Introducing Causality and Traceability in Word-of-Mouth Algorithms

Heath Hohwald, Enrique Frías-Martínez, Manuel Cerbián and Nuria Oliver
Data Mining and User Modeling Group, Telefonica Research
Emilio Vargas 6, 28043, Madrid, Spain
{heath, efm, nuriao}@tid.es

Abstract

Understanding the spread of information in a social network has proven useful in numerous areas, with viral marketing and epidimeology being two of the more prominent ones. Word-of-mouth algorithms are one class of algorithms that have been developed to model how information is verbally spread in a social network. However, two significant limitations of current word-of-mouth algorithms are their inability to: (1) capture *when* communication or contacts take place and (2) explain *where* the information possessed by each individual came from. In this paper, we present a novel algorithm that addresses these drawbacks by considering the temporality of communication and by tracing the spread of influence within a social network. The traces of influence prove useful for the identification of the most important individuals in a social network and for inferring causality. By applying the proposed algorithm to a large set of call detail records (CDRs), we are able to validate the proposed algorithm via simulations of word-of-mouth traces. Our two main findings are that (1) influence is better understood when temporality is added to the model and (2) the spread of information and influence in a network has several statistical invariants.

1 Introduction and Motivation

Word-of-mouth (WoM) or information diffusion algorithms first appeared in the social sciences [1] to model social interaction. This family of algorithms has been successfully used in a variety of areas, including viral marketing [2], epidemiology, and churn prediction [3]. In typical WoM algorithms [1][3], inferring the structure of the 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. 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.

2 Traceable Word-of-mouth Algorithm

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 β 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 *ds* 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 χ. The total influence accumulated by node i, $act(c(i))$, is defined as:

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

2.1 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 β.

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 β and the *influence_transfer* function. A typical value for β 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 functions: a piecewise-linear function and a Gompertz [5] function.

3. 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).

3.1 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:

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

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:

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

3.2 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. DSI can also be defined globally for all the nodes of a network *C* with the function Global Direct Sources of Influence, *GDSI(C.)*

3.3 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. The ISI concept can also be defined globally for all the nodes of a network *C* with the function Global Intermediary Sources of Influence, *GISI(C).*

3.4 Influence Paths (IP)

Influence paths (*IP*) are defined 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. 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 quantifies the number of nodes that the influence has to travel from the activated node to the destination node.

4. Methodology

Two simulations were run in order to model how influece spreads using the CDR traces of 250,000 users over a six month period: (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 (the 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, frequency 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.

5. Results

Figure 1 presents the results for *Exp1* and *Exp2* for GPSI (similar graphs have been obtained for DSI, ISI and IP). 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. The y-axis represents the total energy.

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.

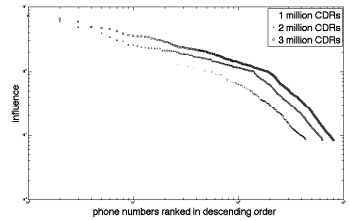
	Degree	Frequency	Duration	MLR
Experiment 1	*0.24*	*0.42*	*0.60*	*0.60*
Experiment 2	*0.24*	*0.24*	*0.29*	*0.30*

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.

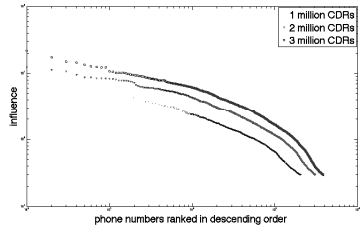
Log-log Rank Plots

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.

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.). Curves were fitted using power law and lognormal baselines. GPSI has in both experiments a lognormal distribution. This could be an indication that the distribution of the originating influence is an invariant, independent 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.



(a)



(b)

Figure 1. Rank plots (log-log) of Global primary source of influence, GPSI, for *Exp1*and (b) for *Exp2* for 1 million calls, 2 million calls and the entire data set.

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.

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