Word-of-Mouth Algorithms: What You Don't Know Will Hurt You

Manuel Cebrian
Telefonica Research / MIT
Cambridge, MA, USA
cebrian@mit.edu

Enrique Frias-Martinez
Telefonica Research
Madrid, Spain
efm@tid.es

ABSTRACT
Word-of-mouth communication 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.

Categories and Subject Descriptors
J.4 [**Social and Behavioral Sciences**]: *Sociology*. B.4.4 [**Input/Output and Data Communications**]: *Simulation*.

General Terms
Performance, Economics, Experimentation, Security, Human Factors.

Keywords
Social Network, Word-of-Mouth, Simulation, Link.

1. INTRODUCTION
Word-of-Mouth algorithms implicitly assume that the social network used for the spreading of influence 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: 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, call from/to competitors, etc. is reflected.

ICMI-MLMI'09, Workshop on Multimodal Sensor-Based Systems and Mobile Phones for Social Computing. November 6, 2009 Cambridge MA, USA.

Figure 1 presents a simplified example of these cases for a telecommunications operator. The figure on the left represents the network that can be reconstructed using phone calls where some of the users identified are some 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 in information spreading and influence.

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 telecommunications 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.

2. 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.

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. 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.

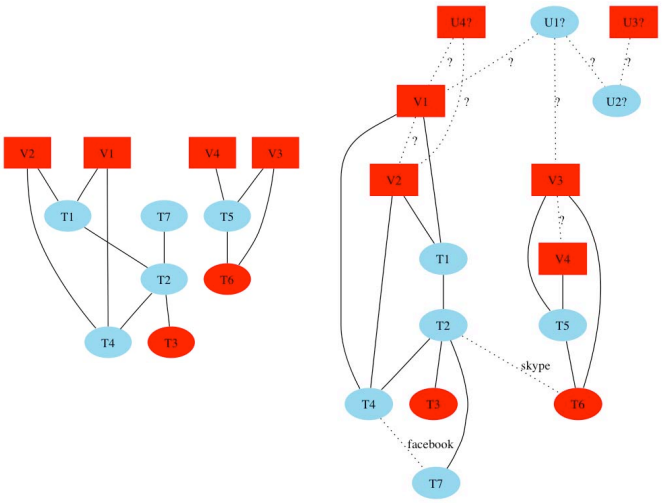

Figure 1. Example of (left) the social network from a telecommunication perspective and (right) the actual social network, with all the extra interactions and missing nodes.

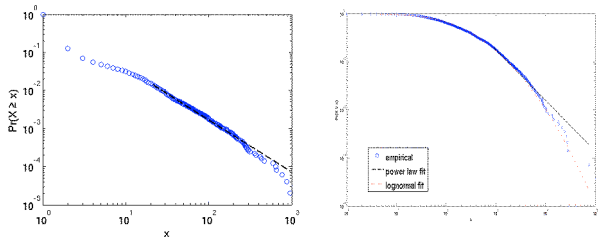

Figure 2. (left) log-log distribution of the degree distribution and (right) of the duration distribution of the original network.

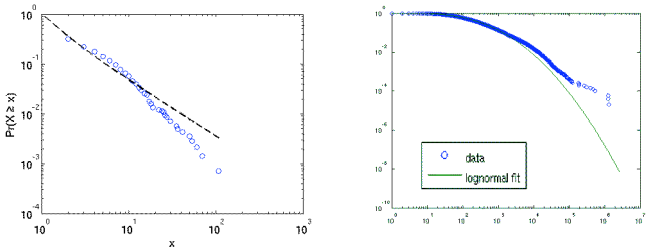

Figure 3. (left) log-log distribution of the degree distribution and (right) of the duration distribution of the sampled network.

A random walk sampling technique proved to be the best to reproduce the original network behavior information spreading wise [1]. 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 2 presents the log-log representation of the distribution degree (left) and the duration distribution (right) of the original network and Figure 3 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 lognormal 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 [3,4] although some of these papers argue that the approximation can be improved with a Double Pareto LogNormal fit.

3. 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, (2) evaluate the impact of missing nodes, and (3) evaluate the impact of missing nodes and missing links.

In order to run these experiments the algorithm described in [2] 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 [2]: 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 [2] 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, GTi the level of energy of node i in GT after the application of the information spreading algorithm and Si the value of energy assign by the algorithm to node i in the S network, the error introduced by missing information is defined as:

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

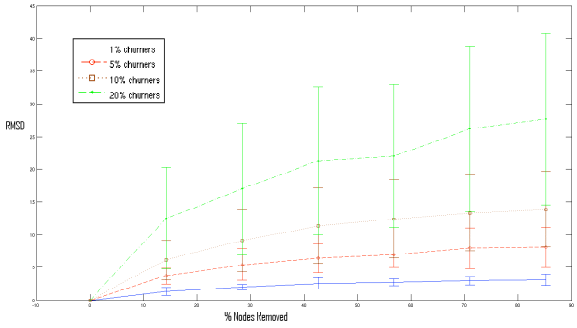

Figure 4. Impact in terms of RMSD, Y axis, of the percentage of missing links for 1%, 5%, 10% and 20% activated nodes (churners).

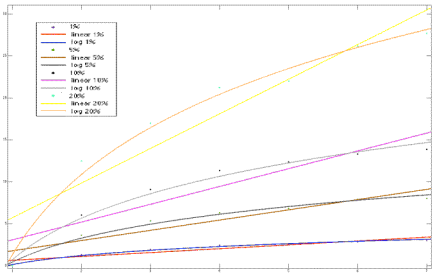

Figure 5. Error fitting of missing links using linear and logarithmical regression

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 higher than 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.

4. RESULTS AND DISCUSSION

Figure 4 presents the impact of missing links in the information spreading algorithm[1]. 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).

[1] The experiments reported in this section consider 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.

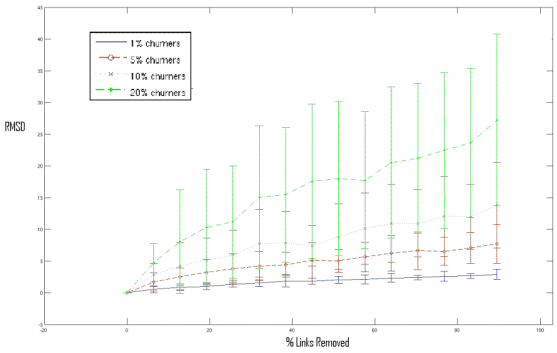

Figure 6. Impact of the percentage of missing nodes for 1%, 5%, 10% and 20% activated nodes (churners).

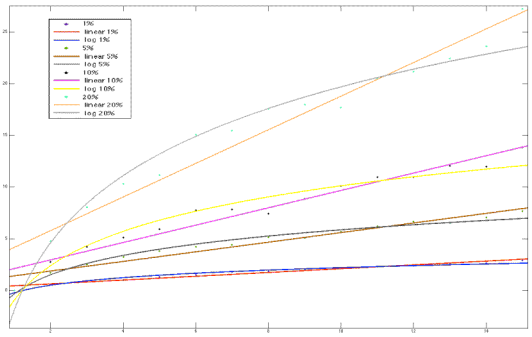

Figure 7. Error fitting of missing nodes using linear and logarithmical regression.

The results in each case are reported with the mean and the standard deviation of the RMSD (Y axis). Figure 5 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 6 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 deviation of the RMSD (Y axis). Figure 7 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.

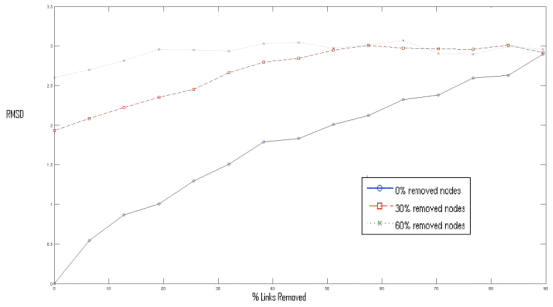

Figure 8. Impact of the percentage of missing links, for 1% activated nodes and 0%, 30% and 60% missing nodes.

Although the previous two experiments present interesting results, they do not reflect a real situation because only nodes or links are missing. Figures 8, 9 and 10 present the error when both situations happen.

Figure 8 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.

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 considerably 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 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

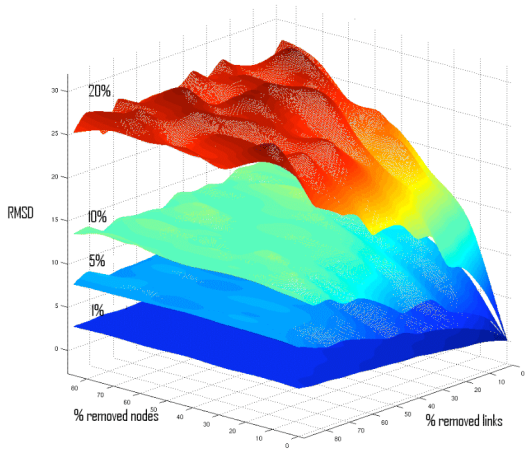

Figure 9. Impact of the percentage of missing links for 20% activated nodes and 0%, 30% and 60% missing nodes.

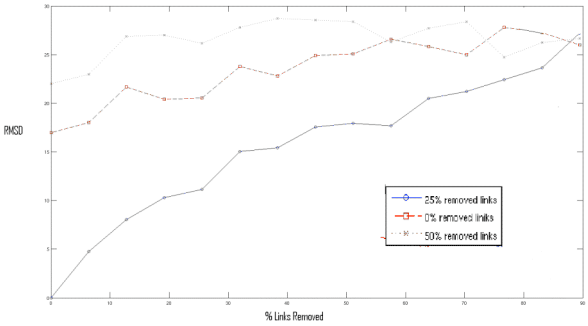

Figure 10. Representation of RMSD for different 1%, 5%, 10% and 20% activated nodes (churners) and different percentages of nodes and links missing (removed).

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

5. REFERENCES

[1] J. Leskovec et al., 'Sampling from Large Graphs', *KDD,* 2006.

[2] K. Dasgupta et al., 'Social Ties and their relevance to churn in mobile telecom network', *EDBT,* 2008.

[3] M. Seshadri et al., 'Mobile Call Graphs: Beyond Power-Law and Lognormal Distributions', *KDD,* 2008.

[4] J-.P. Onnela et al., 'Structure and tie strenghs in mobile communi- cation networks', *New Journal of Physics 9, 2007.*