Predicting the Influencers on Wireless Subscriber Churn

Sara Motahari[1], Taeho Jung[2], Hui Zang[3], Krishna Janakiraman[4], Xiang-Yang Li[2], Kevin Soo Hoo[1]
[1]Sprint Advanced Analysis Lab, Burlingame, CA, 94101
[2]Illinois Institute of Technology, Chicago, IL, 60616
[3]Guavus Inc, San Mateo, CA, 94404
[4]Bandpage, San Francisco, CA, 94107
sara.gatmir-motahari@sprint.com, tjung@hawk.iit.edu, {hui.zang,jkrish,xiangyang.li}@gmail.com, Kevin.SooHoo@sprint.com

***Abstract*—Wireless carriers have various churn models that are mainly based on profiling the customers and assigning churn probabilities to them. Profiling is usually limited to their individual data, such as their subscription history, demographics, usage, etc. However, our analysis of a major wireless carrier data shows that such churn prediction methods do not fully model wireless subscriber churn, and that the subscribers can be influenced by other subscribers' churn in their social network.**

We propose a novel method to identify 'churn influencers', whose influence makes their social contacts churn subsequently. To build our model, we scored the subscribers' influence level in a way that can take current churn models into account. We further used large scale call records to identify social network and communication features that abstract the strong influencers. Using real world churn data, we trained classification tools to classify high influencers with up to ninety nine percent precision.

Keywords-Wireless Subscribers, Churn, Influence prediction

I. INTRODUCTION

Customers of wireless carriers and telecom service providers are mostly contractual customers, which means customer churn rate is a significant factor for wireless carriers. Therefore, service providers and the research community constantly search for new methods to investigate, predict, or prevent customer churn [1] [2] [3] . While wireless carriers have different ways for profiling subscribers and predicting their churn likelihood, here we do not focus on the individual churn prediction problem. Instead, we choose to focus on churn 'influence' from the influencers among a subscriber's social contacts for the following reasons.

First, the influence factor is currently left out of the wireless carriers' churn models while as we'll see, this plays an important role in subscriber churn.

Second, the average monthly churn rate for wireless subscribers is usually between 1.5% and 2.5%. This means a carrier will have tens of thousands of churners each month. Therefore, even if we are given a perfectly correctly predicted churners list for the coming months, the carrier will not have the budget to treat all of them and prevent them all from churning. It would be more cost efficient to treat the 'influencers' assuming that preventing them from churn will also prevent some of their followers from churning.

Our identification of influence and classification of influencers were carried out only based on the data sources that are always available to wireless carriers for billing purposes. This means that first, their social network was made based on their call records (people who they call or receive calls from). Second, no information was collected on the content and subject of their communication with any of their social contacts. Therefore, to identify influence, we looked at the actual churn data to see how a subscriber's churn was followed by subsequent churns of more subscribers. This is different from conventional work on identifying and analyzing influence. Influence propagation in a network is widely assumed to be the equivalent of the information propagation problem [4] [5] [6] . We do not follow this convention for the network of wireless subscribers, because the fact that a subscriber receives a piece of information does not necessarily mean that their actions will be influenced by that.

Looking at the actual churn events enabled us to score the influence of a churner based on the number of subsequent churners in their social network. We further, used a currently available churn probability list (provided by the wireless carrier) to adjust their scores. Then using their call records, we extracted features that model their communication behavior and their social network structure. We analyzed these features and identified the ones that separated the subscribers with a high influence score from the subscribers with a low influence score. Our analysis shows that these features are predictive and can be used to predict future influencers. In fact, training and testing a classifier on the churn data from the wireless carrier showed that by a tradeoff between precision and recall, we can achieve over 99% true positive rate.

Being able to predict the influencers and their level of influence can enable the wireless carriers to select the likely churners to be treated in a smart way as well as enhance their current churn prediction models.

II. PREVIOUS WORK ON CHURN AND INFLUENCE

As mentioned earlier, wireless carriers mainly focus on profiling their customers to assign churn likelihoods to them. The research community has proposed more sophisticated churn prediction algorithms for this purpose. We summarize the more relevant ones below.

A. Churn Prediction Problem

In previous research work, various features extracted from Call Detail Records (CDR) data are typically used as attributes in the machine learning based techniques, and subscribers are classified as churners or non-churners, or they are assigned churn scores.

Yeshwanth et al. [1] analyzed 6 months of CDR collected from one of the major mobile network carriers in India. They used attributes in the CDR data related to usage, spending, refill and interconnection, to predict the churn scores for all subscribers by combining two data mining techniques. They further proposed the abstract idea of influence scores to prioritize predicted churners in their marketing treatment, but implementation or evaluation [2]

Unlike the above work which considers only nodal features as call network attributes, Verbeke et al. [3] proposed a methodology which uses local classifiers, relational classifiers and collective inference procedure to capture more comprehensive information about call data and churn behavior. Without achieving highly promising results, they did consider that churn behavior was not determined only by node-centric attributes but also by pair-wise relational attributes.

Dasgupta et al. [7] exploited social relationship in the churn prediction. They considered the CDR data of one of the largest mobile network carriers in the world for a month, which contain voice calls, SMS and value-added calls and applied an energy-propagation model in their churn prediction. That is, they assume every node gets some churn energy from other churners depending on pair-wise call durations and a system-wide spreading decay factor.

We remind that our goal in this paper is not churn prediction but identification of influence and prediction of influencers. Thus, we also summarize relevant work on influence.

B. Influence Propagation Problem

Influence propagation in a network, which has been getting increasing attention recently, is widely assumed to be the equivalent of the information propagation problem in the network [5] [6] [4] . For commercial purposes, researchers mostly focus on the influence 'maximization' problem. In this problem, each node in a directed social network graph is considered as either active or inactive. An active node can successfully infect his neighbor when certain conditions are met, and those infected neighbors become active. Then, the objective of this problem, with the number of initial active nodes bounded, is to maximize the number of active nodes after the propagation process converges.

Two diffusion models, the Linear Threshold Model and the Independent Cascade Model, are proposed to solve the problem by Kempe et al. [8] who also proved that the influence maximization problems in these two models are both NP-hard, and they proposed a polynomial time approximation algorithm. Subsequently, more relevant work [9] [10] [11] also proposed various mathematical models for this problem, proved the NP-hardness of them and presented approximation algorithms to efficiently find the optimum initial set.

Myers et al. proposed an idea of external influence in [5] which considers the fact the behavior of a node in a network may be affected by factors besides influence from other nodes within the network.

This external influence can be translated to prior churn probabilities in our work, but our model can use any prior likelihood modeling, we use actual marketing churn probabilities to implement and test it, and as we mentioned in the introduction, we do not follow an information or energy propagation model to model the influence. In the next section, we will explain how we identify and score influence, before implementing and validating the model in Section IV.

III. Scoring a Subscriber's Level of Influnce

Similar to the findings of previous work, our study shows that not all subscribers are equal in terms of the impact of their churn. As the real world data presented in Section IV.D. shows, some subscribers are followed by no followers in their social network when they churn, while the churn of some subscribers is followed by quite a few subsequent churns. This fact remains true after adjusting for the total number of their social network contacts and their prior churn probabilities. Therefore, even when the number of subsequent churners is normalized based on the total number of their contacts, still some subscribers have a higher ratio of subsequent churners than expected. In other words, some subscribers are more influential than the others.

Before we can identify such influencers, we need to define a metric to measure a subscribers' 'influence' on their social network. We also need to clarify how their social network or social graph is built, which is what we will explain in Subsection A. Having the graph of churners and influencers in place, we will move forward to scoring their influence level in Subsection B.

A. Construction of the Social Graph

We mentioned that we use the data widely available to wireless carriers, call data, to model their social network. The call graph G = (U; E) is constructed to show the call relationship, where U is the set of nodes and E is the set of edges. Nodes in the graph represent the subscribers of the wireless carrier, and each edge $(u; v) > E$ carries information about the calls initiated from u to v. There exists an edge between u and v if and only if there exists at least one call in each direction between them. In this case v is called a 'social contact' of u and vice versa.

B. *Modeling the Influence*

Suppose that a subscriber, *u*, churns from a network and his churn is followed by a number of subsequent churns among his social contacts. To have a more realistic model of *u*'s influence, we take the following parameters into accounts:

1. The number of *u*'s 'subsequent churners', which is the number of *u*'s social contacts who churned within a given time period after *u*'s churn. Generally, the more subsequent churners *u* has, the higher his influence score.
2. The number of *u*'s social contacts that did follow *u*'s churn. Generally, the more non-followers *u* has, the lower his influence score.
3. The prior churn probability of *u*'s subsequent churners. The churners with a higher prior churn probability give *u* a smaller influence score, compared to the ones with a lower churn probability. This prior churn probability can model any reason for churn besides *u*'s influence including outside-network influence and profile-based churn likelihoods.

Based on the above considerations, the influence score for a subscriber *u* is defined as:

$$\mathrm{I}(u)=\sum\nolimits_{v\epsilon V} (1\text{-}p(V))\ \text{-}\ \sum\nolimits_{v'\epsilon V'} p(V') \quad \mathbf{(1)}$$

where V is the set of *u*'s subsequent churners who churned within a given time period, T, after *u*'s churn; V' is the set of *u*'s social contacts who did not churn after *u*; and $p(v)$ is the prior probability of churn for node *v*.

IV. Predicting the Influencers

Based on the influence scores calculated in the previous section, the top influencers can be selected for treatment to be prevented from churning. However, the above influence score was calculated using the actual churn events, which means in order to calculate this score, we need to wait until a set of subscribers and their subsequent churners churn. Then it is obviously too late for any churn treatment. Thus, we need a way to predict the influence scores or classify the high influencers before they actually churn. This is our goal in this section.

To achieve this goal, we tried two different methods; 1) predicting the value of the influence score using linear regression, and 2) classifying the high influencers. Both methods were carried out using the widely available call data records and the call graph from the previous section. For any pair of nodes, *u* and *v*, the edge between them carries information or 'features' about their calls that models their relationship. We investigated whether these features have any predictive power and can detect high influence scores. After summarizing our data collection and filtering process below, we will explain both methods in subsections *B* and *C*.

A. *Data Sets and Preprocessing of the Data*

We collected Call Details Records (CDR), churn data, and prior churn probabilities for scoring, training, and validation.

1) *Call Detail Records (CDR)*

Call Detail Records contain a record for every outgoing and every incoming call and SMS from/to a subscriber. CDRs for this study were collected from a major wireless carrier in the United States between Nov-01-2011 and Jan-31-2012. They contain many data fields including the caller number, the called number, and date and timing of the call. The anonymized phone numbers identify the nodes in the graph.

For the purpose of scoring the influence in case a subscriber churns, the following subscribers were filtered out and eliminated from the list of the nodes whose influence score was to be calculated ('*u*'s):

- Their status was not active during the study period
- Joined the wireless carrier before November 2011
- Left the wireless carrier before January 2012
- Toll free numbers
- Had family plans and were under the same account with their potential subsequent churners
- Made no calls or one call in a single direction
- Were not a subscriber of the wireless carrier (landline or subscribers of other networks)

Table 1 shows the data set size.

2) *Churn Event Data*

We collected real churner lists from the carrier and retrieved two groups from them. First, subscribers who churned between Feb-01-2012 and Apr-30-2012 (time period T1). Second, the ones who churned between May-01-2012 and July-31-2012 (time period T2). We considered the first group as the ones who churned first (first churners), and the second group as subsequent churners. The reason for selecting such time periods are previous experiments that show followers usually follow their influencers' actions within several months, which is beyond the scope of this paper.

The subscriber churn rate was about two million. We selected the ones that satisfied the above filtering criteria, reducing the data size down to 1,082,606 first churners ('*u*' in Equation 1) and 426,185 subsequent churners ('*v*' in Equation 1). This means most first churners had no subsequent churners. The histogram of the number of subsequent churners per first churner is depicted in Fig. 1.

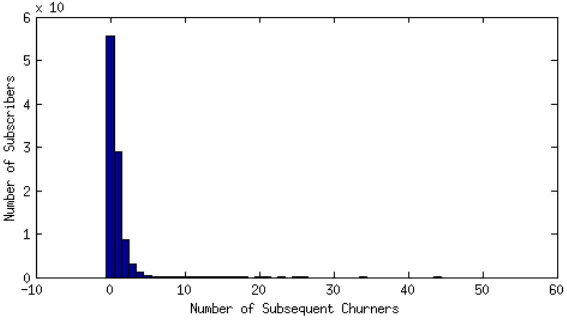

Fig. 1. The Histogram of the Number of Subsequent Churners per First Churner

TABLE 1. DATA SET SIZE

Total number of nodes	270,547,528
Total number of edges	22,120,697,697
Number of first churners after filtering	982,606
Number of subsequent churners after filtering	426,185
Number of subsequent churners with a prior churn probability after filtering	235,319

3) *Prior Churn Probabilties*

This data set was made by- and received from the marketing division within the wireless carrier. Each subscriber was assigned a churn probability for any three months; in particular, between May-01-2012 and July-31-2012 (time period T2). Internal network influence was not considered in calculating this probability. We used this data to model any other churn factor. Thus, it represents $p(v)$ and $p(v')$ for churners and non-churners in Equation 1.

4) *Features*

Features to be used for prediction were extracted and built from Call Detail Records. We hypothesized that high influencers can be distinguished from others by the strength, structure, or nature of their social relationships and network. Therefore, we calculated many features to model the relationship between a pair of subscribers and also the position of a subscriber in her social network. The features can be summarized as the 'number of calls', the 'duration of calls', and the 'time interval between consecutive calls'. However, each of them breaks down to many subcategories, such as 'nodal' and 'edge' features (e.g. the number of calls going to u versus the number of calls between u and v), 'total' and 'average', 'outgoing' and 'incoming', 'weekend' and 'weekday', 'work hours' and 'night hours', 'internal' and 'external' (i.e. customer of the wireless carrier versus customer of other carriers). Considering all different combinations of the above break downs, we obtained over 40 features. All features where then normalized.

Of course some or all of the above features may be uncorrelated with the influence level or have very week predictive power, so we performed a feature selection process that will be explained below.

B. *Method 1: Influence Score Prediction*

As the first step, we evaluated the significance and correlation of the features with respect to the influence score of the first churners. Twenty six features had significant correlations with the influence score. After a multivariate regression analysis, the list was further narrowed down to the following eight variables:

1. Internal degree: nodal degree based on connections to the carrier's *customers*;
2. Internal total calls: the total number of calls made/received to/from the other customers;
3. Internal total duration: The total cumulative duration of calls to/from other customers;
4. Internal mean interval: The average length of the time interval between calls with customers;
5. External degree: nodal degree based on connections to non-customers of the carrier;
6. External total calls;
7. External total duration;
8. External mean interval.

The score prediction for the first churners was carried out through linear regression. The above features formed the independent variables and the influence score, calculated based on equation 1, was the dependent variable. Out of 982,606 first churners, 687,824 were used for training and 296882 were used for test.

We should remind that the goal of this study is to identify the high influencers. Therefore, aside from looking at the rms of the error, we also grouped the subscribers based on their influence score to calculate the misclassification rate for the high influencers. The results will be presented in Subsection *D*.

C. *Method 2: High Influencers Classification*

In the second method, instead of trying to predict the exact influence score, we classified the train set into two groups based on their influence score; (1) low influencers: the influence score equals 0, 1, or 2, and (2) high influencers: influence score is higher than two[1]. Then, we tried to predict the class of a subscriber.

For this purpose, we used the eight features listed above as the independent variables of the classification tool, and the influence class as the dependent variable. Out of 982,606 first churners, 687,824 were used for training and 296882 were used for test. Both Random Forests and single Decision Trees were tested as classification tools. Since internal degree is a dominant factor, the single Decision Tree outperformed the Random Forest. Thus, we'll only present the results of the single tree classification in the next subsection.

D. *Results*

1) *Descriptive Statistics of the Data*

The majority of first churners had zero subsequent churners followed by the ones with one subsequent churner. This is also reflected in their churn scores. Fig. 2 shows the histogram of the churn scores.

Looking at the correlation of the independent variables and the dependent variable, the internal degree (int_degree) appears as a dominant factor, however, all variables have significant correlations with the influence score (p-values smaller than 0.001 for all variables). These correlations are listed in Table 2.

[1] We should mention that the influence score thresholds used for this grouping are arbitrary division and mostly depends on what the marketing unit of the carrier considers as valuable customers to pay attention to. However, further analysis, which we will not discuss in this paper, showed that changing the thresholds of this grouping does not significantly change the prediction results.

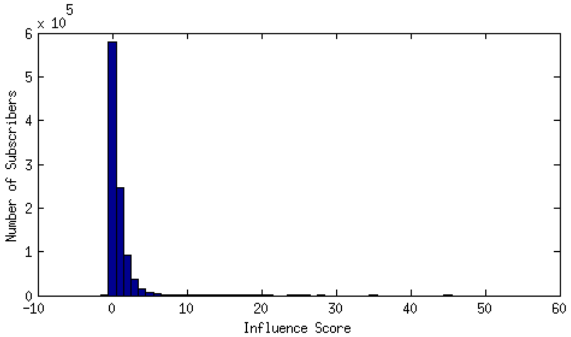

Fig. 2. The Histogram of the Influence Score

TABLE 2. CORRELATIONS BETWEEN THE INDEPENDENT VARIABLES AND THE DEPENDENT VARIABLE

$\rho_{\text{int\_degree,score}}$	0.55
$\rho_{\text{int\_calls,score}}$	0.39
$\rho_{\text{int\_duration,score}}$	0.16
$\rho_{\text{int\_interval,score}}$	-0.12
$\rho_{\text{ext\_degree,score}}$	0.39
$\rho_{\text{ext\_calls,score}}$	0.15
$\rho_{\text{ext\_duration,score}}$	0.06
$\rho_{\text{ext\_interval,score}}$	-0.07

2) *Results of Method 1- Linear Regression*

Our regression model was trained to perform covariance weighted least squares. We tuned the weights to put more emphasis on lowering the false positive rate although it worsens the overall performance. This is because the carries prefer to miss some influencers rather than treating many non-influencers. With this condition, the mean squared error was 0.8, false negative rate was 80% and false positive rate was 1%. This means that although 80% of the high influencers were missed, almost all subscribers who were predicted to be high influencers were truly high influencers. The scores as predicted by the regression are shown in Fig. 3 along with the ground truth. As seen in the figure, the predicted influence score for the high influencers in on average higher than the predicted scores of low influencers.

3) *Results of method 2- Tree Clasification*

For the same reason mentioned for linear regression, the Decision Tree also put double the weight on false positives resulting in a weaker overall performance, but a stronger true positive rate. With this weighting, the false negative rate was almost 50% and the false positive rate was 2%. This means that half of the high influencers were missed, but 98% percent of the selected ones were truly high influencers.

4) *Results for the combination of the two methods*

To further reduce the false positive rate, we selected a subscriber as a high influencer only if she was classified as a high influencer by both the regression model and the Decision Tree. That reduced the false positive rate to 0.4%. Putting the same condition on non-influencer to reduce the false negative rate reduced the false negative rate to 35%. These results are summarized in Table 3.

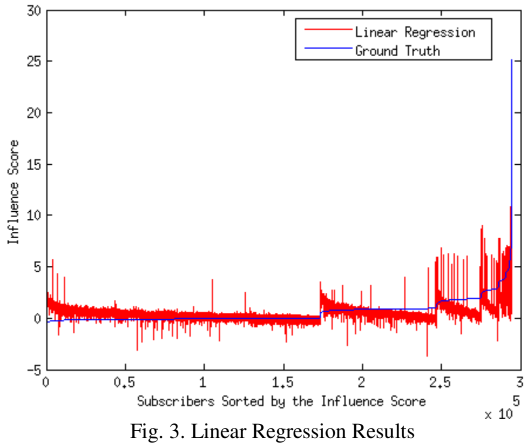

Fig. 3. Linear Regression Results

TABLE 3. MISSCLASSIFICATION RATES FOR CLASSIFYING THE INFLUENCERS

Method	**False Positive Rate**
Linear Regression	0.01 (1%)
Decision Tree	0.02 (2%)
Intersection of the two	0.004 (0.4%)

5) *Lift Curve*

Since the purpose of this prediction was to capture the subscribers who will lead many subsequent churners, we extracted the subscribers that were classified as high influencers and their subsequent churners. We also selected the same number of random subscribers and their subsequent churners. The lift curve depicted in Fig. 4 compares the number of captured subsequent churners. As we see, our prediction method significantly increases the number of subsequent churners per selected first churner compared to random selection.

V. DISCUSSION AND CONCLUSION

In this paper, we explored the issue of churn influencers and their predictability, and focused on the high influencers whose churn will result in the churn of many subsequent churners. We proposed a method to enhance the current churn prediction/treatment method by identifying the high churn influencers in the social network of wireless subscribers.

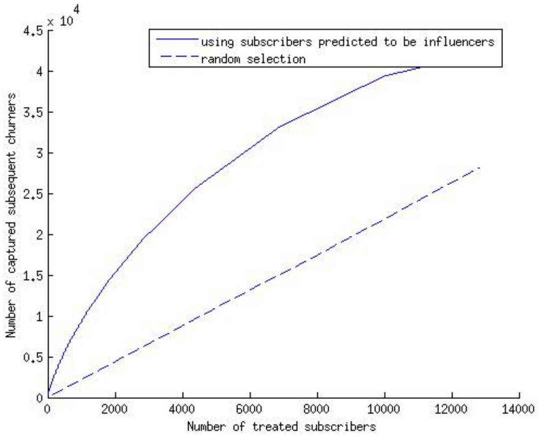

Fig. 4. The Lift Curve

Analyzing real churn event data and call data records of wireless subscribers, we extracted and selected call features that had the ability to predict the high influencers with flexible true positive rates. Our tests and training performed on a major wireless carrier data achieved over 99% true positive rate (precision) with acceptable trade off with the true negative rate (recall). It also significantly increased the number of captured subsequent churners.

This also means that by treating such influencers, the wireless carrier can prevent the churn of subscribers whose churn would be followed by the churn of many subsequent churners. That could save the wireless carrier from having to spend money on treating the subsequent churners as wireless carriers do not have a big enough budget to treat every likely churner. However, before we come to this conclusion, we should note the following points.

- Gaining the great benefit of saving multiple subscribers by treating only one subscriber (the influencer) relies on the assumption that winning the influencer's satisfaction back will eliminate their churn influence on their social network. At this point we only know that the influencer's churn is correlated with the churn of the subsequent churners. However, it may be that even after saving the influencers, the subsequent churners will still churn, for example, because the influencer has already had his adverse impact on them before being treated. The effectiveness of saving the influencer as a means of saving the subsequent churners would be the focus of future work.

- We only analyzed wireless 'call' data records to build the social graph and to extract the predictive features. The advantage of call data is that the data is already available to wireless carriers, and anyone who ones a mobile phone produces records in this data set. Nevertheless, we could have collected SMS data as a secondary or as an alternative data source. SMS records may represent closer and stronger relationship between the subscribers which could result in a richer social graph and more relevant features.

- Wireless carriers want to make the best of their limited churn treatment budget. Generally this budget covers a very small subset of subscribers. Therefore, we focused on lowering the false positive rate so that non-influencers will not be classified as an influencer and treated instead of them. For this purpose, we had a significant trade off on the overall performance and true negative rate of our prediction. Carriers with a higher treatment budget may choose to optimize the overall performance and capture as many influencers as possible at the cost of treating some non-influencers.

While the above points can be considered in future work as alternatives, our study highlighted the gap in current churn prediction methods, offered and enhancing solution by considering and identifying the churn influencers, proposed a framework for prediction of such influencers, and tested the validation of this framework using real world data. Wireless carriers can benefit from such solutions by both enhancing their churn prediction methods and smart selection of likely churners to be treated.

ACKNOWLEDGMENT

We would like to thank Mark Evans for helping us with data collection.

REFERENCES

[1] V. YeshWanth and M. Saravanan. Churn analysis in mobile telecom data using hybrid paradigms. NetMob 2011, 2011.

[2] V. Yeshwanth, V. Raj, and M. Saravanan. Evolutionary churn prediction in mobile networks using hybrid learning. In Twenty-Fourth International FLAIRS Conference, 2011.

[3] W. Verbeke, T. Verbraken, D. Martens, and B. Baesens. Relational learning for customer churn prediction: the complementarity of networked and non-networked classifiers. In NETMOB, 2011.

[4] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. TKDD, 2012.

[5] S. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In SIGKDD, 2012.

[6] S. Ye and S. Wu. Measuring message propagation and social influence on twitter. com. Social Informatics, 2010.

[7] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. Nanavati, and A. Joshi. Social ties and their relevance to churn in mobile telecom networks. In EDBT, 2008.

[8] D. Kempe, J. Kleinberg, and E`. Tardos. Maximizing the spread of influence through a social network. In SIGKDD, 2003.

[9] W. Chen, C. Wang, and Y. Wang. Scalable influence maximization for prevalent viral marketing in large-scale social networks. In SIGKDD, 2010.

[10] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In SIGKDD, 2009.

[11] E. Mossel and S. Roch. On the submodularity of influence in social networks. In STOC, 2007.