Social Ties and their Relevance to Churn in Mobile Telecom Networks

Koustuv Dasgupta, Rahul Singh, Balaji Viswanathan,
Dipanjan Chakraborty, Sougata Mukherjea, Amit A. Nanavati
IBM India Research Lab

Anupam Joshi
University of Maryland Baltimore County
Email: joshi@cs.umbc.edu

ABSTRACT

Social Network Analysis has emerged as a key paradigm in modern sociology, technology, and information sciences. The paradigm stems from the view that the attributes of an individual in a network are less important than their ties (relationships) with other individuals in the network. Exploring the nature and strength of these ties can help understand the structure and dynamics of social networks and explain real-world phenomena, ranging from organizational efficiency to the spread of information and disease.

In this paper, we examine the communication patterns of millions of mobile phone users, allowing us to study the underlying social network in a large-scale communication network. Our primary goal is to address the role of social ties in the formation and growth of groups, or communities, in a mobile network. In particular, we study the evolution of churners in an operator's network spanning over a period of four months. Our analysis explores the propensity of a subscriber to churn out of a service provider's network depending on the number of ties (friends) that have already churned. Based on our findings, we propose a spreading activation-based technique that predicts potential churners by examining the current set of churners and their underlying social network. The efficiency of the prediction is expressed as a lift curve, which indicates the fraction of all churners that can be caught when a certain fraction of subscribers were contacted.

1. INTRODUCTION

In today's extremely challenging business environment, many telecommunications carriers are measuring their success by the size and growth of their profit margins. As a result, carriers are under intense pressure to reduce or eliminate the major threats to these margins which arise from revenue leakage, inaccurate inter-carrier billing, fraud, and churn. Carriers rely on analysis of terabytes of Call Detail Record (CDR) data to help them make business-critical decisions that will positively affect their bottom line. High-end data warehouses and powerful Business Intelligence (BI) solutions are thus becoming essential tools to help carriers meet profit goals. Analyzing and integrating in-depth data enables carriers to reduce revenue leakage and churn, mitigate fraud, optimize network usage and increase profits.

Interestingly, as mobile penetration is increasing and even approaching saturation, the focus of Telecom BI is shifting from customer acquisition to customer retention. It has been estimated that it is much cheaper to retain an existing customer than to acquire a new one [8]. To maintain profitability, telecom service providers must control *churn*, i.e. the loss of subscribers who switch from one carrier to another. For the particular mobile operator we consider, annual churn rates in the prepaid segment average between a significant 50 to 70 percent. This implies that the operator must offer the right incentives, adopt the right marketing strategies, and place network assets appropriately to protect its customers.

Retrieving information from CDR data can provide major business insights for designing such strategies. A CDR contains various details pertaining to each call, e.g. who called whom, when was it made, etc. Based on this information, one can construct a call graph with customer mobile numbers as nodes and the calls as edges. The weight of an edge captures the strength of the relationship (tie) between two nodes. An edge with a high weight (e.g. call frequency or call volume) signifies a strong tie, while an edge with a low weight represents a weak one. Consequently, one can view the call graph as a social network consisting of n actors (nodes) and a relationship $R_{i,j}$ measured on each ordered pair of actors $i, j = 1,\ldots,n$.

We consider the call graph obtained from CDR data of one of the largest mobile operators in the world. Our objective is to explore the local and global structure of the underlying social network in this massive communication graph, and understand the role of social relationships as it pertains to the formation of groups (or communities) in the network. Understanding the structure and dynamics of social groups is a natural goal for network analysis, since such groups tend to be embedded within larger social network structures, growing in a potentially complex fashion [2, 31]. For example, a group that grows through aggressive recruitment of friends by other friends would appear as a subgraph branching out rapidly over time, while a group in which the decision to join depends relatively little on such influence might appear as a collection of disconnected components growing in a motley fashion.

EDBT'08, March 25-30, 2008, Nantes, France.

The central question that we strive to answer is whether the decision of a subscriber to churn out of the operator's network is dependent on the existing members of the community that the subscriber has a relationship with (referred to as *friends*). A social relationship between two friends, in this context, is based on the duration of voice calls, call frequency etc. that are exchanged during a certain period. Our analysis explores the propensity of a subscriber to churn out of a service provider's network depending on the number of friends that have already churned. For example, consider a subscriber Joshua. His friend has recently churned out of the operator's network. What is the probability of Joshua to churn? How would the probability change if Joshua had another friend who is also a churner? Some of these questions have been raised in the context of growth and evolution of communities in online social networks [2,18]. We believe that our analysis is a first of its kind that exploits the underlying social network in a telecom call graph, and interestingly, indicates that social ties play an important role in affecting customer churn. Further, we develop a prediction model that explores the social network of the churners to identify customers susceptible to churn in the near future.

1.1 Challenges and Contribution

The problem of churn prediction has been addressed by academicians as well as BI practitioners. Traditional solutions have used data mining techniques [1,8,24] that create a customer profile from her calling pattern (often described by hundreds of variables), and then predict the probability of churn based on certain attributes of the subscriber. For these customers, there are various data sources available for modeling including historical usage, billing, payment, customer service, application, and credit card data.

However, in our case, the mobile operator was interested in developing a churn prediction model for its prepaid segment, for which there exists very little data except for CDR data. What we could extract from this data, included aggregated call usage information for each customer, along with the call destination numbers, and call frequency and duration for each destination. Thus, any prediction model needed to be purely based on the available link information. Moreover, business rules and data availability restrictions imposed by the operator, required us to use a single month's CDR data to design and validate any prediction technique. Such practical limitations certainly make the problem more challenging, but we will demonstrate how reasonable prediction accuracy can still be achieved using only link information. To do so, we explore a diffusion (or *spreading activation*)-based approach, which is based on the premise that a few key individuals (churners) may lead to strong "word-of-mouth" effects, wherein they influence their friends to churn, who in turn spread the influence to others, and so forth. Such a diffusion process has a long history in social sciences [10, 20, 25] and essentially uses the underlying social network for spread of influences. We explore the use of this diffusion-based technique for identifying potential churners and report the success of the technique in terms of churn prediction accuracy.

Interestingly, a key feature of the proposed technique is that it allows the operator to pro-actively identify potential churners and pursue them for retention, based on "early warnings". For example, a subscriber Joshua can be identified as a potential target as soon as a number of his close friends churn. This is different from existing approaches where typically a customer is flagged when there is noticeable change in his recent usage profile (e.g. reduced spending, prepaid card not recharged etc.) - by which time he might have already decided to churn. Our feedbacks from multiple telecom operators suggest that this capability can provide a key value-add, where social network analytics can complement and enhance existing BI solutions for churn management.

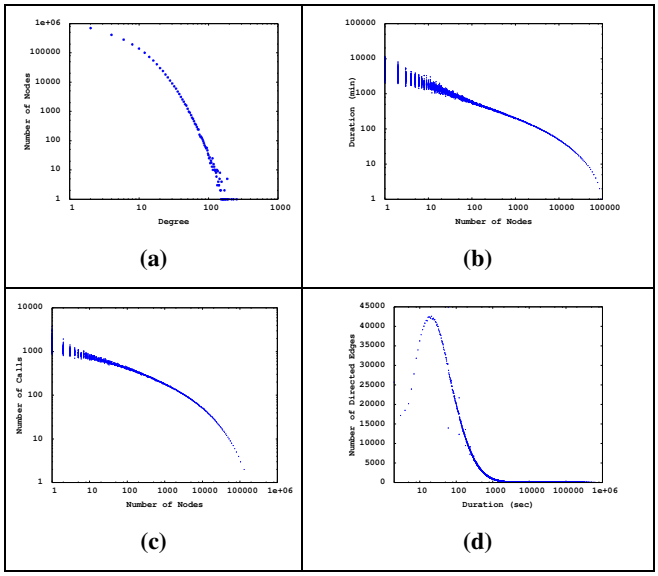

Figure 1. (a) Degree Distribution (b) Call Volume Distribution (c) Call Frequency Distribution (d) Call Duration Distribution of the Mobile call graph

2. DATA SET

We consider the Call Detail Record (CDR) data of one of the largest mobile operators in the world between 1st March and 31st March 2007. The data set is about 60 Gigabytes large and contains detailed information about voice calls, SMS, value-added calls etc. of users. Our analysis is based on a representative region in the operator's network and all intra-region (local) calls made during the specified period.

The raw CDR data contained $3.1X10^6$ nodes and $12.3X10^6$ edges. Calls within 5 seconds are assumed to be accidentally dropped and filtered out. Further, we include a pair of nodes A and B, if and only if, A calls B and B calls A. While a single call between two individuals may not carry much information, *reciprocal* calls of long duration (or high frequency) serve as a signature of a social relationship. Therefore, in order to translate the data into a network representation that captures the characteristics of the underlying communication network, we consider a directed edge $<A,B>$ if there has been at least one reciprocated edge $<B,A>$ between them, i.e., A called B, and B called A. We refer to two individuals as *friends* if they are connected by a pair of reciprocal edges. The weight W_{AB} of a directed edge $<A,B>$ is the aggregate of all calls made by A to B.

During pre-processing, we also excluded the service numbers, e.g. the operator's customer service number, number for retrieving voice mail, and numbers similar to 1-800 numbers. We observed that these numbers greatly skewed the call distributions in the operator's network. The omission of these numbers resulted in the removal of about 450 nodes and $1.2X10^6$ edges. The final

(reciprocal) call graph contained $2.1X10^6$ nodes and $9.3X10^6$ directed edges. Overall, the reciprocal graph contains 32.1 $X10^6$ calls and total call duration of $955X10^3$ hours.

The rest of our analysis is based on this social network. Our objective is two-fold. First, we wish to extract characteristics of customer churn as it relates social influence in the underlying network. Second, we wish to address the fundamental question whether churn can be modeled as a diffusion process [10, 20, 25] that spreads through the network.

3. MEASUREMENTS

We begin by summarizing the overall characteristics of the call graph, referred to as G_{March}. Next, using churner data available from the operator, we highlight the role of social ties (influences) in affecting churn in the prepaid customer segment

3.1 Basic Call Graph Properties

Figure 1 summarizes the basic structural properties of the call graph. As expected, the call graph is found to be characterized by presence of a highly heterogeneous topology, with degree distribution characterized by wide variability and heavy tails. Observing the log-log plot in Fig. 1(a) we can see that degree distribution fits well to a power law distribution. The power law exponent, in specific, is 2.91. The trend implies that most pre-paid customers call a relatively smaller number of people (friends), while a small number of individuals have relationships with a large group of people. Such a skewed distribution is also observed for the (node) call volume and (node) call frequency distributions, as shown in Fig. 1(b) and 1(c), respectively. In Fig. 1(d), we plot the distribution of call durations, obtained from the call duration of each directed edge in the graph. The plot shows that most calls in the mobile network are short-lived, while a few dozens of calls last for hours. Interestingly, the distribution exhibits a peak at around 1 minute. This reflects a caller's tendency to finish a conversation within 60 seconds (which is the pulse rate of the operator for charging voice calls in prepaid segment).

3.2 Analyzing the Churner Community

We next turn to the community of churners in the mobile network. Our analysis is based on the churners between the months of April and July, as provided by the operator, and their observed interactions in the call graph of March. The observation period is a month ahead of the churn period and hence contains a large portion of churner calls, which can be used to approximate the social network(s) of these churners. Table 1 gives the number of churners in different months. Note that, there are quite a few subscribers who have churned but not captured in the CDR data, simply because they did not make or receive calls in March. Since our objective is to gauge the role of social influences w.r.t churn, we evaluate our findings strictly based on churners with CDR data in March.

Table 1: Churner Population during April to July

Month	Churners with CDRs
April	44266
May	42458
June	65796
July	58565

To understand the characteristics of churn behavior and relate it to a diffusion process, we first need to find out whether there is any evidence of influences in affecting a customer to churn. The underlying premise, in this case, is that an individual's probability of adopting a new behavior increases with the number of friends that have already engaged in the behavior—to be specific, the number of friends who have churned in an earlier period (e.g. the previous month). In Fig. 2(a), we show this relationship. We compute the probability $P(k)$ as suggested in [2]. For the churners of May, we consider churners of April. Then we find all triples $<u,C,k>$ such that C is the set of churners, u is a user who has not churned in April, and u has k friends in C. $P(k)$, for a given k, is then the fraction of all such triples $<u,C,k>$, such that u belongs to C in May (and not in April). Similarly, for June churners we can compute $P(k)$ by considering April and May churners, and so on.

Surprisingly, the curves indicate that the probability of churn is significantly influenced by the number of friends who have churned in previous months. In fact, the probabilities increase if an individual has these friends churning over subsequent months, hinting towards a cascading effect of these influences. To gain further insight, we also measured the probability of churning as a function of the internal connectedness of friends who have churned. While the details of the technique are omitted for the sake of brevity (refer to [2] for details), our results reveal that individuals whose churner friends are linked to each other are significantly more likely to churn (in Fig. 2(b)). Stated otherwise, the probability of churn is not only affected by the number of churner friends that one has, but also the local topology connecting these friends. This result is strikingly consistent with the growth of online communities observed in [2]. In fact, it forms the basis of our hypothesis that churn as a "behavior" could be attributed to diffusion models that posit very simple dynamics by which influence is transmitted in a (highly) connected social network.

Role of Strong and Weak Ties in Diffusion. In order to build on this hypothesis, we must explore the role of social ties in driving a global diffusion process.

Figure 3(a) shows the distribution of tie strengths in the mobile network, where *tie strength* is defined as the sum of the weights of the edges $<A,B>$ and $<B,A>$. The tie strengths show wide variability and a heavy-tail, indicating that while majority of ties correspond to a few minutes of air time, a small fraction of users spend hours chatting with each other. Interestingly, the distribution is similarly skewed (w.r.t strong and weak ties), when we consider churner pairs only (Fig. 3(b)).

At this point, we allude to sociological principles that suggest that the strength of a tie could depend only on the *dyad*, i.e. the relationship between two individuals (independent of the network), or alternatively, be dependent on the network i.e. friendship circles, resulting in the importance of the weak ties in connecting communities [5,11]. To understand the implications of this relationship between tie strength and the local network structure, one needs to explore the network's ability to withstand the removal of either strong or weak ties. We measure the relative size of the *giant component*, providing the fraction of nodes that can all reach each others through connected paths as a function of the fraction of removed links *f.* Fig. 3(c) demonstrates the effect of removing links in order of strongest (or weakest) link. We also

measure the relative *topological overlap* of the neighborhood of two users A and B, representing the proportion of their common friends, as $O_{AB} = N_{AB}/((K_A-1)+(K_B-1)-N_{AB})$, where N_{AB} is the number of common neighbors of A and B, and K_A (K_B) denotes the degree of node $A(B)$.[1] Fig. 3(d) demonstrates the effect of removing links in order of strongest (or weakest) overlaps. In both cases, we find that removing ties in rank order of weakest to strongest ties will lead to a sudden disintegration of the network. In contrast, reversing the order shrinks the network without precipitously breaking it apart.

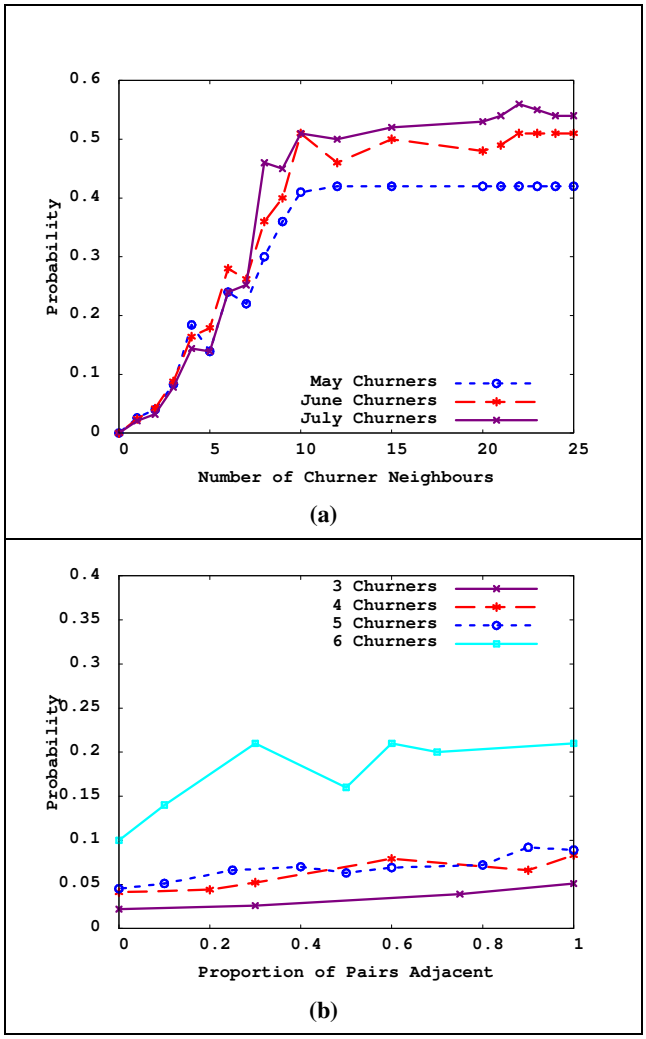

Figure 2. Probability of churning when (a) k friends have already churned (b) adjacent pairs of friends have already churned

This result is broadly consistent with the strength of weak ties hypothesis [5], offering one of its first confirmations in mobile networks. Accordingly, tie strength is driven not only by the individuals involved in the tie, but also by the network structure in the tie's immediate vicinity. Further, given that the strong ties are predominantly within communities, their removal will only locally disintegrate a community, while the removal of the weak links will delete bridges that connect different communities, leading to a network collapse. Further, we believe that the observed local relationship, between network topology and tie strength affects any global information diffusion process (like churn). In fact, we opine that *churn as a behavior can be viewed less as a dyadic phenomenon (affected only by strong churner-churner ties), but more as a diffusion process where both strong and weak ties play a significant role in spreading the influence through the network topology.*

4. PREDICTING CHURNERS IN THE CALL GRAPH

We next discuss how to exploit social ties to identify potential churners in an operator's network. Our approach is as follows. We start with a set of churners (e.g. for April) and their social relationships (ties) captured in the call graph (for March). Using the underlying topology of the call graph, we then initiate a diffusion process with the churners as *seeds*. Effectively, we model a "word-of-mouth" scenario where a churner influences one of his neighbors to churn, from where the influence spreads to some other neighbor, and so on. At the end of the diffusion process, we inspect the amount of influence received by each node. Using a threshold-based technique, a node that is currently not a churner can be declared to be a potential future one, based on the influence that has been accumulated. Finally, we measure the number of correct predictions by tallying with the actual set of churners that were recorded for a subsequent month (e.g. for May). The diffusion model is based on Spreading Activation (SPA) techniques proposed in cognitive psychology and later used for trust metric computations [32]. In essence, SPA is similar to performing a breadth-first search on the call graph $G_{March}=(V,E)$. The basic steps are outlined below:-

Node Activation: During each iterative step i, there is a set of active nodes. Let X be an active node which has associated energy $E(X,i)$ at step i. Intuitively, $E(X,i)$ is the amount of (social) influence[2] transmitted to the node via one or more of its neighbors. A node with high influence has a greater propensity to churn. Let $N(X)$ be the set of neighbors of X. Active nodes for step $i+1$ comprises of nodes which are neighbors of currently active members. Further, a currently active node X transfers a fraction of its energy to each neighbor Y (connected by a directed edge $<X,Y>$), in the process of activating it. The amount of energy that is transferred from X to Y depends on the Spreading Factor d and the Transfer Function F, respectively.

Spreading Factor: SPA starts with a set of active nodes (seed nodes) each having initial energy $E(X,0)$. At each subsequent step i, an active node transfers a portion of its energy $d \cdot E(X,i)$ to its neighbors, while retaining $(1 - d) \cdot E(X,i)$ for itself, where d is the global Spreading Factor. The spreading factor concept is very intuitive and, in fact, very close to real models of energy spreading. Observe that the overall amount of energy in the network does not change over time, i.e. $\sum_X E(X,i) = \sum_{X \in V} E(X,0) = E_0$, for each step i. The spreading factor determines the amount of

[1] If A and B have no common acquaintances we have $O_{AB} = 1$.

[2] The terms "energy" and "influence" are used interchangeably in this context.

importance we wish to associate on the distance of an active node from the initial seed node(s). Low values of d favor influence proximity to the source of injection, while high values allow the influence to also reach nodes which are further away. We discuss the choice of values for d in the next section.

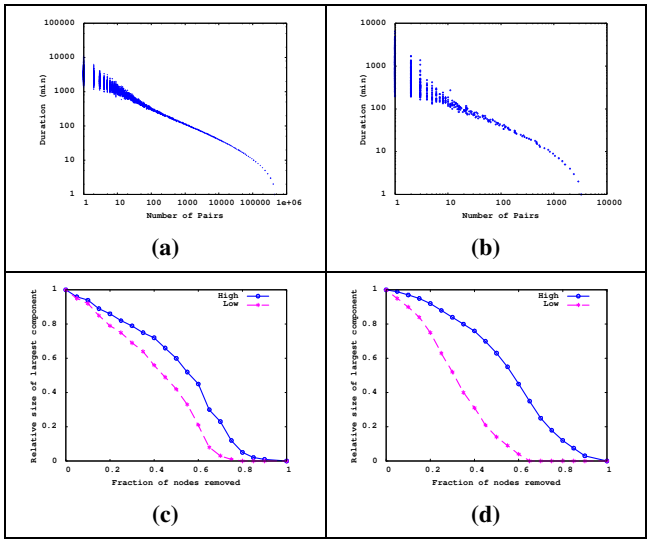

Figure 3. Tie Strength Distribution for (a) All Pairs (b) Churner-Churner Pairs; Stability of call graph w.r.t removal of links based on (c) Tie strengths (d) neighborhood overlap

Energy Distribution: Once a node decides what fraction of energy to distribute, the next step is to decide what fraction of the energy is transferred to each neighbor. This is controlled by a Transfer Function F. In our case, we use a linear edge weight normalization, i.e., the energy distributed along the directed edge <X,Y> depends on its *relative* weight W_{XY} compared to the sum of weights of all outgoing edges of X. In other words, $E(X \rightarrow Y) = d \cdot F \cdot E(X,i)$, where $F = W(X,Y)/\sum_{<X,S> \in E} W(X,S)$. We believe that this definition of the Transfer Function blends intuitively with how influence spreads through a call graph. People may be believed to exert a much stronger influence on those to whom they speak for longer durations of time.

Termination Condition: Since the directed call graph contains cycles, the computation of energy values for all reachable nodes is inherently recursive. Several iterations for all nodes are required in order to make computed information. Suppose $Vi \subseteq V$ represents the set of nodes that have been discovered (activated) until step i. Then the algorithm terminates when both of the following conditions are satisfied:

(a) $Vi = Vi-1$

(b) $\forall X \in Vi : E(X,i+1) - E(X,i) \leq E_T$

i.e. when no new nodes have been activated and when changes in influence w.r.t. prior iteration are not greater than accuracy threshold E_T.

5. EXPERIMENTAL RESULTS

We next proceed to validate our approach using real churner data. We consider the directed call graph G_{March}, with the churners as seed nodes. The weight of each directed edge in the graph is normalized between [0,1], using a function of the base form $F(x) = 2/(1+e^{\wedge}-x) - 1$.

Table 2: SPA Parameters for Churn Prediction

Parameter	Value(s)
Initial Energy E_0	1.0
Spreading Factor d	0.25-0.90
Accuracy Threshold E_T	0.01

Next, we run the iterative SPA routine on this directed graph. After termination, each node in the network accumulates a certain energy value (influence). If not a already a churner, this value reflects the propensity of the node to churn.

Since decision making ultimately requires a "churn" (i.e. likely to churn) or "no churn" (i.e. not likely to churn) prediction, the continuous energy measure must be thresholded to obtain a discrete predicted outcome. To this end, we use a simple threshold-based technique which works as follows: *Fix a threshold T_C. Label a node X as "churn" if its energy is greater than the threshold, else label it as "no churn".*

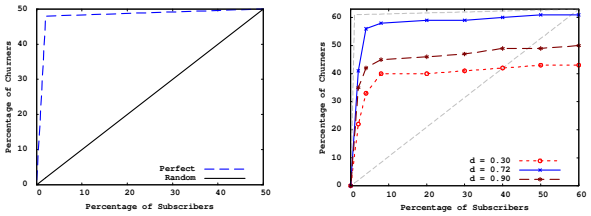

Figure 4. (a) Illustrative Lift Curves (b) Effect of the Spreading Factor on performance of SPA

Having predicted a list of potential churners, we now need to quantify the accuracy of prediction. In the telecommunications industry the outcome is often expressed using a *lift curve*. The lift curve is related to the ROC curve of signal detection theory and the precision-recall curve in information retrieval. The lift curve plots the fraction of all churners having churn probability above the threshold T_C against the fraction of all subscribers having churn probability above the threshold. The lift curve indicates the fraction of all churners can be caught (retained) if a certain fraction of all subscribers were contacted. Note that, an operator's customer services center only has a fixed number of personnel to contact some fraction of all subscribers Hence, the lift curve, which can estimate the fraction of churners that can be caught given limited resources, is very useful. For ease of understanding, we illustrate two sample lift curves in Fig. 4(a) - the lift curve representing perfect discrimination of churners from non-churners (best case), and that representing no discrimination (worst case). In general, the more bowed the curve is to the upper-left corner of the graph, the better the predictor. Note that, in SPA, the fraction of subscribers contacted can be increased (decreased) by setting a low (high) threshold T_C.

For the sake of comparison, we also consider a simple churn prediction heuristic, based on ties strengths, which works as follows: Consider the top K-percentile of churners, in terms of total call duration (incoming and outgoing). For each of these high-volume churners, and for a given variable k, identify the nodes which constitute the top k-percentile of the churner's tie strengths, i.e., the neighbors with whom the churners interact for longest duration. Label each of these neighbors as "churn". We

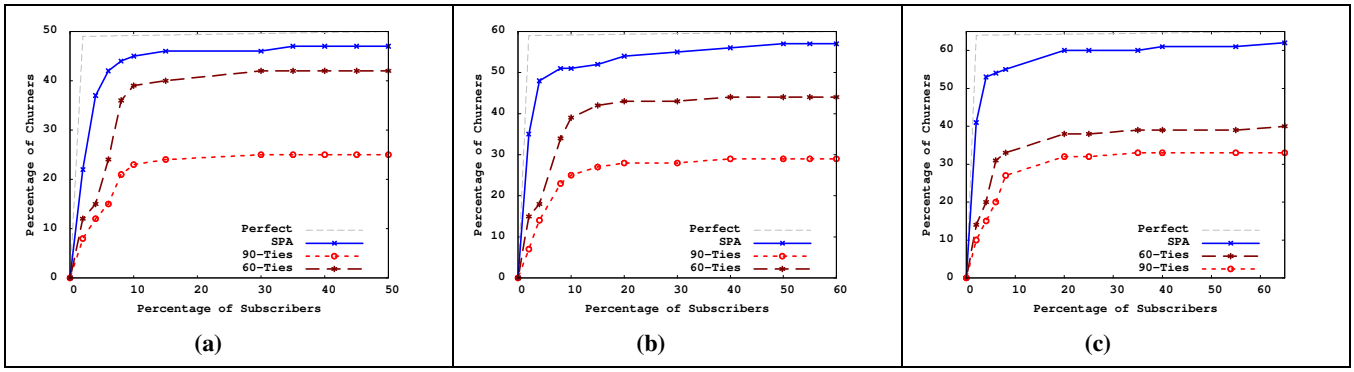

Figure 5: Performance of SPA and K-Ties heuristic for different observation and validation periods

refer to this as the **K-Ties** heuristic. As before, by setting a large (or small) *k,* the heuristic can contact more (or less) subscribers. We report the lift curves obtained by using representative values of *K* and *k* and compare them with SPA. Table 2 lists the parameters used by SPA.

Effect of Spreading Factor: We first try to understand the how the Spreading Factor *d* affects the performance of SPA. As discussed earlier, this determines the diffusion process in the network. A low value of *d* would facilitate rapid spread of the influence. A higher value, on the other hand, would imply that the influence takes a while to spread, often being trapped in highly connected localities (e.g. communities) of the network , before finding an escape to other parts of the network. We present the results for three representative values, i.e. d = 0.3, 0.72. 0.9, in Fig. 4(b). The experiments were performed with April churners as *the observation set*, i.e. churners marked as seed nodes in the call graph. Further, the May-July churners were treated as the *validation set,* i.e. each "churn" prediction made by SPA was validated against the churner logs from these months to determine if the node actually churned in one of these months. Interestingly, we find that while the lift curve improves by using a higher value of spreading factor, it is not beneficial to use very large values of *d.* In fact, we found that the best predictor was obtained by using a value of d = 0.72. For the rest of the experiments, we report results with a spreading factor of 0.72.

Predicting Churners over Time: We present results from three sets of experiment. In each experiment, we compare the performance of SPA with the K-Ties approach. In fact, we consider two instances of the K-Ties approach. In the 60-Ties heuristic, we consider the top 60^{th}–percentile of churners and then predict future churners, based on subscribers who fall within the top k (= 10, 20,….90, 99) percentile of the churner ties. Similarly, for the 90-Ties heuristic, we consider the top 90^{th} – percentile of churners and their strongest ties. In Figs. 5(a), 5(b), and 5(c), we demonstrate the relative performances of SPA and K-Ties approaches for the following sets of experiments, respectively.

a) April churners as observation set, and May churners as validation set in Fig. 5(a).

b) April-May churners as observation set, and June churners as validation set in Fig. 5(b).

c) April-May-June churners as observation set, and July churners as validation set in Fig. 5(c).

From the plots we observe that, SPA consistently outperforms K-Ties heuristic, in terms of the lift curve. This result agrees with our hypothesis that churn as a behavior is affected not only by strong ties between pairs of individuals (in particular, between an existing churner and a potential one), but more importantly, by the network topology and its local relationship with strong and weak ties. There are a number of additional observations that are worth considering. First, we note that the 60-Ties heuristic performs much better than 90-Ties. In fact, we found that 60- Ties performs the best among all values of K (= 10, 20,…,90, 99) that we considered. Further, we observe that the relative differences between the lift curves obtained from SPA and 60(90)-Ties heuristics, increase noticeably as the observation set becomes larger. Intuitively, this points to the underlying social network in the call graph, which grows richer (denser) over time, and can then be exploited to reason about interesting behavioral processes, like churn. Finally, the lift curves saturate beyond a certain point simply due to the inherent limit imposed by the number of ties (and influences) that can be explored, by starting from a set of seed churners.

Taking a closer look at the lift curves, we observe that SPA is generally successful in making correct predictions about 50-60% of future churners, by contacting a relatively small fraction (10-20%) of the subscribers. At the same time, the numbers are not remarkable by themselves and leave scope for improvements. We remind the readers that the main objective of this study was to demonstrate to the operator, how ties in an underlying social network can be used to analyze and predict churn behavior in a telecom network. Going forward, we plan to extract additional CDR information (e.g. inter-region calls, SMS records, etc.), as well as, graph-theoretic properties (e.g. cliques, hubs, and authorities, etc.) that can be incorporated within the SPA to potentially enhance the lift curves.

Hit Rate: We define *hit rate* to be the number of correct "churn" predictions, as a percentage of the total number of nodes labeled "churn". A low hit rate implies a large number of "false positives", and vice versa. We observe (Fig. 6) that the hit rate of all approaches usually reduces as the number of subscribers contacted increase. As expected, 60-Ties has a low hit rate, while SPA (with spreading factor of 0.72) performs best. What is interesting to note is that SPA, with d = 0.9, starts with a high hit rate (influence spreading rapidly through the network), but also suffers from rapidly diminishing returns as more subscribers are contacted. In fact, at one point, it falls even below 60-Ties. This also explains why very large values of d lead to inferior lift curves

in Figs. 5(a)-(c). Finally, as before, we note that hit rate of SPA can be potentially improved by incorporating other properties and/or additional information (from CDR data or other data sources) in the decision making process.

6. DISCUSSIONS

Our results demonstrate that a good prediction accuracy can be achieved by using a simple, diffusion-process that exploits social influences affecting churn. However, there exists scope for further improvements by either optimizing the SPA parameters and/or learning from similar flavored models. In particular, it is imperative to examine how well the notion of social influences can be captured in prevalent classification techniques. In this section, we briefly discuss some of these issues. Most of these form a part of our future deliverables' roadmap for the telecom operator.

To start with, we revisit Fig. 2 that estimates the probability of a customer churning given on a single feature, i.e. the number of friends who have actually churned. While this is a single feature, we can derive a range of other features related to the individuals themselves (extracted from CDR data), as well as features related to social ties in the underlying network. By constructing a decision-tree model, one of the most common classification techniques, one can then estimate the probabilities of an individual to churn. Further, the predictions can be validated using churner information to compute a lift curve.

Table 3 summarizes a broad range of features (attributes) that we have used in our experiments. As mentioned in the table, some of these attributes (i.e. *usage* attributes) are based purely on information extracted from CDR data. The second set of attributes (i.e. *connectivity* attributes) is based on the social ties of a (labeled) individual with existing (labeled) churners. Finally, the *interconnectivity* attributes are derived from the structural ties between these churners. We use the J48 classifier implemented as part of WEKA[3] to obtain the predictions. The WEKA implementation of J48 uses information gain to select attributes while growing the tree. Our data set comprises of nodes in the March call graph, along with their attributes and "churn"/"no-churn" labels. As is common, part of this data is used for training, after which we classify unlabeled data in the test set.

Fig. 7 compares the lift curves obtained from a decision-tree based approach with SPA. Note that, the features described above are intentionally chosen to understand which ones among the activity/structural features are more relevant. For obtaining better accuracy, there are possibly other features that are of less importance for our current purpose. The results show that using a decision-tree technique with only usage attributes, i.e. DT-1, performs the worst. This simply implies that usage information based on prepaid CDR data is highly insufficient to perform any meaningful churn prediction. On the other hand, using the connectivity attributes along with usage attributes (DT-2), improves the lift curve by exploiting knowledge about direct or indirect relationships of an individual with the churner community. What is noticeable is that, adding the interconnectivity attributes, i.e. DT-3, significantly improves the performance of the lift curve. Note that, these features relate exclusively to the structure of the social network among the churners themselves – once again, corroborating the fact that churn depends not only on the relationships of an individuals with churners, but more importantly, on the structural relationships that are present between them in a social network setting. However, a traditional (label) attributes-based classifier loses out to SPA because it fails to adequately learn all the ties in the network neighborhood. To be more precise, links among the unlabeled data (or test set) can provide information that can help with classification. Similarly, links between labeled training data and unlabeled (test) data induce dependencies that should not be ignored.

This leads us to believe that further gains in the prediction accuracy can be potentially achieved by applying link mining techniques, based on collective classification [20,21]. *Link-based classification*, unlike traditional classification, focuses on predicting the category (churn/no-churn) of a node, based not just on its attributes, but on the links it participates in, and on attributes of nodes linked by some path of edges. Going forward, it would be important to measure the efficiency (both in terms of complexity and prediction accuracy) of an iterative link-based classification algorithm (e.g. [21]) for churn prediction in a telecom call graph. We have started to investigate this as part of our current agenda, and plan to implement a pilot solution with the telecom operator in near future.

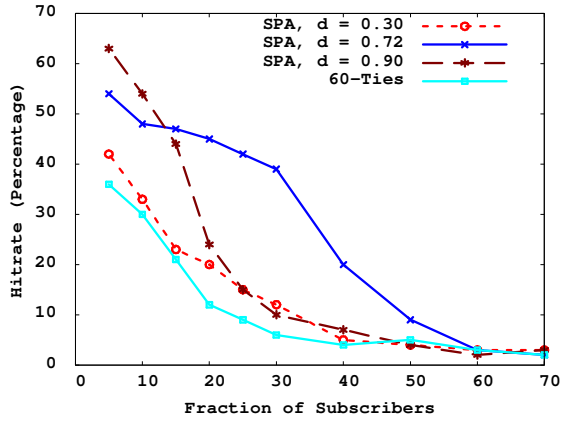

Figure 6. Hit Rates for SPA and K-Ties

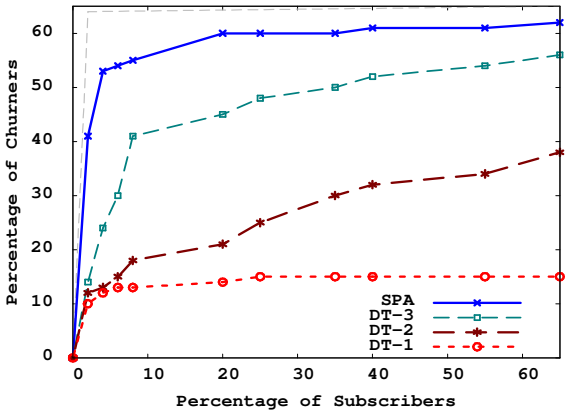

Figure 7. Performance of Decision Tree-based approach compared to SPA

[3] http://www.cs.waikato.ac.nz/ml/weka/

Table 3: Feature set

Feature Set	Feature Name
Usage	Total Call Frequency Number of outgoing calls Number of incoming calls Total Call volume (seconds) Total Incoming call duration (seconds) Total Outgoing call duration (seconds) Number of unique incoming, outgoing edges Number of unique neighbors Number of incoming, outgoing calls to/from a different operator's network Total Incoming call duration from a different operator's network (seconds) Total Outgoing call duration to a different operator's network (seconds) Total Incoming, Outgoing edges to/from a different operator's network Total number of neighbors in a different operator's network Call volume percentage (w.r.t total) to/from a different operator's network Call frequency percentage (w.r.t total) to/from a different operator's network
Connectivity	Total call frequency to/from churner neighbors Total call volume to/from churner neighbors (seconds) Call volume and frequency percentages (w.r.t total) to/from a churner Maximum call volume, frequency to/from a churner neighbor Number of unique churner neighbors Percentage of churner neighbors (w.r.t total neighbors) Number of non-churner neighbors who have churners as neighbors Maximum call volume and frequency with any of these non-churners Call volume and frequency percentages (w.r.t total) to/from these non-churners
Interconnectivity	Number of adjacent pairs in the set of churner friends i.e., connected by an edge Number of pairs in the set of churner friends connected by a path length of 2 Number of pairs of churner friends connected by a path that only includes churners Total call frequency on edges connecting adjacent churner friends Total call volume on edges connecting adjacent churner friends

7. RELATED WORK

Social network analysis (SNA) as a theme has been studied for years. There is a significant amount of work on mining a number of real world graphs (e.g. Internet, email, citation graphs, and email graphs) that are formed by interactions amongst individuals of a social network. One of the main areas of focus has been on *degree power laws*, showing that the set of node degrees has a heavy-tailed distribution [3, 4, 23]. Other properties include the small-world phenomenon, popularly known as "six degrees of separation", which states that real graphs have small (average or effective) diameters [6, 18, 23]. With the web growing, much social network data is available and recent efforts try to leverage social ties for expertise location [15], viral marketing [7,16,27], and online social networking sites [2,18]. The well known link analysis algorithms, such as PageRank [4] and HITS [17], can also be viewed as social network analysis on the web. While initial studies were limited to identifying patterns in static graphs, more recent work has focused on studying the temporal evolution of real-world graphs. Some interesting time-varying properties, related to densification and diameter shrinkage, have been observed for a number of real-world social networks [18], There has been a considerable amount of work devoted to finding (connected) communities in graphs [6,9]. A particular area of interest has been to study how online communities evolve over time [2,18]. Based on the findings, a number of generative models have been proposed to generate graphs that resemble social communities (see [4,31] for a detailed review).

Network models have been widely used to represent relational information (ties) between interacting units. The emphasis has been on random graph models [14] where nodes represent the actors and edges represent relationships between them. Recently, generalized models [30] have been proposed to capture changing relationships over time. In the same spirit, a number of models have been explored to explain how a new influence, idea or epidemic spreads in a social network using the relationships. [10] introduced a cellular automaton based model for simulating the spread of information in a social network. In [11], a model of information diffusion is proposed where a node gets converted when the fraction of its infected neighbors crosses a certain threshold. In a similar vein, [5] showed how actors can exploit the existence of structural holes in order to gain advantage in a competitive or cooperative scenario. The Bass diffusion model [22] and the game-theoretic model [26] are other notable efforts in this area. Recent research has concentrated on how to utilize diffusion-based models for viral marketing [7, 29]. An approximate algorithm for solving the problem of "influence maximization" in this setting has also been proposed in [16].

Trust (distrust) also has an intuitive connotation in social networks. A person can only believe and propagate a piece of information conveyed to it by some other person depending on how much it trusts the source of information. Trust management has been an important research issue stemming from the areas of cryptography and authentication [19]. Computation of trust metrics is central to the issue of trust management. A P2P-based

reputation system called EigenTrust is presented in [14]. In [28], the issue of trust computation is addressed in a semantic web setting. [12] presents a broad taxonomy of schemes through which trust and distrust may propagate in a web of trust. Finally, a spreading-activation based technique is employed for computation of trust metrics in [32].

8. CONCLUSIONS

Social Network Analysis (SNA) has emerged as an important paradigm for studying real-world, complex networks. In this paper, we provide substantial evidence that social relationships play an influential role in affecting churn in the operator's network. We also demonstrate a simple, yet effective, diffusion-based approach that exploits these influence to identify a significant fraction of churners in the network.

Influences, in the current framework, are purely derived from call volumes between individuals. However, there are a number of graph-theoretic properties of nodes (edges) in the network that can be used to guide the diffusion process. For example, it would be interesting to study how the transfer function in SPA can be varied based on node/edge properties, and its effect on the lift curve. A related interesting problem is that of (budget-constrained) churn prevention, where only a subset of potential churners is contacted, based on the value of the churner and cost of contacting the churner. Finally, there are a number of telecom analytics problems including customer segmentation, targeted advertising, and fraud detection, which are worth pursuing using SNA techniques.

9. REFERENCES

[1] W. H. Au, K. C. Chan, X. Yao. A novel evolutionary data mining algorithm with application to churn prediction. *IEEE Transaction on Evolutionary Computation*, 7, 6 (2003), pp. 532– 545.

[2] L. Backstrom, D. Huttenlocher, J. Kleinberg, X. Lan. Group Formation in Large Social Networks: Membership, Growth and Evolution. In *Proceedings of ACM SIGKDD* (Philadelphia, PA, USA, 2006).

[3] A.-L. Barabasi, R. Albert. Emergence of scaling in random networks. *Science*, 286 (1999), pp. 509-512.

[4] S. Brin, L. Page. The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of Seventh International WWW Conference* (Brisbane, Australia, 1998).

[5] R. Burt. Structural Holes: The Social Structure of Competition. Harvard, 1992.

[6] A. Clauset, M.E.J. Newman, C. Moore. Finding community structure in very large networks. *Physical Review E*, 70, 066111 (2004).

[7] P. Domingos, M. Richardson. Mining the Network Value of Customers. In *Proceedings of ACM CIKM* (San Francisco, CA, USA, 2001).

[8] Euler, T. Churn prediction in telecommunications using Miningmart. In *Proceedings of the Workshop on Data Mining and Business* (DMBiz, 2005).

[9] M. Girvan, M. E. J. Newman. Community structure in social and biological networks. In *Proceedings of Natl. Acad. Sci*. 99 (2002), pp. 7821-7826.

[10] J. Goldenberg, B. Libai, E. Muller. Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters,*12:3 (2001), pp.211-223.

[11] M. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6), 1973, pp. 1360-1380.

[12] R. Guha, R. Kumar, P. Raghavan, A. Tomkins. Propagation of Trust and Distrust. In *Proc. of WWW Conference* (New York, NY, USA, 2004).

[13] P. Hoff, A. Raftery, M. Handcock. Latent space Approaches to social network analysis. *Journal of the American Statistical Association*, 97 (2002).

[14] S. Kamvar, M. Schlosser, H. Garcia-Molina. The EigenTrust algorithm for reputation management in P2P networks. In *Proceedings of WWW Conference* (Budapest, Hungary, 2003).

[15] H. Kautz, B. Selman, M. Shah. ReferralWeb: Combining social networks and collaborative filtering. *Communications of the ACM*, 40:3 (1997), pp. 63-66.

[16] D. Kempe, J. Kleinberg, E. Tardos. Maximizing the Spread of Influence through a Social Network. In *Proceedings of ACM CIKM* (Washington, DC, USA, 2003).

[17] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *Proceedings of ACM SODA* (Baltimore, MD, US, 1998).

[18] R. Kumar, J. Novak, A. Tomkins. Structure and Evolution of Online Social Networks. In *Proceedings of ACM SIGKDD* (Philadelphia, PA, USA, 2006).

[19] R. Levien, A. Aiken. Attack-resistant trust metrics for public key certification. In *Proceedings of USENIX Security Symposium* (San Antonio, Texas, USA, 1998).

[20] Q. Lu, L. Getoor. Link-based Classification. In *Proceedings of 20th International Conference on Machine Learning* (pp. 496-503, 2003).

[21] S. Macskassy and F. Provost. Classification in Networked Data: A toolkit and a univariate case study. *Journal of Machine Learning Research.* 8(May):935--983, 2007.

[22] V. Mahajan, E. Muller, F. Bass. New Product Diffusio0 Models in Marketing: A Review and Directions for Research. *Journal of Marketing,* 54:1 (1990) PP. 1-26.

[23] S. Milgram. The small world problem. *Psychology Today*, Vol. 2 (1967), pp. 60-67.

[24] M. Mozer, R. Wolniewicz, D. B. Grimes, E. Johnson, H. Kaushanky. Predicting Subscriber Dissatisfaction and Improving Retention in Wireless Telecommunications Industry. *IEEE Transactions on Neural Networks*, 11 (2000), pp. 690-696.

[25] A.Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasgupta, S. Mukherjea, A. Joshi. On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications. In *Proceedings of ACM CIKM*, (Arlington, VA, USA, 2006).

[26] M.Newman. The structure and function of complex networks. *SIAM Review*, 45 (2003), pp.167–256.

[27] H. Peyton Young. The Diffusion of Innovations in Social Networks. *Santa Fe Institute Working Paper*.02-04-018 (2002).

[28] M. Richardson, R. Agrawal, P. Domingos. Trust management for the semantic web. In *Proceedings of the Second International Semantic Web Conference* (Sanibel Island, FL, USA, 2003).

[29] M. Richardson, P. Domingos. Mining Knowledge-Sharing Sites for Viral Marketing. In *Proceedings of ACM CIKM* (Edmonton, Alberta, Canada, 2002).

[30] P. Sarkar, A. Moore. Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter* Volume 7, Issue 2 (2005) pp: 31 – 40.

[31] Wasserman, Stanley and Katherine Faust. 1997. *Social Network Analysis: Methods and Applications*: Cambridge University Press.

[32] C. Ziegler, G. Lausen. Spreading Activation Models for Trust Propagation. In *Proceedings of the IEEE International Conference on e-Technology, e-Commerce, and e-Service* (Taipei, Taiwan, 2004).