The Impact of Social Affinity on Phone Calling Patterns: Categorizing Social Ties from Call Data Records

Sara Motahari
Sprint Labs
Burlingame, CA 94010
Sara.Gatmir-Motahari@sprint.com

Ole J. Mengshoel
Carnegie Mellon University
Moffett Field, CA 94035
Ole.Mengshoel@sv.cmu.edu

Phyllis Reuther
Sprint Labs
Burlingame, CA 94010
Phyllis.Reuther@sprint.com

Sandeep Appala
Carnegie Mellon University
Moffett Field, CA 94035
Sandeep.Appala@west.cmu.edu

Luca Zoia
Carnegie Mellon University
Moffett Field, CA 94035
Luca.Zoia@west.cmu.edu

Jay Shah
Carnegie Mellon University
Moffett Field, CA 94035
Jay.Shah@west.cmu.edu

ABSTRACT

Social ties defined by phone calls made between people can be grouped to various affinity networks, such as family members, utility network, friends, coworkers, etc. An understanding of call behavior within each social affinity network and the ability to infer the type of a social tie from call patterns is invaluable for various industrial purposes. For example, the telecom industry can use such information for consumer retention, targeted advertising, and customized services. In this paper, we analyze the patterns of 4.3 million phone call data records produced by 360,000 subscribers from two California cities. Our findings can be summarized as follows. We reveal significant differences among different affinity networks in terms of different call attributes. For example, members within the family network generate the highest average number of calls. Despite the differences between the two cities, for a given affinity network they show similar phone call behaviors. We identify specific features that model statistically meaningful changes in call patterns and can be used for prediction and classification of affinity networks, and we also find correlations between the features associated with call behavior. For example, when subscribers call each other after a long time, their calls tend to take longer. This knowledge leads to discussions of proper machine learning classification approaches as well as promising applications in telecom and security.

Categories and Subject Descriptors

G.3 [**Probability and Statistics**]: Statistical Computing; J.4 [**Social and Behavioral Sciences**]: Sociology.

General Terms

Measurement, Experimentation.

Keywords

Call Data, Social Networks, Call Behavior, Social Tie Inference

The 6th SNA-KDD Workshop '12 (SNA-KDD'12), August 12, 2012, Beijing, China.

1. INTRODUCTION

Today's pervasive use of mobile phones produces huge volumes of call data records that can potentially provide significant amount of valuable information about various patterns of human interaction, social relations, mobility, etc. In particular, being able to understand, distinguish and discover social affinities and their nature from people's call data records are invaluable in many practical areas, such as planning advanced marketing strategies, dynamic pricing, customized services, and recommendation services. Telecom providers can use this information for customer retention and targeted advertising. For example, wireless service providers can identify the family members that are not under the same family plan as potential customers, or identify subscribers within a network of non-subscriber friends to predict and prevent churn. The impact of social ties on costumer attraction and retention is already known to telecom providers. For example, previous studies have shown that the number of customers who churn out of a service provider's network depends on the number of their friends that have already churned [1]. The strength of a social tie is also typically proportional to degree of trust [26], and trust plays a crucial role in security. Consequently, this line of research may also have applications to security.

Despite the importance of such information about social relations and call behavior, there is a major gap in previous studies when it comes to the analysis of social network based on phone calls. Previous work has mainly focused on the structure of the obtained social graphs and the communities inside them, such as topological properties [1] [12][8][13]. Some previous studies have also tried to predict the existence of links and social connections [8][3][14]. Such analyses, while essential to our understanding of social networks and call graphs in general, do not provide any detailed insights on how to characterize, categorizes, or classify the type of a social tie from calling patterns.

In this paper, we analyze a set of features that abstract calling patterns between subscribers, and investigate their ability to discriminate between different affinity networks. We collect and process Call Detail Records (CDRs) of a major wireless service provider from two different cities in California: a "rural city" (Modesto) with a relatively small population situated in an agricultural region and an "urban city" (San Francisco) with a larger and more diverse population, located close to Silicon Valley and two other cities (Oakland and San Jose). We assign each subscriber into various types of social ties (see also [21]

[23]) such as family, toll free, utility services or other based on our access to information sources about their registration accounts. We define each social group as an "affinity network." In other words, according to our terminology an affinity network is a group of subscribers in which all the subscribers have a common social tie.

We then identify and calculate various features that model call behavior and patterns such as the frequency, length, timing, and symmetry of calls. The results show that the affinity networks show meaningful and statistically significant differences in terms of their identified features. For example, family members call each other more often, their calls are shorter, and they have a more mutual (outgoing versus incoming) tie compared to other affinity networks. We will see that these distinctive patterns are consistent across the two geographical areas of Modesto and San Francisco despite the differences in population characteristics between the two cities. For example, in Modesto 39.3% of households contain children under the age of 18, while this is the case for 18.4% of households in San Francisco.

We highlight some of the underlying correlations between these features. For example, when subscribers call each other after a long time, they tend to make longer calls. All these finding have implications on predictability and classifiability of social ties, as well as which machine learning techniques to be fruitfully used, as we also discuss in this paper.

After a brief review of previous work in Section 2, we will describe the dataset that was used in this study in Section 3. Section 4 explains our methodology, and Section 5 summarizes the empirical results. Highlights of the results and their implications are discussed in Section 6.

2. SOCIAL NETWORK AND CALL DATA ANALYSIS

Social networks have been analyzed from many perspectives. A number of recent studies have specifically used mobile call graph data to investigate and characterize the social interactions of subscribers [1][12], the evolution of social groups and the adoption of new products and services [6].

The main focus of previous work on social networks (whether from phone calls or other sources of data) has been on the structure of the obtained social graphs, including topological properties, degree distributions, core clusters, strongly connected components, extraction of communities, and community structure identification [1][12][8][13]. In particular, graph partitioning, such as spectral clustering is a popular approach for studying community structure in graphs [16][18][19][21].

Several previous studies have tried to predict the existence of social connections in different contexts, for example, by considering mutual connections on Facebook and social networking sites, or by considering proximity patterns on university campuses and other specific environments [8][3][14]. This issue has also been explored in the form of predicting missing and future links in co-authorship networks [5], phone call graphs [5][4], and simulated social graphs.

The issues of node (or vertex) partitioning [2] and, more recently, link (or edge) partitioning [21][23] have also been studied in network science. When node partitioning is performed, one assigns each node to a partition or class. For edge partitioning, each edge is assigned to a partition or class. The benefit of edge partitioning compared to node partitioning is that it can model situations in which nodes do not neatly separate into disjoint, non-overlapping classes.

As we mentioned, such studies have been essential to our understanding of social networks. However they do not give us information about call patterns with respect to the type of social relations, and do not enable us to infer the affinity networks. This is the focus of this paper. Specifically, we focus on the characteristics of the edges of a social network induced by CDRs instead of focusing on node characteristics as done in most previous work.

3. DATA SET

The data set we have used in our analysis contains CDRs of mobile subscribers in the cities of Modesto and San Francisco, in the state of California, as shown in Figure 1. The geographic locations of the base stations in downtown Modesto and San Francisco were taken as reference points. From each reference point, all the regions within a radius of 20 miles were covered and the data from their base stations were collected.

The CDRs were collected by the telecommunication service provider Sprint. Our data collection methodology resulted in millions of phone calls between subscribers in these two cities from 30 consecutive days in the month of October, 2011.

3.1 Call Detail Records

CDRs are collected from various base stations and stored in a distributed file system warehouse. Each cell phone call of a subscriber is saved as a record that contains the following information:

- Unique subscriber IDs and phone number of the caller;
- Unique subscriber IDs and phone number of the call receiver;
- Date and time of the call's initiation;
- Date and time of the call's end;
- Direction of the call (outgoing versus incoming);
- Switch ID;
- Cell tower ID;
- Sector ID.

From the CDRs, we constructed, as further discussed in Section 4, a social network that represents each mobile user as a vertex and the calls between them as an edge.

3.2 Population Characteristics

Modesto and San Francisco are rural and urban cities respectively. They have different population size and demographics. Below, we summarize some of their characteristics.[1]

3.2.1 Modesto

Modesto has a population of about 200,000 and the population density is 5,423.4 people per square mile (2,094.0/km²). Ninety eight percent of the population lives in a household. There are about 69,000 households, out of which 39.3% (~27,000) have children under the age of 18 living in the house; 48.1% (~33,000) are married couples living together; the rest is a householder (male or female) living alone. The average family size is 3.38. The population is spread out with 26.8% under the age of 18 (~54,000), 10.4% of the population aged 18 to 24 (~20,000); 26.4% of the population aged 25 to 44 (~53,000); 24.7% aged 45 to 64 (~49,000) and 11.7% of the population is 65 years of age or older (~23,000). The median age is 34.2 years. For every 100

[1] Our source is the United States Census for the year 2010.

females there are 95.0 males. For every 100 females age 18 and over, there are 91.5 males.

3.2.2 San Francisco

San Francisco has a population around 805,000 and the population density is 17,160 per square mile (6,632/km²). Ninty seven percent of the population lives in a household. There are about 345,000 households, out of which 18.4% (around 63,000) have children under the age of 18 living in them; 31.6% (around 109,000) are married couples living together; the rest is a householder (male or female) without another present. The average family size is 3.11. The population is spread out with 13.4% under the age of 18 (~107,000), 9.6% of the population aged 18 to 24 (~77,000), 37.5% of the population aged 25 to 44 (~301,000), 25.9% of the population aged 45 to 64 (~301,000), and 13.6% are 65 years of age or older (~109,000). The median age is 38.5 years. For every 100 females there are 102.9 males. For every 100 females age 18 and over, there are 102.8 males.

Figure 1. Areas of interest: San Francisco and Modesto

4. METHODOLOGY

In this section, we explain our methodology of processing and analyzing the CDRs as well as identifying features.

4.1 Anonymization

Aside from the security measures and control of access to subscriber information, the CDRs were anonymized as the originating and destination phone numbers and identification numbers were encrypted using hashing. Hence, no personal identification was accessible. The subscriber numbers tied to a family plan account were also anonymized using hashes. Also, all our results are presented as aggregates and calculated for the overall demographic population, and no individual subscriber is pinpointed for the study.

4.2 Identification of Affinity Networks

As we mentioned, a social affinity network may be a network of friends, family members, coworkers, etc. Obviously, the information about the relation types between subscribers is not provided in the CDRs. Therefore, we had to use outside information to identify the social ties without compromising the privacy of individual and personal accounts. We therefore grouped the social ties into the following affinity types:

Family: for every primary subscriber, other subscribers belonging to the primary subscriber's family plan were identified. If any two subscribers were part of a common family accounts in a family plan, they were identified as family members.

Toll: the numbers starting with 1-800 were categorized as Toll free numbers.

Utility:[2] we considered a pair of subscribers to be part of a utility network if one of the subscribers is a business establishment in Modesto or San Francisco. We created a limited but representative list of business establishments by scraping the Web.

Others: all the other numbers which may include friends, professional colleagues, or any other personal relationship that a subscriber could hold.

Above, we discuss "toll free numbers" (i.e. nodes), but in Table 1 we present "toll free edges." This reflects the fact that the node partitioning into {Family, Toll, Utility, Others}, as introduced above, induces a corresponding edge partitioning. In other words, for each of the affinity edge types, we have a rule that says how it is defined in terms of the affinity node types. For example, if one node is a toll node, then all adjacent (both incoming and outgoing edges) are toll edges. By considering all the edges of one affinity type, in our case {Family, Toll, Utility, Other}, we obtain an affinity network.

We could potentially identify additional affinity networks (e.g. grouping coworkers by accessing email domain and corporate accounts information). However, the use of more information from Sprint or other external sources could compromise the privacy and confidentiality of the subscribers. Therefore, we limited the identified networks to the above ones. Nevertheless, we will see below that the above grouping is enough for pinpointing and analyzing the call patterns and identifying meaningful features, which is the purpose of this study.

4.3 Sampling

We extracted CDRs for a set of subscribers from San Francisco and Modesto and ranked subscribers by their total number of calls. We then removed the outliers (top 10% and bottom 10%) and uniformly sampled 10,000 subscribers and their one-hop connections from this set. We then filtered all the CDRs from the two locations for this set of subscribers. The number of subscriber pairs in each edge partition (affinity network type) from each city is shown in Table 1. The table excludes self-edges, which may be used to reflect calls to voice-mail.

Table 1. Statistics for call data record (CDR) data sets

	Modesto	San Francisco
Number of call data records in the sample	1,966,022 (~1.97 million calls)	2,333,826 (~2.33 million calls)
Number of subscribers including primary and one-hop nodes	203,864	249,591
Number of social ties (edges)	304,053	350,236
Number of family member edges	534	4,220
Number of utility edges	25,232	4,111
Number of toll free edges	19,528	27,618
Number of 'other' edges	258,759	324,287

4.4 Visualization and Feature Identification

We used a visualization tool for getting an intuitive understanding of the characteristics of different affinity networks. The visualization software NetEx [24] shows the vertices and the edges of the social graph using multiple GUI elements and

[2] http://www.yellowpages.com/modesto-ca

configurations. For example, as seen in Figure 2, the thickness of the edges can be varied based on the average number of calls.

Such visualizations enabled us to identify and consider various features that may distinguish between affinity networks. In particular, we observed two categories of features:

1. Graph topological features: For example, in Figure 2(a), nodes 23, 24, 28, and 32 belong to the same family plan and have several mutual contacts. Figure 2(a) also shows a random set of three utility numbers and their one-hop neighbors. The topology resembles an ego centric network where the calls occur between the utility service subscriber and all the subscriber's one-hop neighbors, but no calls occur between their one-hop neighbors. In addition, the utility numbers picked do not have any mutual contacts.
2. Call pattern features: These features relate to the number, frequency, length, and timing of the calls. For example, when the thickness of the line is selected to represent the number of calls, we observe thicker edges between many family members in Figure 2.

Here is how Figure 2(a) was created. All of the nodes belonging to a family plan were selected, and the number of mutual contacts between them was calculated by looking at their individual one-hop neighbors. For this visualization, a set of family nodes with high numbers of mutual contacts and total calls between them were chosen. However, we observed similar graph structures and features across other family sub-networks as well.

The above discussion illustrates the identification of potential candidates for the features that model the differences between call graphs of different affinity networks. We now turn to the discussion of the definition and discussion of the features that we picked for the purpose of this paper.

4.5 Definition of the Features

Before introducing the features that were used, we define how call records and graphs were defined and constructed, respectively.

Definition. A call record represents a call and can be defined as a tuple $\theta = (u, v, t, d)$, where u and v are subscribers, t is the call start time, and d is the call duration. Note that this is a subset of a CDR as introduced in Section 3.1. The order of u and v in θ is important: u is the subscriber initiating the call while v receives the call. A set of call records is denoted Θ.

Among calls between u and v, the outgoing calls from u are defined as

$$\Theta_u(u,v) = \{(u, v, t, d)\ \{|(u, v, t, d) \in \Theta\}\},$$

while the outgoing calls from v are defined as:

$$\Theta_v(u,v) = \{(u, v, t, d)\ \{|(v, u, t, d) \in \Theta\}\}.$$

A call record relation with all calls between u and v can now be defined as $\Theta(u,v) = \Theta_u(u,v) \cup \Theta_v(u,v)$. Below, in our features, we are generally using $\Theta(u,v)$ since call direction generally does not matter except in one case (the engagement feature) where we use both $\Theta_v(u,v)$ and $\Theta_u(u,v)$.

Let $G = (V, E)$ be an undirected graph where V represents the set of all subscribers in our sample Θ, and E represents the set of all undirected edges $\{u, v\}$, with $|\Theta(u,v)| > 1$. We define a social network N that consists of G along with a set of features Φ defined for each edge; $N = (G,\Phi)$. A key point in this paper is that we are, similar to previous work on link partitions [21] and link communities [23], focusing on features defined on edges E rather than features defined on nodes N.

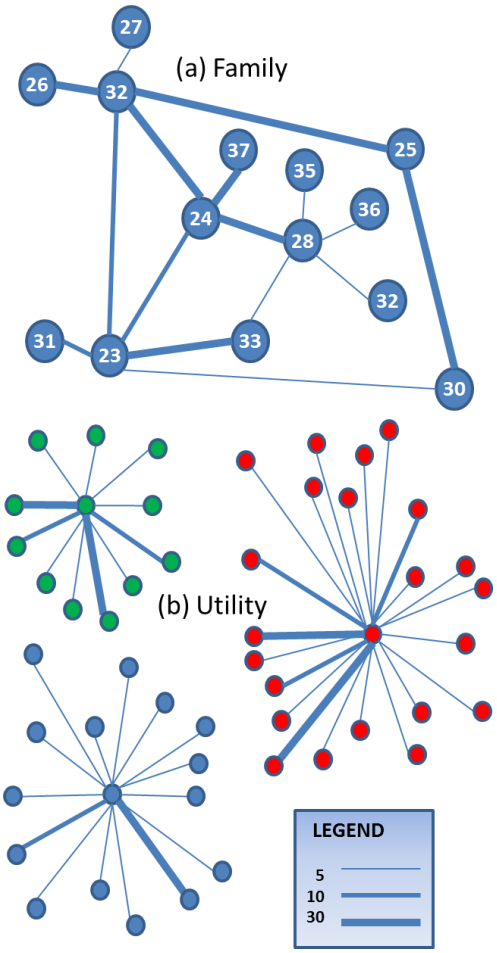

Figure 2. Average total call feature: (a) Family and (b) Utility affinity sub-networks plus one-hop neighbors

4.5.1 Features

Our features are defined for a given pair of subscribers, $\{u, v\} \in E$, and our goal is to characterize the nature of the social tie between u and v through the distribution of these features.

The *total number of calls* ϕ_{nc} between subscribers u and v is defined as $\phi_{nc} = |\Theta(u,v)|$.

The *average call duration* ϕ_{acd} between u and v is defined as: $\phi_{acd}(u,v) = \sum_{(u,v,t,d)\in\Theta(u,v)} d / |\Theta(u,v)|$.

Given a pair of subscribers u and v, we consider the calls between them, arranged in ascending order by their start times. Now, consider for $i \geq 1$ any two consecutive call records (u_i,v_i,t_i,d_i) and $(u_{i+1},v_{i+1},t_{i+1},d_{i+1})$ from $\Theta(u,v)$, and define $n = |\Theta(u,v)|$. The *average inter-call interval* ϕ_{aii} is then defined as:

$$\phi_{aii}(u,v) = \left(\sum_{i=1}^{n-1} (t_{i+1} - t_i)\right)/(n-1).$$

The following feature is based on partitioning $\Theta(u,v)$ into weekend calls $\Theta_{we}(u,v)$ and weekday calls $\Theta_{wd}(u,v)$ such that $\Theta_{wd}(u,v) \cup \Theta_{we}(u,v) = \Theta(u,v)$ and $\Theta_{wd}(u,v) \cap \Theta_{we}(u,v) = \varnothing$.

For the purpose of this paper, we treat Saturday and Sunday as weekend days, and other days as weekdays. The asymmetry between weekend and weekday calls, or *weekday to weekend asymmetry* ϕ_{ww}, is defined with respect to the total number of calls:

$$\phi_{ww}(u,v) = \frac{\| \Theta_{wd}(u,v) | - | \Theta_{we}(u,v) \|}{| \Theta(u,v) |}.$$

If $\phi_{ww}(u,v) = 1$ there is maximal asymmetry, where calls take place on weekends or weekdays only. If $\phi_{ww}(u,v) = 0$, this means minimal asymmetry, with calls evenly distributed.

The *engagement ratio* ϕ_{eg} feature gives a measure of the social interaction between *u* and *v*.

$$\phi_{eg}(u,v) = 1 - \left| \frac{| \Theta_u(u,v) | - | \Theta_v(u,v) |}{\phi_{\mathrm{nc}}(u,v)} \right|$$

The purpose of this feature is to characterize the interaction patterns between subscribers *u* and *v*. If the ratio of engagement is $\phi_{eg}(u,v) = 1$, there is a symmetric interaction between the pair of subscribers. If $\phi_{eg}(u,v) = 0$, there are either only outgoing calls to *v* (from *u*) or incoming calls from *v* (to *u*).

After we calculated features for all edges, we compared the features between different types of affinity networks and for different geographic regions (Modesto and San Francisco). We also explored the underlying correlations between the features. Statistically significant results are presented in Section 5.

4.6 Software Tools and Techniques

As the CDRs for subscribers typically amount to massive data sets, it can be computationally intensive to construct features from them. We used Hadoop, see http://hadoop.apache.org/, which implements the MapReduce parallel computing model [25] in order to process CDRs. Specifically, CDRs were stored in a Hadoop distributed file system and processed by feature construction algorithms written using the MapReduce framework.

Figure 3 illustrates how we used the MapReduce mappers and reducers for our purpose. For every CDR, a mapper function emits a *(key,value)* pair, in our case a pair for each pair of subscribers, and the reducer function performs the aggregation operation needed for the desired features. In other words, the key emitted by the mapper application is the subscriber pair (i.e., the source subscriber phone number and the destination subscriber phone number), the direction of the call (incoming or outgoing), and so forth. For every key (subscriber pair), the corresponding values from the CDR are emitted as a value.

Once the information had been calculated using Hadoop, SQL scripts were written to mine the information accordingly in the database. The Java-based NetEx software [24] was interfaced with the database to present data for visualizations.

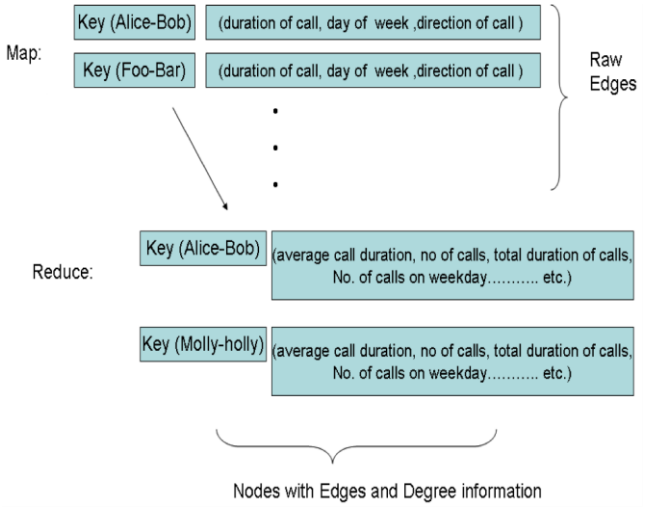

Figure 3. Call data record (CDR) processing using MapReduce and Hadoop

5. RESULTS

We analyzed the call behavior with respect to the above features from several perspectives: 1) the behavioral differences among different affinity networks which is reflected in meaningful variations of their associated features; 2) dependencies between various characteristics of human call behavior; and 3) differences between geographic regions.

5.1 Differences between Affinity Networks

We here present the results of analyzing and comparing the above features with respect to different affinity networks in the two cities. We observed different call patterns reflected in various features as explained in the following subsections.

5.1.1 Total Number of Calls

Most of the subscribers in a family network have very high number of calls ϕ_{nc} between each other. The average value of total number of calls made is the highest between family members ($\phi_{nc} = 38.98$ calls in a month) in Modesto, whereas the average value of the total calls made to utility and toll free numbers have relatively low values of $\phi_{nc} = 2.48$ and $\phi_{nc} = 2.29$ respectively. Analysis of variance shows a significant difference between the family members and the others [Fisher's Score, $F = 2946$, p (p-value for statistical significance) < 0.001].

The average value of total number of calls made to family members in San Francisco is high, too, ($\phi_{nc} = 31.44$) when compared to toll free and utility service numbers which are $\phi_{nc} = 2.26$ and $\phi_{nc} = 2.18$. The difference of the means between family members and others is again statistically significant ($F = 20931$, $p < 0.001$). No meaningful difference was found in the number of family member calls between the two cities. Figure 4 shows the mean value of total number of calls made within different affinity networks in the Modesto and San Francisco regions.

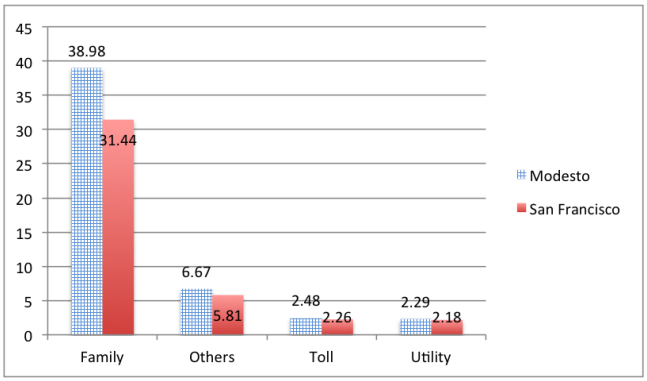

Figure 4. Average number of calls

5.1.2 *Engagement Ratio*

The engagement ratio ϕ_{eg} defines how reciprocal the call pattern between two subscribers is. The higher the engagement ratio, the more balance we observe in the number of outgoing and incoming calls. The average engagement ratios calculated for various affinity networks are shown in Figure 5. The family network has a high engagement ratio of ϕ_{eg} = 0.57 whereas the toll free and utility services have very low values of ϕ_{eg} = 0.05 and ϕ_{eg} = 0.14, respectively, in Modesto. The engagement ratio in San Francisco also shows a similar value for the family network, namely ϕ_{eg} = 0.56, and low values of ϕ_{eg} = 0.03 and ϕ_{eg} = 0.07 respectively for the toll free and utility service networks. An analysis of variance confirms a statistically meaningful difference between family members and other affinity networks (F = *27.3* and p < 0.001 for Modesto, and F = 480 and p < 0.001 for San Francisco).

The above results suggest that family subscribers have a reciprocal social interaction with each other. In contrast, mostly outgoing calls are made to utility and toll free numbers; utility and toll free numbers do not reciprocate equally.

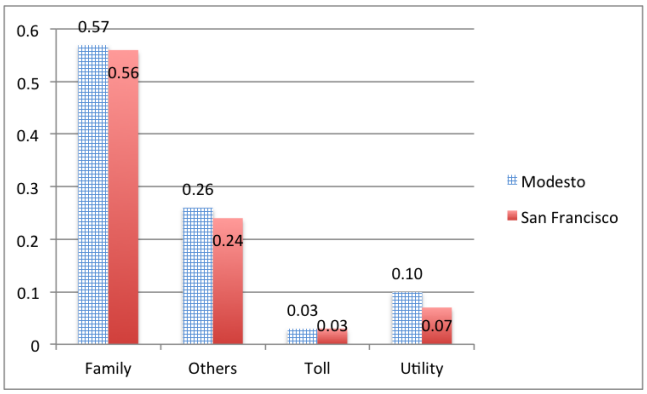

Figure 5. Average engagement ratio

5.1.3 *Weekend versus Weekday*

Figure 6 shows the percentage of the calls that were made on the weekends and Figure 7 shows the asymmetry between weekday and weekend calls for each affinity network. The main difference, perhaps, is observed in the asymmetry of weekday to weekend calls between family members compared to the rest (others, toll, and utility). Toll free and utility numbers are called much more during the weekdays. Family members are more likely to call each other on the weekends (F = 56.7 and p < 0.001 for Modesto, and F =350, p < 0.001 for San Francisco).

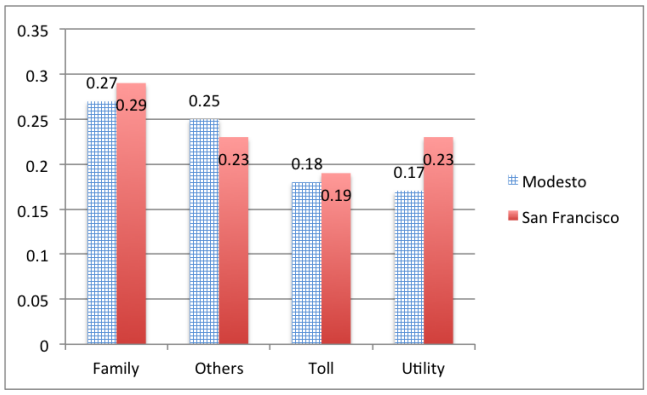

Figure 6. Average weekend percentage of the calls

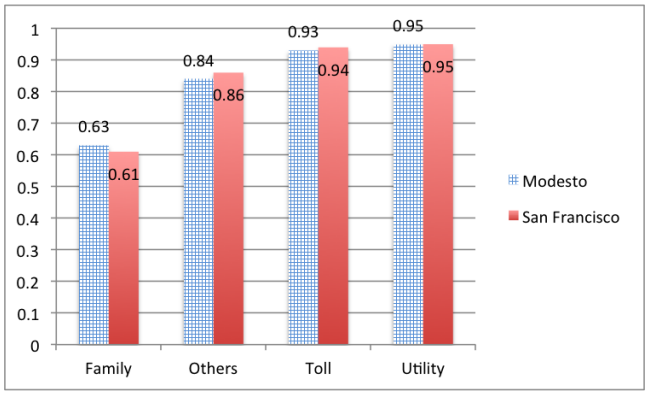

Figure 7. Average weekday to weekend asymmetry

5.1.4 *Call Duration*

Figure 8 reports empirical results for average call duration. The calls occurring between pairs of subscribers in a family network have small average call duration ϕ_{acd} = 40.02 and ϕ_{acd} = 39.97 seconds, respectively, both in Modesto and San Francisco. The utility services and toll free networks have higher average call durations ϕ_{acd} = 87.78 and ϕ_{acd} = 239.82 seconds in Modesto, and ϕ_{acd} = 104.91 and ϕ_{acd} = 261.17 seconds in San Francisco, respectively. Running an Anova analysis, we also observed a statistically meaningful difference between the call duration of the three groups of 1) family members, 2) toll-free/utility network, and 3) others (F = 8.2 and p = 0.004 for Modesto, and F = 1302 and p < 0.001 for San Francisco). This suggests that the calls between family members are quicker and contain short conversations, whereas calls to business establishments and toll-free numbers last for a substantially longer time. The mean values are depicted in Figure 8.

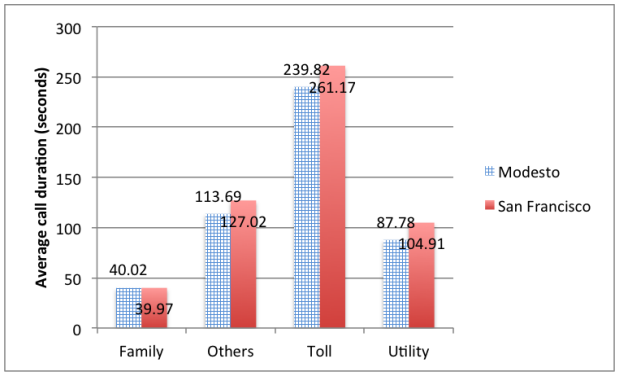

Figure 8. Average call duration

5.1.5 Inter-call Interval

The calls occurring between subscribers in a family network have a smaller average inter-call interval ϕ_{aii} of $\phi_{aii} = 3272$ minutes (~2 days) in Modesto and $\phi_{aii} = 3586$ minutes (~ 2.5 days) in San Francisco when compared to other affinity networks. The utility service networks have the highest inter-call interval of ϕ_{aii} = 8361 minutes (~6 days) and ϕ_{aii} =8589 (~7 days) in Modesto and San Francisco respectively, i.e. they occur very sporadically. This suggests that the subscribers belonging to a family network call each other more frequently compared to how often they call their utility or toll free numbers. The significance of the results were confirmed by an Anova analysis (F = 25.5, p < 0.001 for Modesto, and F = 458 and p < 0.001 for San Francisco). The mean values for the inter-call duration for each affinity network are plotted in Figure 9.

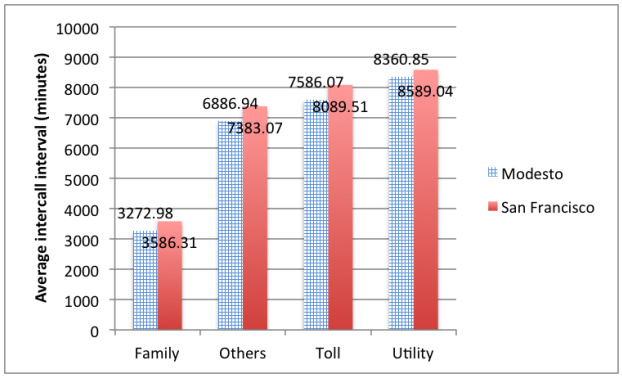

Figure 9. Average inter-call interval

In the next subsection, we will analyze the dependencies between the above features and in Section 6 we will talk about the meanings and implications of the results.

5.2 Correlations between Features

When it comes to feature selection, feature extraction, and the selection of the appropriate machine learning and data mining techniques for classification, it is important to be aware of the correlations among the features. The features that we analyzed in the previous subsection are obviously not independent. We looked at the correlation matrix and investigated the statistically meaningful correlations. The following pairs of features were found to have significant correlations:

- The total number of calls ϕ_{nc} and the average inter-call interval ϕ_{aii} (ρ=-0.48): This correlation is relatively obvious and is due to the definition of these features. For the pairs with more calls during a given time interval, there needs to be more frequent calls (i.e., shorter time intervals).
- The total number of calls ϕ_{nc} and the weekday to weekend asymmetry ϕ_{ww} (ρ=-0.31): The subscriber pairs that make more calls to each other also make more calls on the weekends (meaning a smaller asymmetry ϕ_{ww}). This could be because family members make more calls on the weekends. We talk more about the cause versus effect issue in Section 6, as this discussion applies to most correlations found.
- The average call duration ϕ_{acd} and average inter-call arrival time ϕ_{aii} (ρ=0.21): people who call each other less frequently make longer calls. This effect is illustrated in Figure 10 and Figure 11.
- The total number of calls ϕ_{nc} and the engagement ratio ϕ_{eg} (ρ=0.34): The subscriber pairs that make more calls have a more reciprocal call pattern.
- The weekday to weekend asymmetry ϕ_{ww} and the engagement ratio ϕ_{eg}(ρ=-0.36): The subscriber pairs with more calls on the weekends have more reciprocal call patterns.

The significance of the above correlations also holds for each city separately and is consistent across the two geographical regions. For example, Figures 10 and 11 show the average inter-call interval ϕ_{aii} versus the average call duration ϕ_{acd} in Modesto and San Francisco respectively. We see that in both cities, the calls take longer as the time interval between them increases.

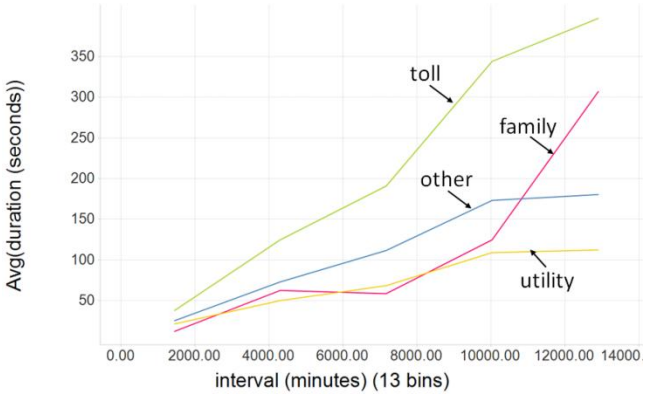

Figure 10. Average call duration (*y*-axis) as a function of inter-call interval (*x*-axis) in Modesto

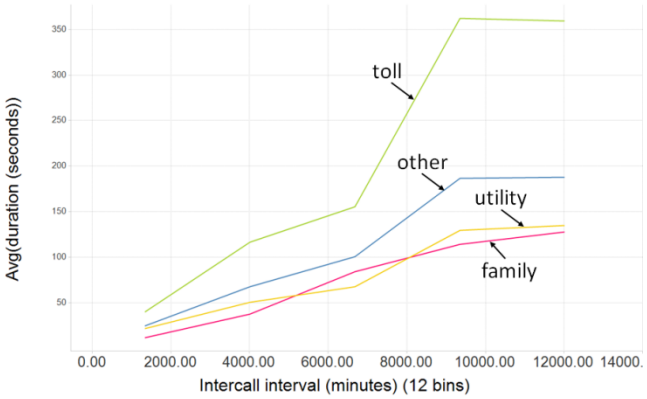

Figure 11. Average call duration (*y*-axis) as a function of inter-call interval (*x*-axis) in San Francisco

6. DISCUSSION AND CONCLUSION

In this study, we aimed to address the current gap in our knowledge of human phone call behavior, one manifestation of ties in social networks. In particular, we have investigated how behaviors vary across different social affinity networks, which could include family members, coworkers, friends, service providers, customers, etc.

We collected and processed Call Detail Records (CDRs) provided by Sprint from the two California cities of Modesto and San Francisco. We built a social graph with the edges (social ties) between two subscribers specified by phone calls between them. We then assigned each subscriber into four groups of social ties: family, toll free, utility services or other based on our access to information sources about their registration accounts. We defined each social group as an "affinity network," in which all the subscribers grouped in an affinity networks have a certain common social tie.

We then identified, calculated, analyzed, and compared various features that model call behavior and patterns such as the frequency, length, timing, and symmetry of the calls. Some key results and their implications can be summarized as follows:

- Call patterns of family members in both cities indicate strong social ties between them, which is reflected in the total number of calls and the frequency of the calls. These two features show statistically significant changes between family members and all other affinity networks. Furthermore, subscribers belonging to a family network in Modesto and San Francisco make on average 77.31% and 75.4%, respectively, of their total calls to each other and not outside the family network; on average only 9.46% and 10.65%, respectively, of their total calls are made to the utility and toll free numbers affinity networks when combined. Family members also call each other very frequently with a small time interval between successive calls, but calls to utility numbers are sporadic as the time taken between successive calls is long.
- Family members show a more reciprocal and mutual call behavior. This is reflected in the strong engagement ratios of 57% and 56% in Modesto and San Francisco respectively. In contrast, most of the calls made to utility numbers are not reciprocated equally (this is measured by weak engagement ratios of 13% and 10% respectively).
- While previous research [16][19] identified differences between communities in terms of personal network topologies and some behavioral characteristics, we found that the nature of social ties and their associated call patterns in Modesto and San Francisco have strong similarities. They also show very similar variations in terms of their related features across different affinity networks.
- We found strong correlations between the features that model call patterns. Some of these correlations are rather intuitive. For example, the number of calls is correctly expected to have a negative correlation with the inter-call arrival time. However, some show interesting call behavior. For example, the subscribers who call each other more frequently make shorter calls. Currently, it is not clear whether such correlations are the cause or the effect of some of the differences across affinity networks. For example, the above correlation could be the cause or the effect of shorter calls among family members. This is a topic for future research.
- Statistically significant features and their correlations have implications for research on prediction and classification of social ties. They mean that given these features and the social graph, we should be able to classify each edge of the graph into a specific social affinity network and infer the type of relationship. However, the dependencies among the features indicate that the classification approach would benefit from being able to capture such dependencies. This appears to benefit, for example, machine learning using random forests and Bayesian networks over simpler approaches such as single decision trees and naïve Bayesian classifiers.

As a limitation of this study, we should note that the affinity network types that we identified and used as our ground truth were inferred and not directly verified. For example, we assumed that family plans must be shared among family members. While this is common sense, we did not verify this assumption. This means that non-family members who share a family plan introduce some noise into the dataset. There are similar limitations associated with the Utility, Toll, and Others edge types.

From an application point of view, there are several opportunities to build on and expand this work. Knowledge about social tie strength and type is invaluable for different research and industrial purposes, especially for telecom providers. For example, customer acquisition can be enhanced by focusing on family members of existing customers that are not under their network; Churn prevention can be enhanced by focusing on friends of churned customers; Search engines and mediating businesses that recommend services to subscribers or customers to services can refine their recommendations and categorize them based on relationship type information, etc. Furthermore, knowledge and modeling of distinct call behavior of different social affinities, even without the affinity prediction capabilities, can be useful in predicting how call patterns and consequently, the load on the network may change in future as a result of predicted acquisitions and churns. There are also applications of tie strength and type in the area of security. For example, strength of social ties is a useful indicator of trust in many real-world relationships [26], and trust plays a crucial role in security applications.

As future work, we intend to a) expand the type of social affinities that we considered and add coworkers, friends, etc. by using more information sources without invading subscriber privacy; b) investigate the impact of the features that are associated with the social graph topology and also location and proximity patterns of the subscribers [3]; c) include additional cities and geographical areas from different parts of the country and the world to have more diverse datasets; and d) apply the proper classification techniques to classify the edges into affinity networks.

7. ACKNOWLEDGMENT

The authors would like to thank Kevin Leduc and Jason Powers, from Sprint Applied Research & Advanced Technology Labs, for providing technical and infrastructure support. We would also like to thank Sprint-Nextel and the Carnegie Mellon University - Silicon Valley community for their support of this work. This work is supported, in part, by NSF award CCF0937044 to Ole J. Mengshoel.

REFERENCES

[1] 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 *Proc. of 15th ACM CIKM*, pages 435–444, 2006.

[2] A. Clauset, C. Moore, M. E. J. Newman, Hierarchical structure and the prediction of missing links in networks, *Nature*, 453, pages 98-101, May 2008.

[3] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg: Inferring social ties from geographic coincidences. *PNAS*, 107, pages 22436-22441, December 2010.

[4] D. Wang, D. Pedreschi, C. Song, F. Giannotti, A. L. Barabasi. Human mobility, social ties, and link prediction. In *Proc. of KDD-11*, pages 1100–1108, 2011.

[5] D. Liben-Nowell and J. Kleinberg, The link prediction problem for social networks. In *Proc. of the 12th International Conference on Information and Knowledge management* (CIKM-03), 2003.

[6] E. Cho, S. A. Myers, J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proc. of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. San Diego, California, USA, August 21-24, 2011.

[7] G. Palla, A.-L. Barabasi, and T. Vicsek. Quantifying social group evolution. *Nature*, 446, pages 664-667, April 2007.

[8] H. Zhang and R. Dantu, Predicting social ties in mobile phone networks. In *Proc. of 2010 IEEE International Conference on Intelligence and Security Informatics*, pages 25-30, 2010.

[9] J. Resig, S. Dawara, C. Homan, and A. Teredesai. Extracting social networks from instant messaging populations. In *Proc. of ACM SIGKDD*, pages 22-25, 2004.

[10] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *PNAS*, 104, pages 7332–7336, May 2007.

[11] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati, and A. Joshi. Social ties and their relevance to churn in mobile telecom networks. In *Proc. of* EDBT-08, pages 668-677, 2008.

[12] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, and J. Leskovec. Mobile call graphs: Beyond power-law and lognormal distributions. In *Proc. of KDD-08*, pages 596-604, 2008.

[13] M. E. J. Newman. Detecting community structure in networks. The European Physical Journal B - Condensed Matter and Complex Systems, 38, pages 321-330, March 2004),

[14] M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature:* 453(7196), pages 779-782, 2008.

[15] N. Eagle, A. Pentland, and D. Lazer. Inferring friendship network structure by using mobile phone data. Proceeding of National Academic Science, 106(36), pages 15274–15278, 2009.

[16] N. Eagle, Y.-A. de Montjoye, and L. M. A. Bettencourt. Community Computing: Comparisons between Rural and Urban Societies using Mobile Phone Data, In *Proc. of CSE-09*, pages 144-150, 2009.

[17] P. Perona and W. T. Freeman. A factorization approach to grouping. In *Proc. of European Conference on Computer Vision*, pages 655–670, 1998.

[18] S. Sarkar and K. L. Boyer. Quantitative measures of changed based on feature organization: Eigenvalues and eigenvectors. *Computer Vision and Image Understanding*, 71, pages 110–136, July 1998.

[19] S. Isaacman, R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky. A tale of two cities. In *Proc. of HotMobile-10*, pages 19-24 2010.

[20] T. Shi, M. Belkin, and B. Yu. Data Spectroscopy: Learning Mixture Models using Eigenspaces of Convolution Operators. In *Proc. of ICML*, 2008.

[21] T. S. Evans and R. Lambiotte. Line Graphs, Link Partitions, and Overlapping Communities. Phys. Rev. E, 80 016105, 2009.

[22] U. V. Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17, pages 395–416, 2007.

[23] Y.-Y. Ahn, J. P. Bagrow, and S. Lehman. Link communities reveal multiscale complexity in networks. *Nature*, 466. pages 761-764, June 2010.

[24] M. Cossalter, O. J. Mengshoel, and T. Selker. Visualizing and understanding large-scale Bayesian networks. In *Proc. of the AAAI-11 Workshop on Scalable Integration of Analytics and Visualization*, pages 12-21, 2011.

[25] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. In *Proc. of the 6th conference on Symposium on Operating Systems Design & Implementation - Volume 6 (OSDI-04)*, San Francisco, CA, 2004.

[26] T. Hyun-Jin Kim, A. Yamada, V. Gligor, J. I. Hong, and A. Perrig. User-Controlled Trust Establishment through Visualization of Tie Strength. Technical report CMU-CyLab-11-014, Carnegie Mellon University, February 2011.