Finding Trendsetters in Information Networks

Diego Saez-Trumper
Universitat Pompeu Fabra
Barcelona, Spain
diego.saez@upf.edu

Giovanni Comarela
UFMG
Belo Horizonte, Brazil
giovannicomarela@dcc.ufmg.br

Virgílio Almeida
UFMG
Belo Horizonte, Brazil
virgilio@dcc.ufmg.br

Ricardo Baeza-Yates
Yahoo! Research
Barcelona, Spain
rbaeza@acm.org

Fabrício Benevenuto
UFOP
Ouro Preto, Brazil
fabricio@dcc.ufmg.br

ABSTRACT

Influential people have an important role in the process of information diffusion. However, there are several ways to be influential, for example, to be the most popular or the first that adopts a new idea. In this paper we present a methodology to find trendsetters in information networks according to a specific topic of interest. Trendsetters are people that adopt and spread new ideas influencing other people before these ideas become popular. At the same time, not all early adopters are trendsetters because only few of them have the ability of propagating their ideas by their social contacts through word-of-mouth. Differently from other influence measures, a trendsetter is not necessarily popular or famous, but the one whose ideas spread over the graph successfully. Other metrics such as node in-degree or even standard Pagerank focus only in the static topology of the network. We propose a ranking strategy that focuses on the ability of some users to push new ideas that will be successful in the future. To that end, we combine temporal attributes of nodes and edges of the network with a Pagerank based algorithm to find the trendsetters for a given topic. To test our algorithm we conduct innovative experiments over a large Twitter dataset. We show that nodes with high in-degree tend to arrive late for new trends, while users in the top of our ranking tend to be early adopters that also influence their social contacts to adopt the new trend.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: information filtering, retrieval models

General Terms

Algorithms, Design, Experimentation

Keywords

Information Networks, Influence, Social Networks

KDD'12, August 12–16, 2012, Beijing, China.

1. INTRODUCTION

The impressive growth of social networking services has made personal contacts and relationships more visible and quantifiable than ever before. Online information networks have been pointed out as places where users influence and are influenced by others, and have become ideal channels for spreading news or innovative ideas [11]. Online social networks have emerged as a popular medium where users discuss about everything, including noteworthy events, giving opinions and expressing sentiments concerning facts and ideas of daily life. Additionally, they pose opportunities for sharing information of local interest. For instance, local businesses actively reach out to their customers by announcing promotions and asking users to propagate them.

Online information networks data offers an opportunity to answer important questions related to information diffusion. Recently, the concept of *Follower Hubs* and *Innovative Hubs* has been borrowed from the economics literature, to describe how a new idea or product propagates over such networks [1]. Follower Hubs are nodes with high in-degree, hence they can deliver content to a larger audience than a normal user. However, previous research [21] shows that Follower Hubs usually have a high threshold for the adoption of new ideas. Here is where the Innovation Hubs are important. Innovation Hubs usually have lower in-degree than Follower Hubs, and also a lower threshold for the adoption of a new idea. Therefore, they have a key role in an information propagation process. In other words, the Followers Hubs are influencers, but the Innovation Hubs are trendsetters. Although these definitions are interesting, in traditional social experiments, it is not easy to identify the different and multiple roles of the participants without restricting the size of the study. Data collected from online information networks allow researchers to carry out detailed studies about dissemination of ideas and information with a large number of participants. Indeed, recent efforts quantified the level of influence of participants on online social networks [8] and proposed techniques to identify those who are likely to spread information to a large audience [16]. While marketing services actively search for potential influencers to promote various items, influencers actively search for innovative ideas and important innovators. In this paper, we address the problem of identifying trendsetters in information networks.

Trendsetters are people that adopt and spread new ideas (trends, fashions) before these ideas become popular. They are not necessarily well known news outlets, celebrities or politicians, but are the ones whose ideas spread widely and successfully through word-of-mouth. To be an innovator, a person needs to be one of the first people to pick up a new or nascent trend, which may be adopted by other members of a social or information network. On the other hand, not all the early adopters are trendsetters because only few of them have the ability of propagating their ideas to their social contacts through word-of-mouth.

To identify trendsetters there are two important aspects that we need to take into account. The first one is the area or topic of the innovator, as people have different levels of expertise on various subjects. For example, marketing services actively search for potential influential people in a specific domain or area to promote certain products or services. Influential people include "cool" teenagers, local leaders, and popular public figures. Thus, it is important to specify topics and themes that define the context where trendsetters will be identified.

Second, it is important to consider time information associated with the posting of innovative ideas. Traditional ranking algorithms on social networks, such as the standard Pagerank algorithm [22] do not consider time information concerned to ideas that become popular. Instead, they consider only aggregate usage statistics and a static network topology. For example, Figure 1 considers that for node X, $t_X = n$ represents that X adopted the trend h in time n. Thus, node G was the first one to adopt h, while node E was the last one to adopt the same trend. Note that, although node G is an innovator, it's information was passed to H but not to the rest of the network. Thus, node G cannot be considered a trendsetter. On the other hand, if we compute the standard Pagerank algorithm using this graph and ignore the time when trend h was adopted, node C would be considered the top ranked node although it has just incoming links from nodes A and E and simply spread it to a larger audience. However, if we pay attention to time, we will see that C adopted the trend before E, and therefore we cannot consider that C received information from E. We can also observe that nodes A and E have the same rank according to Pagerank, despite that A adopted the trend before E. In this example, the top trendsetter is node A because it was the first one to adopt this trend being followed - directly or indirectly - by many other participants of the network, such as nodes C, D, B, and F.

This paper presents a novel approach to identify trendsetters in information networks. Differently with previous work on ranking influential users in social and information networks, we introduce timing information on the social graph to be able to identify persons that spark the process of disseminating ideas that become popular in the network. We propose a robust way to model the dissemination of innovation, representing a topic as a collection of trends, that can be applied in several scenarios. We define a topic-sensitive weighted innovation graph that provides key information to understand who adopted a certain topic that triggered attention of others in the network. We then introduce a Pagerank inspired time-sensitive algorithm to find trendsetters. Next,

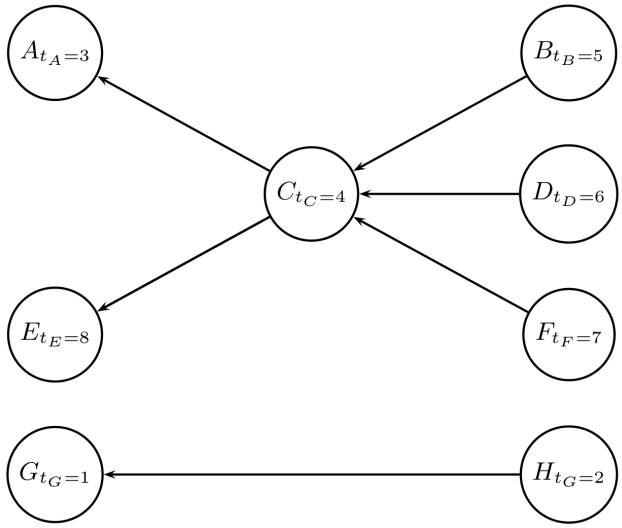

Figure 1: Illustrative example of timing importance: Without considering time information, nodes A and E are symmetric, regardless of whether A adopted the trend first. The edges represents social connections between nodes and the arrows goes opposite to the information flow.

we tested our algorithm using a robust dataset containing the complete snapshot of the Twitter network and containing all tweets from 2006 to the mid-2009. The result shows that the proposed algorithm is able to measure the direct and indirect influence adding also the early adoption as a key feature to be influential. This characteristic is useful to differentiate between trendsetters and other nodes that despite having a large in-degree, adopt the trends only after they became popular.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 presents a formal definition of the trendsetters ranking. Section 4 describes the experimental evaluation and the results obtained. Finally, Section 5 concludes the paper summarizing its contributions and discussing future work.

2. RELATED WORK

The concept of influence has been studied by different disciplines. In the 1950s the social psychologist Solomon Asch published a well known study about the group's influence in individuals decisions [3]. Years later, in 1968, the marketing researcher Frank Bass proposed a model for the adoption of new technology on the market [7]. He considered that there are two types of new adopters: the innovators and the imitators. Bass models the relation between these two types with a differential equation. Other two important influence models based on individual thresholds were proposed at the end of 1970s by the economist Thomas Schelling [26] and by the sociologist Mark Granovetter [14]. Nowadays, the study of influence and information diffusion is a hot topic in the research community. Next, we review and summarize the different approaches related to our work.

Probabilistic Models: The relevance of viral marketing

has inspired studies looking for influential users that can maximize the information propagation in social networks. This problem has been faced by probabilistic models [10], and as a discrete optimization problem [16]. A complementary work in the same direction has been done by [9]. Most recently, a probabilistic approach to find topical authorities in microbloging sites has been proposed in [23]. In our case, we study users' ability of spreading an innovation as a ranking problem.

Group Influence: In [4] the authors shows that Livejournal users and also DBLP authors tends to join new groups when their *friends* joins too. In the same way, [24] has studied the adoption of *hashtags* in Twitter about different topics related with the number of *friends* that have used these tags previously. They show that users wait that their friends use some hashtag before them, and depending on how controversial is the topic, they wait for more friends to follow before they jump in. We have used the *hashtags* categorization done in this work as input for our analysis. However, those works focus in the direct influence of the group (friends or friend of friends) but does not take in account how the indirect influence spreads over the graph.

Early Adopters: The concept of Early Adopters was studied in [6], analyzing data from a popular on-line virtual world (Second Life), discovering that the trendsetters are usually users with few friends, i.e. nodes with low degree, and moreover, they are users that are not too evolved in the game (they play less hours than the average). These are key points for our work, because shows that users that can be considered outsiders in a trivial analysis gain importance when the time factor (to be an early adopter) is considered. We take the concept of early adopter, but we propose a way to differentiate those that create cascade behavior. That means that we are interested in influential early adopters.

Temporal Factor: The importance of the temporal factor in influence studies was remarked in [2]. Studying the different sources of correlation among users actions, the authors proposed three possible explanations to a group of users performing the same action: environmental factors, homophily and influence. An example of an environmental factor is the fact that a group of social media users living in the same city can post about the same event, because that event is taking place in their city. Homophily is similar, but corresponds to intrinsic characteristics of people. That means, that two users can post about the same topic because they have the same interests. The authors point out that we can only talk about social influence when exists a time causality associated to the actions among users. Another consideration about the time factor in information networks is described in [18], showing that considering only the topology is not enough to understand how the information spreads over an information network, because some edges could be *slower* than others and that in many cases the information can go faster through a multi-hop pathway that uses *faster* edges. Other important studies have been conducted taking in account the temporal factor to find the backbone of cascades produced on the Web [12], but they do not propose a ranking function neither model topics, and they only establish a temporal relation among nodes, looking to the most common path in cascades process, creating a influence pathway.

The work of [13] describes leaders as users that take actions that will be imitated by their friends later. They discover, among other things, that there are users that are tribe leaders, meaning that they are imitated in different actions by the same group of friends (the tribe). In this case to be the first is a signal of leadership. On the other hand, in [29] being the last is considered as a signal of expertise. This work studies the relation into a Java Forum among users making questions and giving answers for this programming language. These relations are model as a graph and ranking algorithms - such as PageRank [22] and HITS [17] - are used to find relevant users. Comparing their results with human evaluation they conclude that those algorithms allows to find expert users. HITS also was used to model influence and passivity in social media [25].

Memeshapes: In [28] the authors present the KSC-algorithm to cluster temporal series by their shape, applying it to *internet memes* and Twitter hashtags. They found different kinds of shapes and they explain them by the nature of the sources (such as bloggers, mainstream media, etc). This work is a fundamental input for our work, as we have used the algorithm proposed by them to cluster the trends in our dataset. However, the goal of our work is different as we are not interested in who is talking about a topic when it is popular, but just before it became popular. Differences about how we apply the KSC-algorithm are explained in Section 4.

Pagerank based algorithms: The idea to use Pagerank to evaluate user's influence in Social Media was also developed in [27]. Here, the authors used a topic-sensitive Pagerank extension [15] proposing a *TwitterRank* to evaluate Twitter users. In fact, this work is not focused on influence but in homophily, measuring the similarity among users, avoiding the temporal factor and considering the amount of information that each user posts (i.e. number of tweets) in a given topic to assign importance to each node for this topic. Something similar has been done in [20]. Other rankings on Twitter have been studied by [8], showing that the number of followers (node in-degree) is not necessarily an indicator of influence, naming this fact as: "The Million Follower Fallacy". In [19] they also ranked Twitter users, showing the differences between a node degree based ranking versus the results of Pagerank. However, all those works have been using the social graph without consider the temporal dynamics of communication.

Differences with our approach: Note that previous work has used different definitions of what is an influential user. Different from other works using a ranking based on the network topology, our approach also considers the early adoption as a key to be a trendsetter. To the best of our knowledge, this is the first paper that considers Pagerank and temporal factors to find influential people in information networks. Additionally, our algorithm presents a flexible way to model topics that is adaptable to different scenarios. Also, the influence as a function of time can be easily adjusted with a single parameter. Hence, we believe that our

approach is innovative and complementary with existing approaches.

3. RANKING TRENDSETTERS

This section presents our algorithm to rank trendsetters in an information network according to some topic of interest. We start with basic definitions related to the concept of a *topic*, network graphs, and the interactions among nodes over time. We represent a network as a directed graph $G(N, E)$, where N is the set of nodes and E the set of edges. Each edge is an ordered pair (u, v), $u, v \in N$, representing a relation between u and v. Furthermore, we define $In_G(v) = \{w|\ (w, v) \in E\}$, $Out_G(v) = \{w|\ (v, w) \in E\}$, the incoming and outgoing neighbor sets respectively, and $|S|$ is the cardinality of a set S.

As we are interested in ranking nodes according to a specific topic, we look only at nodes that are related with the topic. The two next definitions formalize what a topic is and how to select the nodes.

Definition 1. We define a topic as a collection of trends related to a specific theme. We denote this collection by $\{h_1, \ldots, h_{n_k}\}$. Each one of the n_k trends could be a word, a phrase, a meme, a tag, an URL, or any other kind of label that can be associated with a node.

Definition 2. We denote $G_k(N_k, E_k)$ as the induced graph of $G(N, E)$ over the topic k. The set N_k is obtained by considering all nodes of N that used at least one trend of k and E_k represent all edges (u, v) such that, if $(u, v) \in E$ and $u, v \in N_k$ then $(u, v) \in E_k$.

As mentioned in the introduction, the timing information is the key to determine social influence. We include this information in the temporal attributes of nodes and edges of $G_k(N_k, E_k)$ in the following definition.

Definition 3. Let $t_i(v)$ be the time when node $v \in N_k$ adopts the trend $h_i \in k$ ($t_i(v) = 0$, if v does not adopt h_i). We define two vectors, $s_1(v)$ (for all $v \in N_k$) and $s_2(u, v)$ (for all $(u, v) \in E_k$), each one with n_k components given respectively by:

$$s_1(v)_i = \begin{cases} 1, & \text{if } t_i(v) > 0, \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

and

$$s_2(u, v)_i = \begin{cases} e^{-\frac{\Delta}{\alpha}}, & \text{if } t_i(v) > 0 \text{ and } t_i(v) < t_i(u), \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

for $i = 1, \ldots, n_k$, where $\Delta = t_i(u) - t_i(v)$ and $\alpha > 0$.

Vector $s_1(v)$ informs if node v adopted (or not) each trend of k, while $s_2(u, v)$ shows if u adopted these trends after v and weights the relation as a function of the period of time between $t_i(u)$ and $t_i(v)$. For a fixed α, if $\Delta \to 0^+$ then $e^{-\frac{\Delta}{\alpha}} \to 1$ and if $\Delta \to +\infty$ then $e^{-\frac{\Delta}{\alpha}} \to 0$. These limits mean that if the node u adopts a trend just after v then $s_1(v)_i$ is very close to $s_2(u, v)_i$, and, on the other hand, if u adopts the trend after a long time, we have that $s_1(v)_i$ and $s_2(u, v)_i$ are very different. The exponential time decay to compute influence has been proposed in previous work related to temporal factors in the web graph [5].

The α parameter allows to control the time window that will be considered to compute $s_2(u, v)$. This settable parameter is useful because it can be adapted to different scenarios. Depending on the nature of the problem, we want to consider that one node is strongly influencing another if the second one imitates the first one in few seconds, and in other cases we want to use a longer span of time. Moreover, for the same problem we could be interested in studying the influence of a short span of time, or a long term influence.

So, when many components of $s_1(v)$ and $s_2(u, v)$ are similar and different from 0, we assume that v has a strong influence over u according to a topic k. Based on the previous definitions we can now define influence:

Definition 4. Let $G_k(N_k, E_k)$ be an induced graph of a network $G(N, E)$ over a topic k with n_k trends. For each $(u, v) \in E_k$ we define the influence of v over u by:

$$I_k^*(u, v) = \left(\frac{s_1(v) \cdot s_2(u, v)}{||s_1(v)|| \times ||s_2(u, v)||} \right) \times \left(\frac{L(s_2(u, v))}{n_k} \right), \quad (3)$$

where the operator $\cdot$ refers to the scalar product, $||x||$ to the Euclidian norm of any vector x, and $L(s_2(u, v))$ to the number of components of $s_2(u, v)$ that are different from 0. If $||s_2(u, v)|| = 0$, we define $I^*(u, v) = 0$. It is important to notice that, by definition, $||s_1(v)|| \neq 0$ for all $v \in N_k$.

Equation 3 is the main outcome of our previous discussion. The first part is given by the cosine similarity between $s_1(v)$ and $s_2(u, v)$, which is close to 1 if u adopted the same trends than v in a reasonable lag of time, and close to 0, otherwise. The second term is the fraction of trends of k that u adopted after v. We use this to indicate that if u adopted more trends influenced by v than by other node z, then the influence of v over u is greater than the influence of z.

One important fact is that u can be influenced to adopt a trend of k by several nodes in $G_k(N_k, E_k)$. So, we normalize $I^*(u, v)$ as follows:

$$I_k(u, v) = \frac{I_k^*(u, v)}{\sum_{w \in Out_{G_k}(u)} I_k^*(u, w)}, \quad (4)$$

noticing that if the denominator of Equation 4 is zero, we define $I_k(u, v)$ as 0.

The next definition presents how we rank trendsetters according to a Pagerank-like algorithm.

Definition 5. The trendsetters (TS) rank of node v in a network $G_k(N_k, E_k)$, denoted by $TS_k(v)$, is given by:

$$TS_k(v) = d\ D_k(v) + (1 - d) \sum_{w \in In_{G_k}(v)} TS_k(w) I_k(w, v), \quad (5)$$

where $0 \leq d \leq 1$ is the damping factor and D_k is a probability distribution over all nodes of $G_k(N_k, E_k)$. In this paper

we consider a uniform $D_k(v) = 1/|N_k|$ for all $v \in N_k$, but this distribution could be topic dependent.

Making an analogy with the random surfer model in the Pagerank algorithm presented in [22] on graph $G_k(N_k, E_k)$ we can analyze Equation 5 in the following way: consider that the surfer is in any node of $G_k(N_k, E_k)$, for example u. With probability $1-d$, the surfer leaves u and goes to other node in $Out_{G_k}(u)$, and with probability d, to any node in N_k. In the first case the node $v \in Out_{G_k}(u)$ will be visited with probability $I_k(u, v)$. So, the node that influences more u has a higher probability to be visited. In the second case, any $v \in N_k$ will be visited with probability $D_k(v)$, reflecting the independent adoption of that topic. Hence, in the steady state the surfer will spend more time in the most influential nodes of the network.

Let us now see an example in Twitter: The Iran election was an important topic during 2009. The main hashtag used to talk about this was #iranelection, and other related tags where #iran and #tehran. So using definition 1, the topic Iran Election could be represented by: $k_{\text{Iran Election}} = [\#\text{iranelection}, \#\text{iran}, \#\text{tehran}]$. Following the methodology proposed previously, we compute the graph for all the nodes that used at least one of the trends. Next, we compute the PageRank(PR), InDegree Rank(ID) and the TS rank for this graph. PR and ID selects @cnnbr (CNN Breaking News) as top user, while TS selects a user named @Lara, self-described as "Reporting from the Middle East for ABC News and Bloomberg Television." This user twitted with the two most popular hashtags in this topic, adopting them before than they became popular (see Figure 2). In other cases related with politics we have also find other activists or reporters "on site" being ranked on the top of TS, while PR and ID selects CNN for all of them.

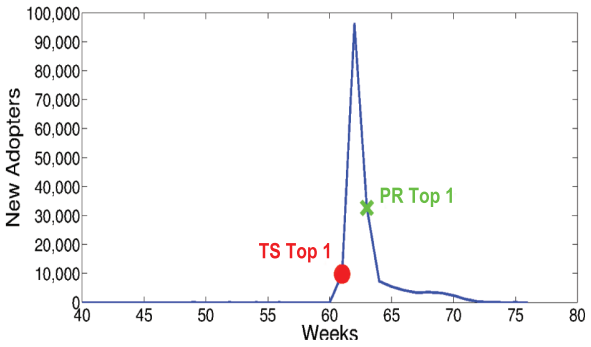

Figure 2: Iran Election Topic Timeline: comparison between top TS and top PR users.

Another interesting example is the idiom #musicmonday, which is one of the most popular in Twitter. Its name is very descriptive because it is used to share music on Mondays. Considering that we have almost the complete Twitter information from its beginning, we can know who invented this tag: *@rubenharris*. This fact should identify that user as influential in this trend. However, if we analyze this topic using PR, that user is ranked in the position 164,970 among 179,119 users. But using TS this user appears in the 4th position. Note that, as we explained before, to be an innovator - and thus an early adopter - a necessary but not sufficient characteristic is to be a trendsetter. TS considers that the most trendsetter for #musicmonday is user *@twtfm*, which is the corporate user of the site http://twt.fm, a company that offers a service to share and search music using Twitter. However this user was the 76th to adopt the trend. Now, note that PR and ID selects the same user as the top one: *@perezhilton*. He is a famous professional blogger, and he has been indicated as one of the users with more followers in Twitter [19]. However, if we check *@perezhilton*, he was the 18,718th user to adopt the trend, therefore we cannot consider he as an innovator. However, because he is prestigious, TS ranked it in the top-100. The example of #musicmonday is useful to understand that TS captures both characteristics that we consider important for a trendsetter: the early adoption and the capacity to spread the trend over the graph.

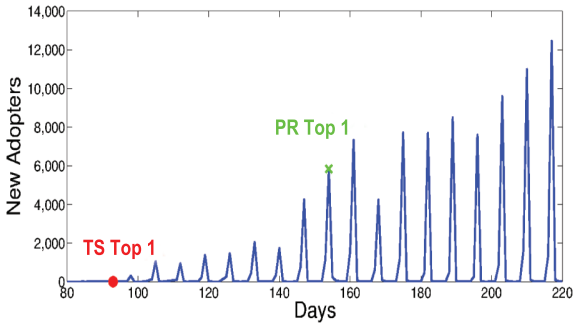

Figure 3: #musicmonday new adopters timeline: comparison between top TS and top PR users.

4. EXPERIMENTAL EVALUATION

In order to test the TS ranking we have conducted a set of experiments over a huge Twitter dataset. We consider the Twitter social graph, where the connections among users are directed. Using the notation described in the previous section the Twitter graph will be $G(N, E)$, where an edge $(u, v) \in E$ means that user u *follows* v. Trends are modeled using *hashtags*, so a topic k is a collection of hashtags, that is $k = [\#tag_1, \ldots, \#tag_{n_k}]$. Next, we create the induced graph G_k considering all the nodes that have posted at least a tweet with one hashtag of k. Over this graph we compute the TS ranking using a time window of one day ($\alpha = 86,400$ seconds) in Equation 2 and $d = 0.2$ in Equation 5, following the original PageRank value. We have tested other values and they do not change the results significantly.

Our hypothesis is that other measures of influence are not suitable to find trendsetters because they tend to favor nodes that do not propose new trends, but follow those that are already popular. To test this, we grouped the trends by different methodologies: first we group them by categories related with topics such as music, sports, movies, etc., next we grouped the new adopters curve shape. In each case we compare the TS ranking with In-Degree (ID) ranking - where nodes are sorted by incoming links; and the traditional Pagerank, PR. We also quantify the followers influenced by the top users of each ranking and we compute how similar are the rankings under study.

4.1 Dataset

We have used a dataset containing almost the total information in Twitter until August 2009. We have over 50 millions users, with all their social connections (*Followers* and *Followees*) and approximately 1.6 billions of Tweets. Note that differently from other works that use a big amount of tweets, we also have the complete social graph, so we do not need to use heuristics to infer it. A detailed description of this dataset can be found in [8].

To select the hashtag for the experiment, we use the classification made by Romero *et al.* [24], where each of the 500 most popular hashtags in their dataset was assigned to a category such as politics, music, or celebrities (see Table 1). From those 500 hashtags, only 370 are mentioned among the 2,000 most popular of our dataset, with #followfriday being the most popular with 3,051,316 mentions and #jemi the least mentioned, with only 1,810 occurrences. The 130 remaining hashtags do not appear or have a very low level of mentions. This is because each dataset was obtained on different dates.

To complete the topic modeling, we looked for other hashtags related with the main one. For the 370 hashtags we searched for others that had a co-occurrence of at least 5% with the main one. This means that each of the 370 topics is modeled by a vector containing the main hashtag and others related to it. For example, the topic modeled with more hasthtags was #realstate, having other 20 related hashtags. Over 74% of the topics were modeled with at least two hashtags. Note that these related hashtags creates the vectors that are described in Definition 2. The categories are used only to facilitate the analysis of the results.

4.2 Adoption before Peak: Categories

Our first approach to answer this question is to analyze the percent of users of each ranking adopting the trend before the peak of adoption. By peak of adoption we refer to the time when a trend had its bigger number of new adopters. To do that, first we obtained the peak of adoption from each of the 370 topics studied. We denote by P_k the peak of adoption of trend k. Next, we have to compare it with the time of adoption of the top-p users of each ranking, where $T(i)_{k_r}$ represents the time in which user i from ranking r adopted topic k. Therefore, if $P_k - T(i)_{k_r} < 0$, this means that user i adopted the trend before the peak. Finally, we grouped all the topics k by their category, and we computed the percentage of users adopting the trend after the peak for each ranking. The results presented have been calculated with $p = 100$, but values from 5 to 1000 do not present significant variations.

Figure 4 shows that in categories such as music, celebrity and idioms, most of the nodes in the top of ID and PR start talking about these topics after the peak, and only in sports and technology they obtain a good performance. In contrast, in 6 of the 9 categories, more than 50% of the TS top users adopted the trend before the peak. One motivation to develop the TS ranking was our intuition that nodes with high in-degree do not propose or push new trends but follow those that are already popular. The performance of ID in the previous experiment tends to confirm this intuition. In order to better understand how the in-degree is

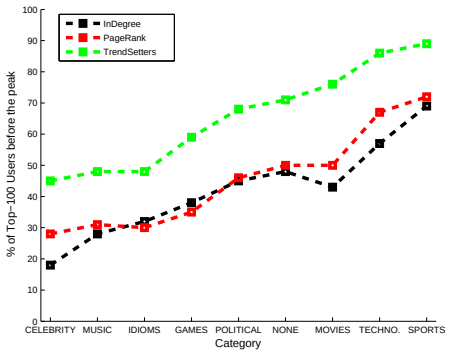

Figure 4: Percentage of top-100 users of each ranking that adopts the trend before the peak.

related with the adoption time, we repeated the previous experiment, but instead of computing only if $P_k - T(i)_{k_r}$ is > 0, we recorded this time span as well as the user i in-degree. Therefore, for each ranking in each trend we have a list of tuples representing the time span and the in-degree of the top-p nodes. Again we group the trends by their category, and at the end we computed the median of the time span and of the in-degree of each ranking for each category. In Figure 5 we plot the time span median in the horizontal axis and the in-degree median in the vertical axis for each ranking. It is clear that top users of TS adopt trends before the peak and they have a smaller in-degree than top users of the other rankings. These results confirm that nodes with high in-degree tend to be *slower* than other nodes.

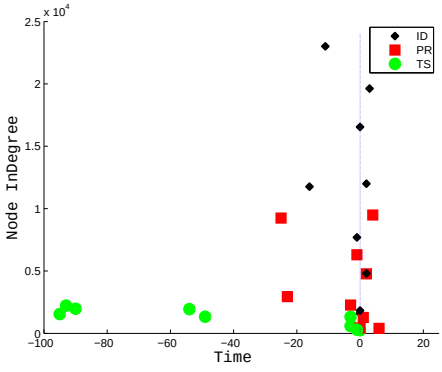

Figure 5: Relation among time span and in-degree for the top-100 users of each ranking in all the categories (peak is at time 0).

4.3 Adoption before Peak: Shapes

The categories by topic are very descriptive, but the nature of the TS ranking suggests that the quality is also related with the shape of the curve of adoption. For this reason we grouped the trends by their curve of new adopters. For this aim we have used the KSC-algorithm [28], that receives as input a set of time series, and gives as an output a classification by shape and the centroids of each cluster, providing a visual representation. Note that, unlike that paper, we are interested in the adoption of the trend, that is, the first

Table 1: Summary of categories. Note that some hashtags could belong to more than one Topic.

Category	#Topics	Example of Hashtags	#Tweets
Celebrity	16	#michaeljackson, #niley	1,036,101
Games	13	#mafiawars, #ps3 #	2,556,437
Idioms	35	#musicmonday,#followfriday	7,882,209
Movies	29	#heroes,#tv	1,769,945
Music	33	#lastfm, #musicmonday	2,785,522
None	153	#quotes, #sale	2,227,971
Political	39	#honduras, #Iranelection,	8,156,786
Sports	27	#soccer, #rugby	1,914,061
Technology	41	#twitter,#android	7,459,471
Total	370	-	41,442,741

time that the hashtag is mentioned by a user. Additionally, we are interested in all the popular trends, not only in those with a short duration. Hence, our time discretization is done by days, not hours.

To apply the KSC-algorithm we have created a time series to represent the curve of new adopters from each trend, considering that the time of adoption is the first mention of any of the hashtags in the topic (repeated or later mentions are not taken in account). We have created time series of 128 elements, where each element represents a day. Next, we align the peaks of all the time series at 2/3, that is position 86. This is different than the 1/3 peak centering used in [28]. Our reason to move the peak to the right is because we are more interested in what happens before the peak, rather than later. The last step, was to select the number of clusters K to use. We tried with values from 2 to 12, finding that with more than 4 clusters we found only small variations of the 4 main clusters. Figure 6 shows the shapes of the 4 clusters obtained. Next, we repeatead the calculations described in the previous subsection to find the time span and the median of time adoption of the top users of each ranking but now grouped by cluster. Finally, we plot these points over the curves.

Table 2: Number of topics in KSC Clusters.

#Cluster	1	2	3	4
#Topics	91	115	**128**	36

Table 2 show that most of the topics corresponds to clusters with a clear peak of adoption such as cluster 3. In figure 6 we can see that TS appears clearly before the peak in all the clusters. In contrast, the top users of ID and PR only appears before the peak in cluster 2, that is, in the cluster with the less pronounced peak. Specially interesting are the results for cluster 1, where TS appears over a little first peak before the largest one. This suggests that TS is detecting a topic that will be potentially interesting in the future.

4.4 Influenced Followers Ratio

Now we try to understand how many of the social contacts were influenced by the top users of each ranking, that is, how many of their total followers adopt the trends after them. To evaluate this, we create a simple indicator that we call *Influenced Followers* ratio for a topic k, $IF_k(v)$, defined as the fraction of followers of v that adopted at least one trend of the topic k after v.

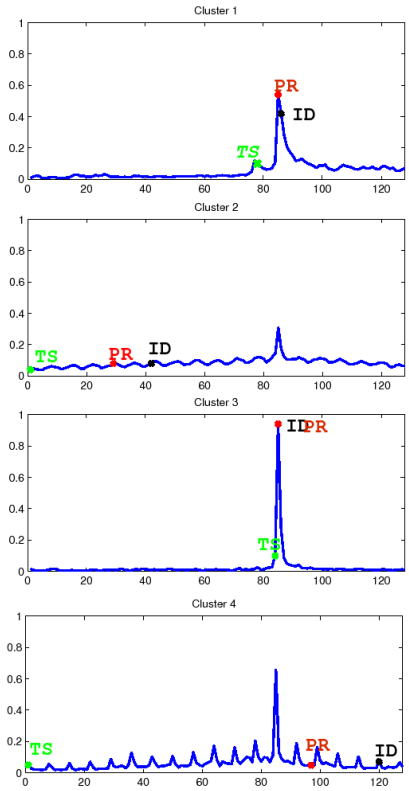

Figure 6: KSC clusters. Each ranking is represented with the median of time deviation with respect to the peak in each cluster.

Table 3 shows that TS top users have a bigger IF ratio than for PR and ID. It is interesting to note that in the category Political, the TS rankings obtain the best performance, and in all of them is always over 0.06. Note that in 7 of the 9 categories TS is one order of magnitude better than ID and almost doubles PR. This confirms that TS users influence more their social contacts than other rankings.

4.5 Ranking Similarities

Whereas one of the main features of TS ranking is to capture the early adoption behavior, makes sense to compare it with an Early Adoption EA ranking. That is, a ranking where

Table 3: Influenced Followers (IF) ratio for top-100 users of each ranking.

Category	(%)ID	(%)PR	**(%)TS**
Political	0.013	0.084	**0.174**
Celebrity	0.015	0.089	**0.148**
Music	0.013	0.096	**0.160**
Games	0.022	0.058	**0.115**
Sports	0.004	0.054	**0.098**
Idioms	0.001	0.034	**0.088**
None	0.011	0.001	**0.085**
Technology	0.006	0.054	**0.078**
Movies	0.006	0.043	**0.067**

the top one will be the first adopter and the next position will be assigned by the adoption time.

To compare rankings we use the Kendall Rank Correlation Coefficient τ. This coefficient gives an idea of the agreement between two rankings. It varies in the interval $[-1, 1]$, where 1 means total agreement and -1 means that one ranking is the reverse of the other.

For this experiment we use the trends with more mentions in each category and then we compute the average among trends for the four rankings: EA, TS, ID and PR.

Table 4 shows that PR and ID tend to be similar but completely different from EA. TS is not too similar with any of them but presents a nice balance among them. This results shows that TS has the ability to mix different characteristics of the other rankings. It also shows that not all the early adopters are trendsetters.

Table 4: Kendall τ comparison among rankings.

	EA	PR	ID
EA	-	-	-
PR	0.11	-	-
ID	0.09	0.74	-
TS	0.37	0.56	0.48

4.6 Ranking with Partial Information

Previous results show that TS give high scores to the early adopters. Considering this we can conjecture that probably it is not necessary to use the information about all users to find the top ones. To answer this question we conducted the following experiment: first we selected the topic with more users for each category to use it as representative of this category. Next, we ordered the users by adoption time and then we compute the ranking considering only the first 10% of them, then 20%, and so on in increments of 10%. Next, we compared it with the final ranking (i.e. with the 100% of users). For all the trends, we were able to find the top-one user, considering only the initial 10%. Moreover, we found 7 of the top 10 users, and more than 70 of the top-100 users if we consider the 20% of initial users. In contrast, PR could find the top-one with the 10% only in one case, requiring more than 50% of the users in the other cases. Moreover, PR required over 60% of users, to find at least 7 of the top-10 users.

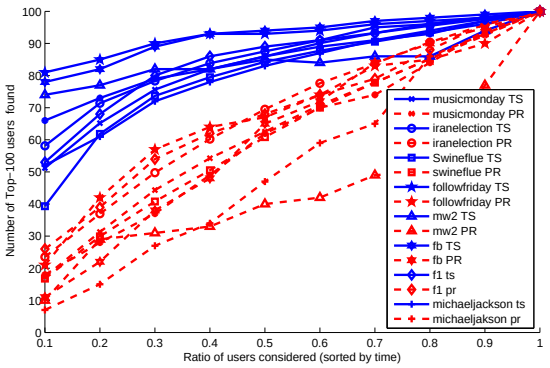

Figure 7: Number of top-100 users found using a only a fraction of total users sorted by time. A Comparison between PR and TS in three trends.

Figure 7 shows the total number of top-100 found considering different fractions of users. These results suggest that TS is able to find the most influential users faster than PR. This behavior can be explained considering that TS ranks on the top many early adopters, unlike PR that is more sensitive to the arrival of a node with high in-degree at anytime that they arrive. The time decay used makes TS less sensitive if those nodes arrive late.

5. CONCLUSIONS

We have proposed a robust algorithm to rank trendsetters, presenting the problem of the spread of an innovation as a ranking problem considering the temporal factor. We have also presented a sound and extensible way to model topics and influence allowing to run this algorithm in different contexts. Although we have conducted experiments only on Twitter, the problem formulation makes it possible to apply this algorithm to other information networks. We have also presented an innovative methodology to evaluate our algorithm, using different types of classifications such as categories and curve shapes.

One important finding is that users with high in-degree do not propose the ideas that became popular, as usually they adopt them when they are already popular. This confirms the importance of developing new techniques such as TS to find the users that create or early adopt these trends. This appear to be critical in topics related with celebrities, music, idioms and politics.

The results presented in Section 4.6 highlight two important advantages of TS over other algorithms. First, the possibility to find a big fraction of trendsetters requiring only the first 10% of trend adopters, something very useful in real-time scenarios. Second, the differences in the behavior along the time of TS against PR results could be explained because when nodes with a high in-degree adopt a trend late, the time decay function reduces their impact in the final rank.

In future work, using machine learning techniques we will compare the trendsetters with other users that appear to be similar but that do not achieve success, trying to identify

the key characteristics of trendsetters. We will also test our algorithms in other data sets as they can be used in any social network. We also want to explore the impact of the parameters α and d as well as other functions to model the temporal behavior.

Acknowledgements

This work has been partially funded by the HIPERGRAPH project (TIN2009-14560-C03-01) from the Spanish Economy and Competitiveness Ministry, and partially funded by the Brazilian National Institute of Science and Technology for the Web (MCT/CNPq/INCT grant number 573871/2008-6).

6. REFERENCES

[1] P. Adams. *Grouped: How Small Groups of Friends Are the Key to Influence on the Social Web*. Voices That Matter. Pearson Education, 2011.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proc. of Int'l conf. on Knowledge discovery and data mining*, KDD'08, pages 7–15, NY, USA, 2008. ACM.

[3] S. E. Asch. Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow, editor, *Groups, Leadership, and Men*, pages 177–190. Carnegie Press, Pittsburgh, PA, 1951.

[4] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *Proc.of Int'l conf. on Knowledge discovery and data mining*, KDD'06, pages 44–54, NY,USA, 2006. ACM.

[5] R. Baeza-Yates and E. Davis. Web page ranking using link attributes. In *Proc.of Int'l conf. World Wide Web*, WWW'04, pages 328–329, NY, USA, 2004. ACM.

[6] E. Bakshy, B. Karrer, and L. Adamic. Social influence and the diffusion of user-created content. In *Proc.of conf. on Electronic commerce*, EC '09, pages 325–334, NY, USA, 2009. ACM.

[7] F. Bass. A new product growth for model consumer durables. *Management Sciences*, 15(1):215–227, 1969.

[8] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *Proc.of the 4th Int'l AAAI conf. on Weblogs and Social Media (ICWSM)*, Washington DC, USA, 2010.

[9] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *Proc. Int'l conf. on Knowledge discovery and data mining*, KDD '09, pages 199–208, NY, USA, 2009. ACM.

[10] P. Domingos and M. Richardson. Mining the network value of customers. In *Proc.of Int'l conf. on Knowledge discovery and data mining*, KDD '01, pages 57–66, NY, USA, 2001. ACM.

[11] *A special report on social networking, The Economist*, Jan 2010. http://tinyurl.com/ylxtsek. Accessed 07/2011.

[12] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. *Networks*, Proceeding:1019–1028, 2010.

[13] A. Goyal, F. Bonchi, and L. Lakshmanan. Discovering leaders from community actions. In *Proc. of Int'l conf. on Information and knowledge management*, CIKM '08, pages 499–508, NY, USA, 2008. ACM.

[14] M. Granovetter. Threshold Models of Collective Behavior. *Amercan Journal of Sociology*, 15:1420 – 1443, 1978.

[15] T. Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE Transactions on Knowledge and Data Engineering*, 15:784–796, 2003.

[16] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *Proc.of Int'l conf. on Knowledge discovery and data mining*, KDD'03, pages 137–146, NY, USA, 2003. ACM.

[17] J. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46:604–632, September 1999.

[18] G. Kossinets, J. Kleinberg, and D. Watts. The structure of information pathways in a social communication network. In *Proc. Int'l conf. on Knowledge discovery and data mining*, KDD '08, pages 435–443, NY, USA, 2008. ACM.

[19] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc.of Int'l conf. on World wide web*, WWW'10, pages 591–600, NY, USA, 2010. ACM.

[20] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang. Mining topic-level influence in heterogeneous networks. In *Proc.of Int'l conf. on Information and knowledge management*, CIKM '10, pages 199–208, NY, USA, 2010. ACM.

[21] G. Moore. *Crossing the Chasm*. HarperBusiness, revised edition, Sept. 2002.

[22] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November 1999.

[23] A. Pal and S. Counts. Identifying topical authorities in microblogs. In *ACM WSDM*, pages 45–54, 2011.

[24] D. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proc. of Int'l conf. on World wide web*, WWW'11, pages 695–704, NY, USA, 2011. ACM.

[25] W. Romero, D.and Galuba, S. Asur, and B. Huberman. Influence and passivity in social media. In *ECML/PKDD (3)*, pages 18–33, 2011.

[26] T. Schelling. *Micromotives and Macrobehavior*. Norton, 1978.

[27] J. Weng, E. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proc.of Int'l conf. on Web search and data mining*, WSDM '10, pages 261–270, NY, USA, 2010. ACM.

[28] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *Proc.of Int'l conf. on Web search and data mining*, WSDM'11, pages 177–186, NY, USA, 2011. ACM.

[29] J. Zhang, M. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *Proc.of the 16th Int'l conf. on World Wide Web*, WWW '07, pages 221–230, NY, USA, 2007. ACM.