Information Diffusion in Online Social Networks: A Survey

Adrien Guille[1] Hakim Hacid[2] Cécile Favre[1] Djamel A. Zighed[1,3]
[1]ERIC Lab, Lyon 2 University, France
{firstname.lastname}@univ-lyon2.fr
[2]Bell Labs France, Alcatel-Lucent, France
hakim.hacid@alcatel-lucent.com
[3]Institute of Human Science, Lyon 2 University, France
abdelkader.zighed@ish-lyon.cnrs.fr

ABSTRACT

Online social networks play a major role in the spread of information at very large scale. A lot of effort have been made in order to understand this phenomenon, ranging from popular topic detection to information diffusion modeling, including influential spreaders identification. In this article, we present a survey of representative methods dealing with these issues and propose a taxonomy that summarizes the state-of-the-art. The objective is to provide a comprehensive analysis and guide of existing efforts around information diffusion in social networks. This survey is intended to help researchers in quickly understanding existing works and possible improvements to bring.

1. INTRODUCTION

Online social networks allow hundreds of millions of Internet users worldwide to produce and consume content. They provide access to a very vast source of information on an unprecedented scale. Online social networks play a major role in the diffusion of information by increasing the spread of novel information and diverse viewpoints [3]. They have proved to be very powerful in many situations, like Facebook during the 2010 Arab spring [22] or Twitter during the 2008 U.S. presidential elections [23] for instance. Given the impact of online social networks on society, the recent focus is on extracting valuable information from this huge amount of data. Events, issues, interests, *etc.* happen and evolve very quickly in social networks and their capture, understanding, visualization, and prediction are becoming critical expectations from both end-users and researchers. This is motivated by the fact that understanding the dynamics of these networks may help in better following events (*e.g.* analyzing revolutionary waves), solving issues (*e.g.* preventing terrorist attacks, anticipating natural hazards), optimizing business performance (*e.g.* optimizing social marketing campaigns), *etc.* Therefore researchers have in recent years developed a variety of techniques and models to capture information diffusion in online social networks, analyze it, extract knowledge from it and predict it.

Information diffusion is a vast research domain and has attracted research interests from many fields, such as physics, biology, *etc.* The diffusion of innovation over a network is one of the original reasons for studying networks and the spread of disease among a population has been studied for centuries. As computer scientists, we focus here on the particular case of information diffusion in online social networks, that raises the following questions : (i) *which pieces of information or topics are popular and diffuse the most*, (ii) *how, why and through which paths information is diffusing, and will be diffused in the future*, (iii) *which members of the network play important roles in the spreading process?*

The main goal of this paper is to review developments regarding these issues in order to provide a simplified view of the field. With this in mind, we point out strengths and weaknesses of existing approaches and structure them in a taxonomy. This study is designed to serve as guidelines for scientists and practitioners who intend to design new methods in this area. This also will be helpful for developers who intend to apply existing techniques on specific problems since we present a library of existing approaches in this area.

The rest of this paper is organized as follows. In Section 2 we detail online social networks basic characteristics and information diffusion properties. In Section 3 we present methods to detect topics of interest in social networks using information diffusion properties. Then we discuss how to model in-

formation diffusion and detail both explanatory and predictive models in Section 4. Next, we present methods to identify influential information spreaders in Section 5. In the last section we summarize the reviewed methods in a taxonomy, discuss their shortcomings and indicate open questions.

2. BASICS OF ONLINE SOCIAL NETWORKS AND INFORMATION DIFFUSION

An online social network (*OSN*) results from the use of a dedicated web-service, often referred to as *social network site* (*SNS*), that allows its users to (i) create a profile page and publish messages, and (ii) explicitly connect to other users thus creating social relationships. *De facto*, an OSN can be described as a user-generated content system that permits its users to communicate and share information.

An OSN is formally represented by a graph, where nodes are users and edges are relationships that can be either directed or not depending on how the SNS manages relationships. More precisely, it depends on whether it allows connecting in an unilateral (*e.g.* Twitter social model of *following*) or bilateral (*e.g.* Facebook social model of *friendship*) manner. Messages are the main information vehicle in such services. Users publish messages to share or forward various kinds of information, such as product recommendations, political opinions, ideas, *etc.* A message is described by (i) a text, (ii) an author, (iii) a time-stamp and optionally, (iv) the set of people (called "mentioned users" in the social networking jargon) to whom the message is specifically targeted. Figure 1 shows an OSN represented by a directed graph enriched by the messages published by its four members. An arc $e = (u_x, u_y)$ means that the user "u_x" is exposed to the messages published by "u_y". This representation reveals that, for example, the user named "u_1" is exposed to the content shared by "u_2" and "u_3". It also indicates that no one receives the messages written by "u_4".

DEFINITION 1 (Topic). *A coherent set of semantically related terms that express a single argument. In practice, we find three interpretations of this definition: (i) a set S of terms, with $|S| = 1$, e.g. {"obama"} (ii) a set S of terms, with $|S| > 1$, e.g. {"obama", "visit", "china"} and (iii) a probability distribution over a set S of terms.*

Every piece of information can be transformed into a topic [6, 30] using one of the common formalisms detailed in Definition 1. Globally, the content produced by the members of an OSN is a stream of messages. Figure 2 represents the stream produced by the members of the network depicted in the previous example. That stream can be viewed as a sequence of decisions (*i.e.* whether to adopt a certain topic or not), with later people watching the actions of earlier people. Therefore, individuals are influenced by the actions taken by others. This effect is known as *social influence* [2], and is defined as follows:

DEFINITION 2 (Social Influence). *A social phenomenon that individuals can undergo or exert, also called imitation, translating the fact that actions of a user can induce his connections to behave in a similar way. Influence appears explicitly when someone "retweets" someone else for example.*

DEFINITION 3 (Herd behavior). *A social behavior occurring when a sequence of individuals make an identical action, not necessarily ignoring their private information signals.*

DEFINITION 4 (Information Cascade). *A behavior of information adoption by people in a social network resulting from the fact that people ignore their own information signals and make decisions from inferences based on earlier people's actions.*

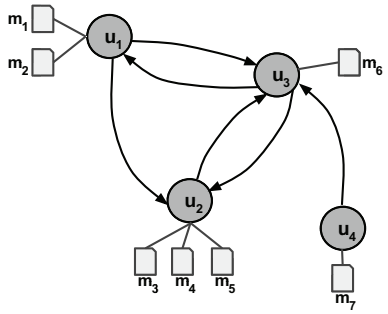

Figure 1: An example of OSN enriched by users' messages. Users are denoted u_i and messages m_j. An arc (u_x, u_y) means that u_x is exposed to the messages published by u_y.

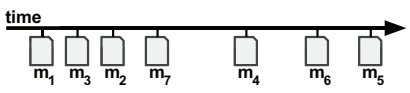

Figure 2: The stream of messages produced by the members of the network depicted on Figure 1.

Based on the social influence effect, information can spread across the network through the principles of *herd behavior* and *informational cascade* which we define respectively in Definition 3 and 4. In this context, some topics can become extremely popular, spread worldwide, and contribute to new trends. Eventually, the ingredients of an information diffusion process taking place in an OSN can be summarized as follows: (i) a piece of information carried by messages, (ii) spreads along the edges of the network according to particular mechanics, (iii) depending on specific properties of the edges and nodes. In the following sections, we will discuss these different aspects with the most relevant recent work related to them as well as an analysis of weaknesses, strength, and possible improvements for each aspect.

3. DETECTING POPULAR TOPICS

One of the main tasks when studying information diffusion is to develop automatic means to provide a global view of the topics that are popular over time or will become popular, and animate the network. This involves extracting "tables of content" to sum up discussions, recommending popular topics to users, or predicting future popular topics.

Traditional topic detection techniques developed to analyze static corpora are not adapted to message streams generated by OSNs. In order to efficiently detect topics in textual streams, it has been suggested to focus on bursts. In his seminal work, Kleinberg [26] proposes a state machine to model the arrival times of documents in a stream in order to identify bursts, assuming that all the documents belong to the same topic. Leskovec *et al.* [27] show that the temporal dynamics of the most popular topics in social media are indeed made up of a succession of rising and falling patterns of popularity, in other words, successive bursts of popularity. Figure 3 shows a typical example of the temporal dynamics of top topics in OSNs.

DEFINITION 5 (BURSTY TOPIC). *A behavior associated to a topic within a time interval in which it has been extensively treated but rarely before and after.*

In the following, we detail methods designed to detect topics that have drawn bursts of interest, *i.e. bursty topics* (see Definition 5), from a stream of topically diverse messages.

All approaches detailed hereafter rely on the computation of some frequencies and work on discrete data. Therefore they require the stream of messages to be discretized. This is done by transforming the raw continuous data into a sequence of collection of messages published during equally sized time slices. This principle is illustrated on Figure 4, which shows a possible discretization of the stream previously depicted in Figure 2. This pre-processing step is not trivial since it defines the granularity of the topic detection. A very fine discretization (*i.e.* short time-slices) will allow to detect topics that were popular during short periods whereas a discretization using longer time-slices will not.

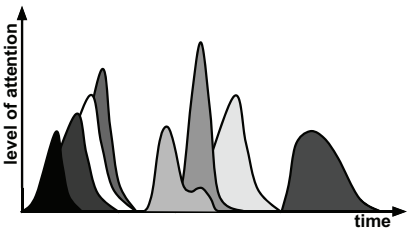

Figure 3: Temporal dynamics of popular topics. Each shade of gray represents a topic.

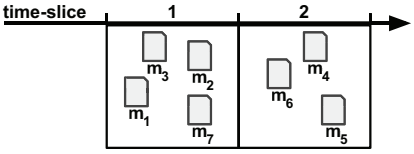

Figure 4: A possible discretization of the stream of messages shown on Figure 2.

Shamma *et al.* [46] propose a simple model, *PT* (*i.e. Peaky Topics*) , similar to the classical tf-idf model [44] in the sense that it is based on a normalized term frequency metric. In order to quantify the overall term usage, they consider each time slice as a pseudo-document composed of all the messages in the corresponding collection. The normalized term frequency ntf is defined as follows: $ntf_{t,i} = \frac{tf_{t,i}}{cf_t}$, where $tf_{t,i}$ is the frequency of term t at the i^{th} time slice and cf_t is the frequency of term t in the whole message stream. Using that metric, bursty topics defined as single terms are ranked. However, some terms can be polysemous or ambiguous and a single term doesn't seem to be enough to clearly identify a topic. Therefore, more sophisticated methods have been developed.

AlSumait *et al.* [1] propose an online topic model, more precisely, a non-Markov on-line LDA Gibbs sampler topic model, called *OLDA*. Basically, LDA (*i.e.* Latent Dirichlet Allocation [4]) is a statistical generative model that relies on a hierarchical Bayesian network that relates words and mes-

sages through latent topics. The generative process behind is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. The idea of *OLDA* is to incrementally update the topic model at each time slice using the previously generated model as a prior and the corresponding collection of messages to guide the learning of the new generative process. This method builds an evolutionary matrix for each topic that captures the evolution of the topic over time and thus permits to detect bursty topics.

Cataldi *et al.* [6] propose the *TSTE* method (*i.e.* Temporal and Social Terms Evaluation) that considers both temporal and social properties of the stream of messages. To this end, they develop a five-step process that firstly formalize the messages content as vectors of terms with their relative frequencies computed by using the augmented normalized term frequency [43]. Then, the authority of the active authors is assessed using their relationships and the Page Rank algorithm [35]. It allows to model the life cycle of each term on the basis of a biological metaphor, which is based on the calculation of values of nutrition and energy that leverage the users authority. Using supervised or unsupervised techniques, rooted in the calculation of a critical drop value based on the energy, the proposed method can identify most bursty terms. Finally, a solution is provided to define bursty topics as sets of terms using a co-occurence based metric.

These methods identify particular topics that have drawn bursts of interest in the past. Lu *et al.* [40] develop a method that permits predicting which topics will draw attention in the near future. Authors propose to adapt a technical analysis indicator primary used for stock price study, namely *MACD* (*i.e.* Moving Average Convergence Divergence), to identify bursty topics, defined as a single term. The principle of *MACD* is to turn two trend-following indicators, precisely a short period and a longer period moving average of terms frequency, into a momentum oscillator. The trend momentum is obtained by calculating the difference between the long and the shorter moving averages. Authors give two simple rules to identify when the trends of a term will rise: (i) when the value of the trend momentum changes from negative to positive, the topic is beginning to rise; (ii) when the value changes from positive to negative, the level of attention given to the topic is falling.

The above methods are based on the detection of unusual term frequencies in exchanged messages to detect interesting topics in OSNs. However, more and more frequently, OSNs users publish non-textual content such as URL, pictures or videos. To deal with non-textual content, Takahashi *et al.* [47] propose to use mentions contained in messages to identify bursty topics, instead of focusing on the textual content. Mentioning is a social practice used to explicitly target messages and eventually engage discussion. For that, they develop a method that combines a *mentioning anomaly score* and a change-point detection technique based on *SDNML* (*i.e.* Sequentially Discounting Normalized Maximum Likelihood). The anomaly is calculated with respect to the standard mentioning behavior of each user, which is estimated by a probability model.

Table 1 summarizes the surveyed methods according to four axes. The table is structured according to four main criteria that allow for a quick comparison: (i) how is a topic defined, (ii) which dimensions are incorporated into each method, (iii) which types of content each method can handle, and (iv) either the method detects actual bursts or predicts them. It should be noted that the table is not intended to express any preference regarding one method or another, but rather to present a global comparison.

reference	**topic definition**			**dimension(s)**		**content type**		**task type**	
	single term	set of terms	distribution	content	social	textual	non-textual	observation	prediction
PT	x			x		x		x	
OLDA			x	x		x		x	
TSTE		x		x	x	x		x	
SDNML	x				x	x	x	x	
MACD	x			x		x			x

Table 1: Summary of topic detection approaches w.r.t topic definition, incorporated dimensions, handled content and the task.

4. MODELING INFORMATION DIFFUSION

Modeling how information spreads is of outstanding interest for stopping the spread of viruses, analyzing how misinformation spread, *etc.* In this section, we first give the basics of diffusion modeling

and then detail the different models proposed to capture or predict spreading processes in OSNs.

DEFINITION 6 (Activation Sequence). *An ordered set of nodes capturing the order in which the nodes of the network adopted a piece of information.*

DEFINITION 7 (Spreading Cascade). *A directed tree having as a root the first node of the activation sequence. The tree captures the influence between nodes (branches represent who transmitted the information to whom) and unfolds in the same order as the activation sequence.*

The diffusion process is characterized by two aspects: its structure, *i.e.* the diffusion graph that transcribes who influenced whom, and its temporal dynamics, *i.e.* the evolution of the diffusion rate which is defined as the amount of nodes that adopts the piece of information over time. The simplest way to describe the spreading process is to consider that a node can be either activated (*i.e.* has received the information and tries to propagate it) or not. Thus, the propagation process can be viewed as a successive activation of nodes throughout the network, called *activation sequence*, defined in Definition 6.

Usually, models developed in the context of OSNs assume that people are only influenced by actions taken by their connections. To put it differently, they consider that an OSN is a closed world and assume that information spreads because of informational cascades. That is why the path followed by a piece of information in the network (*i.e.* the diffusion graph) is often referred to as the *spreading cascade*, defined in Definition 7. Activation sequences are simply extracted from data by collecting messages dealing with the studied information, *i.e.* topic, and ordering them according to the time axis. This principle is illustrated in Figure 5. It provides knowledge about where and when a piece of information propagated but not how and why did it propagate. Therefore, there is a need for models that can capture and predict the hidden mechanism underlying diffusion. We can distinguish two categories of models in this scope: (i) explanatory models and (ii) predictive models. In the following, we detail these two categories and analyze some representative efforts in both of them.

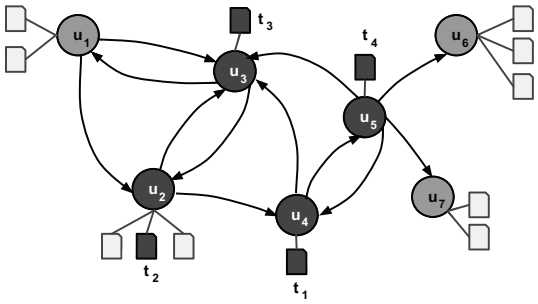

Figure 5: An OSN in which darker nodes took part in the diffusion process of a particular information. The activation sequence can be extracted using the time at which the messages were published: $[u_4; u_2; u_3; u_5]$, with $t_1 < t_2 < t_3 < t_4$.

4.1 Explanatory Models

The aim of explanatory models is to infer the underlying spreading cascade, given a complete activation sequence. These models make it possible to retrace the path taken by a piece of information and are very useful to understand how information propagated.

Gomez *et al.* [15] propose to explore correlations in nodes infections times to infer the structure of the spreading cascade and assume that activated nodes influence each of their neighbors independently with some probability. Thus, the probability that one node had transmitted information to another is decreasing in the difference of their activation time. They develop *NETINF*, an iterative algorithm based on submodular function optimization for finding the spreading cascade that maximizes the likelihood of observed data.

Gomez *et al.* [14] extend *NETINF* and propose to model the diffusion process as a spatially discrete network of continuous, conditionally independent temporal processes occurring at different rates. The likelihood of a node infecting another at a given time is modeled via a probability density function depending on infection times and the transmission rate between the two nodes. The proposed algorithm, *NETRATE*, infers pairwise transmission rates and the graph of diffusion by formulating and solving a convex maximum likelihood problem [9].

These methods consider that the underlying network remains static over time. This is not a satisfying assumption, since the topology of OSNs evolves very quickly, both in terms of edges creation and deletion. For that reason, Gomez *et al.* [16] extend *NETRATE* and propose a time-varying inference algorithm, *INFOPATH*, that uses stochastic gradients to provide on-line estimates of the structure and temporal dynamics of a network that changes over time.

In addition, because of technical and crawling API limitations, there is a *data acquisition bottle-*

reference	network		inferred properties			supports missing data
	static	dynamic	pairwaise transmission probability	pairwaise transmission rate	cascade properties	
NETINF	x		x		x	
NETRATE	x		x	x	x	
INFOPATH	x	x	x	x	x	
k-tree model	x				x	x

Table 2: Summary of explanatory models w.r.t the nature of the underlying network, inferred properties and the ability of the method to work with incomplete data.

neck potentially responsible for missing data. To overcome this issue, one approach is to crawl data as efficiently as possible. Choudhury *et al.* [7] analysed how the data sampling strategy impacts the discovery of information diffusion in social media. Based on experimentations on Twitter data, they concluded that sampling methods that consider both network topology and users' attributes such as activity and localisation allow to capture information diffusion with lower error in comparison to naive strategies, like random or activity-only based sampling. Another approach is to develop specific models that assume that data are missing. Sadikov *et al.* [41] develop a method based on a *k*-tree model designed to, given only a fraction of the complete activation sequence, estimate the properties of the complete spreading cascade, such as its size or depth.

We summarize the surveyed explanatory models in Table 2. In the following, we detail the second category of models, namely, predictive models.

4.2 Predictive Models

These models aim at predicting how a specific diffusion process would unfold in a given network, from temporal and/or spatial points of view by learning from past diffusion traces. We classify existing models into two development axes, graph and non-graph based approaches.

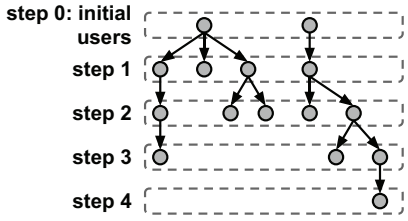

Figure 6: A spreading process modeled by Independent Cascades in four steps.

4.2.1 Graph based approaches

There are two seminal models in this category, namely *Independent Cascades* (*IC*) [13] and *Linear Threshold* (*LT*) [17]. They assume the existence of a static graph structure underlying the diffusion and focus on the structure of the process. They are based on a directed graph where each node can be activated or not with a monotonicity assumption, i.e. activated nodes cannot 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, commonly named *early adopters* [37]:

DEFINITION 8 (Early Adopters). *A set of users who are the first to adopt a piece of information and then trigger its diffusion.*

In the case of *IC*, for each iteration, the newly activated nodes try once to activate their neighbors with the probability defined on the edge joining them. In the case of *LT*, at each iteration, the inactive nodes are activated by their activated neighbors if the sum of influence degrees exceeds their own influence threshold. Successful activations are effective at the next iteration. In both cases, 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. An example of spreading process modeled with *IC* is given by Figure 6. We detail hereafter models arising from those approaches and adapted to OSNs.

Galuba *et al.* [11] propose to use the *LT* model to predict the graph of diffusion, having already observed the beginning of the process. Their model relies on parameters such as information virality, pairwise users degree of influence and user probability of adopting any information. The *LT* model

is fitted on the data describing the beginning of the diffusion process by optimizing the parameters using the gradient ascent method. However, LT can't reproduce realistic temporal dynamics.

Saito *et al.* [42] relax the synchronicity assumption of traditional IC and LT graph-based models by proposing asynchronous extensions. Named $AsIC$ and $AsLT$ (*i.e.* asynchronous independent cascades and asynchronous linear threshold), they proceed iteratively along a continuous time axis and require the same parameters as their synchronous counterparts plus a time-delay parameter on each edge of the graph. Models parameters are defined in a parametric way and authors provide a method to learn the functional dependency of the model parameters from nodes attributes. They formulate the task as a maximum likelihood estimation problem and an update algorithm that guarantees the convergence is derived. However, they only experimented with synthetic data and don't provide a practical solution.

Guille *et al.* [19] also model the propagation process as asynchronous independent cascades. They develop the *T-BaSIC* model (*i.e.* Time-Based Asynchronous Independent Cascades), which parameters aren't fixed numerical values but functions depending on time. The model parameters are estimated from social, semantic and temporal nodes' features using logistic regression.

4.2.2 Non-graph based approaches

Non-graph based approaches do not assume the existence of a specific graph structure and have been mainly developed to model epidemiological processes. They classify nodes into several classes (*i.e.* states) and focus on the evolution of the proportions of nodes in each class. *SIR* and *SIS* are the two seminal models [21, 34], where S stands for "susceptible", I for "infected" (*i.e.* adopted the information) and R for recovered (*i.e.* refractory). In both cases, nodes in the S class switch to the I class with a fixed probability β. Then, in the case of SIS, nodes in the I class switch to the S class with a fixed probability γ, whereas in the case of SIR they permanently switch to the R class. The percentage of nodes in each class is expressed by simple differential equations. Both models assume that every node has the same probability to be connected to another and thus connections inside the population are made at random.

Leskovec *et al.* [28] propose a simple and intuitive SIS model that requires a single parameter, β. It assumes that all nodes have the same probability β to adopt the information and nodes that have adopted the information become susceptible at the next time-step (*i.e.* $\gamma = 1$). This is a strong assumption since in real-world social networks, influence is not evenly distributed between all nodes and it is necessary to develop more complex modeling that take into account this characteristic.

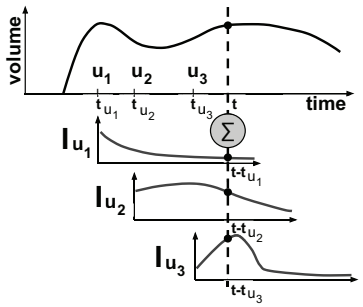

Figure 7: LIM forecasts the rate of diffusion by summing the influence functions of a given set of early adopters. Here, the early adopters are u_1, u_2 and u_3 whose respective influence functions are Iu_1, Iu_2 and Iu_3.

Yang *et al.* [50] start from the assumption that the diffusion of information is governed by the influence of individual nodes. The method focuses on predicting the temporal dynamics of information diffusion, under the form of a time-series describing the rate of diffusion of a piece of information, *i.e.* the volume of nodes that adopt the information through time. They develop a Linear Influence model (LIM), where the influence functions of individual nodes govern the overall rate of diffusion. The influence functions are represented in a non-parametric way and are estimated by solving a non-negative least squares problem using the Reflective Newton Method [8]. Figure 7 illustrates how LIM forecasts the rate of diffusion from a set of early adopters and their activation time.

Wang *et al.* [48] propose a Partial Differential Equation (PDE) based model to predict the diffusion of an information injected in the network by a given node. More precisely, a diffusive logistic equation model is used to predict both topological and temporal dynamics. Here, the topology of the network is considered only in term of the distance from each node to the source node. The dynamics of the process is given by a logistic equation that models the density of influenced users at a given distance of the source and at a given time. That definition of the network topology allows to formulate the problem simply, as for classical non-graph based methods while integrating some spatial knowledge. The

reference	dimension(s)			basis		mathematical modeling	
	social	time	content	graph based	non-graph based	parametric	non-parametric
LT-based	x		x	x		x	
AsIC, AsLT	n/a	n/a	n/a	x		x	
T-BaSIC	x	x	x	x		x	
SIS-based		x			x	x	
LIM	x	x			x		x
PDE	x	x			x	x	

Table 3: Summary of diffusion prediction methods, distinguishing graph and non-graph based approaches w.r.t incorporated dimensions and mathematical modeling.

parameters of the model are estimated using the Cubic Spline Interpolation method [12].

We summarize the surveyed predictive models in Table 3. In the following section, we discuss the role of nodes in the propagation process and how to identify influential spreaders.

5. IDENTIFYING INFLUENTIAL INFORMATION SPREADERS

Identifying the most influential spreaders in a network is critical for ensuring efficient diffusion of information. For instance, a social media campaign can be optimized by targeting influential individuals who can trigger large cascades of further adoptions. This section presents briefly some methods that illustrate the various possible ways to measure the relative importance and influence of each node in an online social network.

DEFINITION 9 (K-Core). *Let G be a graph. If H is a sub-graph of G, $\sigma(H)$ will denote the minimum degree of H. Thus each node of H is adjacent to at least $\sigma(H)$ other nodes of H. If H is a maximal connected (induced) sub-graph of G with $\sigma(H) >= k$, we say that H is a k-core of G [45].*

Kitsak *et al.* [25] show that the best spreaders are not necessarily the most connected people in the network. They find that the most efficient spreaders are those located within the *core* of the network as identified by the k-core decomposition analysis [45], as defined in Definition 9. Basically, the principle of the k-core decomposition is to assign a core index k_s to each node such that nodes with the lowest values are located at the periphery of the network while nodes with the highest values are located in the center of the network. The innermost nodes thus forms the core of the network. Brown *et al.* [5] observe that the results of the k-shell decomposition on Twitter network are highly skewed. Therefore they propose a modified algorithm that uses a logarithmic mapping, in order to produce fewer and more meaningful k-shell values.

Cataldi *et al.* [6] propose to use the well known *PageRank* algorithm [35] to assess the distribution of influence throughout the network. The *PageRank* value of a given node is proportional to the probability of visiting that node in a random walk of the social network, where the set of states of the random walk is the set of nodes.

The methods we have just described only exploit the topology of the network, and ignore other important properties, such as nodes' features and the way they process information. Starting from the observation that most OSNs members are passive information consumers, Romero *et al.* [38] develop a graph-based approach similar to the well known *HITS* algorithm, *IP* (*i.e. Influence-Passivity*), that assigns a relative influence and a passivity score to every users based on the ratio at which they forward information. However, no individual can be a universal influencer, and influential members of the network tend to be influential only in one or some specific domains of knowledge. Therefore, Pal *et al.* [36] develop a non-graph based, topic-sensitive method. To do so, they define a set of nodal and topical features for characterizing the network members. Using probabilistic clustering over this feature space, they rank nodes with a within-cluster ranking procedure to identify the most influential and authoritative people for a given topic. Weng *et al.* [49] also develop a topic-sensitive version of the Page Rank algorithm dedicated to Twitter, TwitterRank.

Kempe *et al.* [24] adopt a different approach and propose to use the *IC* and *LT* models (previously described in Section 4.2.1) to tackle the influence maximization problem. This problem asks, for a parameter k, to find a k-node set of maximum influence in the network. The influence of a given set of nodes corresponds to the number of activated nodes at the end of the diffusion process according

reference	graph based	incorporated dimension(s)	
		users' features	topic
k-shell decomposition	x		
log k-shell decomposition	x		
PageRank	x		
Topic-sensitive PageRank	x		x
IP	x	x	
Topical Authorities		x	x
k-node set	x		

Table 4: Summary of influential spreaders identification methods distinguishing graph and non-graph based approaches w.r.t incorporated dimensions.

to *IC* or *LT*, using this set as the set of initially activated nodes. They provide an approximation for this optimization problem using a greedy hill-climbing strategy based on submodular functions.

The surveyed influence assessment methods are summarized in Table 4.

6. DISCUSSION

In this article, we surveyed representative and state-of-the-art methods related to information diffusion analysis in online social networks, ranging from popular topic detection to diffusion modeling techniques, including methods for identifying influential spreaders. Figure 8 presents the taxonomy of the various approaches employed to address these issues. Hereafter we provide a discussion regarding their shortcomings and related open problems.

6.1 Detecting Popular Topics

The detection of popular topics from the stream of messages produced by the members of an OSN relies on the identification of *bursts*. There are mainly two ways to detect such patterns, by analyzing (i) term frequency or (ii) social interaction frequency. In this area, the following challenges certainly need to be addressed:

Topic definition and scalability. It is obvious that not all methods define a topic in the same way. For instance *Peaky Topics* simply assimilates a topic to a word. It has the advantage to be a low complexity solution, however, the produced result is of little interest. In contrast, *OLDA* defines a topic as a distribution over a set of words but in turn has a high complexity, which prevents it from being applied at large scale. Consequently, there is a need for new methods that could produce intelligible results while preserving efficiency. We identify two possible ways to do so, through: (i) the conception of new scalable algorithms, or (ii) improved implementations of the algorithms using, *e.g.* distributed systems (such as Hadoop).

Social dimension. Furthermore, popular topic detection could be improved by leveraging burstiness and people authority, as does *TSTE*, which relies on the *PageRank* algorithm. However, that possibility remains ill explored so far.

Data complexity. Currently the focus is set on the textual content exchanged in social networks. However, more and more often, users exchange other types of data such as images, videos, URLs pointing to those objects or Web pages, *etc.* This situation has to be fully considered and integrated at the heart of the efforts carried out to provide a complete solution for topic detection.

6.2 Modeling Information Diffusion

We distinguish two types of models, explanatory and predictive. Concerning predictive models, on the one hand there are non-graph based methods, that are limited by the fact that they ignore the topology of the network and only forecast the evolution of the rate at which information globally diffuses. On the other hand, there are graph based approaches that are able to predict who will influence whom. However, they cannot be used when the network is unknown or implicit. Although a lot of effort have been performed in this area, generally speaking, there is a need to consider more realistic constraints when studying information diffusion. In particular, the following issues have to be dealt with:

DEFINITION 10 (Closed World). *The closed world assumption holds that information can only propagate from node to node via the network edges and that nodes cannot be influenced by external sources.*

Closed world assumption. The major observation about modeling information diffusion is certainly that all the described approaches work under a closed world assumption, defined in Definition 10. In other words, they assume that people can only be influenced by other members of the network and that information spreads because of informational cascades. However, most observed spreading pro-

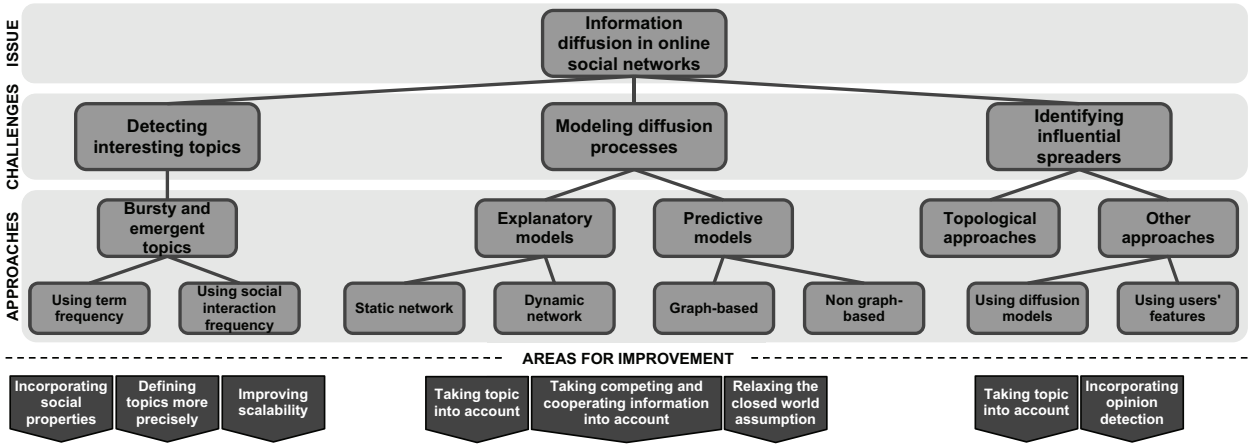

Figure 8: The above taxonomy presents the three main research challenges arising from information diffusion in online social networks and the related types of approaches, annotated with areas for improvement.

cesses in OSNs do not rely solely on social influence. The closed-world assumption is proven incorrect in recent work on Twitter done by Myers *et al.* [32] in which authors observe that information tends to jump across the network. The study shows that only 71% of the information volume in Twitter is due to internal influence and the remaining 29% can be attributed to external events and influence. Consequently they provide a model capable of quantifying the level of external exposure and influence using hazard functions [10]. To relax this assumption, one way would be to align users' profiles across multiple social networking sites. In this way, it would be possible to observe the information diffusion among various platforms simultaneously (subject to the availability of data). Some work tend to address this type of problems by proposing to de-anonymize the social networks [33].

Cooperating and competing diffusion processes. In addition, the described studies rely on the assumption that diffusion processes are independent, *i.e.* each information spreads in isolation. Myers *et al.* [31] argue that spreading processes cooperate and compete. Competing contagions decrease each other's probability of diffusion, while cooperating ones help each other in being adopted. They propose a model that quantifies how different spreading cascades interact with each other. It predicts diffusion probabilities that are on average 71% more or less than the diffusion probability would be for a purely independent diffusion process. We believe that models have to consider and incorporate this knowledge.

Topic-sensitive modeling. Furthermore, it is important for predictive models to be topic-sensitive. Romero *et al.* [39] have studied Twitter and found significant differences in the mechanics of information diffusion across topics. More particularly, they have observed that information dealing with politically controversial topics are particularly persistent, with repeated exposures continuing to have unusually large marginal effects on adoption, which validates the *complex contagion principle* that stipulates that repeated exposures to an idea are particularly crucial when the idea is controversial or contentious.

Dynamic networks. Finally, it is important to note that OSNs are highly dynamic structures. Nonetheless most of the existing work rely on the assumption that the network remains static over time. Integrating link prediction could be a basis to improve prediction accuracy. A more complete review of literature on this topic can be found in [20].

6.3 Identifying Influential Spreaders

There are various ways to tackle this issue, ranging from pure topological approaches, such as *k-shell decomposition* or *HITS* to textual clustering based approaches, including hybrid methods, such as *IP* which combines the *HITS* algorithm with nodes' features. As mentioned previously, there is no such thing as a universal influencer and therefore topic-sensitive methods have also been developed.

Opinion detection. The notion of influence is strongly linked to the notion of opinion. Numerous studies on this issue have emerged in recent years, aiming at automatically detecting opinions or sentiment from corpus of data. We believe that

it might be interesting to include this kind of work in the context of information diffusion. Work dealing with the diffusion of opinions themselves have emerged [29] and it seems that there is an interest to couple these approaches.

6.4 Applications

Even if there are a lot of contributions in the domain of online social networks dynamics analysis, we can remark that implementations are rarely provided for re-use. What is more, available implementations require different formatting of the input data and are written using various programming languages, which makes it hard to evaluate or compare existing techniques. *SONDY* [18] intends to facilitate the implementation and distribution of techniques for online social networks data mining. It is an open-source tool that provides data preprocessing functionalities and implements some of the methods reviewed in this paper for topic detection and influential spreaders identification. It features a user-friendly interface and proposes visualizations for topic trends and network structure.

7. REFERENCES

[1] L. AlSumait, D. Barbará, and C. Domeniconi. On-line lda: Adaptive topic models for mining text streams with applications to topic detection and tracking. In *ICDM '08*, pages 3–12, 2008.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD '08*, pages 7–15, 2008.

[3] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic. The role of social networks in information diffusion. In *WWW '12*, pages 519–528, 2012.

[4] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] P. Brown and J. Feng. Measuring user influence on Twitter using modified k-shell decomposition. In *ICWSM '11 Workshops*, pages 18–23, 2011.

[6] M. Cataldi, L. Di Caro, and C. Schifanella. Emerging topic detection on Twitter based on temporal and social terms evaluation. In *MDMKDD '10*, pages 4–13, 2010.

[7] M. D. Choudhury, Y.-R. Lin, H. Sundaram, K. S. Candan, L. Xie, and A. Kelliher. How does the data sampling strategy impact the discovery of information diffusion in social media? In *ICWSM '10*, pages 34–41, 2010.

[8] T. F. Coleman and Y. Li. A reflective newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM J. on Optimization*, 6(4):1040–1058, Apr. 1996.

[9] I. CVX Research. CVX: Matlab software for disciplined convex programming, version 2.0 beta. `http://cvxr.com/cvx`, sep 2012.

[10] R. C. Elandt-Johnson and N. L. Johnson. *Survival Models and Data Analysis*. John Wiley and Sons, 1980/1999.

[11] W. Galuba, K. Aberer, D. Chakraborty, Z. Despotovic, and W. Kellerer. Outtweeting the twitterers - predicting information cascades in microblogs. In *WOSN '10*, pages 3–11, 2010.

[12] C. F. Gerald and P. O. Wheatley. *Applied numerical analysis with MAPLE; 7th ed.* Addison-Wesley, Reading, MA, 2004.

[13] 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.

[14] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the temporal dynamics of diffusion networks. In *ICML '11*, pages 561–568, 2011.

[15] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *KDD '10*, pages 1019–1028, 2010.

[16] M. Gomez-Rodriguez, J. Leskovec, and B. Schökopf. Structure and dynamics of information pathways in online media. In *WSDM '13*, pages 23–32, 2013.

[17] M. Granovetter. Threshold models of collective behavior. *American journal of sociology*, pages 1420–1443, 1978.

[18] A. Guille, C. Favre, H. Hacid, and D. Zighed. Sondy: An open source platform for social dynamics mining and analysis. In *SIGMOD '13*, (demonstration) 2013.

[19] A. Guille and H. Hacid. A predictive model for the temporal dynamics of information diffusion in online social networks. In *WWW '12 Companion*, pages 1145–1152, 2012.

[20] M. A. Hasan and M. J. Zaki. A survey of link prediction in social networks. In *Social Network Data Analytics*, pages 243–275. Springer, 2011.

[21] H. W. Hethcote. The mathematics of infectious diseases. *SIAM REVIEW*, 42(4):599–653, 2000.

[22] P. N. Howard and A. Duffy. Opening closed

regimes, what was the role of social media during the arab spring? *Project on Information Technology and Political Islam*, pages 1–30, 2011.
[23] A. Hughes and L. Palen. Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260, 2009.
[24] D. Kempe. Maximizing the spread of influence through a social network. In *KDD '03*, pages 137–146, 2003.
[25] M. Kitsak, L. Gallos, S. Havlin, F. Liljeros, L. Muchnik, H. Stanley, and H. Makse. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11):888–893, Aug 2010.
[26] J. Kleinberg. Bursty and hierarchical structure in streams. In *KDD '02*, pages 91–101, 2002.
[27] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD '09*, pages 497–506, 2009.
[28] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst. Cascading behavior in large blog graphs. In *SDM '07*, pages 551–556, (short paper) 2007.
[29] L. Li, A. Scaglione, A. Swami, and Q. Zhao. Phase transition in opinion diffusion in social networks. In *ICASSP '12*, pages 3073–3076, 2012.
[30] J. Makkonen, H. Ahonen-Myka, and M. Salmenkivi. Simple semantics in topic detection and tracking. *Inf. Retr.*, 7(3-4):347–368, Sept. 2004.
[31] S. Myers and J. Leskovec. Clash of the contagions: Cooperation and competition in information diffusion. In *ICDM '12*, pages 539–548, 2012.
[32] S. A. Myers, C. Zhu, and J. Leskovec. Information diffusion and external influence in networks. In *KDD '12*, pages 33–41, 2012.
[33] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *SP '09*, pages 173–187, 2009.
[34] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.
[35] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. In *WWW '98*, pages 161–172, 1998.
[36] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *WSDM '11*, pages 45–54, 2011.
[37] E. M. Rogers. *Diffusion of Innovations, 5th Edition*. Free Press, 5th edition, aug 2003.
[38] D. Romero, W. Galuba, S. Asur, and B. Huberman. Influence and passivity in social media. In *ECML/PKDD '11*, pages 18–33, 2011.
[39] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on Twitter. In *WWW '11*, pages 695–704, 2011.
[40] L. Rong and Y. Qing. Trends analysis of news topics on Twitter. *International Journal of Machine Learning and Computing*, 2(3):327–332, 2012.
[41] E. Sadikov, M. Medina, J. Leskovec, and H. Garcia-Molina. Correcting for missing data in information cascades. In *WSDM '11*, pages 55–64, 2011.
[42] K. Saito, K. Ohara, Y. Yamagishi, M. Kimura, and H. Motoda. Learning diffusion probability based on node attributes in social networks. In *ISMIS '11*, pages 153–162, 2011.
[43] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523, 1988.
[44] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1986.
[45] S. B. Seidman. Network structure and minimum degree. *Social Networks*, 5(3):269 – 287, 1983.
[46] D. A. Shamma, L. Kennedy, and E. F. Churchill. Peaks and persistence: modeling the shape of microblog conversations. In *CSCW '11*, pages 355–358, (short paper) 2011.
[47] T. Takahashi, R. Tomioka, and K. Yamanishi. Discovering emerging topics in social streams via link anomaly detection. In *ICDM '11*, pages 1230–1235, 2011.
[48] F. Wang, H. Wang, and K. Xu. Diffusive logistic model towards predicting information diffusion in online social networks. In *ICDCS '12 Workshops*, pages 133–139, 2012.
[49] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: finding topic-sensitive influential twitterers. In *WSDM '10*, pages 261–270, 2010.
[50] J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *ICDM '10*, pages 599–608, 2010.