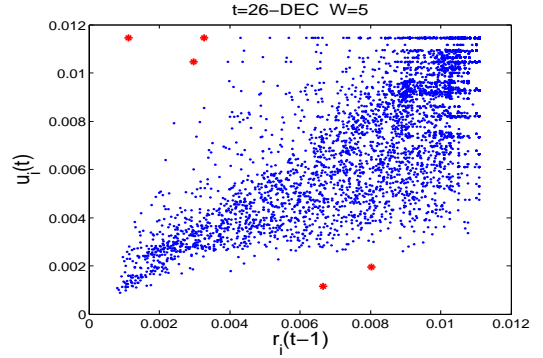


Figure 13: Scatter plot $u(t)$ versus $r(t-1)$ of nodes on December 26th. Each blue dot depicts a node. Nodes far away from the diagonal change in “behavior” the most (top 5 marked with red stars).

Similarly, Figure 14 shows the amount of change ratio $\frac{|u_i(t) - r_i(t-1)|}{r_i(t-1)}$ (%) for 10K nodes (bottom row shows the values in sorted order). Again, the same top 5 nodes as in Figure 13 are marked in red.

Since the data does not contain ground truth of anomalies, in Figure 15 we plot the time series (inweights versus days) of these top 5 nodes marked in Figures 13 and 14 (each row for each node). Here we observe that, three of the nodes (rows 1, 4 and 5) have no activity on the week of Dec 26th. This is flagged because they are observed to have some activity over the previous weeks. On the other hand the other two nodes (rows 2 and 3) have the opposite behavior: they start receiving SMSs during Christmas. We also observe that these two sets of nodes lie in the different halves of the diagonal in Figure 13, also indicating an opposite change in their behaviors.

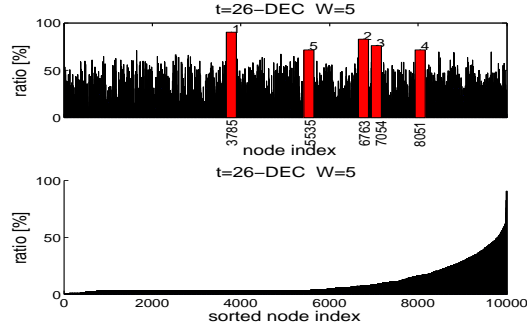


Figure 14: (top) Change ratios (%) of top 10K nodes in $u(t)$ and $r(t-1)$. Each bar depicts a node (top 5 with highest change ratio is shown in red). (bottom) Ratio values sorted.

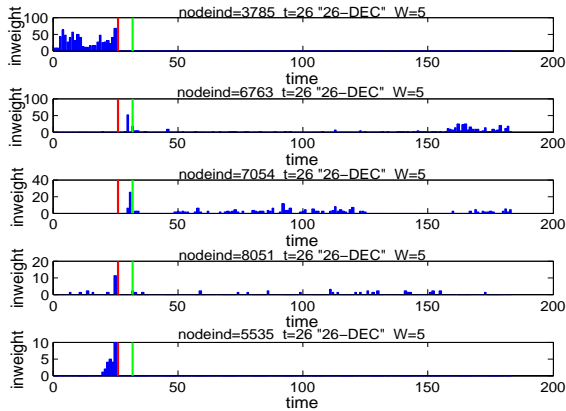


Figure 15: Time series of inweight values of top 5 nodes with highest deviation in “eigenbehavior” marked in Figures 13 and 14. Beginning and end of week December 26th is marked with red and green vertical bars on the time line, respectively.

5. CONCLUSIONS

In this paper we studied a large network of mobile phone users that grows and changes over time. Our study consists of three major parts: (1) Analysis of structural properties; (2) Prediction of tie persistence; and (3) Event and change-point detection. Our findings and conclusions can be summarized as follows:

1. Structure analysis

- The SMS network contains far less reciprocal ties than the phonecall network, pointing to spam messengers. Also, the weights (total duration, number of phonecalls and SMSs) on reciprocated ties are usually even and small in both networks. Reciprocity patterns can be used to spot outliers as users with low reciprocity (many non-reciprocated links) and pairs with unbalanced reciprocity (i calls j far more than j calls i) are suspicious.
- Degree of a node and average degree of its neighbors exhibit assortative mixing on average. In other words, users with high number of contacts tend to connect to other users with high number of contacts.

- Total node strength grows super-linearly (power-law) by increasing degree. That is, the more number of contacts users have indicates that super-linearly more amount of time they will spend on the phone.
- Tie strength between a given pair of users increases by increasing neighborhood overlap on average. In other words, users who have more common neighbors tend to exhibit stronger ties, i.e. spend more time communicating on the phone.

2. Tie persistence prediction

- Local network attributes such as clustering coefficient and tie attributes such as reciprocity help to predict whether ties will persist in the future.
- Our prediction results using logistic regression show that tie attributes give better accuracy than node attributes and both types of attributes yields the best prediction accuracy.
- Regression techniques give better accuracy than rule based techniques.

3. Event detection

- We used an “eigenbehavior”-based method on the time-series of users and considered the amount of change in their “eigenbehaviors” to flag change-points in time.
- Realistic anomaly detection is difficult with unlabeled data, but our results have demonstrated that we were able to detect events that coincide with major holidays and festivals in our data.
- Our method can also be reverse-engineered to spot the top users who contribute to the changes the most.

6. REFERENCES

- [1] N. Eagle. Behavioral inference across cultures: Using telephones as a cultural lens. *IEEE Intelligent Systems*, 23:62–64, 2008.
- [2] N. Eagle, A. Pentland, and D. Lazer. Inferring social network structure using mobile phone data. *Proceedings of the National Academy of Sciences (PNAS)*, 106:15274–15278, 2009.
- [3] C. A. Hidalgo and C. Rodriguez-Sickert. The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387, 2008.
- [4] T. IDÉ and H. KASHIMA. Eigenspace-based anomaly detection in computer systems. In *KDD*, pages 440–449, 2004.
- [5] A. A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, and A. Joshi. On the structural properties of massive telecom call graphs: findings and implications. In *CIKM*, pages 435–444, 2006.
- [6] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, M. A. de Menezes, K. Kaski, A.-L. Barabasi, and J. Kertesz. Analysis of a large-scale weighted network of one-to-one human communication, 2007.
- [7] J. P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, and A. L. Barabasi. Structure and tie strengths in mobile communication networks, 2006.
- [8] S. Papadimitriou, J. Sun, and C. Faloutsos. Streaming pattern discovery in multiple time-series. In *VLDB*, pages 697–708, 2005.
- [9] J. Sun, C. Faloutsos, S. Papadimitriou, and P. S. Yu. Graphscope: parameter-free mining of large time-evolving graphs. In *KDD*, pages 687–696, 2007.
- [10] Q. Ye, B. Wu, L. Suo, T. Zhu, C. Han, and B. Wang. Telecomvis: Exploring temporal communities in telecom networks. In *ECML PKDD*, pages 755–758, 2009.