A Predictive Model for the Temporal Dynamics of Information Diffusion in Online Social Networks

Adrien Guille
ERIC, Université Lumière Lyon 2
5 av. Pierre Mendès-France, 69676 Bron, France
Adrien.Guille@univ-lyon2.fr

Hakim Hacid
Bell Labs France
Route de Villejust, 91620 Nozay, France
Hakim.Hacid@alcatel-lucent.com

ABSTRACT

Today, online social networks have become powerful tools for the spread of information. They facilitate the rapid and large-scale propagation of content and the consequences of an information – whether it is favorable or not to someone, false or true – can then take considerable proportions. Therefore it is essential to provide means to analyze the phenomenon of information dissemination in such networks. Many recent studies have addressed the modeling of the process of information diffusion, from a topological point of view and in a theoretical perspective, but we still know little about the factors involved in it. With the assumption that the dynamics of the spreading process at the macroscopic level is explained by interactions at microscopic level between pairs of users and the topology of their interconnections, we propose a practical solution which aims to predict the temporal dynamics of diffusion in social networks. Our approach is based on machine learning techniques and the inference of time-dependent diffusion probabilities from a multidimensional analysis of individual behaviors. Experimental results on a real dataset extracted from Twitter show the interest and effectiveness of the proposed approach as well as interesting recommendations for future investigation.

Categories and Subject Descriptors

I.6.5 [**Computing Methodologies**]: Simulation and Modeling—*Model Development*

General Terms

Experimentation

Keywords

Information Diffusion, Machine Learning, Online Social Networks

1. INTRODUCTION

The "Social Web" is the latest evolution of the Web where information is generated very quickly, consumed by millions of users, and updated quickly by others through commenting, replying, transferring, etc. This is practiced by people who differ in culture, knowledge, background, ideology, etc. Moreover, received information generally comes from several channels and goes through different channels if sent. This is amplified by the social networking phenomenon where people can easily socialize, discuss, organize or comment on events, share photos, etc. This makes the information move from one location to another, from one node of the network to another or to another community, etc. This is the well known phenomenon of information diffusion or propagation, which has been, and is still, of interest for the research community [15, 14, 2, 8, 13].

Propagation has been studied in areas like epidemiology for centuries, for understanding the diffusion processes in **complex systems** such as virus propagation considering certain conditions. Most of the current efforts in information propagation widely reuse those of epidemiology as a basis to tackle the propagation in other environments. With the emergence of Web 2.0, and especially social networks, the mechanisms of information diffusion have become more complex, because of the following reasons: (i) Modern digital social networks are very large, making the existing models inefficient in this area and even meaningless; (ii) There is a wide diversity in users profiles and the dynamics in their updates and changes; (iii) There is not yet a clear understanding of the laws governing digital social networks making it difficult to provide a clear model capturing this phenomena [1]. Thus, there is a need for models that better approximate the propagation of information in social networks while providing results close to real situations or capturing a part of the whole diffusion picture.

Understanding, capturing, and being able to predict such phenomenon can be helpful for several areas such as marketing, security, and Web search. For instance, if we consider the case of marketing, it may be useful to know which are the features that control the process of diffusing information when it's created to, e.g. better advertise a product or to better protect it against attacks on the network. The marketing may also benefit from information such as how many initial users to start with in a marketing campaign (budget optimization), how much time to wait between actions, etc. In the case of security, criminal investigators generally need to understand the information flow between, e.g. members of a given community, to extract hints regarding possible guilt or innocence of a person or a group of persons [3]. This is clearly an observation phase where the user wants to understand the route that information took and possible links. Finally, as Web search evolves, if we consider the case of subscriptions to feeds, a propagation prediction model may be useful for the user to, e.g. subscribe to the most interesting topic according to its expected growth (in

WWW 2012 Companion, April 16–20, 2012, Lyon, France.
ACM 978-1-4503-1230-1/12/04.

addition to his interests). This reflects a more active usage of the diffusion prediction.

We define information diffusion as the process by which a piece of information, e.g., a message, is spread between entities, i.e. users in the case of social networks, potentially receptive to that piece, in a closed environment, i.e. ignoring external effects. A diffusion can be associated with both topological properties, such as scale and range [14], and temporal properties. The problem we intend to solve is the following: having a closed environment in which users interact through a social network, how can we model this environment to capture and predict some properties of the diffusion process? In our case, we consider the temporal dynamics of the diffusion as the main aspect to target. For availability reasons, we use the social network Twitter to develop and discuss our proposal.

Most of the existing models in this area consider information propagation from a theoretical perspective. In fact, many assumptions are generally made to restrict the problem and propose theoretical solutions which are generally not applicable in real situations. Moreover, and as we will see it in the related work section, most of the existing work rely on a restricted features space to build a model. In this paper, we consider the problem from a more practical perspective and we propose an approach that uses machine learning to build a model that captures and predicts information propagation in social networks. Our model is built on the assumption that the propagation of information in a social network relies on an explicit graph connecting the users and is explained by micro-interactions between pairs of nodes, according to local properties. We rely then on a statistical analysis of the behaviour of individuals instead of a global analysis of the graph. The contributions of this paper are the following:

1. An analytical discussion that highlights some of the features which may be useful for capturing the diffusion process. This step has been performed using a dataset and some discussion threads on Twitter. It enabled us to understand the overall process of information diffusion in a real social network, and thus extracting some features for the model.
2. A set of concrete features and a model that captures and predicts information diffusion. This model relies on: (i) three dimensions (semantic, social, and time). We show how this combination is of interest for predicting the propagation. (ii) The consideration of local behaviours of users instead of global information on the network.
3. An experimental study providing insights regarding the values of some parameters of the model as well as the correlation between them. We also provide, thanks to certain measures, a global view on the performances of the proposed model. Due to the difficulty of identifying meaningful and verifiable threads in the considered dataset, we limit this step to specific and meaningful use-cases.

The rest of this paper is organized as follows: Section 2 reviews some related work and discusses their relation to ours. Section 3 describes the analysis we have performed on a real dataset to understand the features that participate in the diffusion process and the possible links between them. Section 4 discusses the proposed model in detail. In Section 5, a set of experiments is described to provide some insights about the different possible values of the model as well as a general estimation of the performance of the model. We conclude and provide some future work in Section 6.

2. RELATED WORK

In this section we review two categories of related work: (i) models for diffusion and (ii) studies about information diffusion in social networks.

A social network can be modeled as a graph, in which information spreads through the publication of messages in various parts of this structure. By posting a message dealing with a specific information, the author takes an active part in its diffusion and the corresponding node in the modelisation is said to be "activated". The propagation process can then be viewed as an ordered sequence of activation. This definition allows for the analogy between information propagation and the spread of disease in a population or innovation diffusion in a network and most of the current efforts in information propagation widely reuse this work.

2.1 Models of diffusion in networks

The propagation of viruses has been studied for centuries and the diffusion of innovation is one of the original reasons for studying networks. Therefore, there is much literature regarding these two areas and this work is an important basis.

Works on innovation diffusion focus on the topology of the process and follow either *Independent Cascades* (IC) [5] or *Linear Threshold* (LT) [6] model. They are based on a directed graph where each node can be active or inactive, with a monotonicity assumption, i.e. active nodes can not deactivate. The IC model requires a diffusion probability to be associated to each edge whereas LT requires an influence degree to be defined on each edge and an influence threshold for each node. For both models, the diffusion process proceeds iteratively in a synchronous way along a discrete time-axis, starting from a set of initially activated nodes. In the case of IC, for each iteration, the newly activated nodes try once to activate their neighbors with the probability defined on the arc joining them. If the transmission succeeds, the distant node becomes active at the next iteration. In the case of LT, at each iteration, the inactive nodes are influenced by their active neighbors with a force equal to the sum of the weights of the respective arcs. If this sum exceeds the threshold, the node becomes active at the next iteration. The process ends when no new transmission is possible, i.e. no neighboring node can be contacted. These two mechanisms reflect two different points of view: IC is sender-centric while LT is receiver-centric. Both models have the inconvenience to proceed in a synchronous way along a discrete time-axis, which doesn't suit what is observed in real social networks. In order to make these models more adapted to this particular context, Saito et al. recently proposed asynchronous extensions of these models, namely *AsIC* and *AsLT* [12], that use a continuous time-axis and require a time-delay parameter on each edge of the graph.

Works on virus propagation are interested in the dynamics of the process and focus on the repartition of the population of nodes into several classes. The two most common models are *SIR* and *SIS*, where nodes in the S class are "susceptible" to catch the disease with a probability β. "Infected" nodes are in the I class and have a probability γ to recover. In the case of SIS, nodes who recover become susceptible again while in the case of SIR, nodes stay in class R, i.e. "recovered". The percentage of nodes in each class is given by simple differential equations. Both models are fully-mixed, which means that every node has the same probability to be connected to another and thus connections inside the population are made at random. But the topology of the nodes' relations is very important in social networks and thus the assumptions made by these model are unrealistic.

2.2 Information diffusion in social networks

Various studies in the context of social networks have been conducted to predict properties of the information spreading process. By its objective, i.e. predict the temporal dynamics, the work of Yang and Leskovek [15] is certainly the most related to our proposal. They studied the diffusion of hashtags in Twitter and proposed a model based on the assumption that the influence of a node depends of how many other nodes it influenced in the past. However, there is a substantial difference with our work because this approach is non-graphical (author consider the network to be implicit) and doesn't study nodes attributes. Leskovec et al. [9] proposed another model adapted to diffusion in the blogosphere and similar to SIS. Bakshy et al. [2] proposed a graphical approach that aims to predict the size of the cascade generated by the diffusion of a URL in Twitter graph of followers, starting with a given initial user. This model relies on a regression tree and some social attributes and the past influence of the initial user only. The influence of the initial user is approximated by counting the number of implicit paths of diffusion (inferred from the follower graph) in which he was involved in the past. Galuba et al. [4] also studied the diffusion of URLs in Twitter and proposed to use the LT model to predict which users will predict which URL. Yang and Counts [14] adopted a graphical approach based on survival analysis to study the impact of several attributes from both users and content to predict the size of the cascade generated by the spread of a topic in Twitter. They focused on the explicit path of diffusion expressed by mentions in tweets (i.e. targeted tweets containing "@username") that they represented as a particular case of IC with only a single initial user.

3. PRELIMINARIES

In this section we describe the dataset used in this paper and present the results of a study (details are not reported here, due to page limitation) we conducted to identify the dimensions needed to capture the diffusion process.

3.1 Dataset

We have selected a dataset for this step according to three main criteria:

- Scale: the dataset needs to be large enough to be statistically significant;
- Completeness: all ties and friendships in the network should be observed;
- Realism: the social content should be extracted from real and various interactions.

Thus, we use a 467 million Twitter posts dataset from 20 million users covering a 7 months period from June 1, 2009 to December 31, 2009 gathered by Yang and Leskovec [16] as well as the topology of the network extracted by Kwak et al. [7] (1.47 billion following links) which includes the users observed by Yang and Leskovec and their followers. It may well be considered that this dataset meets all criteria.

3.1.1 Data preparation

We extract several distinct subsets of users with the following method: from an initial seed user, we fetch all his followers distant from at most 2 hops. We then connect the nodes with their following relationships. In each set of users we study the diffusion of information for various topics and construct the spreading cascades. The inference of the cascades relies on the assumption that the last follower to mention a topic before a given node is the one who influenced him. We decompose each edge of the propagation cascade in an instance of diffusion that we associate to statistical informations about the two nodes and the topic. For each instance of this kind, we also generate an instance of non-diffusion between the same source node and another follower that didn't reposted the information.

3.2 Analysis of the diffusion process

Three essential dimensions emerge from the analysis we performed. Firstly, by extracting distinct sub-communities we have seen that they had different characteristics, both structurally (in terms of density, clustering coefficient, etc.) and in terms of activity. This highlights the importance of taking into account the topology of the network and the social dimension. Then we found that information had a different impact depending on its topic, which demonstrates the need for integrating a semantic layer to the model. Finally, we have seen that users are more or less active (i.e. receptive to information) depending on time (at different scales: days, weeks, exceptional periods). To study how user activity is distributed over a day, we gather the timestamps of all the tweets published by each user, partition a day into 6 periods of 4 hours, aggregate timestamps by period and normalize the values. Thus we know the percentage of tweets a user emitted during each period. Figure 1 shows how many observations (instances of diffusion or non-diffusion) were made function of the usual activity intensity of the destination node at the time of day at which the source posted the information. It is clearly visible that the probability that a diffusion occurs between two nodes at a moment of the day when the destination node is usually totally inactive is significantly low.

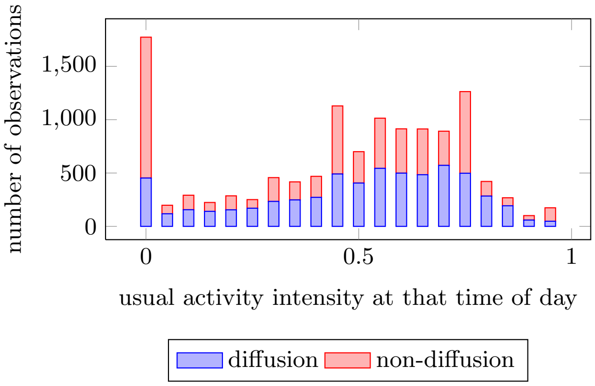

Figure 1: Number of observations versus activity intensity of the destination node.

This parameter is stronger than it might look. Indeed, all of the current diffusion models in social networks based on a concrete graph structure work in a synchronous manner.

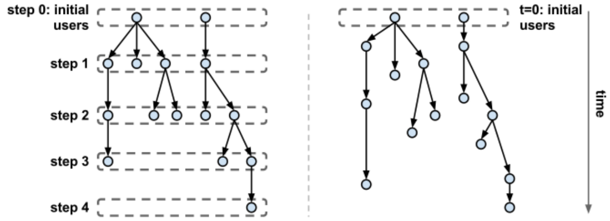

Figure 2: Comparison of a cascade modeled by IC (left) and AsIC (right).

Therefore, the temporal aspect of the diffusion is ignored or at least very superficial and thus the estimation they produce is less reliable.

Furthermore, this analysis raises other interesting points, such as competition between informations and topics correlation. Still, we haven't yet investigated these aspects and didn't integrate them in the proposal at this stage, as discussed in the next section.

4. PROPOSED APPROACH

Our approach models the diffusion of a topic as cascades and thus adopts a sender-centric vision. We use the Twitter follower graph as a basis and consider three dimensions: (i) semantics, (ii) social, and (iii) time. To make use of the third dimension, we leverage the AsIC meta-model, an extension for a continuous time axis of the broadly used IC model by adding a time-delay parameter on each edge. This allows us to model the propagation as an asynchronous process and therefore capture the temporal dynamics more accurately, as shown on Figure 2. The diffusion probabilities are defined on each edge of the graph from local properties representing the three dimensions and a model produced using machine learning techniques.

This proposal is generic in that its attributes are. Indeed, even if they are described in the following with the Twitter terminology, all social networking sites are based on an explicit graph (whether it is based on the notion of subscription or friendship) and allow the publication of global and targeted messages. Nevertheless, the model coefficients are related to the characteristics of each platform (e.g. reciprocity is lower in Twitter social graph than in Facebook graph) and need to be adjusted accordingly. This is the objective of the learning step.

4.1 Notations

We consider a social network represented by a set U of users interacting through messages M (all the tweets of the environment in our case). The set of interactions generated by a user $u \in U$ is denoted by M_u. We distinguish between general messages, i.e. sent to all users of the network, and the directed messages denoted $\mathcal{D}_u \subset M_u$ which are the messages of the user which are targeted for a specific user. Different standards are used when communicating in Twitter, and the one which is of a particular interest in our case is the mentioning practice. Thus, the set of users who are mentioned in the messages of a user $u \in U$ is denoted with $\mathcal{M}_u$. Inversely, the set of users who mentioned the user $u \in U$ in their messages is denoted with $\bar{\mathcal{M}}_u$. We also denote with $t\mathcal{M}^u$ all the messages which have mentioned a user $u \in U$. Then, we denote with $K = \{k_1, k_2, ..., k_p\}$ the set of all keywords used in all the interactions of the network and $K_u \subset K$ the set of keywords included in the interactions generated by a user $u \in U$. Finally, we consider a set of topics $\mathcal{C} = \{c_1, c_2, ..., c_p\}$. For each topic may be associated one or more keywords $k_i \in K$, i.e. $c_j = \{\cup_{i=1}^{l} k_i, k_i \in K, i \leq |K|\}$.

4.2 Model description

The computation of a probability relies on three dimensions: social, semantic, and temporal. We denote $p_{u_1 u_2}(i,t)$ the diffusion probability of information i (described by its topic c_i) at time t between users u_1 (sender) and u_2 (receiver). The attributes we derive from these dimension are either numerical values varying between 0 and 1 or boolean values translated into integer values (i.e., 0 or 1). Their calculation is based on user activity for a month.

Social dimension: This dimension intends mainly to capture the different properties of the social network (i.e., nodes and arcs) and the relations between them which may impact the diffusion process in such networks. Five properties are captured from this perspective as described below.

- *Activity (I):* an activity index expresses users' relative activity. The activity is computed as the average amount of tweets emitted per hour bounded by 1. For a user u, the formula is as follows :

$$\mathrm{I}(u) = \begin{cases} \frac{|M_u|}{\epsilon} & \text{if } |M_u| < \epsilon \\ 1 & \text{Otherwise} \end{cases} \quad (1)$$

  with $\epsilon = 30.4 \times 24$ for the hourly frequency.
- *Social homogeneity (H):* a social homogeneity index for u_1 and u_2 reflects the similarity of the sets of users they talk to. It is computed with the Jaccard coefficient as shown in the following formula.

$$\mathrm{H}(u_1, u_2) = \frac{|\mathcal{M}_{u_1} \cap \mathcal{M}_{u_2}|}{|\mathcal{M}_{u_1} \cup \mathcal{M}_{u_2}|} \quad (2)$$

- The ratio of directed tweets for each user (dTR) which provides an idea about the ability of a given user to distribute a content for other users and how this is done (i.e., to specific users or to general communities).

$$\mathrm{dTR}(u) = \begin{cases} \frac{|\mathcal{D}_u|}{|M_u|} & \text{if } |M_u| > 0 \\ 0 & \text{Otherwise} \end{cases} \quad (3)$$

- A boolean value for each user regarding the mentioning behaviour to capture the existence of a social relation between users.

$$\mathrm{hM}(u_1, u_2) = \begin{cases} 1 & \text{if } u_2 \in \mathcal{M}_{u_1} \\ 0 & \text{Otherwise} \end{cases} \quad (4)$$

- The mention rate (mR) of each user represents their popularity. The higher the value, the more the user is cited in tweets and thus receive more tweets.

$$\mathrm{mR}(u) = \begin{cases} \frac{|t\mathcal{M}^u|}{\mu} & \text{if } |t\mathcal{M}^u| < \mu \\ 1 & \text{Otherwise} \end{cases} \quad (5)$$

  Based on our empirical observation of the distribution of the mention rates we have chosen $\mu = 200$.

Semantic/Topical dimension: Beyond the structure of the network, we consider the exchanged content to better understand and capture the reasons of the diffusion. We currently consider only one feature that indicates if the user employed at least one of the keyword of the exchanged content in his

Social network	# of users	# of tweets	# of following edges
1	24,571	303,564	1,928,999
2	44,410	469,775	4,398,953
3	11,614	169,689	308,849
4	29,625	226,753	2,507,768

Table 1: Properties of the four experimental social networks

past tweets. It is under the form of a boolean value.

$$\mathrm{hK}(u,i) = \begin{cases} 1 & \text{if } (K_u \cap C_i \neq \emptyset) \\ 0 & \text{Otherwise} \end{cases} \qquad (6)$$

Temporal dimension: This dimension intends to capture the dynamics of the network to be incorporated in the model. As discussed in the previous sections, this is an important aspect to respect as much as possible the nature of social networks. In order to obtain a significant distribution even for the less active users, we consider a day as a partition of 6 blocks of 4 hours. So we compute the fraction of tweets the user emitted during each block and fill a 6-dimensional vector noted V: $\sum_{i=0}^{5} V^i = 1$. It allows us to get the proportion of activity for each user at a time of day t.

$$\mathrm{A}(u,t) = V_u^{t'} \text{ where } t' = \lfloor \frac{t}{4} \rfloor \qquad (7)$$

4.3 Probability inference

Once the representation space is set, we consider the task of finding a suitable model for inferring the diffusion probabilities. To do so, we adopt a machine learning based approach and use the methodology described in Section 3.1.1 to prepare the data. We compute the attributes of more than 100,000 distinct users divided into 4 social networks (see Table 1 for their description), according to their activity in November 2009, and generate instances of the binary class {diffusion,non-diffusion} based on the propagation of 6 topics in each each network during December 2009. Thus, each instance observed in December is described by the attributes of the two involved nodes in November.

We test three algorithms on those data: a *C4.5 decision tree*, a *Linear and Multilayer (1 hidden layer with 14 cells) Perceptron*, and a *Bayesian Logistic Regression*. We define the supervised classification task: $P(Y|V)$, with Y ={diffusion, non-diffusion} and V the 13-dimensional vector of attributes. The results of a cross-validation are illustrated in Table 2.

Classifier parameters	Correctly classified instances
C4.5	91%
Linear Perceptron	85%
Multilayer Perceptron	86%
Bayesian Logistic Regression	85%

Table 2: Classifiers performances on a 5 folds cross-validation

At first glance, we see that all classifiers obtain an error rate lower than 15%. We also see that the Linear Perceptron has almost the same performance as the Perceptron with 14 hidden layers. This suggests that diffusion probability can be seen as a linear combination of the variables. The decision tree obtains the lowest error rate, but its model is more specific because of the partitioning algorithm on which it is based. We then focus on the linear Perceptron and the Bayesian logistic regression equations and compare their normalized coefficients on Figure 3. It reveals a common tendency in the importance they accord to the attributes. In particular, the two classifiers highlight an important emphasis on the social homogeneity. This is caused by the distribution of the values taken by this attribute. In fact, its mean is 0.004 and its standard deviation is 0.02 while the other attributes have an average mean of 0.268. However, the Logistic Regression has a more leveled aspect than the Perceptron. For that reason, we decided to infer the diffusion probabilities with the model produced by the Bayesian Logistic Regression.

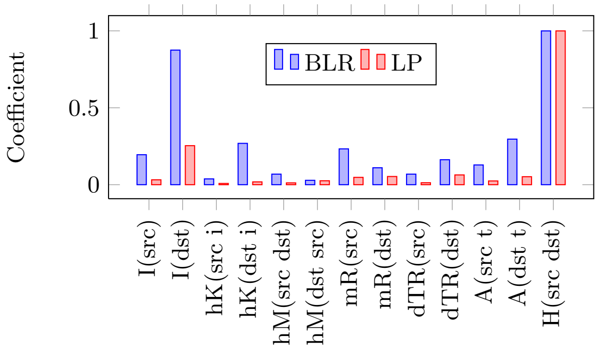

Figure 3: Comparison of the normalized values of the coefficients in linear perceptron and Bayesian logistic regression equations.

The logistic regression assumes a parametric form for the distribution $P(Y|V)$. The parametric model used by the logistic regression is as follows (as defined in [10]):

$$P(Y = \text{diffusion}|V) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{13} w_i V_i)} \qquad (8)$$

$$P(Y = \text{non-diffusion}|V) = \frac{\exp(w_0 + \sum_{i=1}^{13} w_i V_i)}{1 + \exp(w_0 + \sum_{i=1}^{13} w_i V_i)} \qquad (9)$$

The model learned with the Bayesian logistic regression is implemented in the form of a prediction engine. The algorithm uses a fake clock to simulate the course of days, used by the parameter $A(u,t)$. Thus, the engine produces a time-serie of the volume of tweets generated each day by the diffusion of the topic inside the studied social network.

The engine requires four parameters: (i) a social network, (ii) a topic, (iii) a set of $n << |U|$ initially informed users, and (iv) a formalization of the delay parameter r_{u_1,u_2}. The social network represents the group of people which is intended to be studied, described by its graph and the users' attributes. The topic is an information translated by a set of keywords which is expected to be transported by the interactions. The idea is that the user of the engine is expected to provide a set of semantically related keywords which may refer to this information from his perspective. The next parameter which defines the set of initial informed users, consists of the number of early adopters and is expected to somehow control the speed of the diffusion and its importance. Finally, the last parameter defines the delay of transmission between two users.

5. USING THE MODEL FOR PREDICTION

In this section, we attempt to study the performances of the above model for predicting the temporal dynamic of information spread in social networks which relies on the implementation of the model as a prediction engine, as described before.

5.1 Experimental setup

For each example of diffusion in December 2009, we identify the keywords, the n first distinct users involved in the diffusion for each experimental network and ask the engine to predict it, based on the attributes computed in November. We consider the experiments from two different perspectives: (i) the study of the values of the different parameters of the model, w.r.t. certain conditions, and (ii) the evaluation of the model precision. Finally, we present in the next section the results we obtained in the two first experimental social networks described in Table 1.

5.2 Parameters study

The main idea behind this type of tests is to consider different values for each evaluated parameter and analyze the evolution of the overall dynamics according to each value in order to extract the optimal values as well as understanding potential correlations which may exist between those considered parameters. First of all, we set the time delay parameter of the model, r_{u_1,u_2}. Since we have defined it as $r_{u_1,u_2} = (1 - I(u_2)) \times \sigma$, the objective is to provide an approximation for σ. Thus, for each predicted time-serie with a given value of σ, we compute an Euclidian distance to its corresponding real time-serie. We repeat the process for different topics and different values of σ. Figure 4 shows the evolution of the euclidian distance w.r.t. different values of σ. We can see that the distance, i.e. the difference between the predicted dynamic and the real one, is minimal around $\sigma = 10$. Outside this value, it is large. For this reason, we have set the value of σ to 10 in all the remaining experiments in this paper.

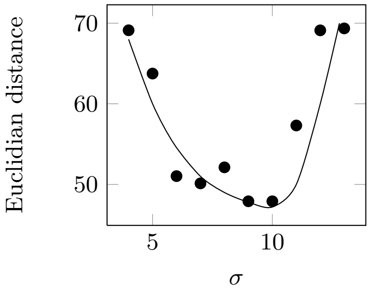

Figure 4: Euclidian distance vs σ.

Figure 5 shows the comparison between real data and the predicting result about an information dealing with the release date of the new iPhone in the two social networks. The x axis represents time units in days and the y axis represents the activity level with tweets volume as unit. The blue plot with circles corresponds to the real activity while the red boxed plot corresponds to the predicted activity. We have set the value of $n = 8$ for the first network and $n = 5$ for the second. It appears from the figure that each social network exhibits a particular activity pattern which is properly captured by the model. However the volume is not correctly estimated by the model. Several explanations are possible, first, the capture of messages dealing with other information about the iPhone can artificially increase the volume of tweets; also, even if we suppose that most of the users are informed via internal interactions, some can also get the information outside of Twitter (i.e. two-step theory) and locally reinforce the diffusion. Second, it reveals a gap in our modeling. More particularly, we envision the existence of an amplification factor related to a macroscopic phenomenon, such as social imitation for instance.

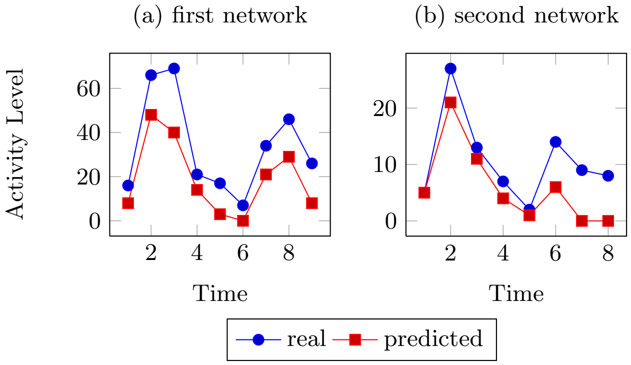

Figure 5: Comparison of real and predicted activity about the iPhone in the two test-sets

To go further in the understanding of the pattern observed on Figure 5(a), we give in Figure 6 the volume of transmitters and stiflers, two categories of users defined by Nevokee et al. [11], across time for that simulation. Stiflers are the people who receive information but don't transmit it for a any given reason, e.g. they have already received it or they don't see the interest to share it. It shows that information reaches the most active users first and then dies out by reaching more and more stiflers. In other words, the diffusion process depends on the ratio of transmitters and stiflers and depending on the quantity of one or another, the information keeps on spreading or is blocked. It should be noted that the two peaks are also visible on this figure and the amount of transmitters/stiflers captures well the observation.

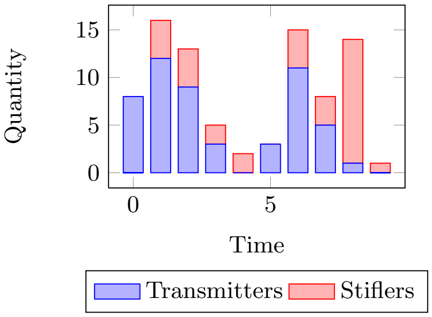

Figure 6: Amount of transmitters and stiflers.

Figure 7 shows the results for the diffusion of an information about another topic, the acquisition of a startup by Google in December 2009, with $n = 11$ for the first network and $n = 14$ for the other. Although we observe two noticeable peaks for all the other diffusion processes, we only observe one in the first network. To understand this, we study the evolution of N, the total amount of people reached by the information, according to n in that set of users, represented in Figure 8(a). We observe that the diffusion rate is stable for n between 6 and 11 then we observe an important outbreak from $n = 12$. Therefore, as one can see in Figure 8(b), when running the prediction engine with $n = 14$, we observe a stronger activity level and two main peaks while the predicted model is completely non realistic compared to the real observations. By considering these observations, it comes out that the optimal values of n span from 6 to 11 in this case. As a matter of fact, n depends on the information and the social network but for the total 24 predictions we ran (6 predictions for the 4 experimental networks), the optimal value of n varied between 5 and 20.

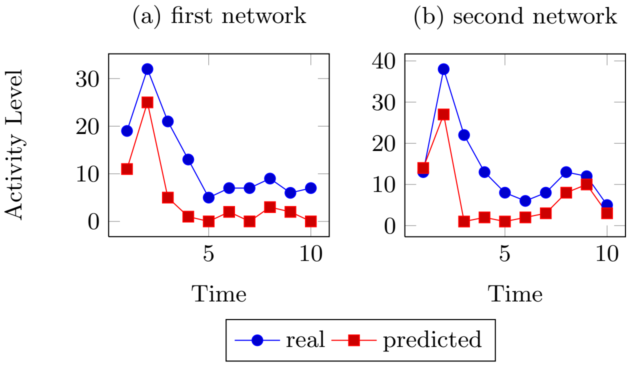

Figure 7: Comparison of real and predicted activity about Google in the two networks

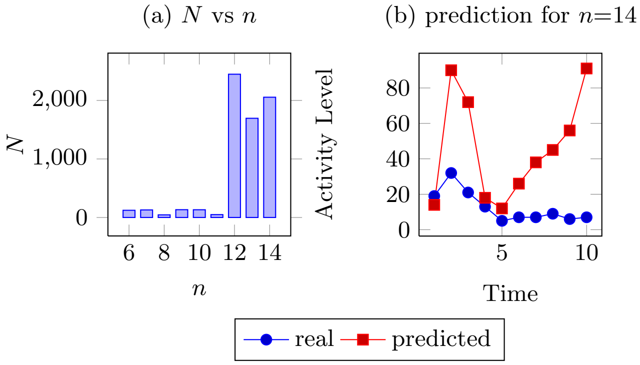

Figure 8: Study of the epidemic threshold for the diffusion of an information about Google in the first network

6. CONCLUSION AND FUTURE WORK

Starting from the hypothesis that information propagation is governed by the structure of the social and local behaviors of the involved entities, we proposed in this paper a concrete model that captures this process using Twitter as an experimental social network. A set of features, resulting from a deep observation of a real dataset and belonging to three dimensions – social, semantic, and time – are incorporated in the model. This model relies on the *AsIC* principle and is based on machine learning techniques, i.e. a Bayesian logistic regression, to infer time-dependent diffusion probabilities between nodes of the network. A set of experiments has been performed which provided some insights regarding the possible values of the parameters of the model as well as a general overview of the behavior of the model. The results showed mainly that the model predicts well the dynamic of the diffusion (our initial objective) but fails in accurately predicting the volume of tweets generated by the propagation.

Also, the experimental study enabled us to observe a typical pattern of the temporal dynamic of the diffusion process with two main peaks of activity. Globally, there is a quick outbreak of the information at the beginning, then the activity seems to die out before another peak of activity occurs. While the results showed that our hypothesis (i.e., "the dynamics of the spreading process at the macroscopic level is explained by interactions at microscopic level between pairs of users and the topology of their interconnections") was correct, they also showed the need to take into account broader factors at the social (e.g. global imitation phenomenon [8]) and topic (e.g. the virality [4] of a topic) levels. Finally, we need to improve the preparation of the data so it can be automated since we performed this step mainly manually and finding threads was a complicated and time consuming task.

7. REFERENCES

[1] R. Albert and A. L. Barabási. Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74, 2002.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. WSDM'11, 2011.

[3] A. Bennamane, H. Hacid, A. Ansiaux, and A. Cagnati. Visual analysis of implicit social networks for suspicious behavior detection. DASFAA'11, 2011.

[4] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. WOSN'10, 2010.

[5] J. Goldenberg, B. Libai, and E. Muller. Talk of the network: A complex systems look at the underlying process of word-of-mouth. *Marketing Letters*, 2001.

[6] D. Kempe. Maximizing the spread of influence through a social network. KDD'03, 2003.

[7] H. Kwak, C. Lee, H. Park, and S. Moon. What is Twitter, a social network or a news media? WWW'10, 2010.

[8] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. KDD'09, 2009.

[9] J. Leskovec, M. McGlohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs: Patterns and a model. SDM'07, 2007.

[10] T. M. Mitchell. *Machine learning*. McGraw Hill series in computer science. 1997.

[11] M. Nekovee, Y. Moreno, G. Bianconi, and M. Marsili. Theory of rumour spreading in complex social networks. *Physica A: Statistical Mechanics and its Applications*, 2007.

[12] K. Saito, M. Kimura, K. Ohara, and H. Motoda.

Selecting information diffusion models over social networks for behavioral analysis. PKDD'10, 2010.

[13] D. Wang, Z. Wen, H. Tong, C.-Y. Lin, C. Song, and A.-L. Barabási. Information spreading in context. WWW'11, 2011.

[14] J. Yang and S. Counts. Predicting the speed, scale, and range of information diffusion in twitter. ICWSM'10, 2010.

[15] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. ICDM'10, 2010.

[16] J. Yang and J. Leskovec. Patterns of temporal variation in online media. WSDM'11, 2011.