# *How did you get to know that?* A Traceable Word-of-Mouth Algorithm

**Manuel Cebrián, Enrique Frías-Martínez, Heath Hohwald, Ruben Lara and Nuria Oliver**
Data Mining and User Modeling Group, Telefonica Research
Emilio Vargas 6, 28043, Madrid, Spain
{manuelc, efm, heath, rubenlh, nuriao}@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 virus spreading. A family of algorithms, generally known as information spreading algorithms or word-of-mouth algorithms, has been developed to characterize such behavior. However, they have limitations, including the inability to: (1) capture *when* the communications or contacts take place and (2) explain *where* the influence comes from. These drawbacks have limited the studies about how the spreading of influence takes place in social networks. In this paper, we present a new word-of-mouth algorithm that considers the temporality of the communications and keeps track of how influence travels over the social network. We validate the proposed algorithm via simulations of word-of-mouth traces on call detailed records, in order to model how influence spreads. Our results indicate that (1) static factors of social networks are not enough to model influence and (2) there seems to be statistical invariants of how influence spreads in a network.

## 1 Introduction and Motivation

Word-of-mouth (WoM) or information diffusion algorithms appeared in the social sciences [1] and are based on the idea of using a social interaction network to model the flow of information. This family of algorithms has been successfully used in a variety of areas, including viral marketing [2], epidemiology, and churn prediction [3]. In each of these applications, the concept of information that passes from one node to another when an interaction takes place has different semantic implications: in viral marketing and churn prediction, it is seen as influence, whereas in epidemiology, it is the viral charge. Note that in this paper, the concepts of energy, information or influence spread are used with the same semantics.

In typical WoM algorithms [1][3], inferring the structure of the social network and modeling the diffusion of information are considered two different problems that are solved using different algorithms: first, a social network is constructed, followed by an information spreading algorithm, such that the *order* with which nodes interchange information or influence is not considered. However, in the case of viral spread (*e.g.* marketing, human or computer viruses, etc.), *when* the interactions take place is very important, because an individual will propagate information only if he or she has previously received it [4]. Also, in most of the applications where WoM algorithms are applied it is relevant to know who is responsible for each node's activation, *i.e.* the causality of the influence. Therefore, the algorithms should be able to identify how the energy that activates a node reached that node. For example, causality identifies which nodes are responsible for the churn of a node in churn prediction or for the acquisition of a product in viral marketing. This analysis characterizes the importance –from an information spread perspective-- of each node in the network. Traditionally, the identification of important nodes has been tackled with the concept of social leaders or alpha users [5]. Nevertheless, the concept of an alpha user is typically seen as a static value defined by the architecture of the associated social network. The algorithm proposed in this paper models not only the importance of the nodes of a network, but also the dynamic aspects of information spread. The contributions of this paper are twofold: (1) we propose a novel information spreading algorithm that considers the order in which interactions take place and models how a node receives influence from its neighbors; and (2) we validate the algorithm by means of simulations on a real cell phone network.

## 2 Related Work

Generally, the most popular WoM algorithms [1][3]can be summarized in the following steps:

1. Activated nodes are given an activation value (typically 1) while non-activated nodes are given a value of 0.
2. The set of nodes that are activated transfer part of their energy to neighboring nodes, modulated by a *spreading* or propagation *factor* that indicates which part of the energy is transferred and by a *distribution function* that indicates the percentage of energy that is transferred to the neighboring nodes.
3. Step 2 is repeated until the variation of the energy in the nodes is below a threshold.
4. Once the energy distribution has converged, the nodes with associated energy above a threshold are considered to be "infected".

Note how these algorithms distribute the original energy of the network among all the nodes until the level of energy stabilizes. This implies that the activated nodes, and in general any node that has received energy, lose part of that energy during an interaction. However, this assumption might not necessarily be appropriate for modeling the spread of information. If the energy level represents an influence capacity, the fact that someone gets in contact with someone else does not imply that the original person loses his/her influence. The literature includes studies that manifest the importance of influence in social networks. The work of Dasgupta *et al.* [3] presents the use of an activation spreading algorithm for churn prediction. Lahiri *et al.* [4] measure 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 [1]. Blondel *et al.* [5] show that users within the same social network tend to use the same set of tags in Flickr, thus highlighting the effect of influence.

In our work, we identify the correlation between the level of influence of a node and its static factors (degree, duration, etc.). Our results indicate that the correlation does not fully explain the level of influence. Therefore, we extend our study to model temporality, causality and how influence spreads over the network. Our results show that the spread of influence is determined by parameters that are invariant regarding, among other factors, the set of interactions.

# 3 Traceable Word-of-mouth Algorithm

## 3.1 Notation and Data Structures

The set of $N$ nodes of a network $C$ is defined by $C=\{c_1,...,c_N\}$ or $C=\{c(1),...,c(N)\}$. Each node $C(i)$ has an associated data structure that specifies its initial influence (if any), denoted by $T_i$ or $T(i)$ $i=1...N$. The algorithm uses two data inputs: (1) a set of interactions between nodes and (2) a set of active nodes. The set of interactions is defined by a set of time-ordered vectors $k=1...M$:

$$(src_k, dst_k, len_k) / src_k \in C, dst_k \in C, len_k \in \Re^+ \quad (1)$$

where, $src_k$ and $dst_k$ are the source and destination nodes of interaction $k$, respectively; and $len_k$ is the length of the interaction $k$, typically measured in seconds.

Initially, nodes are classified into two sets: (1) *active* nodes, with $T_i=\{(\beta,\{\})\}$, and where $\beta$ represents their initial influence; and (2) *inactive* nodes. The output of the algorithm consists of $T_i$, $i=1...N$, where each $T_i$ is updated to represent each node's influence and its trace according to the set of previous interactions. After a set of interactions has taken place, $T_i$ is defined as a time-sorted list of influence tuples:

$$T_i = [t_i^1, t_i^2, ...], i = 1..N \quad (2)$$

where each influence tuple ($t_i^j$) of $T_i$ contains a load of influence and its path:

$$t_i^j = (load_i^j, path_i^j), i = 1..N, j = 1..|T_i| \quad (3)$$

The tuple represents an interaction in which a *load* of influence was transmitted from the source node $i$ to the destination node $j$. The *path* represents *how* that influence was transmitted:

$$path_i^j = \{dst, c_i^j(2), c_i^j(3), ..., active\_node\} \quad (4)$$

Where $c_i^j(2), c_i^j(3)\ldots$ are the intermediate nodes that have transferred the influence from the *active_node* to the *dst* node (the set of interactions is ordered in reverse time). The first element of the path, *dst*, can be referred to as $path_i^j(1)$ while the last element can be referenced as $path_i^j(|path_i^j|)$, where $|\chi|$ indicates the length of vector $\chi$. The total influence accumulated by node $i$, $act(c(i))$, is defined as:

$$act(c_i) = \sum_{j=1...|T_i|} load_i^j \quad (5)$$

## 3.2 Algorithm

Figure 1 presents the proposed algorithm to compute the evolution of $T_n$. With each interaction, the source nodes that have an influence greater that 0 transfer influence to the destination nodes, according to the *influence_transfer* function, annotating the path of the transfer in the process. Source nodes do not lose influence in each interaction and destination nodes will only receive influence until their accumulated influence equals $\beta$.

**for** $k=1...M$ *do*
  **if** ($act(src_k)=0$ **or** $act(dst_k)>\beta$)
    **next interaction** ($k=k+1$)
  **else**
    $d=influence\_transfer(len_k)$
    **for** $j=1...|T(src_k)|$
      $T(dst_k) = [T(dst_k), (d \times t^j_{src(k)}(load), (src_k, t^j_{src(k)}(path)))]$
    **end for**
  **end if**
**end for**

**Figure 1.** Algorithm to compute the trace of influence.

The two parameters that need to be defined in the proposed algorithm are $\beta$ and the *influence_transfer* function. A typical value for $\beta$ used in WoM algorithms is 1 [3]. The *influence_transfer* function is a function that considers the length of the interaction between the source node and the destination node and transfers a proportional amount of influence. We have experimented with two influence transfer functions:

(1) A piecewise-linear function:

$$influence\_transfer(len) = \begin{cases} 0, len < 60 \\ len/3600, 60 < len < 3600 \\ 1, len > 3600 \end{cases} \quad (6)$$

(2)A Gompertz function [6]:

$$inluence\_transfer(len) = \begin{cases} 0, len < 60 \\ e^{be^{clen}}, 60 < len < 3600 \\ 1, len > 3600 \end{cases} \quad (7)$$

with *b=-2* and *c=1/600*. The parameters of the two influence transfer functions have been defined according to the characteristics of the interaction data: 53% of calls are less than 1 minute long and 99% of calls are less than 1 hour long, with 46% of the calls between 1 and 60 minutes.

## 4. Characterization of Influence

We propose four concepts to characterize the spread of influence: (1) primary source of influence (PSI); (2) direct source of influence (DSI); (3) intermediary sources of influence (ISI) and (4) influence paths (IP).

### 4.2 Primary Sources of Influence (PSI)

The primary sources of influence of node *A, PSI(A),* are the set of nodes where the energy received by *A* originated from, indicating for each originating node the total amount of energy transferred. Formally, they are defined as:

$$\begin{aligned} &PSI(A) = \{S_i, \varepsilon_i\}_{i=1...|S|}, S \subset activated, \varepsilon_i \in \Re^+ \\ &S = \bigcup_{i=1}^{|T_A|} path_A^i(|path_A^i|) \\ &\varepsilon_i = \sum load_A^j / j = 1..|T_A| \,\&\, path_A^j(|path_A^j|) = S_i \end{aligned} \quad (8)$$

where *S* stores the originating nodes, which will always be a subset of the *active* nodes. The originating nodes of influence of a node *A* are the last elements of each path of $T_A$--see Eq. 4. The union set operator only includes the nodes once, in case of repetitions. The energy transferred by each originating node $S_i$ is obtained as the sum of the loads of each path of $T_A$ where the last element is $S_i$. *PSI* can also be defined globally for all the nodes of a network. The Global Primary Sources of Influence of a network *C*, *GPSI(C)*, is defined as the set of nodes where the energy received by any node of the network originated from. Formally:

$$\begin{aligned} &GPSI(C) = \{activated_n, \varepsilon_n\}_{n=1...|activated|} \\ &\varepsilon_n = \sum_{i=1...N} \sum_{j=1...|T_i|} load_i^j / path_i^j(|path_i^j|) = activated_n \end{aligned} \quad (9)$$

### 4.3 Direct Sources of Influence (DSI)

The direct sources of influence of node *A*, *DSI(A),* are the set of nodes that *directly* transmitted energy to *A* (*i.e.,* in one hop), indicating for each direct node the total amount of energy transferred. Formally:

$$\begin{aligned} &DSI(A) = \{S_i, \varepsilon_i\}_{i=1...|S|}, S \subset C, \varepsilon_i \in \Re^+ \\ &S = \bigcup_{i=1}^{|T_A|} path_A^i(1) \\ &\varepsilon_i = \sum load_A^j / j = 1..|T_A| \,\&\, path_A^j(1) = S_i \end{aligned} \quad (10)$$

The only difference between PSI and DSI is that in the case of DSI, the direct node is defined as the first node presented in any path of $T_A$. DSI can also be defined globally for all the nodes of a network. Formally, the Global Direct Sources of Influence, *GDSI(C)*, of a network *C* are given by:

$$\begin{aligned} &GDSI(C) = \{S_i, \varepsilon_i\}_{i=1...|S|}, S \subseteq C, \varepsilon_i \in \Re^+ \\ &S = \bigcup_{n=1}^{N} \bigcup_{j=1}^{|T_n|} path_n^j(1) \\ &\varepsilon_i = \sum_{n=1...N} \sum_{j=1...|T_n|} load_n^j / path_n^j(1) = S_i \end{aligned} \quad (11)$$

### 4.4 Intermediary Sources of Influence (ISI)

The intermediary sources of influence of node *A*, *ISI(A),* are the set of nodes used to transmit the influence from its origin to *A*, *excluding* the source of the influence and the direct influence node, with the total amount of energy transmitted by each intermediary node. Formally, *TSI(A)* is defined as:

$$\begin{aligned} &TSI(A) = \{S_i, \varepsilon_i\}_{i=1..|S|}, S \subset C, \varepsilon_i \in \Re^+ \\ &S = \bigcup_{j=1}^{|T_A|} \bigcup_{i=2}^{|path_A^j|-1} path_A^j(i) \\ &\varepsilon_i = \sum_{j=1...|T_A|} load_A^j / S_i \in \bigcup_{k=2}^{|path_A^j|-1} path_A^j(k) \end{aligned} \quad (12)$$

The formulation in this case is more complex because the intermediary nodes are between the direct node and the originating node, thus the double union to define that set. The ISI concept can also be defined globally for all the nodes of a network. Formally:

$$\begin{aligned} &GTSI(C) = \{S_i, \varepsilon_i\}_{i=1...|S|}, S \subseteq C, \varepsilon_i \in \Re^+ \\ &S = \bigcup_{n=1}^{N} \bigcup_{j=1}^{|T_n|} \bigcup_{i=2}^{|path_A^j|-1} path_n^j(i) \\ &\varepsilon_i = \sum_{n=1...N} \sum_{j=1...|T_n|} load_n^j / S_i \in \bigcup_{k=2}^{|path_n^j|-1} path_n^j(k) \end{aligned} \quad (13)$$

### 4.5 Influence Paths (IP)

Influence paths are defined globally for a network *C* as the set of paths used to transmit influence from active nodes to destination nodes, with the value of total influence transmitted. Formally, the influence paths of network *C*, *IP(C)* are given by:

$$\begin{aligned} &IP(C) = \{P_i, \varepsilon_i\}_{i=1..|P|}, \varepsilon_i \in \Re^+ \\ &P = \bigcup_{n=1}^{N} \bigcup_{j=1}^{|T_n|} path_n^j \\ &\varepsilon_i = \sum_{n=1...N} \sum_{j=1...|T_n|} load_n^j / P_i = path_n^j \end{aligned} \quad (14)$$

The set of paths *P* is obtained by joining all possible paths from all nodes. A global Length Path (LP) measure can be defined as the length of each one of the paths in *IP(C)*, where the length is given in number of nodes, i.e. LP quanti-

fies the number of nodes that the influence has to travel from the activated node to the destination node.

## 5. Experimental Results

In this section, we simulate influence being transferred between individuals by applying the proposed algorithm on Call Detail Records (CDR) data. It is a simulation because the underlying assumption in our experiments is that phone calls longer than certain duration imply the propagation of influence from the caller to the callee, although there is no hard evidence of that in the data.

### 5.1 Data Set

Cell phone call data in the form of CDRs (Call Detail Records) were obtained for a number of users close to 250,000 over a period of six months. From all the information contained in a CDR, only the originating encrypted number, the destination encrypted number, the time and date of the call, and the duration of the call were considered. Calls were used to create a static social network. Figure 2 presents the log-log representation of the degree distribution (left) and the distribution of call duration (right) of the network. The degree distribution has a power law fitting with $\alpha$=2.3 and the call duration distribution has a lognormal behavior with $\mu$=5.02 $\sigma$=1.77. These values are similar to the values reported in the literature [7][8].

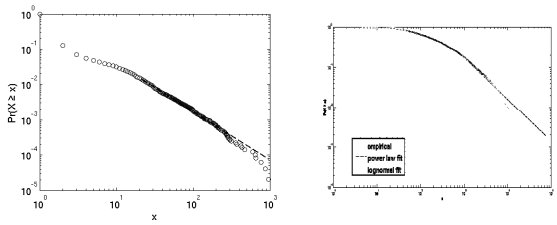



**Figure 2.** (lefy) log-log distribution of the degree distribution and (right) of the duration distribution of the original network.

### 5.2 Methodology

Two simulations were run in order to model how influece spreads: (1) *Experiment 1 (Exp1),* considers 1% of randomly chosen activated nodes, uses the first month of the data and a linear influence transfer function; and (2) *Experiment 2 (Exp2),* considers that 5% of the nodes are activated, where the nodes are selected in this case using a random walk [9], uses a different month of data (fourth month) and a Gompertz influence transfer function. The proposed algorithm was run for each experiment, producing two sets of $T_i$. Next, we computed the correlation between the final level of energy in each node and the degree, frecuency of calls and total duration of calls for the same node. In addition, we computed and plotted in a log-log scale the ranked GPSI, GDSI, GISI, IP and LP functions as they are relevant for modeling the spread of influence in the network.

### 5.3 Results

Figures 3 and 4 summarize the results for *Exp1* and *Exp2*. The heads of the distributions represent nodes that have a lot of influence, while the tails include nodes that play a minor role in spreading the influence. Each figure presents the log-log rank plot of the nodes, where the x-axis contains the number of nodes (phone numbers) in decreasing order of energy and the y-axis corresponds to: global primary source of influence (GPSI) (Fig. 3a and Fig. 4a for *Exp.1* and *Exp.2* respectivelly), the global direct source of influence (GDSI) (Fig. 3b and Fig. 4b), the global intermediary source of influence (GISI) (Fig. 3c and Fig. 4c), the influence paths (IP) (Fig. 3d and Fig. 4d) and the length paths (LP) (Fig. 3e and Fig. 4e). Each graph presents the results after 1 million calls, 2 million calls and all the calls. The y-axis represents the total energy, except for the LP plot where it represents the length in number of nodes.

*Correlation Coefficients*

Table 1 presents the correlation coefficients between the final level of influence of each node in *Exp1* and *Exp2* and: (1) degree, (2) frequency of calls, (3) total duration of the calls and (4) multiple linear regression considering degree, frequency and duration of calls, where we report the coefficient of determination.

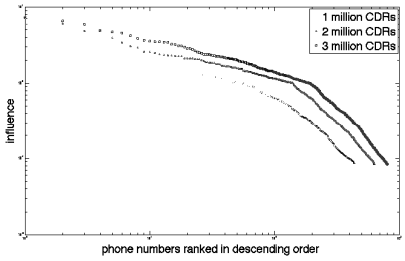
**Table I.** Correlation between influence and degree, frequency, call duration and their combination for the first and second experiment.

| | Degree | Frequency | Duration | MLR |
|---|---|---|---|---|
| **Experiment 1** | *0.24* | *0.42* | *0.60* | *0.60* |
| **Experiment 2** | *0.24* | *0.24* | *0.29* | *0.30* |

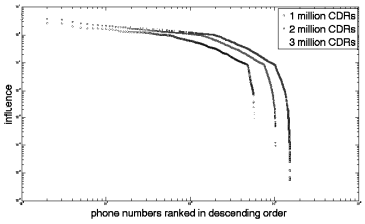
Note that duration is the variable that best justifies the influence received by a node, as much as considering the three parameters together via the MLR. This result is expected due to the role played by duration in the *influence_transfer* function of our model. However, duration can only express as much as 30% of variation in *Exp2* and 60% in *Exp1*, which implies that the rest of the variation is caused by other factors (*e.g.* order of interactions, temporality, nature of the link between each node, nature of the node, etc.). Our results strongly suggest that there is more to the spreading of influence than what is captured by the standard –static-- metrics such as degree, frequency and call duration.

*Log-log Rank Plots*

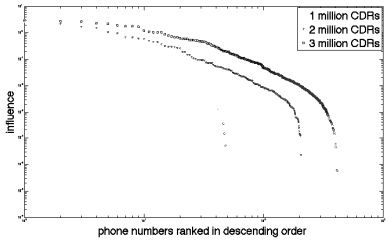
In the plots produced by both experiments, it can be observed that the behavior of the nodes does not *significantly* change when we vary the number of phone calls considered. The curves are basically the same, shifted up and to the right because of the increase in the total influence transmitted over time, but their statistical behavior remains the same. This does not mean that the nodes that are in the head of the distribution at 1 million interactions are still at the head of the distribution later on, but that the relative importance of the nodes that are in the head compared to those at the tail of the distribution remains constant.



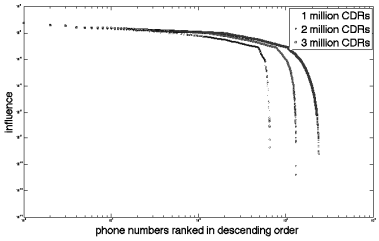

(a)



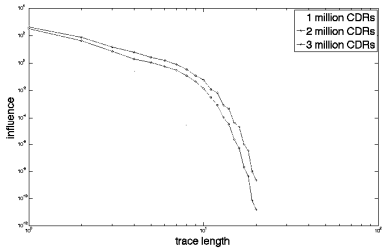

(b)




(c)

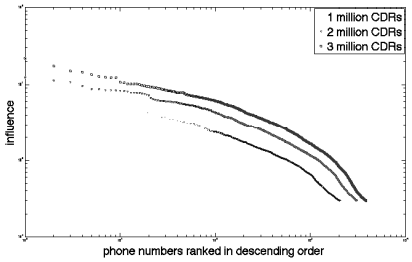



(d)

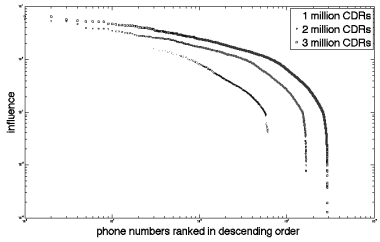



(e)

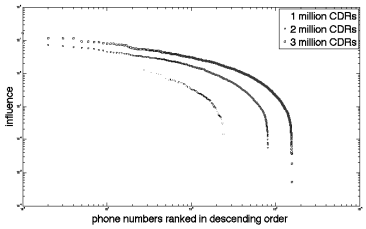
**Figure 3.** Rank plots (log-log) of: a) Global primary source of influence, GPSI; b) global direct source of influence, GDSI; c) global transmitting source of influence, GISI; d) influence paths, IP; and e) length of paths, LP, for *Exp1* for 1 million calls, 2 million calls and the entire data set.



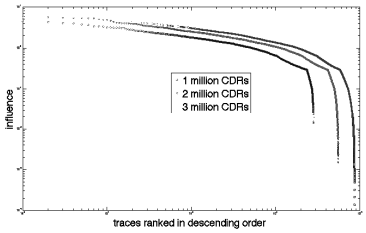

(a)



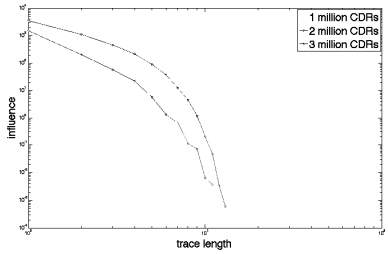

(b)




(c)




(d)




(e)

**Figure 4.** Rank plots (log-log) of: a) global primary source of influence, GPSI; b) global direct source of influence, GDSI; c) global transmitting source of influence, GISI; d) influence paths, IP; and e) length of paths for *Exp2* for 1 million calls, 2 million calls and the whole set.

These plots are very valuable for identifying the importance of each node in the network. For example, GPSI orders the nodes where more energy originates from and GISI orders the nodes by the role they play in transferring energy. Identifying these nodes is fundamental for many social network applications (*e.g.,* churn prediction, marketing, epidemics, etc.). Table II and Table III present the fitting parameters for power law and lognormal distributions of the survival functions obtained after processing all the interactions. The software used was Clauset's *et al*. [10] algorithm for power law fitting and the MATLAB statistical toolbox for the lognormal fit.

**Table II.** Lognormal and power law fitting parameters for each plot for the first experiment.

| | Power Law | | Lognormal | |
|---|---|---|---|---|
| | $\alpha$ | *xmin* | $\mu$ | $\sigma$ |
| **GPSI** | *3.14* | *1* | *-0.07(*)* | *1.03(*)* |
| **GDSI** | *1.85 (*)* | *0.09 (*)* | *-3.71* | *2.26* |
| **GISI** | *1.48 (*)* | *0 (*)* | *-4.73* | *2.60* |
| **IP** | *4.55* | *1* | *-4.92(*)* | *4.79(*)* |
| **LP** | *1.39(*)* | *0.78(*)* | *-5.08* | *7.54* |

**Table III.** Lognormal and power law fitting parameters for each plot for the second experiment.

| | Power Law | | Lognormal | |
|---|---|---|---|---|
| | $\alpha$ | *xmin* | $\mu$ | $\sigma$ |
| **GPSI** | *3.14* | *2.7* | *-0.04(*)* | *0.85(*)* |
| **GDSI** | *3.45* | *2.98* | *-0.99 (*)* | *1.45 (*)* |
| **GISI** | *3.39* | *1.21* | *-0.97 (*)* | *1.15 (*)* |
| **IP** | *4.55* | *1* | *-4.92(*)* | *4.79(*)* |
| **LP** | *1.57(*)* | *1.78(*)* | *1.63* | *4.9* |

In the tables, $\alpha$ is the exponent of the power law and *xmin* the value where the fitting starts. The lognormal distribution is characterized by $\mu$ and $\sigma$ parameters. An asterisk indicates best fit in terms of root mean squared error.

An analysis of the results indicates that GPSI has in both cases a lognormal distribution. This could be an indication that the distribution of the originating influence is an invariant, independently of other factors. Similarly, LP, the length of the paths, has in both cases a power law distribution with similar parameters. This fact indicates that preferential attachment behavior might also hold true for the length of the traces that describe the influence received. Conversely, IP, the set of influence paths, has a lognormal distribution and exhibits similar behavior in both experiments. It is interesting to note that for *Exp1* the maximum trace length is 20 and the average trace length is 1.28, whereas in *Exp2* the maximum trace length is 13 and the average path length is 1.6. Also, in both cases there seems to be an upper bound in the length of the path close to 20. In theory, the lengths of the paths could grow as new phone calls are made. However, this increase might not be very significant as the lognormal has small probability mass in the tail. Finally, GDSI and GTSI are modeled by different distributions in each experiment: while in *Exp1* they both follow a power law distribution, in *Exp2* they have a lognormal distribution. This difference is probably caused by the fact that each experiment used a different set of interaction data.

## 6. Conclusions

WoM algorithms have been used successfully in a variety of applications; nevertheless, typical approaches consider the social network as a static element. This fact implies that the temporality in which interactions take place is ignored. In this paper, we have introduced a novel WoM algorithm that considers the order of the interactions between nodes to spread influence. As a result, it is possible to track the influence of a node to see where each "piece of influence" came from, using the *PSI(A), DSI(A), ISI(A)* and *IP(A)* functions. The ability to trace influence also opens the possibility of modeling how influence spreads over the network. We have used two different experimental settings with the proposed algorithm, in order to find invariants regarding influence spread. Our preliminary results show that while GPSI, IP and LP seem to be invariant with respect to the influence model, GDSI and GISI depend on the set of interactions used to model the network. We have also shown that static metrics such as node degree, phone call duration and frequency (which can be interpreted as edge weights) are not sufficient to fully explain the spread of influence in a social network, suggesting that there are other –temporal-- factors that should be considered.

## References


[1] 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.

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

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

[4] 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.

[5] V. Blondel, C. de kerchove, E. Huens: Social Leaders in Graphs. Lecture notes in Control and Information Sciences, Vol. 341, 2006.

[6] H Gruber . Competition and innovation the diffusion of mobile telecommunications in Central and Eastern Europe. Information Economics and Policy, 2001 – Elsevier.

[7] 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.

[8] 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.

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

[10] A. Clauset, C. R. Shalizi and M. E. J. Newman, "Power-law distributions in empirical data." SIAM Review, to appear (2009).