Influence Propagation in Social Networks: A Data Mining Perspective

Francesco Bonchi*

***Abstract*—With the success of online social networks and microblogs such as Facebook, Flickr and Twitter, the phenomenon of influence exerted by users of such platforms on other users, and how it propagates in the network, has recently attracted the interest of computer scientists, information technologists, and marketing specialists. One of the key problems in this area is the identification of influential users, by targeting whom certain desirable marketing outcomes can be achieved. In this article we take a data mining perspective and we discuss what (and how) can be learned from the available traces of past propagations. While doing this we provide a brief overview of some recent progresses in this area and discuss some open problems.**

By no means this article must be intended as an exhaustive survey: it is instead (admittedly) a rather biased and personal perspective of the author on the topic of influence propagation in social networks.

***Index Terms*—Social Networks, Social Influence, Viral Marketing, Influence Maximization.**

I. On social influence and viral marketing

The study of the spread of influence through a social network has a long history in the social sciences. The first investigations focused on the adoption of medical [1] and agricultural innovations [2]. Later marketing researchers have investigated the "word-of-mouth" diffusion process for *viral marketing* applications [3], [4], [5], [6].

The basic assumption is that when users see their social contacts performing an action they may decide to perform the action themselves. In truth, when users perform an action, they may have any one of a number of reasons for doing so: they may have heard of it outside of the online social network and may have decided it is worthwhile; the action may be very popular (e.g., buying an iPhone 4S may be such an action); or they may be genuinely influenced by seeing their social contacts perform that action [7]. The literature on these topics in social sciences is wide, and reviewing it is beyond the scope of this article.

The idea behind viral marketing is that by targeting the most influential users in the network we can activate a chain-reaction of influence driven by word-of-mouth, in such a way that with a very small marketing cost we can actually reach a very large portion of the network. Selecting these key users in a wide graph is an interesting learning task that has received a great deal of attention in the last years (for surveys see [8] and Chapter 19 of [9]).

*This article summarizes, extends, and complements the keynote that the author gave at WI/IAT2011 conference, whose slides are available at: `www.francescobonchi.com/wi2011.pdf`

F. Bonchi is with Yahoo! Research, Barcelona, Spain.
E-mail: bonchi@yahoo-inc.com

Other applications include personalized recommendations [10], [11] and feed ranking in social networks [12], [13]. Besides, patterns of influence can be taken as a sign of user trust and exploited for computing trust propagation [14], [15], [16], [17] in large networks and in P2P systems. Analyzing the spread of influence in social networks is also useful to understand how information propagates, and more in general it is related to the fields of epidemics and innovation adoption. With the explosion of microblogging platforms, such as Twitter, the analysis of influence and information propagation in these social media is gaining further popularity [18], [19], [20], [21].

Many of the applications mentioned above essentially assume that social influence exists as a real phenomenon. However several authors have challenged the fact that, regardless the existence of correlation between users behavior with their social context [22], this can be really credited to social influence. Even in the cases where some social influence can be observed, it is not always clear whether this can really propagate and drive viral cascades.

Watts challenges the very notion of influential users that are often assumed in viral marketing papers [23], [24], [25], [19]. Other researchers have focussed on the important problem of distinguishing real social influence from *homophily* and other external factors [26], [27], [28], [29]. Homophily is a term coined by sociologists in the 1950s to explain the tendency of individuals to associate and bond with similar others. This is usually expressed by the famous adage "birds of a feather flock together". Homophily assumes *selection*, i.e., the fact that it is the similarity between users to breed connections [27].

Anagnostopoulos *et al.* [26] develop techniques (e.g., *shuffle test* and *edge-reversal* test) to separate influence from correlation, showing that in Flickr, while there is substantial social correlation in tagging behavior, such correlation cannot be attributed to influence.

However other researchers have instead found evidence of social influence. Some popular (and somehow controversial [30]) findings are due to Christakis and Fowler [31] that report effects of social influence over the spread of obesity (and smoking, alcohol consumption, and other unhealthy – yet pleasant – habits). Crandall *et al.* [27] also propose a framework to analyze the interactions between social influence and homophily. Their empirical analysis over Wikipedia editors social network and LiveJournal blogspace confirms that there exists a feedback effect between users similarity and social influence, and that combining features based on social ties and similarity is more predictive of future behavior than either social influence or similarity features alone, showing that both social influence and one's own interests are drivers of future

behavior and that they operate in relatively independent ways.

Cha *et al.* [32] present a data analysis of how picture popularity is distributed across the Flickr social network, and characterize the role played by social links in information propagation. Their analysis provides empirical evidence that the social links are the dominant method of information propagation, accounting for more than 50% of the spread of favorite-marked pictures. Moreover, they show that information spreading is limited to individuals who are within close proximity of the uploaders, and that spreading takes a long time at each hop, oppositely to the common expectations about the quick and wide spread of word-of-mouth effect.

Leskovec *et al.* show patterns of influence by studying person-to-person recommendation for purchasing books and videos, finding conditions under which such recommendations are successful [33], [34]. Hill *et al.* [35], analyze the adoption of a new telecommunications service and show that it is possible to predict with a certain confidence whether customers will sign up for a new calling plan once one of their phone contacts does the same.

These are just few examples among many studies reporting some evidence of social influence. In this article we do not aim at providing an exhaustive survey, nor we dare entering the debate on the existence of social influence at the philosophical/sociological level. We do not even discuss further how to distinguish between social influence, homophily and other factors, although we agree that it is an interesting research problem. Instead, we prefer to take an algorithmic and data mining perspective, focussing on available data and on developing learning frameworks for social influence analysis.

Once sociologists had to infer and reconstruct social networks by tracking people relations in the real world. This is obviously a challenging and costly task, even to produce moderately sized social networks. Fortunately nowadays, thanks to the success of online social networks, we can collect very large graphs of explicitly declared social relations. Moreover, and maybe more importantly, we can collect information about the users of these online social networks performing some actions (e.g., post messages, pictures, or videos, buy, comment, link, rate, share, like, retweet) and the time at which such actions are performed. Therefore we can track real propagations in social networks. If we observe in the data user v performing an action a at time t, and user u, which is a "friend" of v, performing the same action shortly after, say at time $t + \Delta$, then we can think that action a propagated from v to u. If we observe this happening frequently enough, for many different actions, then we can safely conclude that user v is indeed exerting some influence on u.

In the rest of this article we will focus on this kind of data, i.e., a database of past propagations in a social network. We will emphasize that when analyzing social influence, it is important to consider this data and not only the structure of the social graph. Moreover, as this database of propagations might be potentially huge, we will highlight the need for devising clever algorithms that, by exploiting some incrementality property, can perform the needed computation with as few scans of the database as possible.

II. Influence Maximization

Suppose we are given a social network, that is a graph whose nodes are users and links represent social relations among the users. Suppose we are also given the estimates of reciprocal influence between individuals connected in the network, and suppose that we want to push a new product in the market. The mining problem of *influence maximization* is the following: given such a network with influence estimates, how should one select the set of initial users so that they eventually influence the largest number of users in the social network. This problem has received a good deal of attention by the data mining research community in the last decade.

The first to consider the propagation of influence and the problem of identification of influential users by a data mining perspective are Domingos and Richardson [36], [37]. They model the problem by means of *Markov random fields* and provide heuristics for choosing the users to target. In particular, the marketing objective function to maximize is the global expected lift in profit, that is, intuitively, the difference between the expected profit obtained by employing a marketing strategy and the expected profit obtained using no strategy at all [38]. A Markov random field is an undirected graphical model representing the joint distribution over a set of random variables, where vertices are variables, and edges represent dependencies between variables. It is adopted in the context of influence propagation by modelling only the final state of the network at convergence as one large global set of interdependent random variables.

Kempe *et al.* [39] tackle roughly the same problem as a problem in discrete optimization, obtaining provable approximation guarantees in several preexisting models coming from mathematical sociology. In particular their work focuses on two fundamental propagation models, named *Linear Threshold Model* (LT) and *Independent Cascade Model* (IC). In both these models, at a given timestamp, each node is either active (an adopter of the innovation, or a customer which already purchased the product) or inactive, and each node's tendency to become active increases monotonically as more of its neighbors become active. An active node never becomes inactive again. Time unfolds deterministically in discrete steps. As time unfolds, more and more of neighbors of an inactive node u become active, eventually making u become active, and u's decision may in turn trigger further decisions by nodes to which u is connected.

In the IC model, when a node v first becomes active, say at time t, it is considered contagious. It has one chance of influencing each inactive neighbor u with probability $p_{v,u}$, independently of the history thus far. If the tentative succeeds, u becomes active at time $t + 1$. The probability $p_{v,u}$, that can be considered as the strength of the influence of v over u.

In the LT model, each node u is influenced by each neighbor v according to a weight $p_{v,u}$, such that the sum of incoming weights to u is no more than 1. Each node u chooses a threshold θ_u uniformly at random from $[0, 1]$. At any timestamp t, if the total weight from the active neighbors of an inactive node u is at least θ_u, then u becomes active at timestamp $t + 1$.

In both the models, the process repeats until no new node becomes active. Given a propagation model m (e.g., IC or LT) and an initial seed set $S \subseteq V$, the expected number of active nodes at the end of the process is the *expected (influence) spread* of S, denoted by $\sigma_m(S)$. Then the *influence maximization problem* is defined as follows: given a directed and edge-weighted social graph $G = (V, E, p)$, a propagation model m, and a number $k \leq |V|$, find a set $S \subseteq V$, $|S| = k$, such that $\sigma_m(S)$ is maximum.

Under both the IC and LT propagation models, this problem is **NP**-hard [39]. Kempe *et al.*, however, showed that the function $\sigma_m(S)$ is *monotone* and *submodular*. Monotonicity says as the set of activated nodes grows, the likelihood of a node getting activated should not decrease. In other words, $S \subseteq T$ implies $\sigma_m(S) \leq \sigma_m(T)$. Submodularity intuitively says that the probability for an active node to activate some inactive node u does not increase if more nodes have already attempted to activate u (u is, so to say, more "marketing saturated"). This is also called *"the law of diminishing returns"*. More precisely, $\sigma_m(S \cup \{w\}) - \sigma_m(S) \geq \sigma_m(T \cup \{w\}) - \sigma_m(T)$ whenever $S \subseteq T$.

Thanks to these two properties we can have a simple greedy algorithm (see Algorithm 1), which provides an approximation guarantee. In fact, for any monotone submodular function f with $f(\emptyset) = 0$, the problem of finding a set S of size k such that $f(S)$ is maximum, can be approximated to within a factor of $(1 - 1/e)$ by the greedy algorithm, as shown in an old result by Nemhauser *et al.* [40]. This result carries over to the influence maximization problem [39], meaning that the seed set we produce by means of Algorithm 1 is guaranteed to have an expected spread $> 63\%$ of the expected spread of the optimal seed set.

Although simple, Algorithm 1 is computationally prohibitive. The complex step of the greedy algorithm is in line 3, where we select the node that provides the largest marginal gain $\sigma_m(S \cup \{v\}) - \sigma_m(S)$ with respect to the expected spread of the current seed set S. Indeed, computing the expected spread of given set of nodes is $\#\mathbf{P}$-hard under both the IC model [41], [13] and the LT model [42]. In their paper, Kempe *et al.* run Monte Carlo (MC) simulations of the propagation model for sufficiently many times to obtain an accurate estimate of the expected spread. In particular, they show that for any $\phi > 0$, there is a $\delta > 0$ such that by using $(1+\delta)$-approximate values of the expected spread, we obtain a $(1-1/e-\phi)$-approximation for the influence maximization problem. However, running many propagation simulations (Kempe *et al.* report $10,000$ trials for each estimation in their experiments) is practically unfeasible on very large real-world social networks. Therefore, following [39] many researchers have focussed on developing methods for improving the efficiency and scalability of influence maximization algorithms, as discussed next.

Leskovec *et al.* [43] study the propagation problem by a different perspective namely *outbreak detection*: how to select nodes in a network in order to detect as quickly as possible the spread of a virus? They present a general methodology for near optimal sensor placement in these and related problems. They also prove that the influence maximization problem of [39] is a special case of their more general problem definition. By exploiting submodularity they develop an efficient algorithm based on a "lazy-forward" optimization in selecting new seeds, achieving near optimal placements, while being 700 times faster than the simple greedy algorithm.

Algorithm 1 Greedy alg. for influence maximization [39]

Require: G, k, σ_m
Ensure: seed set S
1: $S \leftarrow \emptyset$
2: **while** $|S| < k$ **do**
3: $\quad u \leftarrow \arg\max_{w \in V \setminus S}(\sigma_m(S \cup \{w\}) - \sigma_m(S))$;
4: $\quad S \leftarrow S \cup \{u\}$

Regardless of this big improvement over the basic greedy algorithm, their method still face serious scalability problems as shown in [44]. In that paper, Chen *et al.* improve the efficiency of the greedy algorithm and propose new degree discount heuristics that produce influence spread close to that of the greedy algorithm but much more efficiently.

In their following work Chen *et al.* [41] propose scalable heuristics to estimate coverage of a set under the IC model by considering Maximum Influence Paths (MIP). A MIP between a pair of nodes (v, u) is the path with the maximum propagation probability from v to u. The idea is to restrict the influence propagation through the MIPs. Based on this, the authors propose two models: *maximum influence arborescence* (MIA) model and its extension, the *prefix excluding MIA* (PMIA) model.

Very recently, Chen *et al.* [42] proposed a scalable heuristic for the LT model. They observe that, while computing the expected spread (or coverage) is $\#\mathbf{P}$-hard in general graphs, it can be computed in linear time in DAGs (directed acyclic graphs). They exploit this property by constructing local DAGs (LDAG) for every node in the graph. A LDAG for user u contains the nodes that have significant influence over u (more than a given threshold θ). Based on this idea, they propose a heuristic called LDAG which provides close approximation to Algorithm 1 and is highly scalable.

III. Propagation traces

In most of the literature on influence maximization (as the set of papers discussed above), the directed link-weighted social graph is assumed as input to the problem. Probably due to the difficulties in finding real propagation traces, researchers have simply given for granted that we can learn the links probabilities (or weights) from some available past propagation data, without addressing how to actually do that (with the exception of few articles described in the next section). This way they have been able to just focus on developing algorithms for the problem which takes the already-weighted graph as input.

However, in order to run experiments, the edge influence weights/probabilities are needed. Thus researches have often assumed some trivial model of links probabilities for their experiments. For instance, for the IC model often experiments are conducted assuming *uniform* link probabilities (e.g., all

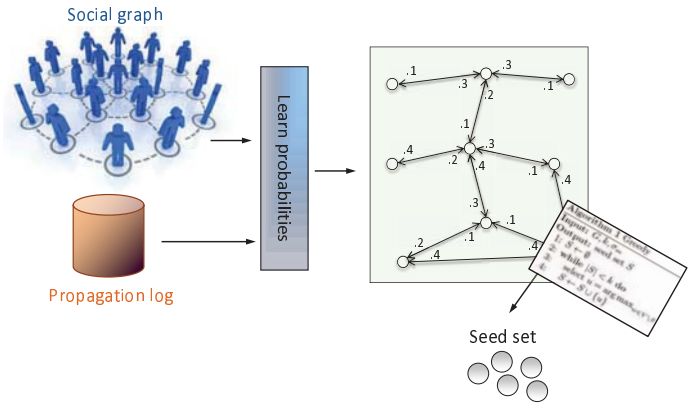

Fig. 1. The standard influence maximization process.

links have probability $p = 0.01$), or the *trivalency* (*TV*) model where link probabilities are selected uniformly at random from the set $\{0.1, 0.01, 0.001\}$, or assuming the *weighted cascade* (*WC*) model, that is $p(u, v) = 1/d_v$ where d_v represent the in-degree of v (see e.g., [39], [41]).

These experiments usually are aimed at showing that a newly proposed heuristic select a seed set S much more efficiently than Algorithm 1, without losing too much in terms of expected spread achieved $\sigma_m(S)$.

In a recent paper Goyal *et al.* [45] have compared the different outcomes of the greedy Algorithm 1 under the IC model, when adopting different ways of assigning probabilities. In particular, they have compared the trivial models discussed above with influence probabilities learned from past propagation traces. This is done by means of two experiments on real-world datasets.

In the first experiment the overlap of the seed sets extracted under the different settings is measured. In the second experiment, the log of past propagations is divided in training and test set, where the training set is used for learning the probabilities. Then for each propagation in the test set, the set of users that are the first to participate in the propagation among their friends, i.e., the set of "initiators" of the action, is considered as the seed set, and the actual spread, i.e., the size of the propagation in the test set, is what the various methods have to predict.

The outcome of this experimentation is that: (i) the seed sets extracted under different probabilities settings are very different (with empty or very small intersection) , and (ii) the method based on learned probabilities outperforms the trivial methods of assigning probabilities in terms of accuracy in predicting the spread. The conclusion is hence that it is extremely important to exploit available past propagation traces to learn the probabilities.

In Figure 1, we summarize the standard process followed in influence maximization making explicit the phase of learning the link probabilities. The process starts with the (unweighted) social graph and a log of past action propagations that say when each user performed an action. The log is used to estimate influence probabilities among the nodes. This produces the directed link-weighted graph which is then given as input to the greedy algorithm to produce the seed set using MC simulations.

We can consider the propagation log to be a relational table with schema $(user\_ID, action\_ID, time)$. We say that an action propagates from node u to node v whenever u and v are socially linked (have an edge in the social graph), and u performs the action before v. In this case we can also assume that u contributes in influencing v to perform that action. From this perspective, an action propagation can be seen as a flow, i.e., a directed subgraph, over the underlying social network. It is worth noting, that such a flow is a DAG: it is directed, each node can have zero or more parents, and cycles are impossible due to the time constraint. Therefore, another way to consider the propagation log is as a database (a set) of DAGs, where each DAG is an instance of the social graph.

In the rest of this article we will always consider the same input consisting of two pieces: (1) the social graph, and (2) the log of past propagations. We will see how different problems and approaches can be defined based on this input.

IV. Learning the influence probabilities

Saito *et al.* [46] were the first to study how to learn the probabilities for the IC model from a set of past propagations. They neatly formalize the likelihood maximization problem and then apply Expectation Maximization (EM) to solve it.

However, their theoretical formulation has some limitations when it comes to practice. One main issue is that they assume as input propagations that have the same shape as they were generated by the IC model itself. This means that an input propagation trace is a sequence of sets of users $D_0, \ldots, D_n$, corresponding to the sets of users activated in the corresponding discrete time steps of the IC propagation. Moreover for each node $u \in D_i$ it must exists a neighbor v of u such that $v \in D_{i-1}$. This is obviously not the case in real-world propagation traces, and some pre-processing is needed to close this gap between the model and the real data (as discussed in [47], [45]).

Another practical limitation of the EM-based method is discussed by Goyal *et al.* [45]. Empirically they found that the seed nodes picked by the greedy algorithm – with the IC model and probabilities learned with the EM-based method [46] – are all nodes which perform a very small number of actions, often just one action, and should not be considered as high influential nodes. For instance, Goyal *et al.* [45] report that in one experiment the first seed selected is a node that in the propagation traces appears only once, i.e., it performs only one action. But this action propagates to 20 of its neighbors. As a result, the EM-based method ends up assigning probability 1.0 to the edges from that node to all its 20 neighbors, making it a high influence node, so much influential that it results being picked as the first seed by the greedy algorithm. Obviously, in reality, such node cannot be considered as a highly influential node since its influence is not statistically significant.

Finally, another practical limit of the EM-based method is its scalability, as it needs to update the influence probability associated to each edge in each iteration.

Goyal *et al.* also studied the problem of learning influence probabilities [48], but under a different model, i.e., an instance

of the *General Threshold Model* (or the equivalent *General Cascade Model* [39]). They extended this model by making influence probabilities decay with time. Indeed it has been observed by various researchers in various domains and on real data, that the probability of influence propagation decays exponentially on time. This means that if u is going to re-do an action (e.g., re-tweet a post) of v, this is likely going to happen shortly after v has performed the action, or never.

Goyal *et al.* [48] propose three classes of influence probabilities models. The first class of models assumes the influence probabilities are static and do not change with time. The second class of models assumes they are continuous functions of time. In the experiments it turns out that time-aware models are by far more accurate, but they are very expensive to learn on large data sets, because they are not incremental. Thus, the authors propose an approximation, known as Discrete Time models, where the joint influence probabilities can be computed incrementally and thus efficiently.

Their results give evidence that Discrete Time models are as accurate as continuous time ones, while being order of magnitude faster to compute, thus representing a good trade-off between accuracy and efficiency.

As the propagation log might be potentially huge, Goyal *et al.* pay particular attention in minimizing the number of scans of the propagations needed. In particular, they devise algorithms that can learn all the models in no more than two scans.

In that work, factors such as the *influenceability* of a specific user, or how influence-driven is a certain action are also investigated.

Finally, the authors show that their methods can also be used to predict *whether* a user will perform an action and *when* with high accuracy, and the precision is higher for user which have an high influenceability score.

V. Direct mining approaches

So far we have followed the standard approach to the influence maximization problem as depicted in Figure 1. First use a log of past propagations to learn edge-wise influence probability, then recombine these probabilities together by means of a MC simulation, in order to estimate the expected spread of a set of nodes.

Recently new approaches emerged trying to mine directly the two pieces of input (the social graph and the propagation log) in order to build a model of the influence spread of a set of nodes, avoiding the approach based on influence probability learning and MC simulation.

Goyal *et al.* [45] take a different perspective on the definition of the expected spread $\sigma_m(S)$, which is the objective function of the influence maximization problem. Note that both the IC and LT models discussed previously are probabilistic in nature. In the IC model, coin flips decide whether an active node will succeed in activating its peers. In the LT model it is the node threshold chosen uniformly at random, together with the influence weights of active neighbors, that decides whether a node becomes active.

Under both models, we can think of a propagation trace as a *possible world*, i.e., a possible outcome of a set of probabilistic choices. Given a propagation model and a directed and edge-weighted social graph $G = (V, E, p)$, let $\mathbb{G}$ denote the set of all possible worlds. Independently of the model m chosen, the expected spread $\sigma_m(S)$ can be written as:

$$\sigma_m(S) = \sum_{X \in \mathbb{G}} Pr[X] \cdot \sigma_m^X(S) \quad (1)$$

where $\sigma_m^X(S)$ is the number of nodes reachable from S in the possible world X. The number of possible worlds is clearly exponential, thus the standard approach (MC simulations) is to sample a possible world $X \in \mathbb{G}$, compute $\sigma_m^X(S)$, and repeat until the number of sampled worlds is large enough.

We now rewrite Eq. (1), obtaining a different perspective. Let $path(S, u)$ be an indicator random variable that is 1 if there exists a directed path from the set S to u and 0 otherwise. Moreover let $path_X(S, u)$ denote the value of the random variable in a possible world $X \in \mathbb{G}$. Then we have:

$$\sigma_m^X(S) = \sum_{u \in V} path_X(S, u) \quad (2)$$

Substituting in (1) and rearranging the terms we have:

$$\sigma_m(S) = \sum_{u \in V} \sum_{X \in \mathbb{G}} Pr[X]\, path_X(S, u) \quad (3)$$

The value of a random variable averaged over all possible worlds is, by definition, its expectation. Moreover the expectation of an indicator random variable is simply the probability of the positive event.

$$\sigma_m(S) = \sum_{u \in V} E[path(S, u)] = \sum_{u \in V} Pr[path(S, u) = 1] \quad (4)$$

That is, the expected spread of a set S is the sum over each node $u \in V$, of the probability of the node u getting activated given that S is the initial seed set.

While the standard approach samples possible worlds from the perspective of Eq. (1), Goyal *et al.* [45] observe that real propagation traces are similar to possible worlds, except they are *"real available worlds"*. Thus they approach the computation of influence spread from the perspective of Eq. (4), i.e., estimate directly $Pr[path(S, u) = 1]$ using the propagation traces available in the propagation log.

In order to estimate $Pr[path(S, u) = 1]$ using available propagation traces, it is natural to interpret such quantity as the fraction of the actions initiated by S that propagated to u, given that S is the seed set. More precisely, we could estimate this probability as

$$\frac{|\{a \in \mathcal{A} | initiate(a, S) \,\&\, \exists t : (u, a, t) \in \mathbb{L}\}|}{|\{a \in \mathcal{A} | initiate(a, S)\}|}$$

where $\mathbb{L}$ denotes the propagation log, and $initiate(a, S)$ is true iff S is precisely the set of initiators of action a. Unfortunately, this approach suffers from a *sparsity issue* which is intrinsic to the influence maximization problem.

Consider for instance a node x which is a very influential user for half of the network, and another node y which is a very influential user for the other half of the network. Their union $\{x, y\}$ is likely to be a very good seed set, but we can not estimate its spread by using the fraction of the actions

containing $\{x, y\}$, because we might not have any propagation in the data with $\{x, y\}$ as the actual seed set.

Summarizing, if we need to estimate $Pr[path(S, u) = 1]$ for any set S and node u, we will need an enormous number of propagation traces corresponding to various combinations, where each trace has as its initiator set precisely the required node set S. It is clearly impractical to find a real-world action log where this can be realized (unless somebody sets up a large scale human-based experiment, where many propagations are started with the desired seed sets). It should be noted that this sparsity issue, is also the reason why it is impractical to compare two different influence maximization methods on the basis of a ground truth.

To overcome this obstacle, the authors propose a "u-centric" perspective to the estimation of $Pr[path(S, u) = 1]$: they scan the propagation log and each time they observe u performing an action they distribute "credits" to the possible influencers of a node u, retracing backwards the propagation network. This model is named *"credit distribution"* model.

Another direct mining approach, although totally different from the credit distribution model, and *not* aimed at solving the influence maximization problem was proposed by Goyal *et al.* few years ago in [49], [50]. In these papers they propose a framework based on the discovery of *frequent pattern of influence*, by mining the social graph and the propagation log. The goal is to identify the "leaders" and their "tribes" of followers in a social network.

Inspired by frequent pattern mining and association rules mining [51], Goyal *et al.*, define the notion of leadership based on how frequently a user exhibits influential behavior. In particular a user u is considered *leader* w.r.t. an action a provided u performed a and within a chosen time bound after u performed a, a sufficient number of other users performed a. Moreover these other users must be reachable from u thus capturing the role social ties may have played. If a user is found to act as a leader for sufficiently many actions, then it is considered a leader.

A stronger notion of leadership might be based on requiring that w.r.t. each of a class of actions of interest, the set of influenced users are the same. To distinguish from the notion of leader above, Goyal *et al.* refer to this notion as *tribe leader*, meaning the user leads a fixed set of users (tribe) w.r.t. a set of actions. Clearly, tribe leaders are leaders but not vice versa.

Other constraints are added to the framework. The influence emanating from some leaders may be "subsumed" by others. Therefore, in order to rule out such cases Goyal *et al.* introduce the concept of *genuineness*. Finally, similarly to association rules mining, also the constraint of *confidence* is included in the framework.

As observed before, the propagation log might potentially be very large, the algorithmic solution must always try to minimize the number of scans of the propagation log needed. This is fundamental to achieve efficiency. In both the "credit distribution" model [45], and the "leaders and tribes" framework [49], [50], Goyal *et al.* develops algorithms that scan the propagation log only once.

VI. Sparsification of Influence Networks

In this section we review another interesting problem defined over the same input: (1) the social graph, and (2) the log of past propagations.

Given these two pieces of input, assuming the IC propagation model, and assuming to have learned the edge influence probabilities, Mathioudakis *et al.* [47] study the problem of selecting the k most important links in the model, i.e., the set of k links that maximize the likelihood of the observed propagations. Here k might be an input parameter specified by the data analyst, or alternatively k might be set automatically following common model-selection practice. Mathioudakis *et al.* show that the problem is **NP**-hard to approximate within any multiplicative factor. However, they show that the problem can be decomposed into a number of subproblems equal to the number of the nodes in the network, in particular by looking for a sparsification for the in-degree of each node. Thanks to this observation they obtain a dynamic programming algorithm which delivers the optimal solution. Although exponential, the search space of this algorithm is typically much smaller than the brute force one, but still impracticable for graphs having nodes with a large in-degree.

Therefore Mathioudakis *et al.* devise a greedy algorithm named Spine (*Sparsification of influence networks*), that achieves efficiency with little loss in quality.

Spine is structured in two phases. During the first phase it selects a set of arcs D_0 that yields a log-likelihood larger than $-\infty$. This is done by means of a greedy approximation algorithm for the `Hitting Set` **NP**-hard problem. During the second phase, it greedily seeks a solution of maximum log-likelihood, i.e., at each step the arc that offers the largest increase in log-likelihood is added to the solution set.

The second phase has an approximation guarantee. In fact, while log-likelihood is negative, and not equal to zero for an empty solution, if we consider the gain in log-likelihood w.r.t. the base solution D_0 as our objective function, and we seek a solution of size $k - |D_0|$, then we have a monotone, positive and submodular function g, having $g(\emptyset) = 0$, for which we can apply again the result of Nemhauser *et al.* [40]. Therefore, the solution returned by the Spine algorithm is guaranteed to be "close" to the optimal among the subnetworks that include the set of arcs D_0.

Sparsification is a fundamental operation that can have countless applications. Its main feature is that by keeping only the most important edges, it essentially highlights the backbone of influence and information propagations in social networks. Sparsifying separately different information topics can help highlighting the different backbone of, e.g., sport or politics. Sparsification can be used for feed ranking [13], i.e., ranking the most interesting feeds for a user. Using the backbone as representative of a group of propagations, can be used for modeling and prototype-based clustering of propagations. Finally, as shown by Mathioudakis *et al.* [47], sparsification can be used as simple data-reduction pre-processing before solving the influence maximization problem. In particular, in their experiments Mathioudakis *et al.* show that by applying Spine as preprocessor, and keeping only half of the links,

Algorithm 1 can achieve essentially the same influence spread σ_m that it would achieve on the whole network, while being an order of magnitude faster.

Another similar problem is tackled by Gomez-Rodriguez *et al.* [52], [53], that assume that the propagations are known, but the network is not. In particular, they assume that connections between nodes cannot be observed, and they use observed traces of activity to infer a sparse, "hidden" network of information diffusion.

Serrano *et al.* [54], as well as Foti *et al.* [55], focus on weighted networks and select edges that represent statistically significant deviations with respect to a null model.

VII. Concluding remarks and open problems

We have provided a brief, partial, and biased survey on the topic of social influence and how it propagates in social networks, mainly focussing on the problem of influence maximization for viral marketing. We have emphasized that while most of the literature has been focussing only on the social graph, it is very important to exploit available traces of past propagations. Finally, we have highlighted the importance of devising clever algorithms to minimize the number of scans of the propagations log.

Although this topic has received a great deal of attention in the last years, many problems remain more or less open.

Learning the strength of the influence exerted from a user of a social network on another user, is a relevant task whose importance goes beyond the mere influence maximization process as depicted in Figure 1. Although some effort has been devoted to investigating this problem (as partially reviewed in Section IV), there is still plenty of room for improving the models and the algorithms for such a learning task.

One important aspect, only touched in [48] is to consider the different levels of user influenceability, as well as the different level of action virality, in the theory of viral marketing and influence propagation. Another extremely important factor is the temporal dimension: nevertheless the role of time in viral marketing is still largely (and surprisingly) unexplored.

We have seen that direct mining methods, as those ones described in Section V, are promising both for what concerns the accuracy and the efficiency in modeling the spread of social influence. In the next years we expect to see more models of this kind.

In a recent paper, Bakshy *et al.* [19] challenge the vision of word-of-mouth propagations that are driven disproportionately by a small number of key influencers. Instead they claim that word-of-mouth diffusion can only be harnessed reliably by targeting large numbers of potential influencers, thereby capturing average effects. From this perspective the "leaders and tribes" framework [49], [50] might be an appealing basic brick to build more complex solutions (as it often happens with frequent local patterns which are not very interesting *per se*, but that are very useful to build global models). It would be interesting to see how tribe leaders extracted with the framework of [49], [50] perform when used as seed set in the influence maximization process. Another appealing idea is to use these small tribes as basic units to build larger communities, thus moving towards *community detection based on influence/information propagation*.

The influence maximization problem as defined by Kempe et al. [39] assumes that there is only one player introducing only a product in the market. However, in the real world, is more likely the case where multiple players are competing with comparable products over the same market. Just think about consumers technologies such as videogame consoles (X-Box Vs. Playstation) or reflex digital cameras (Canon Vs. Nikon): as the adoption of these consumers technologies is not free, it is very unlikely that the average consumer will adopt both competing products. Thus is makes sense to formulate the influence maximization problem in terms of mutually exclusive and competitive products. While there are two papers that have tackled this problem independently and concurrently in 2007 [56], [57], their contribution is mostly theoretical and leaves plenty of room for developing more concrete analysis and methods.

One important aspect largely left uncovered in the current literature is the fact that some people are more likely to buy a product than others, e.g., teenagers are more likely to buy videogames than seniors. Similarly, a user which is influential w.r.t. classic rock music, is not very likely to be influential for what concerns techno music too. These considerations highlight the need of, (1) methods that can take benefit of additional information associated to the nodes (the users) of a social network (e.g., demographics, behavioral information), and (2), methods to incorporate topic modeling in the influence analysis. While some preliminary work in this direction exists [58], [18], [59], we believe that the synergy of topic modeling and influence analysis is still in its infancy, and we expect this to become an hot research area in the next years.

Mining influence propagations data for applications such as viral marketing has non-trivial privacy issues. Studying the privacy threats associated to these mining activities and devising methods respectful of the privacy of the social networks users are important problems.

Finally, the main open challenge in our opinion is that the influence maximization problem, as defined by Kempe et al. [39] and as reviewed in this article, is still an ideal problem: how to make it actionable in the real world? Propagation models, e.g., the IC and LT models reviewed in Section II (but many more exist in the literature), make many assumptions: which of these assumptions are more realistic and which are less? Which propagation model does better describe the real-world? We need to develop techniques and benchmarks for comparing different propagation models and the associated influence maximization methods on the basis of ground-truth.

Acknowledgements

I wish to thank Amit Goyal and Laks V. S. Lakshmanan which are my main collaborators in the research on the topic of influence propagation and the co-authors of most of the papers discussed in this article. I would also like to thank Michael Mathioudakis and my colleagues at Yahoo! Research Barcelona: Carlos Castillo, Aris Gionis, and Antti Ukkonen.

I wish to thank Paolo Boldi for helpful discussions and detailed comments on an earlier version of this manuscript.

Finally I would like to thank the chairs and organizers of WI-IAT 2011 (www.wi-iat-2011.org) conference for inviting me to give a keynote, as well as the editors of the IEEE Intelligent Informatics Bulletin for inviting me to summarize the keynote in this article. My research on influence propagation is partially supported by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037, "Social Media" (www.cenitsocialmedia.es).

REFERENCES

[1] J. Coleman, H. Menzel, and E. Katz, *Medical Innovations: A Diffusion Study*. Bobbs Merrill, 1966.

[2] T. Valente, *Network Models of the Diffusion of Innovations*. Hampton Press, 1955.

[3] F. Bass, "A new product growth model for consumer durables," *Management Science*, vol. 15, pp. 215–227, 1969.

[4] 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*, vol. 12, no. 3, pp. 211–223, 2001.

[5] V. Mahajan, E. Muller, and F. Bass, "New product diffusion models in marketing: A review and directions for research," *Journal of Marketing*, vol. 54, no. 1, pp. 1–26, 1990.

[6] S. Jurvetson, "What exactly is viral marketing?" *Red Herring*, vol. 78, pp. 110–112, 2000.

[7] N. E. Friedkin, *A Structural Theory of Social Influence*. Cambridge University Press, 1998.

[8] J. Wortman, "Viral marketing and the diffusion of trends on social networks," University of Pennsylvania, Tech. Rep. Technical Report MS-CIS-08-19, May 2008.

[9] D. Easley and J. Kleinberg, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.

[10] X. Song, B. L. Tseng, C.-Y. Lin, and M.-T. Sun, "Personalized recommendation driven by information flow," in *Proc. of the 29th ACM SIGIR Int. Conf. on Research and development in information retrieval (SIGIR'06)*, 2006.

[11] X. Song, Y. Chi, K. Hino, and B. L. Tseng, "Information flow modeling based on diffusion rate for prediction and ranking," in *Proc. of the 16th Int. Conf. on World Wide Web (WWW'07)*, 2007.

[12] J. J. Samper, P. A. Castillo, L. Araujo, and J. J. M. Guervós, "Nectarss, an rss feed ranking system that implicitly learns user preferences," *CoRR*, vol. abs/cs/0610019, 2006.

[13] D. Ienco, F. Bonchi, and C. Castillo, "The meme ranking problem: Maximizing microblogging virality," in *Proc. of the SIASP workshop at ICDM'10*, 2010.

[14] R. Guha, R. Kumar, P. Raghavan, and A. Tomkins, "Propagation of trust and distrust," in *Proc. of the 13th Int. Conf. on World Wide Web (WWW'04)*, 2004.

[15] C.-N. Ziegler and G. Lausen, "Propagation models for trust and distrust in social networks," *Information Systems Frontiers*, vol. 7, no. 4-5, pp. 337–358, 2005.

[16] J. Golbeck and J. Hendler, "Inferring binary trust relationships in web-based social networks," *ACM Trans. Internet Technol.*, vol. 6, no. 4, pp. 497–529, 2006.

[17] M. Taherian, M. Amini, and R. Jalili, "Trust inference in web-based social networks using resistive networks," in *Proc. of the 2008 Third Int. Conf. on Internet and Web Applications and Services (ICIW'08)*, 2008.

[18] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proc. of the Third Int. Conf. on Web Search and Web Data Mining (WSDM'10)*, 2010.

[19] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, "Everyone's an influencer: quantifying influence on twitter," in *Proc. of the Forth Int. Conf. on Web Search and Web Data Mining (WSDM'11)*, 2011.

[20] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," in *Proc. of the 20th Int. Conf. on World Wide Web (WWW'11)*, 2011.

[21] D. M. Romero, B. Meeder, and J. M. Kleinberg, "Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter," in *Proc. of the 20th Int. Conf. on World Wide Web (WWW'11)*, 2011.

[22] P. Singla and M. Richardson, "Yes, there is a correlation: - from social networks to personal behavior on the web," in *Proc. of the 17th Int. Conf. on World Wide Web (WWW '08)*, 2008.

[23] D. Watts and P. Dodds, "Influential, networks, and public opinion formation," *Journal of Consumer Research*, vol. 34, no. 4, pp. 441–458, 2007.

[24] D. Watts, "Challenging the influentials hypothesis," *WOMMA Measuring Word of Mouth, Volume 3*, pp. 201–211, 2007.

[25] D. Watts and J. Peretti, "Viral marketing for the real world," *Harvard Business Review*, pp. 22–23, May 2007.

[26] A. Anagnostopoulos, R. Kumar, and M. Mahdian, "Influence and correlation in social networks," in *Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'08)*, 2008.

[27] D. J. Crandall, D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, and S. Suri, "Feedback effects between similarity and social influence in online communities," in *Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'08)*, 2008.

[28] S. Aral, L. Muchnik, and A. Sundararajan, "Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks," *Proc. of the National Academy of Sciences*, vol. 106, no. 51, pp. 21 544–21 549, 2009.

[29] T. L. Fond and J. Neville, "Randomization tests for distinguishing social influence and homophily effects," in *Proc. of the 19th Int. Conf. on World Wide Web (WWW'10)*, 2010.

[30] R. Lyons, "The spread of evidence-poor medicine via flawed social-network analysis," *Statistics, Politics, and Policy*, vol. 2, no. 1, 2011.

[31] N. A. Christakis and J. H. Fowler, "The spread of obesity in a large social network over 32 years," *The New England Journal of Medicine*, vol. 357(4), pp. 370–379, 2007.

[32] M. Cha, A. Mislove, and P. K. Gummadi, "A measurement-driven analysis of information propagation in the flickr social network," in *Proc. of the 18th Int. Conf. on World Wide Web (WWW'09)*, 2009.

[33] J. Leskovec, A. Singh, and J. M. Kleinberg, "Patterns of influence in a recommendation network," in *Proc. of the 10th Pacific-Asia Conf. on Knowledge Discovery and Data Mining,(PAKDD'06)*, 2006.

[34] J. Leskovec, L. A. Adamic, and B. A. Huberman, "The dynamics of viral marketing," *TWEB*, vol. 1, no. 1, 2007.

[35] S. Hill, F. Provost, and C. Volinsky, "Network-based marketing: Identifying likely adopters via consumer networks," *Statistical Science*, vol. 21, no. 2, pp. 256–276, 2006.

[36] P. Domingos and M. Richardson, "Mining the network value of customers," in *Proc. of the Seventh ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'01)*, 2001.

[37] M. Richardson and P. Domingos, "Mining knowledge-sharing sites for viral marketing," in *Proc. of the Eighth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'02)*, 2002.

[38] D. M. Chickering and D. Heckerman, "A decision theoretic approach to targeted advertising," in *Proc. of the 16th Conf. in Uncertainty in Artificial Intelligence (UAI'00)*, 2000.

[39] D. Kempe, J. M. Kleinberg, and É. Tardos, "Maximizing the spread of influence through a social network," in *Proc. of the Ninth ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'03)*, 2003.

[40] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher, "An analysis of approximations for maximizing submodular set functions - i," *Mathematical Programming*, vol. 14, no. 1, pp. 265–294, 1978.

[41] W. Chen, C. Wang, and Y. Wang, "Scalable influence maximization for prevalent viral marketing in large-scale social networks," in *Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'10)*, 2010.

[42] W. Chen, Y. Yuan, and L. Zhang, "Scalable influence maximization in social networks under the linear threshold model," in *Proc. of the 10th IEEE Int. Conf. on Data Mining (ICDM'10)*, 2010.

[43] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. S. Glance, "Cost-effective outbreak detection in networks," in *Proc. of the 13th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'07)*, 2007.

[44] W. Chen, Y. Wang, and S. Yang, "Efficient influence maximization in social networks," in *Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'09)*, 2009.

[45] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "A data-based approach to social influence maximization," *PVLDB*, vol. 5, no. 1, pp. 73–84, 2011.

[46] K. Saito, R. Nakano, and M. Kimura, "Prediction of information diffusion probabilities for independent cascade model," in *Proc. of the 12th Int. Conf. on Knowledge-Based Intelligent Information and Engineering Systems (KES'08)*, 2008.

[47] M. Mathioudakis, F. Bonchi, C. Castillo, A. Gionis, and A. Ukkonen, "Sparsification of influence networks," in *Proc. of the 17th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'11)*, 2011.
[48] A. Goyal, F. Bonchi, and L. V. S. Lakshmanan, "Learning influence probabilities in social networks," in *Third ACM Int. Conf. on Web Search and Data Mining (WSDM'10)*, 2010.
[49] ——, "Discovering leaders from community actions," in *Proc. of the 2008 ACM Conf. on Information and Knowledge Management (CIKM 2008)*, 2008.
[50] A. Goyal, B.-W. On, F. Bonchi, and L. V. S. Lakshmanan, "Gurumine: A pattern mining system for discovering leaders and tribes," in *Proc. of the 25th IEEE Int. Conf. on Data Engineering (ICDE'09)*, 2009.
[51] R. Agrawal, T. Imielinski, and A. N. Swami, "Mining association rules between sets of items in large databases," in *Proc. of the 1993 ACM SIGMOD Int. Conf. on Management of Data (SIGMOD'93)*, 1993.
[52] M. Gomez-Rodriguez, J. Leskovec, and A. Krause, "Inferring networks of diffusion and influence," in *Proc. of the 16th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'10)*, 2010.
[53] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf, "Uncovering the temporal dynamics of diffusion networks," in *Proc. of the 28th Int. Conf. on Machine Learning (ICML'11)*, 2011.
[54] M. A. Serrano, M. Boguñá, and A. Vespignani, "Extracting the multiscale backbone of complex weighted networks," *Proc. of the National Academy of Sciences*, vol. 106, no. 16, pp. 6483–6488, 2009.
[55] N. J. Foti, J. M. Hughes, and D. N. Rockmore, "Nonparametric sparsification of complex multiscale networks," *PLoS ONE*, vol. 6, no. 2, 2011.
[56] S. Bharathi, D. Kempe, and M. Salek, "Competitive influence maximization in social networks," in *Proc. of the Third Int. Workshop on Internet and Network Economics (WINE'07)*, 2007.
[57] T. Carnes, C. Nagarajan, S. M. Wild, and A. van Zuylen, "Maximizing influence in a competitive social network: a follower's perspective," in *Proc. of the 9th Int. Conf. on Electronic Commerce (ICEC'07)*, 2007.
[58] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *Proc. of the 15th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'09)*, 2009.
[59] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influence in heterogeneous networks," in *Proc. of the 19th ACM Conf. on Information and Knowledge Management (CIKM'10)*, 2010.