Peer and Authority Pressure in Information-Propagation Models*

Aris Anagnostopoulos[1], George Brova[2], and Evimaria Terzi[2]

[1] Department of Computer and System Sciences, Sapienza University of Rome,
aris@dis.uniroma1.it
[2] Computer Science Department, Boston University,
gbrova@bu.com, evimaria@cs.bu.com

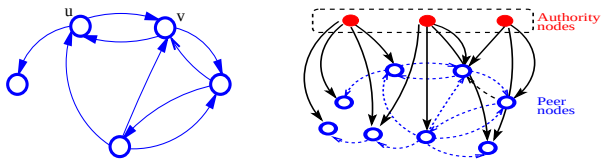
Abstract. Existing models of information diffusion assume that *peer influence* is the main reason for the observed propagation patterns. In this paper, we examine the role of *authority pressure* on the observed information cascades. We model this intuition by characterizing some nodes in the network as "authority" nodes. These are nodes that can influence large number of peers, while themselves cannot be influenced by peers. We propose a model that associates with every item two parameters that quantify the impact of the peer and the authority pressure on the item's propagation. Given a network and the observed diffusion patterns of the item, we learn these parameters from the data and characterize the item as peer- or authority-propagated. We also develop a randomization test that evaluates the statistical significance of our findings and makes our item characterization robust to noise. Our experiments with real data from online media and scientific-collaboration networks indicate that there is a strong signal of authority pressure in these networks.

1 Introduction

Most of the existing models of information propagation in social networks focus on understanding the role of *peer influence* on the observed propagation patterns. [3, 4, 7, 9, 13]. We use the term *peer models* to collectively refer to all such information-propagation models. In peer models, as more neighbors (or peers) of a node adopt an information item, it becomes more probable that the node itself adopts the same item. For example, users adopt a particular instant-messenger software because their friends use the same software; using the same platform makes communication among friends more convenient.

Figure 1(a) depicts a small network of peers and their connections. A directed link from node u to v denotes that v can be influenced by u. The key characteristic of peer models is that all nodes are treated on an equal footing. That is, each node can equally well influence its neighbors or be influenced by them. However, the strength of each agent's influence on others is not the same in reality. For example, mass media can strongly affect the opinions of individu-

* The research leading to these results has received funding from the EU FP7 Project N. 255403 – SNAPS, from the NSF award #1017529, and from a gift from Microsoft.



(a) Peer-only information propagation

(b) Peer and authority information propagation

Fig. 1. Influence graph of a network consisting of *peer nodes* (Figure 1(a)) and *peer and authority nodes* (Figure 1(b))

als, whereas the influence of any one individual on the mass media is most likely infinitesimal. This distinction is often modeled with the use of edge weights.

In this paper we focus on this distinction and we apply a simple way to model it: we posit that some agents are *authorities* (such as the mass media). These nodes have high visibility in the network and they typically influence a large number of non-authority nodes. We call these latter nodes the *peer* nodes. Peers freely exchange information among themselves and therefore there are influence links between them. Peers have no influence on authorities. That is, an influence link that joins an authority to a peer is *unidirectional* from the authority to the peer. In our model, we also ignore the influence that one authority node might have on another. At a global level, the network of authorities and peers looks as in Figure 1(b). That is, peers and authorities are clustered among themselves, and there are only directed one-way links from authorities to peers.

The existence of authority nodes allows us to incorporate the *authority influence* (or *pressure*) into the classic information-propagation models. Given a network of peers and authorities and the observed propagation patterns of different information items (e.g., products, trends, or fads) our goal is to develop a framework that allows us to categorize the items as authority- or peer-propagated. To do so, we define a model that associates every propagated item with two parameters that quantify the effect that authority and peer pressure has played on the item's propagation. Given data about the adoption of the item by the nodes of a network we develop a maximum-likelihood framework for learning the parameters of the item and use them to characterize the nature of its propagation. Furthermore, we develop a randomization test, which we call the *time-shuffle* test. This test allows us to evaluate the statistical significance of our findings and increase our confidence that our findings are not a result of noise in the input data. Our extensive experiments on real data from online media and collaboration networks reveal the following interesting finding. In online social-media networks, where the propagated items are news memes, there is evidence of authority-based propagation. On the other hand, in a collaboration network of scientists, where the items that propagate are themes, there is evidence that peer influence governs the observed propagation patterns to a stronger degree than authority pressure.

The main contribution of this paper lies in the introduction of authority pressure as a part of the information-propagation process. Quantifying the effect

that peer and authority pressure plays in the diffusion of information items will give us a better understanding of the underpinnings of viral markets. At the same time, our proposed methodology will allow for the development of new types of recommendation systems, advertisement strategies and election campaigns. For example, authority nodes are better advertisement targets for authority-propagated products. On the other hand, election campaign slogans, might gain popularity due to the network effect and therefore be advertised accordingly.

Roadmap: The rest of the paper is organized as follows: In Section 2 we give a brief overview of the related work. Sections 3 and 4 give an overview of our methodology for incorporating authority pressure into the peer models. We show an extensive experimental evaluation of our framework in Section 5 and we conclude the paper in Section 6.

2 Related Work

Despite the large amount of work on peer models and on identification of authority nodes, to our knowledge, our work is the first attempt to combine peer and authority pressure into a single information-propagation model. Also, contrary to the goal of identifying authority nodes, our goal is to classify propagated trends as being peer- or authority-propagated.

One of the first models to capture peer influence was by Bass [5], who defined a simple model for product adoption. While the model does not take into account the network structure, it manages to capture some commonly-observed phenomena, such as the existence of a "tipping point." More recent models such as the linear-threshold model [10, 11] or the cascade model [11] introduce the dependence of influence on the set of peers, and since then there has been a large number of generalizations.

In a series of papers based on the analysis of medical data and offline social networks Christakis, Fowler, and colleagues showed the existence of peer influence on social behavior and emotions, such as obesity, alcoholism, hapiness, depression, loneliness [6, 8, 14]. An important characteristic in these analyses is the performance of statistical tests through modifying the social graph to provide evidence for peer influence. It was found that in general influence can extend up to three degrees of separation. Around the same time, Anagnostopoulos et al. [3] and Aral et al. [4], provided evidence that a lot of the correlated behavior among peers can be attributed to other factors such as *homophily*, the tendency of individuals to associate and form ties with similar others. The *time-shuffle* test that we apply later is a randomization test used in [3] to rule out influence effects from peers. Although clearly related, the above work is only complementary to ours: none of the above papers considers authorities as a factor that determines the propagation of information.

Recently, there have been many studies related to the spreading of ideas, news and opinions in the blogosphere. Authors refer to all these propagated items as *memes*. Gomez-Rodriguez et al. [9] try to infer who influences whom based on the time information over a large set of different memes. Contrary to our work where the underlying network is part of the input, Gomez-Rodrigues

et al. assume that the network structure is unknown. In fact, their goal is to discover this hidden network and the key assumption of the method is that a node only gets influenced by its neighbors. Therefore, they do not account for authority influence.

More recently, Yang and Leskovec [16] applied a nonparametric modeling approach to learn the direct or indirect influence of a set of nodes (e.g. news sites) to other blogs or tweets. Although one can consider the discovered set of nodes as authority nodes, the model of Yang and Leskovec does not take into account the network of peers. Our work is mostly focused on the interaction and the separation of peer and authority influence within a social-network ecosystem.

Recent work by Wu et al. [15] focuses on classifying twitter users as "elite" and "ordinary"; elite users (e.g., celebrities, media sources, or organizations) are those with large influence on the rest of the users. Exploiting the twitter-data characteristics the authors discover that a very small fraction of the population (0.05%) is responsible for the generation of half of the content in twitter. Although related, the focus of our paper is rather different: our goal is not to identify the authorities and the peers of the network. Rather, we want to classify the trends as those that are being authority-propagated versus those being peer-propagated.

Related in spirit is also the work of Amatriain et al. [2]; their setting and their techniques, however, are entirely different than ours: they consider the problem of collaborative filtering and they compare the information obtained by consulting "experts" as opposed to "nearest neighbors" (i.e., nodes similar to the node under consideration). The motivation for that work is the fact that data on nearest neighbors is often sparse and noisy, as opposed to the more global information of experts.

3 Peer and Authority Models

An information-propagation network can be represented by a directed graph. The graph consists of a set of n nodes, denoted by V. We refer to these nodes as *peers* (or agents). These nodes are organized in a directed graph $G = (V, E)$. The edges of the graph represent the ability of a node to influence another node. That is, a directed link from node u to node v, $(u \rightarrow v)$ denotes that node u can influence node v. We call the graph G the *peer influence graph*. Given a node u we refer to all the nodes that can influence u, that is, the nodes that have directed links to u as the *peers* or *neighbors* of u.

In addition to the n peer nodes, our model assumes the existence of N globally accepted *authorities*, represented by the set A. Every authority $a \in A$ has the potential to influence *all* the nodes in V. Intuitively, this means that there are directed influence edges from every authority in $a \in A$ to every peer $v \in V$; we use F to represent the directed edges from authorities to peers. For simplicity we assume that there are no edges amongst authorities. We refer to the graph $H = (V \cup A, E \cup E_A)$ as the *extended influence graph*.

Fashion trends, news items or research ideas propagate amongst peers and authorities. We collectively refer to all the propagated trends as *information*

items (or simply *items*). We call the nodes (peers or authorities) that have adopted a particular item *active* and the nodes that have not adopted the same item as *inactive*.

We assume that the propagation of every item happens in discrete time steps; we assume that we have a limited observation period from timestamp 1 to timestamp T. At every point in time $t \in \{1, \ldots, T\}$, each inactive node u decides whether to become active. The probability that an inactive node u becomes active is a function $P(x, y)$ of the number x of peers that can influence u that are already active and the number y of active authorities. In principle, function P can be any function that is increasing in both x and y. As we will see in the next section, we will focus on a simple function that fits our purposes.

4 Methodology

In this section, we present our methodology for measuring peer and authority pressure in information propagation. Based on that we offer a characterization of trends as *peer-* or *authority-propagated* trends. Peer-propagated trends are those whose observed propagation patterns can be largely explained due to peer pressure. Authority-propagated trends are those that have been spread mostly due to authority influence.

We start in Section 4.1 by explaining how logistic regression can be used to quantify the extent of peer and authority pressure. In Section 4.2 we define a randomization test that we use in order to quantify the statistical significance of the logistic regression results.

4.1 Measuring Social Influence

The discussion below focuses on a single propagated item. Assume that at some point in time, there are y active authorities. At this point in time, a node with x active peers becomes active with probability $P(x, y)$. As it is usually the case [3], we use the logistic function to model the dependence of the probability $P(x, y)$ as a function of the independent variables x and y. That is,

$$P(x, y) = \frac{e^{\alpha \ln(x+1)+\beta \ln(y+1)+\gamma}}{1 + e^{\alpha \ln(x+1)+\beta \ln(y+1)+\gamma}}, \tag{1}$$

where α, β and γ are the coefficients of the logistic function. The values of α and β capture respectively the strength of peer and authority pressure in the propagation of item i. More specifically α, β take values in $\mathbb{R}$. Large values of α provide evidence for peer influence in the propagation of item i. Large values of β provide evidence for authority influence in the propagation of i. For every item i, we call α the *peer coefficient* and β the *authority coefficient* of i. Parameter γ models the impact of factors other than peer and authority pressure in the propagation of the item. For example, the effect of random chance is encoded in the value of the parameter γ. We call γ the *externality coefficient* since it quantifies the effect of external parameters.

The logit function of probability $P(x, y)$ (Equation (1)) gives

$$\ln\left(\frac{P(x,y)}{1-P(x,y)}\right) = \alpha \ln(x+1) + \beta \ln(y+1) + \gamma. \quad (2)$$

We estimate α, β and γ using maximum likelihood logistic regression. More specifically, for each $t = 1, 2, \ldots, T$ let $N(x, y, t)$ be the number of users who at the beginning of time t had x active neighbors and they themselves became active at time t when y authorities were active. Similarly, let $\overline{N}(x, y, t)$ be the number of users who at the beginning of time t had x active neighbors, but did not become active themselves at time t when y authorities were also active. Finally, let $N(x, y) = \sum_t N(x, y, t)$ and $\overline{N}(x, y) = \sum_t \overline{X}(x, y, t)$. Then, the maximum-likelihood estimation of parameters α, β are those that maximize the likelihood of the data at time t, namely,

$$\prod_{x,y} P(x,y)^{N(x,y)} \left(1 - P(x,y)\right)^{\overline{N}(x,y)}. \quad (3)$$

While in general there is no closed form solution for the above maximum likelihood estimation problem, there are many software packages that can solve such a problem quite efficiently. For our experiments, we have used Matlab's statistics toolbox.

We apply this analysis to every propagated item and thus obtain the maximum-likelihood estimates of the peer and authority coefficients for each one of them.

4.2 Randomization Test

One way of inferring whether item i's propagation is better explained due to peer or authority influence is to obtain the maximum-likelihood estimates of peer and authority coefficients (α, β) and conclude the following: if $\alpha > \beta$, then i is a peer-propagated item. Otherwise, if $\beta > \alpha$ then i is an authority-propagated item. Although this is might be a reasonable approach towards the categorization of the item, the question of how much larger should the value of α (resp. β) be in order to characterize i as peer- (resp. authority-) propagated item. Even if $\alpha >> \beta$ (or vice versa), we still need to verify that this result is due to strong evidence in the data.

In order to reach conclusions based on strong evidence in the data, we devise a randomization test which we call the *time-shuffle* test. Let H be the input influence graph and let D be the dataset that associates every node in $V \cup A$ that becomes active with its activation time. The time-shuffle test permutes the activation times of the nodes in D. In this way, a randomized version D' of D is obtained. Note that D' contains the same nodes as D (those that eventually become active), however the activation times are permuted.

Assume that the maximum-likelihood estimation method for input $\langle H, D\rangle$ estimates the peer and authority coefficients (α, β). Also, denote by $(\alpha(D'), \beta(D'))$ the peer and authority coefficients computed running maximum-likelihood estimation on input $\langle H, D'\rangle$. Let $\mathcal{D}$ be the set of all possible randomized versions

that can be created from the input dataset D via the time-shuffle test. Then, we define the *strength of peer influence* S_α to be the fraction of randomized datasets $D' \in \mathcal{D}$ for which $\alpha > \alpha(D')$. Therefore,

$$S_\alpha = \Pr_{\mathcal{D}} \left(\alpha > \alpha(D') \right). \tag{4}$$

Note that the probability is taken over all possible randomized versions $D' \in \mathcal{D}$ of the original dataset D.

Similarly, we define the *strength of authority influence* S_β, to be the fraction of randomized datasets D' for which $\beta > \beta(D')$. Therefore,

$$S_\beta = \Pr_{\mathcal{D}} \left(\beta > \beta(D') \right). \tag{5}$$

Both the peer and the authority strengths take values in $[0, 1]$; the larger the value of the peer (resp. authority) strength the stronger the evidence of peer influence (resp. authority influence) in the data.

5 Experimental Results

In this section, we present our experimental evaluation both on real and synthetic data. Our results on real data coming from online social media and computer-science collaboration networks reveal the following interesting findings: In online social-media networks the fit of our model indicates a stronger presence of authority pressure, as opposed to the scientific collaboration network that we examine. Our results on synthetically-generated data show that our methods recover the authority and peer coefficients accurately and efficiently.

5.1 Datasets and Implementation

We experiment with the following real-world datasets:

The `MemeTracker` dataset [12].[3] The original dataset tracks commonly used memes across online (mostly political) blogs and news sources. The dataset contains information about the time a particular meme appeared on a given webpage as well as links from and to each listed webpage.

In order to analyze the data with our approach, we assume that each meme is an information item that propagates through the network of peers. The influence graph $G = (V, E)$ consists of directed relationships between the blog sites in the dataset. That is, the peer nodes (the set V) are the blog sites in the dataset. In our version of the data we only consider blogs from wordpress.com and blogspot.com since their URL structure makes it easy to identify the same blogger across posts. There is a directed link (infuence) from blog b to blog b' if there exist at least one hyperlink from b' to b. That is, b' refers to b and therefore b can influence b'. Such directed links constitute the edges in E.

The authority nodes A in our dataset are the news-media sites available in the original dataset. In total, the processed dataset consists of 123,400 blog sites, 13,095 news sites and 124,694 directed links between the blogs (edges).

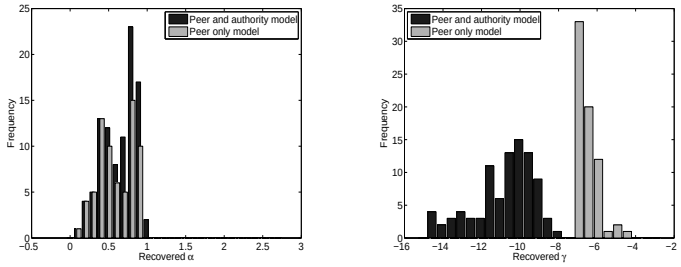
[3] The dataset is available at `http://snap.stanford.edu/data/memetracker9.html`.

Although the dataset contains 71,568 memes, their occurrence follows a power-law distribution and many memes occur very infrequently. In our experiments, we only experiment with the set of 100 most frequently-appearing memes. We denote this set of memes by M_F. For every meme $m \in M_F$ we construct a different extended influence graph. The set of authorities for this meme A_m is the subset of the top-50 authorities in A that have most frequently used this particular meme.

The `Bibsonomy` dataset [1]. BibSonomy is an online system with which individuals can bookmark and tag publications for easy sharing and retrieval. In this dataset, the influence graph $G = (V, E)$ consists of peers that are scientists. There is a link between two scientists if they have co-authored at least three papers together. The influence links in this case are bidirectional since any of the co-authors can influence each other. The items that propagate in the network are tags associated with publications. A node is active with respect to a particular tag if at least one of the node's publications has been associated with the tag. For a given tag t, the set of authorities associated with this tag, A_t, are the top-20 authors with the largest number of papers tagged with t. These authors are part of the extended influence graph of tag t, but not part of the original influence graph.

For our experiments, we have selected papers from conferences. There are a total of 62,932 authors, 9,486 links and 229 tags. Again, we experiment with the top-100 most frequent tags.

Implementation. Our implementation consists of two parts. For each item, we first count the number of users who were active and inactive at each time period; that is, we evaluate the matrices $N(x, y)$ and $\overline{N}(x, y)$ in Equation (3). Then, we run the maximum likelihood regression to find the best estimates for α, β, and γ in Equation (1). We ran all experiments on a AMD Opteron running at 2.4GHz. Our unoptimized MATLAB code processes one meme from the `MemeTracker` dataset in about 204 seconds. On average, the counting step requires 96% of this total running time. The rest 4% is the time required to run the regression step. For the `Bibsonomy` dataset, the average total time spent on a tag is 38 seconds. Again, 95% of this time is spent on counting and 5% on regression.



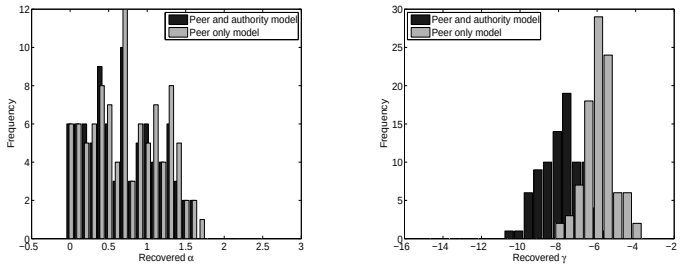
(a) α coefficients (b) γ coefficients

Fig. 2. `MemeTracker` dataset. Histogram of the values of the peer and externality coefficients α and γ recovered for the pure peer ($\beta = 0$) and the integrated peer and authority model.

5.2 Gain of Authority Integration

The goal of our first experiment is to demonstrate that the integration of authority influence in the information-propagation models can help to explain observed phenomena that peer models had left unexplained. For this we use the `MemeTracker` and the `Bibsonomy` datasets to learn the parameters α, β and γ for each of the propagated items. At the same time, we use the *peer-only* version of our model by setting $\beta = 0$ and learn the parameters α' and γ' for each one of the propagated items. This way, the peer-only model does not attempt to distinguish authority influence, and is similar to the models currently found in the literature.

The results for the `MemeTracker` dataset are shown in Figure 2. More specifically, Figure 2(a) shows the histogram of the recovered values of α and α' we obtained. The two histograms show that the distribution of the values of the peer coefficient we obtain using the two models are very similar. On the other hand, the histogram of the values of the externality coefficients obtained for the two models (shown in Figure 2(b)) are rather distinct. In this latter pair of histograms, we can see that the values of the externality coefficient obtained in the peer-only model are larger than the corresponding values we obtain using our integrated peer and authority model. This indicates that the peer-only model could only explain a certain portion of the observed propagation patterns associating the unexplained patterns to random effects. The addition of an authority parameter explains a larger portion of the observed data, attributing much less of the observations to random factors. The results for the `Bibsonomy` dataset (Figure 3) indicate the same trend.

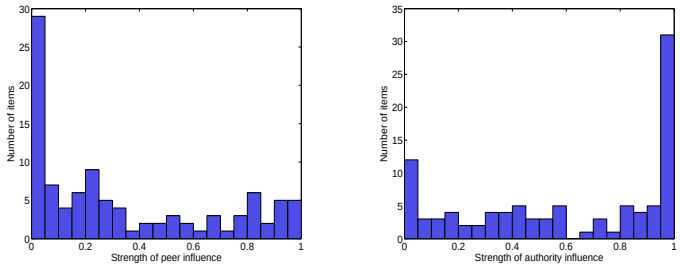


(a) α coefficients (b) γ coefficients

Fig. 3. `Bibsonomy` dataset. Histogram of the values of the peer and externality coefficients α and γ recovered for the pure peer ($\beta = 0$) and the integrated peer and authority model.

5.3 Analyzing the `MemeTracker` Dataset

In this section, we show the results of our analysis for the `MemeTracker` dataset. The results show that the majority of the memes we consider are authority-propagated. This means that the bloggers adopt memes by authoritative online news media sites more than by their fellow bloggers.



(a) Peer influence (b) Authority influence

Fig. 4. `MemeTracker` dataset. Frequency distribution of the recovered strength of peer and authority influence.

The above result is illustrated in Figure 4. These histograms show the number of memes that have a particular strength of peer (Figure 4(a)) and authority influence (Figure 4(b)). We obtain these results by estimating the strength of peer and authority influence using 100 random instances of the influence graph generated by the time-shuffle test (see Section 4.2). The two historams shown in Figures 4(a) and 4(b) indicate that for most of the memes in the dataset, authority pressure is a stronger factor affecting their propagation compared to peer influence. More specifically, the percentage of memes with peer strength greater than 0.8 is only 18% while the percentage of memes with authority strength greater than 0.8 is 45%. Also, 46% of memes have peer strength below 0.2, while only 22% of of memes have authority strength below 0.2.

Fig. 5. `MemeTracker` dataset. Peer strength (x-axis) and authority strength (y-axis) of the top-100 most frequent memes. The size of the circles indicate is proportional to the frequency of the meme.

We demonstrate some anecdotal examples of peer- and authority-propagated memes in Figure 5. The plot is a two-dimensional scatterplot of the peer and

authority strength of each one of the top-100 most frequent memes. The size of the marker associated with each meme is proportional to the meme's frequency.

Table 1. `MemeTracker` dataset. Examples of memes with different peer and authority strengths. Bucketization was done using equi-depth histograms.

Group 1: Top-5 frequent memes with low peer and low authority strength.
1. life liberty and the pursuit of happiness
2. hi how are you doing today
3. so who are you voting for
4. are you kidding me
5. of course not
Group 2: Top-5 frequent memes with high peer and high authority strength.
1. joe the plumber
2. this is from the widows the orphans and those who were killed in iraq
3. our opponent is someone who sees america it seems as being so imperfect imperfect enough that he's palling around with terrorists who would target their own country
4. yes we can yes we can
5. i guess a small-town mayor is sort of like a community organizer except that you have actual responsibilities
Group 3: Top-5 frequent memes with low peer and high authority strength.
1. i need to see what's on the other side i know there's something better down the road
2. i don't know what to do
3. oh my god oh my god
4. how will you fix the current war on drugs in america and will there be any chance of decriminalizing marijuana
5. i barack hussein obama do solemnly swear
Group 4: Top-5 frequent memes with high peer and low authority strength.
1. we're in this moment and if we fail to do the right thing heaven help us
2. if you know what i mean
3. what what are you talking about
4. i think we should all be fair and balanced don't you
5. our national leaders are sending u s soldiers on a task that is from god

A lot of the memes in the lower left, that is, memes with both low peer and authority strength, are commonly seen phrases that are arguably not subject to social influence. The memes in this category tend to be short and generic. For example, "are you kidding me" and "of course not" were placed in this category. The meme "life liberty and the pursuit of happiness" is also placed in the same category. Although this last quote is not as generic as the others, it still does not allude to any specific event or controversial political topic. The low social correlation attributed to these memes indicates that they sporadically appear

in the graph, without relation to each other. Using an equidepth histogram we extract the the top-5 most frequent memes with low peer and low authority strength and show them in Group 1 of Table 1.

Diagonally opposite, in the upper right part of the plot, are the memes with high peer and high authority strength. These are particularly widely-spread quotes that were pertinent to the 2008 U.S. Presidential Election, and that frequently appeared in both online news media sites and blog posts. Examples of memes in this category include "joe the plumber" and President Obama's slogan, "yes we can." Finally, the meme "this is from the widows the orphans and those who were killed in iraq" is also in this category. This is a reference to the much-discussed incident where an Iraqi journalist threw a shoe at President Bush. The top-5 most frequent memes with high peer and authority strengths are also shown in Group 2 of Table 1. Comparing the memes in Groups 1 and 2 in Tables 1, one can verify that, on average, the quotes with high peer and high authority strength are much longer and more specific than those with low peer and low authority strengths. As observed before, exceptions to this trend are the presidential campaign memes "joe the plumber" and "yes we can".

Memes with low peer and high authority strength (left upper part of the scatterplot in Figure 5) tend to contain quotes of public figures, or refer to events that were covered by the news media and were then referenced in blogs. One example is "I barack hussein obama do solemnly swear," the first line of the inaugural oath. The inauguration was covered by the media, so the quotes originated in news sites and the bloggers began to discuss it immediately after. Typically, memes in this group all occur within a short period of time. In contrast, memes with both high peer and high authority influence are more likely to gradually gain momentum. The top-5 most frequent memes with low peer and high authority strength are also shown in Group 3 of Table 1.

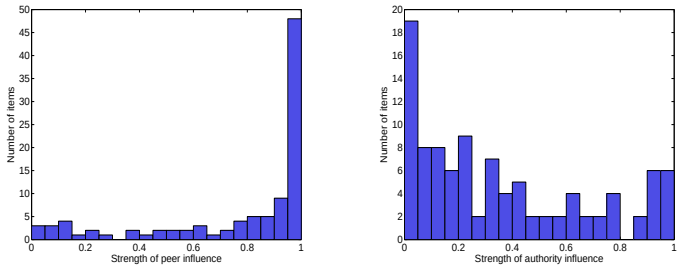
We expect that memes with high peer and low authority strength (right lower part of the scatterplot in Figure 5) are mostly phrases that are not present in the mainstream media, but are very popular within the world of bloggers. An example of such a meme, as extracted by our analysis, is "mark my words it will not be six months before the world tests barack obama like they did john kennedy." This is a quote by Joe Biden that generates many more high-ranked blog results than news sites on a Google search. Another example is "i think we should all be fair and balanced don't you," attributed to Senator Schumer in an interview on Fox News, which was not covered by mainstream media but was an active topic of discussion for bloggers. The top-5 most frequent memes with high peer and low authority strength are also shown in Group 4 of Table 1.

5.4 Analyzing the `Bibsonomy` Dataset

In this section, we show the results of our analysis for the `Bibsonomy` dataset. The results show that the majority of the items we consider here are peer-propagated. Recall that in the case of the `Bibsonomy` dataset the propagated items are tags associated with papers written by the scientists forming the collaboration network. One should interpret tags as research topics or themes. Our

findings indicate that when choosing a research direction, scientists are more likely to be influenced by people they collaborated with rather than experts in their field.

The above result is illustrated in Figure 6. These histograms show the number of tags that have a particular strength of peer (Figure 6(a)) and authority influence (Figure 6(b)). We obtain these results by estimating the strength of peer and authority influence using 100 random dataset instances generated by the time-shuffle test (see Section 4.2).



(a) Peer influence (b) Authority influence

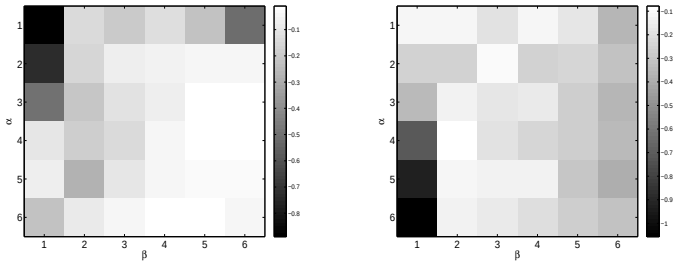
Fig. 6. `Bibsonomy` dataset. Frequency distribution of the recovered strength of peer and authority influence.

Overall, we observe stronger peer influence than authority influence in the `Bibsonomy` dataset, as illustrated in Figures 6(a) and 6(b). The percentage of tags with peer strength greater than 0.8 is 67% while the percentage of tags with authority strength greater than 0.8 is only 15%. Also, 41% of the tags have authority strength below 0.2, while only 11% of the tags have peer strength below 0.2.

5.5 Experiments on Synthetic Data

We conclude our experimental evaluation by showing the performance of our model using synthetically generated data. The sole purpose of the experiments reported here is to demonstrate that in such data the maximum-likelihood parameter estimation and the time-shuffle randomization test lead to conclusions that are consistent with the data-generation process.

Accuracy of Recovery: To verify that the recovery for α and β is accurate, we randomly generate a synthetic power-law graph, and simulate the propagation of an item over this graph using the logistic function with predetermined values for α, β, and γ. In particular, we used the Barabasi model to generate graphs with 10000 peers and 50 authorities. We used $\alpha, \beta \in \{1, 2, 3, 4, 5, 6\}$, and $\gamma = -10$. Higher values for α and β cause all the nodes to become active almost immediately, so it becomes very difficult to observe how the items propagate. To quantify the accuracy with which a parameter x is recovered, we define the relative recovery error. If x is the value of the coefficient used in the generation



(a) of peer coefficient α (b) Authority coefficient β

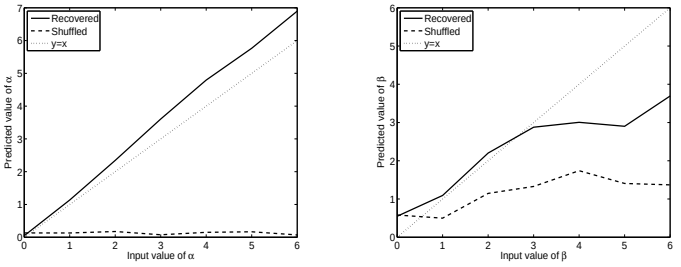
Fig. 7. Synthetic data. Relative error for the peer coefficient α and the authority coefficient β.

process and $\hat{x}$ is the recovered value, then the relative error is given by

$$\text{RelErr}(x, \hat{x}) = \frac{|x - \hat{x}|}{x}.$$

The relative eror takes values in the range $[0, \infty)$, where a smaller value indicates better accuracy of the maximum-likelihood estimation method.

Figure 7 shows the relative recovery errors for different sets of values for α and β in this simulated graph, where darker colors represent smaller relative errors. In most cases, for both the peer and the authority coefficients, the relative error is below 0.2 indicating that the recovered values are very close to the ones used in the data-generation process.



(a) Recovered α for fixed $\beta = 2$ (b) Recovered β for fixed $\alpha = 2$

Fig. 8. Synthetic data. Recovering the peer coefficient α and the authority coefficient β for the real data and the data after time-shuffle randomization.

Time-Shuffle Test on Synthetic Data: Figures 8(a) and 8(b) show the recovered value of peer and authority coefficient respectively, as a function of the value of the same parameter used in the data-generation process. One can observe that in both cases the estimated value of the parameter is very close to the input parameter. Visually this is represented by the proximity of the *recovered* curve to the $y = x$ curve; curve $y = x$ represents ideal, errorless recovery. Second, the recovered values follow the trend of the values used in the data-generation process. That is, as the values of α and β used in the data generation increase,

the recovered values of α and β also increase. This indicates that even if the maximum-likelihood estimation does not recover the data-generation parameters exactly, it correctly identifies the types of influence (peer or authority).

In addition to the recovered values of peer and authority coefficient we also report the average of the corresponding parameters obtained in 100 randomized instances of the originally generated dataset; the randomized instances are all generated using the time-shuffle test. These averages are reported as the dashed line in both Figures 8(a) and 8(b). We observe that the values of peer and authority coefficients obtained for these randomized datasets are consistently smaller than the corresponding recovered and actual values. This means that the time-shuffle test is consistently effective in identifying both peer and authority influence. Observe that as the values of α and β parameters used in the data-generation process increase, the difference between the average randomized value of the parameter and the recovered value increases. This suggests that as the dependence of the propagation on a particular type of influence (peer or authority) becomes larger, it becomes easier to identify the strength of the corresponding influence type using the time-shuffle test.

6 Conclusions

Given the adoption patterns of network nodes with respect to a particular item, we have proposed a model for deciding whether peer or authority pressure played a central role in its propagation. For this, we have considered an information-propagation model where the probability of a node adopting an item depends on two parameters: (a) the number of the node's neighbors that have already adopted the item and (b) the number of authority nodes that appear to have the item. In other words, our model extends traditional peer-propagation models with the concept of authorities that can globally influence the network. We developed a maximum-likelihood framework for quantifying the effect of peer and authority influence in the propagation of a particular item and we used this framework for the analysis of real-life networks. We find that accounting for authority influence helps to explain more of the signal which many previous models classified as noise. Our experimental results indicate that different types of networks demonstrate different propagation patterns. The propagation of memes in online media seems to be largely affected by authority nodes (e.g., news-media sites). On the other hand, there is not evidence for authority pressure in the propagation of research trends within scientific collaboration networks.

There is a set of open research questions that arise from our study. First, various generalizations could fit in our framework: peers or authorities could influence authorites, nodes or edges could have different weights indicatings stronger/weaker influence pressures, and so on. More importantly, while our methods compare peer and authority influence, it would be interesting to account for selection effects [3, 4] that might affect the values of the coefficients of our model. Such a study can give a stronger signal about the exact source of influence in the observed data. Furthermore, in this paper we have considered that the set of authority nodes are predefined. It would be interesting to see

whether the maximum-likelihood framework we have developed can be extended to automatically identify the authority nodes, or whether some other approach (e.g., one based on the HITS algorithm).

References

1. Knowledge and Data Engineering Group, University of Kassel, Benchmark Folksonomy Data from BibSonomy. version of June 30th, 2007.
2. X. Amatriain, N. Lathia, J. M. Pujol, H. Kwak, and N. Oliver. The wisdom of the few: A collaborative filtering approach based on expert opinions from the web. In *SIGIR*, 2009.
3. A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD*, 2008.
4. S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proceedings of the National Academy of Sciences, PNAS*, 106(51), 2009.
5. F. M. Bass. A new product growth model for consumer durables. *Management Science*, 15:215–227, 1969.
6. J. T. Cacioppo, J. H. Fowler, and N. A. Christakis. Alone in the crowd: The structure and spread of loneliness in a large social network. *Journal of Personality and Social Psychology*, 97(6):977–991, 2009.
7. N. Christakis and J. Fowler. *Connected: The surprising power of our social networks and how they shape our lives*. Back Bay Books, 2010.
8. J. H. Fowler and N. A. Christakis. The dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the framingham heart study. *British Medical Journal*, 337, 2008.
9. M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *KDD*, 2010.
10. M. Granovetter. Threshold models of collective behavior. *The American Journal of Sociology*, 83:1420–1443, 1978.
11. D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.
12. J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *KDD*, 2009.
13. J.-P. Onnela and F. Reed-Tsochas. Spontaneous emergence of social influence in online systems. *Proceedings of the National Academy of Sciences, PNAS*, 2010.
14. J. N. Rosenquist, J. H. Fowler, and N. A. Christakis. Social network determinants of depression. *Molecular Psychiatry*, 16(3):273–281, 2010.
15. S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *WWW*, pages 705–714, 2011.
16. J. Yang and J. Leskovec. Modeling information diffusion in implicit networks. In *ICDM*, 2010.