Word-of-Mouth algorithms: What you don't know will hurt you

Abstract

Word-of-mouth communications has been shown to play a key role in a variety of environments such as viral marketing and churn prediction. A family of algorithms, generally known as information spreading algorithms has been developed to model such pervasive behavior. Although these algorithms have produced good results, in general, they do not consider that the social network reconstructed to model the environment of an individual is limited by the information available. In this paper we study how the missing information (in the form of missing nodes and/or missing links) affects the spread of information in the well-known Dasgupta et al. (2008) algorithm. The results indicate that the error made grows logarithmically with the amount of information (links, nodes or both) unknown.

1 Introduction

"Word-of-Mouth" algorithms or information diffusion algorithms originally appeared in social sciences [Goldenberg et al., 2001] and are based on the idea of using a social interaction network to model the flow of information and influence. The concept groups a variety of algorithms that model the pervasive word of mouth behavior and are typically based on the spreading activation method used in cognitive psychology. These family of algorithms has been successfully used in a variety of areas, including viral marketing [Richardson et al., 2002], churn prediction in telecommunications networks [Dasgupta et al., 2008], information retrieval and to model some behaviors such as trust [Ziegler et al., 2004] and spread of epidemics.

Nevertheless "word-of-mouth" algorithms implicitly assume that the social network used for the spreading of influences is completely known, i.e. they do not model or consider the error introduced by missing nodes and missing links. In general, when applying these algorithms to real scenarios the information known is to some extent limited, for example: (1) for churn prediction the social network reconstructed is the one provided by the calls of clients, but no other interaction, such as face to face communication, IP phones, instant messenger, etc. is reflected; and (2) when modeling viral marketing the information is typically collected from review-product sites [Richardson et al., 2002], but again, no other influences and/or players are captured by the network.

Because of the nature of the algorithms, which use the structure of the network to spread influence, this limited view of the network should impact the final results, mainly because: (1) a given individual will receive direct and indirect influence from a variety of (even random) individuals from which no information from the data source used is available or if some information is available, it is very limited; for example, in a telecommunication network a client can receive influence from another individual face to face, which will not be reflected in the Call Detailed Records (CDRs), or form a phone call of an individual that uses a different carrier, from which a limited amount of information is known; and (2) even between the individuals included in a dataset not all possible links are necessarily known; for example in a telecommunication network the data will capture the phone calls between individuals, but no other type of interaction such as social interaction or IP phone calls, i.e. the lack of an edge between two individuals in a CDR-generated graph does not imply that there is no influence between those two individuals by another means of communication.

Figure 1 presents a simplified example of these cases for a telecommunication operator. The figure on the left represents the network that can be reconstructed using phone calls where some of the users identified are part of the network's service (circles) and others are from other networks (squares). From these users only part of the links are known. The figure on the right shows the same graph but with the missing links (web 2.0 services for example) and nodes from which no information is available in the original data, but that will play a role when information spreading and influence.

These concepts of missing links and missing nodes are related to some extent to the work done in the area of weak ties. Weak ties refer to the fact that individuals are often influenced by other individuals with whom they have tenuous or even random relations [Goldenber et al., 2001], while strong ties are defined by an individual's personal network. In the environments in which information spreading algorithms are applied, the strong ties defined by the information used will be reconstructed, nevertheless not all weak ties will be reflected in the network reconstructed as a result of data filtering. There also would be weak ties and strong ties originating from other sources of interaction that will not be reflected. It has been shown that the influence of the weak ties can be as strong as the influence of strong ties in word-of-mouth modeling [Goldenberg et al., 2001]. It is then clear that if no information of weak ties or ties originating from other data resources is considered in the networks an error

of the estimation will be introduced by information spreading algorithms.

The goal of this paper is to measure the impact of missing information (links and/or nodes) when applying diffusion information algorithms and to model these errors in the context of a telecommunication network. In the context in which information spreading algorithms are applied, and considering that the final value of energy of the nodes is used to make predictions, this error should be considered when reporting the final results, especially for sensitive applications such as churn prediction or the spread of epidemics.

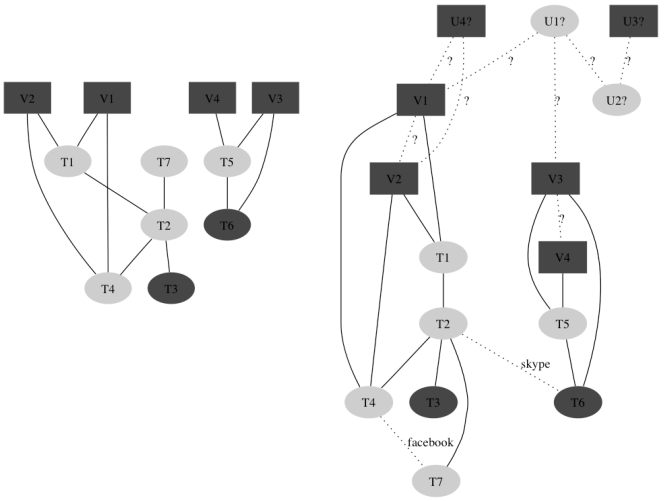

Figure 1. Example of (left) the social network from a telecommunication perspective and (right) the actual social network, with all the extra interactions and missing nodes.

2 Related Work

Although there are a variety of information spreading algorithms in the literature [Dasgupta et al., 2008; Goldenberg et al., 2001], in general all of them are based on the spreading activation method used in cognitive psychology to model the fan out effect. These algorithms are based on the following (simplified) steps:

1. Nodes that are activated are given an activation value (typically one) while non-activated nodes are given a value of 0. This activation value represents the influence energy that the node can spread over the network. In the case of viral marketing it represents the energy (influence) of people that have bought a product, in the case of churn prediction it represents the energy (influence) of the people that have "churned".
2. The set of nodes that are activated transfer part of their energy to neighboring nodes considering a spreading or propagation factor which indicates which part of the energy is transferred and a distribution function which indicates the percentage of energy that is transferred to the neighboring nodes. In viral marketing or churn modeling the spreading factor is a variable, while the distribution function is given by the weight of each link that leaves the activated node.
3. Step 2 is repeated until the variation of the energy in the nodes is below a threshold and no new nodes are activated.
4. Once the energy distribution has converged, the nodes over a given threshold of energy are considered to be "infected". In the case of churn prediction this set of nodes are considered to be at risk of switching to another carrier, while in viral marketing these are considered to be susceptible to buying a product.

The work presented in Dasgupta et al. (2008) presents the use of a spreading activation algorithm for demonstrating the relevance of the social network of a client for churn prediction. The results indicate the influence of the social network of a user in deciding to churn. Nevertheless this churn application of a spreading activation algorithm, and, to the best of our knowledge, the application of any activation algorithm does not consider information that may exist outside the dataset used to construct the network.

The work of Lahiri et al. (2008) measures how changes produced by the evolution in time of dynamic networks impact the accuracy of the prediction of the spread of the Independent Cascade Model [Goldenberg et al., 2001]). The authors focus more on the total size of the spread (how many nodes are affected), whereas we focus on how the spread was, i.e. which particular nodes are affected. Another difference with the present work is that the Independent Cascade Model is inherently stochastic whereas Dasgupta's et al. model is deterministic.

In this paper we will consider Dasgupta et al.´s (2008) algorithm to study which error is actually introduced by the spreading activation algorithm in a telecommunication network considering that some of the information (links, nodes or both) is missing.

3 Data Set

Cell phone call data (CDR: Call Detail Records) from a single carrier was obtained for a number of users close to 50,000. The data was collected from a neighborhood of a major city over a period of six months. The originating number and the destination number of the CDR were both encrypted. From all the information contained in a CDR only the originating encrypted number, the destination encrypted number, the aggregate duration of the calls and the frequency of calls were considered for the study. The data set contained information only for voice calls between users (no SMS or other forms of communication were used). The sample included only residential customers (no business or corporate cell phones), and only calls above 1 second and that had no errors when the called finished were considered. Only aggregated information about the phone calls of each individual was considered

Calls were used to create a network with directed edges. Two nodes X and Y were linked if there was a phone call between X and Y, where the origin and destination of the phone call define the orientation of the edge. Each link is given a weight, normalized in seconds, defined by the total duration of the phone calls from X to Y over the entire 6 months. Note that two nodes can be linked by two edges one in each direction and with a different weight. A total number of links close to 120,000 define the network.

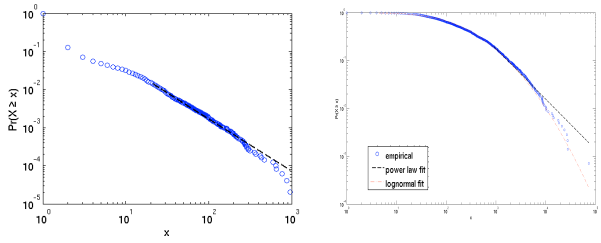

Figure 1. (left) log-log distribution of the degree distribution and (right) of the duration distribution of the original network.

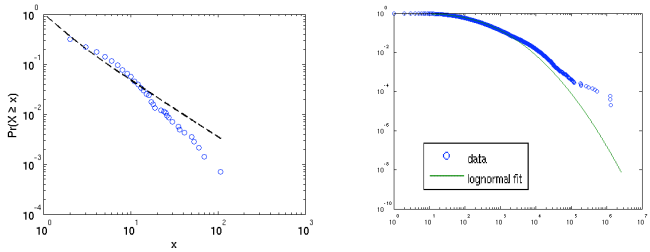

Figure 2. (left) log-log distribution of the degree distribution and (right) of the duration distribution of the sampled network.

Due to the computational complexity of the experiments that will be described in the next section, the fist step was to extract a representative sub-network from the original data. A random walk sampling technique proved to be best to reproduce the original network behavior information spreading wise [Leskovec et al. 2006; Becchetti et al., 2006]. The mechanism started at a random node and followed the edges at random until a given number of nodes were collected. All existing edges between those nodes were added to the network. The resulting sample sub-network contained 1,408 nodes and 3,910 edges.

Figure 1 presents the log-log representation of the distribution degree (left) and the duration distribution (right) of the original network and Figure 2 of the sampled network. In both cases, the degree distribution has a power law fitting with a slope of 2.3 in the original network and of 2.1 in the sample network. Also the duration distribution has a log-normal behavior in both cases, with μ=5.02 σ=1.77 in the original network and μ =5.53 and σ=1.67 in the sample network. These values indicate that the sampled network has two important t macroscopic statistical properties similar to the original network. Also, these values are in agreement with other values reported in the literature for characterizing cell phone telecommunication networks [Seshadri et al., 2008; Onnela et al., 2007], although some of these papers argue that the approximation can be improved with a Double Pareto LogNormal fit [Seshadri et al., 2008].

4 Methodology

Three experiments were run to measure the impact of missing information in information spreading algorithms in the context of a telecommunications network: (1) evaluate the impact of missing links, i.e. links not captured by the telecommunication networks such as physical interactions, IP phones, social network websites, phone calls made with other competitor networks, etc.; (2) evaluate the impact of missing nodes, representing the fact that a telecommunication company only sees its own part of the global telecommunication network, and (3) evaluate the impact of missing nodes and missing links.

In order to run these experiments the algorithm described in Dasgupta et al., (2008) was used with the sampled network presented in the previous section. As with any other information spreading algorithm there are some parameters that need to be defined, and in this case we have used the ones recommended in Dasgupta et al., (2008): nodes that churn are assigned an energy value of 100, nodes that are not churners are assigned a energy of 0, the propagation factor is 0.25, the spreading stops when the relative change of influence in each node is below 1% and the spreading factor (the weight on the links) is defined by the total time talked.

To run the experiments, initially a fixed percentage of random activated users (in the telecommunication context it would mean users that have churned) are considered. After that, the information spreading algorithm is run over the original sampled network. In the end each node of the network will have an energy level, and we consider this distribution of energy the Ground Truth (GT). Note that in Dasgupta's et al. (2008) they know which nodes are initially activated (the churners) whereas we just choose some random nodes.

After the GT has been obtained, a given number of elements are randomly deleted from the original network: (a) for the first experiment we randomly delete a percentage of existing links; (b) for the second experiment we delete all the links from a randomly selected percentage of nodes, thus isolating the nodes – which for the error computation is exactly the same as deleting the nodes; and (c) in the third experiment we delete a percentage of links and a percentage of nodes. The resulting network is called *S*. Once the selected information has been deleted, the spreading algorithm is run over *S*, which will assign a given energy to each one of the nodes.

The error in the information spreading algorithm is measured as the root mean squared deviation (RMSD) obtained from subtracting the final value of energy assigned by the algorithm to each node of the *GT* network from each corresponding node of the S network. Being $N=1,408$ the number of nodes of the *GT* and *S* networks, GT_i the level of energy of node i in *GT* after the application of the information spreading algorithm and S_i the value of energy assign by the

algorithm to node *i* in the *S* network, the error introduced by missing information is defined as:

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\|GT_i - S_i\right\|} \qquad (1)$$

The error is presented as the difference in energy instead of as the error in the number of activated nodes. In general, in spreading algorithms the final prediction (churners in churn prediction, infection in spreading of viruses, propensity to buy a product in viral marketing, etc.) are identified as those having a final value of energy bigger that a threshold. To avoid considering this threshold when measuring the error, whose definition may me arbitrary depending on the particular application, we focused on the difference of the final value of energy between the energy of the individual nodes between *GT* and *S*.

In order to avoid possible artifacts from the randomly activated nodes or from the information randomly deleted (links, nodes or both) each experiment was run 100 times, and the final RMSD was reported with a mean and a standard deviation.

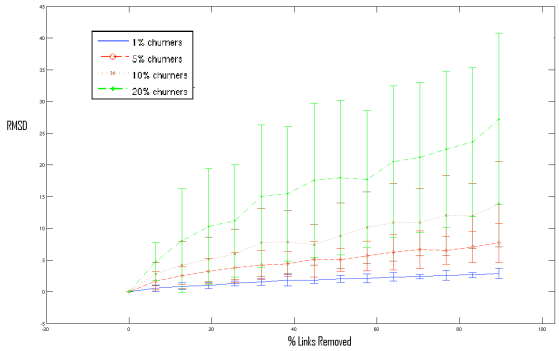

Figure 3. Impact in terms of RMSD, Y axis, of the percentage of missing links for 1%, 5%, 10% and 20% activated nodes (churners).

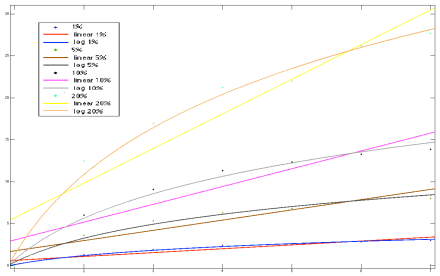

Figure 4. Error fitting of missing links using linear and logarithmical regression.

5 Results Analysis

Figure 3 presents the impact of missing links in the information spreading algorithm. The experiment was run considering 1%, 5%, 10% and 20% of randomly activated users, which are represented by each one of the curves. Note that an activated user is the one who can spread some information to its neighbors. For each one of these cases experiments were run 100 times, from 0% of missing links to 90% of missing links in 5% increases (X axis). The results in each case are reported with the mean and the standard deviation of the RMSD (Y axis). Figure 4 presents the function approximation (linear and logarithmical) that best fits the error. In all cases the best fit is a logarithmic curve, having a smaller SSE (sum of squared errors) than the linear regression. It can be seen that for a number of activated users ranging from 1% to 5%, the missing links introduce a RMSD error in the range 0-5%, but higher percentages of activated users have a higher error which increases logarithmically with the percentage of missing links. Also the RMSD variance increases with the number of missing links.

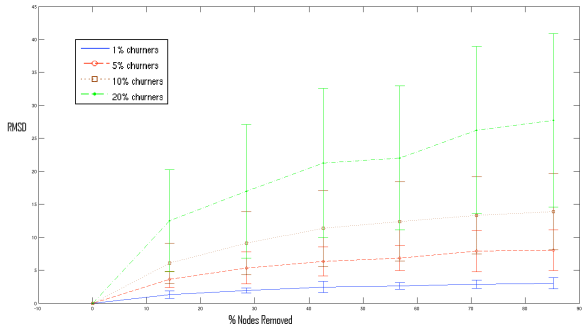

Figure 5. Impact of the percentage of missing nodes for 1%, 5%, 10% and 20% activated nodes (churners).

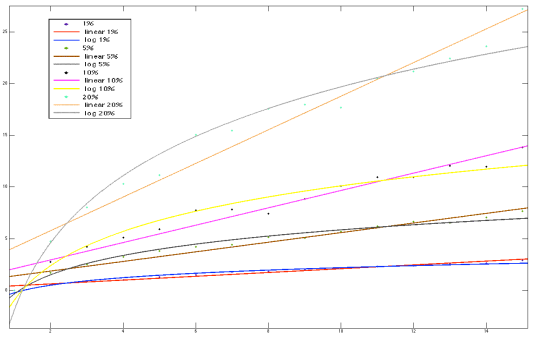

Figure 6. Error fitting of missing nodes using linear and logarithmical regression.

Figure 5 presents the impact of missing nodes in the information spreading algorithm. The experiment was run considering 1%, 5%, 10% and 20% of randomly activated users, which are represented by each one of the curves. For each one of these cases experiments were run 100 times from 0% to 85% in 15% increases (X axis). The results in each case are reported with the mean and the standard de-

viation of the RMSD (Y axis). Figure 6 presents the regression (linear and logarithmical) that best fits the error. As in the previous case there is a logarithmical behavior of the error, which increases with the number of missing nodes. Also the RMSD variance increases with the number of missing nodes.

Although the previous two experiments present interesting results, they do not reflect a real situation because only nodes or links are missing. Figures 7, 8 and 9 present the error when both situations happen.

Figure 7 presents the RMSD error when considering 1% of activated users, for 0% removed nodes, 30% removed nodes and 60% removed nodes (each one of the curves) for a percentage of links removed that evolves from 0% to 90% once the nodes have already been removed (X axis). Figure 8 presents the same experiment but when the number of activated nodes is 20%. In both cases the curve corresponding to 0% removed nodes correspond to the curves presented in Figure 3.

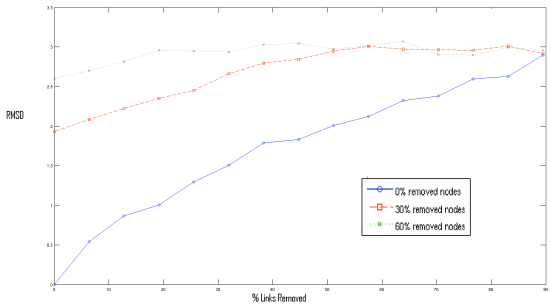

Figure 7. Impact of the percentage of missing links, for 1% activated nodes and 0%, 30% and 60% missing nodes.

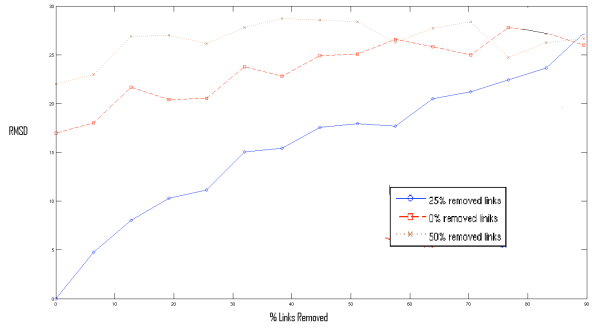

Figure 8. Impact of the percentage of missing links for 20% activated nodes and 0%, 30% and 60% missing nodes.

Both figures indicate how an increase in the number of activated nodes implies an increment of the RMSE of the final energy of the networks. As for the percentage of missing nodes, the error model increases logarithmically with the number of missing nodes. From a practical perspective these results imply that if the number of users that buy a product or the number of users that churn is high, the RMSD, even if the number of missing nodes is low, will be considerable high and the predictions made by the information spreading algorithm should be corrected with other information of the individuals not originating from their social network.

Thus if churn ratio is high then prediction models based solely on word of mouth algorithms may not be accurate enough for sensitive applications. Therefore complementary information about the individual behavior in the form of user models as well as link prediction algorithms may be necessary.

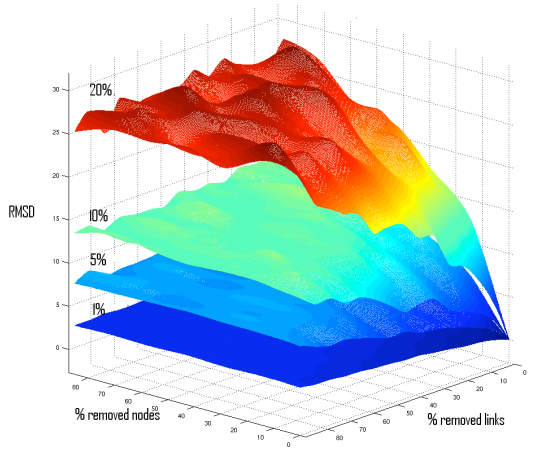

Figure 9. Representation of RMSD for different 1%, 5%, 10% and 20% activated nodes (churners) and different percentages of nodes and links missing (removed).

Figure 9 presents the RMSE considering a missing links (ranging from 0% to 90%, X axis)), missing nodes (ranging from 0% to 90%, Y axis) and a percentage of activated users of 1%, 5%, 10% and 20% (each one of the surfaces, Z axis). These results confirm the previous findings, the logarithmic behaviour of the RMSD error with the percentage of links and percentage of nodes missing, and the increase in the error when the percentage of activated users increases. For a reduced number of activated users, 1%, the error for the prediction is not relevant, nevertheless for higher values the error will have a negative impact in the prediction, as the 5% active users curve already shows.

Figure 10 also shows that the RMSD error introduced is defined by a logarithmic surface where the variables are the percentage of links and nodes missing. This figure can be used to estimate the error introduced by the algorithm in the energy distribution process. The number of activated users is in general known, i.e. number of churners in a specific period, number of users that have contracted a virus or number of users that have acquired a product. The number of missing nodes can also be to some extent estimated. For example, in a telecommunication network the number of clients is known, and the total number of users with phone (potential clients) is also known. The percentage of links missing is much harder to estimate. Thus in general the estimation of the RMSD error introduced is defined by a loga-

rithmic curve where the variable is the percentage of links missing.

5.1 Other Experiments

The experiments reported in the previous section considered directed edges weighted by the total amount of time that two clients had a contact. Nevertheless other weightings can be used for characterizing links. The same set of experiments described in the previous section were run using frequency of calls as weights, and just plain connectivity (no weights) obtaining very similar results. Also, presented results are consistent with our experiments on sparser networks.

6 Conclusions

This paper has presented RMSD experimental results of the error introduced by a particular information spreading algorithm considering that the network used has missing information in the form of nodes and links. From an application perspective the study highlights the fact that for any application only partial information will be known. Because the influence of missing ties and nodes is as relevant as the influence of known ties and nodes, an error in the distribution of the influence (energy) will be introduced. This fact implies an error in the prediction made by the algorithm. The results, summarized in Figure 10, indicate a logarithmic error in the number of nodes and the number of links missing. Figure 10 can be used to estimate the error introduced considering that the number of activated nodes is known, and the total number of missing nodes may be roughly estimated. While missing links can be to some extent estimated [Liben-Nowell et al. 2007; Newman et al. 2008], the source and destination of missing nodes are much more difficult to estimate and further work in this area is needed.

From a more general perspective, another conclusion of our study is that although information spreading algorithms are a powerful tool for modeling behaviors such as churning, spreading of viruses and viral marketing, the predictions obtained, if a relevant part of the information is missing or if there is an elevated number of activated nodes, specially for those nodes that have a final value of energy close to the threshold used can be incorrect. In order to make an improved prediction other sources of individual information should complement the "word-of-mouth" approach. In the case of churn prediction some examples would be calling patterns [Wei et al., 2002] or complaint data [Hadden et al., 2006].

The results that have been presented have been obtained from a telecommunications network and may not generalize to all types of network. Further experimentation is needed to understand how the error introduced depends on the type or the architecture of the network in order to estimate a generic error model introduced by information spreading algorithms.

References

[Becchetti et al., 2006] L. Becchetti, C. Castillo, D. Donato. A Comparison of Sampling Techniques for Web Graph Characterization. In *Proc. of LinkKDD* 2006.

[Dasgupta et. al., 2008] K. Dasgupta, R. Singh, B. Viswanathan, S. Mukherjea, A. Nanavati and A.Joshi. Social Ties and their relevance to churn in mobile telecom networks. In *Proc. of EDBT 2008*, pp. 668-667.

[Goldenberg et al., 2001] J. Goldenberg, B. Libai, E. Muller. Talk of the Network: A Complex System Look at the underlying process of Word-of-Mouth. *Marketting Letters 12(3)*, pp. 211-223.

[Hadden et al., 2006] J. Hadden, A. Tiwari, R. Roy and D. Ruta. Churn Prediction using Complaints Data. In *Proc. World Academy of Science, Engineering and Technology* Vol. 13, May 2006.

[Lahiri et al. 2008] M. Lahiri, A.S. Maiya, R. Sulo, Habiba, T.Y. Berger-Wolf. The Impact of Structural Changes on Predictions of Diffusion in Networks. *ICDM Workshop on Analysis of Dynamic Networks*. December 2008.

[Leskovec et al. 2006] J. Leskovec, C. Faloutsos. Sampling from Large Graphs. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[Liben-Nowell et al. 2007] Liben-Nowell, D. and Kleinberg, J. The Link-Prediction Problem for Social Networks. *Journal of the American Society for Information Science and Technology*, Vol. 58, No. 7 . pp. 1019, 2007.

[Newman et al. 2008] Hierarchical structure and the prediction of missing links in networks, Aaron Clauset, Cristopher Moore, and M. E. J. Newman, Nature 453, 98–101 (2008).

[Onnela et al., 2007] J.P. Onnela, J. Saramaaki, J. Hyvoven, G.Szabo, M. Argollo, K. Kaski and A.L. Barabasi. Structure and tie strenghs in mobile communication networks. *New Journal of Physics* 9, 2007.

[Richardson et al., 2002] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *Proc. ACM CIKM*, Edmonton, Alberta, Canada 2002.

[Seshadri et al., 2008] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos and J. Leskovec. Mobile Call Graphs: Beyond Power-Law and Lognormal Distributions. In *Proc. KDD 2008*, pp. 596-604.

[Wei et al., 2002] C. Wei and I. Chiu. Turning telecommunications calla details to churn prediction: A Data Mining approach. *Expert Systems with Applications*, 23 pp. 103-112.

[Ziegler et al., 2004] C. Ziegler and G. Lausen. Spreading activation models for trust propagation. In *Proc. IEEE Int. Conf. On e-technology, e-commerce and e-service*. Taipei, Taiwan 2004.