ProfileRank: Finding Relevant Content and Influential Users based on Information Diffusion *

Arlei Silva*, Sara Guimarães,
Wagner Meira Jr.
Universidade Federal de Minas Gerais
{arlei,sara,meira}@dcc.ufmg.br

Mohammed Zaki
Rensselaer Polytechnic Institute
zaki@cs.rpi.edu

ABSTRACT

Understanding information diffusion processes that take place on the Web, specially in social media, is a fundamental step towards the design of effective information diffusion mechanisms, recommendation systems, and viral marketing/advertising campaigns. Two key concepts in information diffusion are influence and relevance. Influence is the ability to popularize content in an online community. To this end, influentials introduce and propagate relevant content, in the sense that such content satisfies the information needs of a significant portion of this community.

In this paper, we study the problem of identifying influential users and relevant content in information diffusion data. We propose ProfileRank, a new information diffusion model based on random walks over a user-content graph. ProfileRank is a PageRank inspired model that exploits the principle that relevant content is created and propagated by influential users and influential users create relevant content. A convenient property of ProfileRank is that it can be adapted to provide personalized recommendations.

Experimental results demonstrate that ProfileRank makes accurate recommendations, outperforming baseline techniques. We also illustrate relevant content and influential users discovered using ProfileRank. Our analysis shows that ProfileRank scores are more correlated with content diffusion than with the network structure. We also show that our new modeling is more efficient than PageRank to perform these calculations.

Categories and Subject Descriptors: H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval − *Retrieval models*

General Terms: Algorithms, Experimentation

Keywords: Influence, Relevance, Information diffusion

1. INTRODUCTION

Powered by the remarkable success of Twitter, Facebook, Youtube, and the blogosphere, social media is taking over traditional media as the major platform for content distribution. The combination of user-generated content and online social networks is the engine behind this revolution in the way people share news, videos, memes, opinions, and ideas in general. As a consequence, understanding how users consume and propagate content in information diffusion processes is a fundamental step towards the design of effective information diffusion mechanisms, recommendation systems, and viral marketing/advertising campaigns on the Web.

Two key concepts in information diffusion are *influence* and *relevance*. In social networks, influence can be defined as the capacity to affect the behavior of others [10]. However, in information diffusion scenarios, influence is usually a measure of the ability of popularizing information. In other words, an *influential* is someone who propagates information widely, producing large diffusion cascades [19]. Relevance is a relationship between a user and a piece of information, in the sense that relevant information satisfies a user's information needs/interests, being a fundamental concept also in information retrieval and recommender systems [2, 23].

This work focuses on the link between user influence and information relevance in information diffusion data, which describe how users create and propagate information across time. As we are interested in the diffusion of content (e.g., news, videos) on the Web, we use the terms 'content' and 'information' interchangeably. From a content producer's perspective, we can measure influence as the reach of the content a user introduces to the online community. Furthermore, because users are expected to consume content that is relevant to them, influentials can be seen as users who produce content that is relevant to a significant portion of the community. An implication of this semantic connection between user influence and content relevance is that these measures may be computed by leveraging individual content relevance assessments described as diffusion data. In this paper, we present ProfileRank, a random walk based information diffusion model that computes user influence and content relevance using information diffusion data.

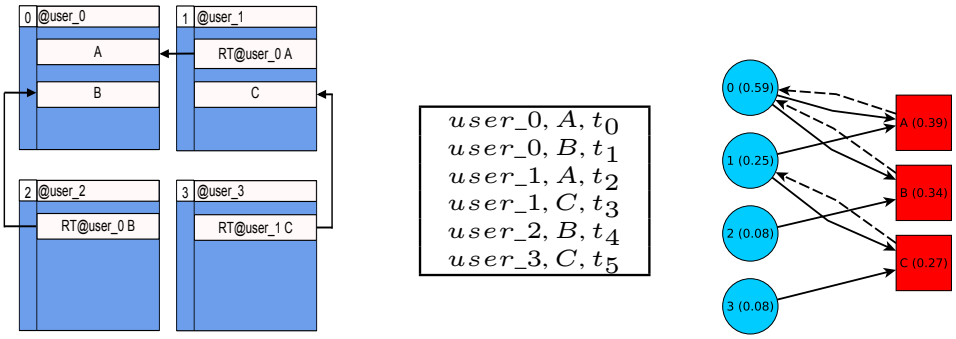
ProfileRank is based on the principle that relevant content is created and propagated by influential users and influential users create relevant content. If we consider Twitter as an information diffusion platform and tweets as content propagated through retweets, ProfileRank can be intuitively described in terms of the behavior of a *random tweeter* (or *twitterer*) that navigates through Twitter profiles by clicking on random tweets (or retweets from these same tweets). Every click on a tweet leads the random tweeter to the profile of the original author of the tweet. We measure user influence as the frequency with which the random tweeter visits a given profile. Likewise, content relevance is measured as the frequency with which the random tweeter clicks on a tweet and its retweets.

Figure 1 depicts the application of ProfileRank to Twitter data using an illustrative example. A set of profiles, with their respective tweets, is shown in Figure 1a. In Figure 1b, we represent these tweets as information diffusion data, which describes user-content

*This work was partially supported by CNPq, CAPES, FINEP, FAPEMIG, InWeb, NSF-EMT-0829835 and NSF-CCF-1240646.

*Now at University of California, Santa Barbara

SNAKDD '13 Chicago, Illinois USA



(a) Twitter (b) Diffusion data (c) Diffusion model

Figure 1: Modeling diffusion data using ProfileRank. (a) Illustrative Twitter dataset with 4 users (user_0, user_1, user_2, user_3) and 3 tweets (A, B, C). (b) Diffusion data. (c) ProfileRank diffusion model. Circles and squares represent users and content, respectively. Solid arrows connect the user to each content he has created or propagated, and dashed arrows link the piece of content to its creator. ProfileRank measures user influence and content relevance using random walks over a user-content graph. Scores are shown inside each vertex.

associations through timestamps in the interval $[t_0, t_5]$. Figure 1c shows how we model diffusion data as a user-content bipartite graph. Based on random walks over this graph, ProfileRank computes user influence and content relevance using a formulation that resembles the PageRank [22] and HITS [16] algorithms. In fact, ProfileRank is a PageRank inspired algorithm for the identification of influential users and relevant content from information diffusion data.

Readers who are familiar with PageRank may find it easier to see ProfileRank as an adaptation of the idea of PageRank to understand influence relations in social networks such as Twitter. However, instead of using the Twitter social interactions, like TwitterRank [30], ProfileRank is based on content propagation. Our approach is motivated by previous work, which has shown that a user's number of followers is not a good measure of her capacity to propagate content on Twitter[5]. Nevertheless, besides Twitter, ProfileRank can also be applied in the analysis of diffusion data in other scenarios, such as social networks and the blogosphere.

An interesting property of ProfileRank is that it can be personalized in order to compute user influence and content relevance from the perspective of a particular user. Hence, we can apply ProfileRank to content and user recommendation tasks. Due to the absence of ground truth information on user influence and content relevance (e.g., a universally accepted ranking of relevant tweets and influential Twitter users), we perform this task as a means to evaluate ProfileRank. Our premise is that the effectiveness of ProfileRank in recommendation problems is an evidence of its effectiveness in the discovery of global influential users and relevant content.

ProfileRank does not rely on the semantics of the content. This property is specially suitable for scenarios where modeling the content is a challenge (e.g., image and video propagation). Moreover, ProfileRank can easily incorporate a keyword-based filtering or a topic modeling technique in case textual features are available.

We summarize the main contributions of this paper as follows:

- **Integrated view on influence and relevance:** We introduce an integrated view on user influence and content relevance with implications to the analysis of information diffusion.
- **Model for content relevance and user influence:** We present ProfileRank, which computes content relevance and user influence based on random walks over a user-content graph.
- **Evaluation of the proposed model:** We evaluate ProfileRank using Twitter and MemeTracker data, showing that it makes effective content and user recommendations, and discussing relevant content and influential users discovered.

2. RELATED WORK

Social influence and information diffusion. The popularization of internet-based communication and interactivity platforms have resulted in a tremendous interest in understanding social interactions at large scale [29]. In particular, the characterization of the dynamics of information propagation in social media applications, such as blogs [13] and other online communities [6, 12, 19], has supported several discoveries regarding the potential of viral marketing strategies and the role played by influentials – people who affect the behavior of the others. Due to its growing popularity, Twitter has become a standard social laboratory for understanding information diffusion and influence [5, 17], but some of these studies have been focused on straightforward influence measures, such as a user's number of followers. Our effort is similar to [30] and [24] in the sense that we study new models for identifying influentials based on diffusion traces. However, while TwitterRank [30] and Influence-Passivity [24] rely on the social network structure in order to identify influential users, ProfileRank measures influence and relevance based only on diffusion data, i.e., user-content associations over time. In fact, TwitterRank is a variation of PageRank [22] over the follower network induced by a particular topic. Similarly, in [26], the authors provide a way to use topic modeling and network structure in order to find the most influential nodes in a particular topic. In Section 5, we compare ProfileRank against Pagerank, which is equivalent to TwitterRank when topics are not considered. To the best of our knowledge, ProfileRank is the first model to integrate user influence and content relevance.

Content search and recommendation. Information retrieval [2] and recommender systems [23] research has lead to the development of several strategies for identifying relevant content on the Web. As the volume of user-generated content on the Web has increased, specially in social media, search and recommendation mechanisms have become increasingly necessary in various applications, such as Twitter [8] and Youtube [3]. ProfileRank evaluates both the global and personalized relevance of content based on its diffusion through users. Therefore, we study the effectiveness of ProfileRank as a tweet and meme recommender system in Section 5.2. Different from most previous works on content recommendation for Twitter and similar platforms [7, 15], ProfileRank does not rely on social network or textual information in order to identify relevant content to users. For that reason, it can be compared to traditional recommendation approaches based on user-content relationships, such as collaborative-filtering techniques [23].

Link prediction in social networks. The link prediction problem in social networks consists of inferring new interactions, given a snapshot of the network [21]. In directed networks, where relationships are not reciprocal, social interactions can depict influence interactions [25]. For instance, Twitter users receive the tweets from those users they follow and a user's number of followers has been considered a measure of influence [5]. As a consequence, the discovery of influence interactions can be seen as an instance of link prediction in social networks. Since ProfileRank measures the degree of influence of one user over another, we apply it to the prediction of follower relationships on Twitter, a problem we call user recommendation (see Section 5.3). More specifically, we are interested in predicting these relationships based solely on content and time information – unlike typical studies on user recommendation on Twitter that assume the availability of a snapshot of the network [15, 14]. This "cold start" version of the link prediction problem [18] is of special interest when the influence network is unobserved, such as in blogs and news media [11]. We show that ProfileRank outperforms baseline strategies on Twitter data.

Identifying authorities and computing relevance scores between nodes in hyperlinked environments. ProfileRank is a PageRank [22, 9] inspired algorithm over a directed user-content bipartite graph. Similar to HITS [16], PageRank is a link analysis algorithm designed to measure the importance of a node in a hyperlinked environment, such as the Web. PageRank is based on random walks in graphs, an idea also applied in the computation of relevance scores between nodes [27, 28] and in content recommendation [3]. Nevertheless, our work is the first that applies random walks in a directed user-content bipartite graph to measure user influence and content relevance in diffusion data. We should emphasize that our model is supported by a formulation that is different from those of PageRank and HITS, as detailed in Section 4.2.

3. RELEVANCE AND INFLUENCE BASED ON INFORMATION DIFFUSION DATA

This section provides definitions for the main concepts employed in this work.

3.1 Information Diffusion Data

We call information diffusion data a sequence of occurrences of content. Each occurrence of a piece of content is defined as a tuple in the form $< u, c, t >$, where u is a user from the set of users U, c is a piece of content from the content set C, and t is a timestamp. For a given tuple $< u, c, t >$, we say that the user u propagated c at time t. Therefore, information diffusion data describes associations between users and content across time. Using this notation, we define an information diffusion dataset as a triple $D = (U, C, T)$.

In Figure 1, we give an example of how information diffusion data can be extracted from Twitter, which is a very popular microblogging system integrated to a social network. Twitter's social network is defined by the follower interactions. If a user u_1 follows another user u_2, tweets from u_2 are seen by u_1. In our model, each Twitter user is represented as a user $u \in U$. Moreover, different types of content, such as tweets (i.e., text messages) and URLs, are represented by the content set C. It is important to notice that we do not assume that the follower network is available. Moreover, we are interested in models that do not take textual information into consideration. In our particular example, we consider tweets as content. For instance, we generate the tuple $< user\_0, A, t_0 >$ because user $user\_0$ posted the tweet A at time t_0 and generate the tuple $< user\_1, A, t_2 >$ due to the fact that the same tweet A was retweeted by user $user\_1$ at time t_2. Information diffusion data can be extracted in several other scenarios, specially social media applications, such as blogs and social networks.

3.2 Measuring Content Relevance and User Influence

The problem studied in this paper is measuring content relevance and user influence based on information diffusion data. In this section, we discuss these problems in more detail.

We define the relevance of a piece of content $c \in C$ as a function $r(c)$, which we call a *content relevance function*. In a similar way, we define the influence of a user $u \in U$ as a function $i(u)$, which we call a *user influence function*. Both of these definitions are based on a given information diffusion dataset $D = (U, C, T)$.

A content relevance function simply gives a global relevance value for a given piece of content based on an information diffusion dataset. In a similar fashion, a user influence function gives a global influence value for a given user. In order to illustrate the meaning of the content relevance and user influence functions, let's consider again Twitter as an information diffusion platform. The relevance of a tweet, according to our definition, can be seen as the overall capacity of this tweet to satisfy the users' information needs. Similarly, user influence is a measure of the capacity of a Twitter user to reach the Twitter audience.

3.3 Personalized Relevance and Influence for Content and User Recommendation

In Section 3.2, we gave definitions for a content relevance and a user influence function. We have emphasized that these definitions are global in the sense that they do not provide relevance and influence measures for a particular user, which is the common case in recommendation tasks [23]. We define personalized versions of a content relevance and a user influence function as follows.

Given an information diffusion dataset $D = (U, C, T)$, a *personalized content relevance function* gives the relevance $r(c, u)$ of a piece of content $c \in C$ for a user $u \in U$. In a similar way, a *personalized user influence function* gives the influence $r(u, v)$ of a user $u \in U$ over a user $v \in U$ based on D.

A personalized content relevance function estimates the relevance of a piece of content for a particular user based on information diffusion data. Therefore, using such a function, we may recommend content to users according to its relevance. Given a pair of users (u,v), a personalized user influence function gives u's degree of influence over v. While the personalized content relevance function may support content recommendation, personalized user influence is useful for user recommendation.

An important difference between the personalized content relevance and user influence functions and their global formulations, is that the former are easier to evaluate. In other words, while there is no ground-truth information on globally relevant content and influential users, personalized content relevance and user influence functions can be evaluated on recommendation problems using historic data. In this paper, we evaluate how our information diffusion model performs as a content and user recommender system. This evaluation gives evidence of the effectiveness of our model in providing global content relevance and user influence measures.

4. PROFILERANK: A NEW INFORMATION DIFFUSION MODEL

In this section, we describe ProfileRank, which is a new model for content relevance and user influence based on diffusion data.

4.1 General Principle

ProfileRank is based on an integrated view of user influence and content relevance in information diffusion. It was designed according to the following principle:

A piece of content is relevant if it is created and propagated by

influential users, and a user is influential if she creates relevant content.

As a consequence, information diffusion data enables an elegant circular definition of content relevance and user influence, which resembles ranking algorithms for information retrieval, such as PageRank [22, 9] and HITS [16].

Given an information diffusion dataset, which is a set of associations between users and content across time, how can we assess content relevance and user influence? Answering this question is the main target of this work. User-content associations represent individual relevance evaluations. Otherwise stated, whenever a user propagates a given piece of content, we may assume that this content is somehow relevant to such user. However, leveraging these low-level relevance evaluations to overall relevance and influence measures is a challenging problem.

Our model can be easily described based on the behavior of a random tweeter, following the idea of the *random surfer* usually applied in the description of PageRank. Our random tweeter starts from a random profile and keeps clicking on tweets and retweets at random. The random tweeter is redirected to the profile of the original author of a tweet by clicking on it. The relevance of a tweet is the relative frequency that the random tweeter clicks on a tweet, or one of its retweets. Moreover, the frequency that the random tweeter visits a user's profile is a measure of this user's influence. Figure 1a shows the edges through which our fictitious random tweeter navigates for a illustrative set of tweets and retweets.

The proposed general principle supports the definition of global relevance and influence functions, as described in Section 3.2. However, it is straightforward to reformulate this principle in order to support personalized relevance and influence functions.

A piece of content c is relevant to a user u if it is created and propagated by users that are influential to u and a user v is influential to u if v creates content that is relevant to u.

We can also describe this principle based on the behavior of a random tweeter. Nevertheless, instead of starting the navigation from a random Twitter profile, the random tweeter starts from the profile of the user for which the content relevance and the user influence evaluations are being personalized.

4.2 Information Diffusion Model

ProfileRank is a model for information diffusion that computes user influence and content relevance based on a bipartite directed graph that describes the flow of information among users.

An *information diffusion graph* is a bipartite graph $G(U, C, F, E)$, where U is the user set, C is the content set, and E and F are sets of edges that associate users to content and the other way around, respectively. For each user $u \in U$ and piece of content $c \in C$, there is a directed edge $(u, c) \in E$ if the user u has created or propagated the content c and a directed edge $(c, u) \in F$ if u created c.

Figure 1c presents the bipartite graph built from the data shown in Figure 1b. Edges in E give relevance to content based on the influence of users who propagated it. Moreover, edges in F give influence to users according to the relevance of the content they create. We materialize the two principles described in the last section as random walks through G.

The bipartite graph G can be represented by a user-content matrix M and a content-user matrix L. The matrix $M = (m_{i,j})$ is a $|U| \times |C|$ matrix where $m_{i,j} = 1/q_i$ and q_i is the number of pieces of content the user u_i has created or propagated. Moreover, $L = (l_{i,j})$ is a $|C| \times |U|$ matrix where $l_{i,j} = 1$ if the user u_j created the piece of content c_i and $l_{i,j} = 0$, otherwise. Based on M and L, content relevance and user influence can be defined as:

$$\mathbf{r} = \mathbf{i}M \qquad\qquad \mathbf{i} = \mathbf{r}L$$

where $\mathbf{r}$ is a content relevance vector (i.e., $\mathbf{r}[j]$ is the relevance of the content c_j) and $\mathbf{i}$ is a user influence vector (i.e., $\mathbf{i}[j]$ is the influence of the user u_j). In this definition, we assume that we already have one of the vectors ($\mathbf{r}$ or $\mathbf{i}$) in order to compute the other one, which is not a realistic case. Nevertheless, $\mathbf{r}$ and $\mathbf{i}$ can be computed recursively as follows:

$$\mathbf{r}^{(k)} = \mathbf{r}^{(k-1)}LM \qquad\qquad \mathbf{i}^{(k)} = \mathbf{i}^{(k-1)}ML$$

where $k \geq 0$ and $\mathbf{r}^{(0)}$ and $\mathbf{i}^{(0)}$ are uniform vectors (i.e., vectors where all values are equal and sum to 1).

Like PageRank, this formulation presents two important possible issues: *dangling users* and *buckets*. A dangling user never propagates content from other users. Considering the random tweeter metaphor, the tweeter will get stuck whenever a dangling user u is reached because it will not be able to leave u's profile. We solve this problem by creating an edge (u, c) from every dangling user to a ghost piece of content c and adding an edge (c, u) from the ghost content to every user $u \in U$. In the graph shown in Figure 1c, $user\_0$ is a dangling user, but we do not add a ghost piece of content in this example for clarity.

A bucket is a strongly connected subgraph of the bipartite graph. When the random tweeter reaches a bucket, it is not able to leave it. In order to prevent this problem, we define a damping factor d in our model. This factor determines a small probability that the random tweeter will teleport from the current to a random profile. We add the damping factor d to the definition of $\mathbf{r}$ and $\mathbf{i}$ as follows:

$$\mathbf{r}^{(k)} = d\mathbf{r}^{(k-1)}LM + (1-d)\mathbf{u}$$
$$\mathbf{i}^{(k)} = d\mathbf{i}^{(k-1)}ML + (1-d)\mathbf{u}$$

where $\mathbf{u}$ is a uniform vector. We can reformulate these equations to obtain their exact solutions in a non-recursive fashion:

$$\mathbf{r} = (1-d)\mathbf{u}(I - dLM)^{-1} \quad (1)$$
$$\mathbf{i} = (1-d)\mathbf{u}(I - dML)^{-1} \quad (2)$$

where I is an identity matrix. In the next section, we discuss why this formulation is not computationally efficient. Two more important questions to be addressed immediately are: (1) Do these equations have a solution? and (2) Are these solutions unique?

Equations 1 and 2 have at least one solution because ML and LM are product of stochastic matrices, and thus they are stochastic themselves. The solutions not only exist, but they are unique, once Equations 1 and 2 are written with matrix inverses, and by the definition of a matrix inverse, $\mathbf{r}$ and $\mathbf{i}$ are unique if we assume that a inverse always exists, and it does in this case because $I - dML$ is a diagonally dominant matrix.

In Figure 1c, we give the values of user influence and content relevance computed by ProfileRank using a damping factor of 0.85. The most influential user is $user\_0$ ($i(user\_0) = 0.59$) because the two pieces of content produced by $user\_0$ (A and B) are propagated by two users ($user\_1$ and $user\_2$, respectively). The content produced by $user\_1$ is propagated by a single user ($user\_3$), and thus $user\_1$ is less influential than $user\_0$. The most relevant piece of content is A ($r(A) = 0.39$) because it was created and propagated by influential users ($user\_0$ and $user\_1$).

Devising personalized values of content relevance $r(c, u)$ and user influence $i(u, v)$ using ProfileRank is straightforward. In these scenarios, instead of starting from an arbitrary user, we assume that the random tweeter starts from a specific profile for which the model is being personalized. In the same way, instead of jumping to a random profile with a non-zero probability, the random tweeter

Dataset	content	#users	#pieces of content	#propagations	period	source
TW-CARS	tweet	529,630	369,287	1,368,080	12/31/2011 - 01/31/2012	Twitter
TW-SOCCER	tweet	837,559	3,485,313	958,144	11/19/2010 - 02/11/2011	Twitter
TW-ELECTIONS	tweet	3,860,251	4,067,221	15,844,788	12/27/2011 - 07/31/2012	Twitter
TW-LARGE	tweet	17,069,982	476,553,560	71,835,017	06/01/2009 - 12/31/2009	Twitter
MEME	meme	96,608,034	210,999,824	126,905,936	08/01/2008 - 09/31/2008	MemeTracker

Table 1: Information diffusion datasets.

Dataset	edge	#edges	source
TW-SOCCER	follower-followee	269,217,548	Twitter
TW-LARGE	follower-followee	1,470,000,000	Twitter

Table 2: Network datasets.

always jumps back to the original profile according to the damping factor. This behavior can be induced by substituting the uniform vector $\mathbf{u}$ by a vector $\mathbf{1}_j$, which is a vector with all elements equal to 0, except the position j that is set to 1, where u_j is the user for which the model is being personalized.

ProfileRank computes user influence and content relevance based on a user-content bipartite directed graph, instead of the (non-bipartite) directed graph employed by PageRank and HITS. ProfileRank's graph is represented by two matrices, a user-content (M) and a content-user (L) matrix, and enables the computation of different score functions (influence and relevance) for different types of nodes (users and content) based on diffusion data. Given a diffusion graph G, one may apply traditional Pagerank for computing user and content scores based on a single $|C \cup U| \times |C \cup U|$ matrix. We compare the performance of these two approaches, showing that ProfileRank is more efficient, in Section 5. The main difference between HITS and ProfileRank is that users don't gain influence by propagating content, but only when they are the authors of the content being propagated. Therefore, HITS cannot be applied to compute ProfileRank scores.

4.3 Efficient Solution using the Power Method

In the last section, we described the equations that define user's influence and content relevance in our model. In order to apply this model in real settings, we need to solve such equations efficiently. In real information diffusion data, the matrices M and L are likely to be very large and sparse. Therefore, an efficient solution for our model must take these properties into consideration.

As shown in Equations 1 and 2, we can compute the vectors $\mathbf{r}$ and $\mathbf{i}$ by inverting a $|U|\times|U|$ matrix and a $|C|\times|C|$ matrix, respectively. Since the fastest matrix inversion algorithm known has complexity $O(n^{2.373})$ [31], computing the exact values of $\mathbf{r}$ and $\mathbf{i}$ may not be feasible in real settings. However, the power method [4], which is a fast iteration method to compute the dominant eigenpair of a matrix, can be applied to compute $\mathbf{r}$ and $\mathbf{i}$ efficiently. Besides applying the power method, we make use of sparse representations of the matrices M and L in order to reduce the amount of memory and the execution time required by ProfileRank.

5. EXPERIMENTAL EVALUATION

This section presents an empirical evaluation of ProfileRank. In all the experiments, we set the number of iterations and damping factor of ProfileRank to 10 and 0.85, respectively. We chose this number of iterations since it leads to an 1-norm error of less than 10^{-6} for all datasets described in Section 5.1. Also, when performing the experiments in Sections 5.2 and 5.3, the results converged in about 5 iterations, considering the metrics being used for the results (AUC and BEP). We omitted the results of such experiments due to lack of space. Finally, we provide a Python implementation of ProfileRank as open-source[1].

5.1 Datasets

We gathered 7 real datasets in order to evaluate several aspects of ProfileRank. Most of these datasets were obtained from Twitter [17], which is a popular micro-blogging service integrated to a social network. We considered tweets as content, Twitter users as users, and retweets as content propagation. We also employ a dataset of news phrases over the Web from MemeTracker [20]. In this case, a meme (or phrase) is a piece of content, URLs (i.e., information sources) are represented as users, and propagations result from further occurrences of a given meme on the Web.

Table 1 shows relevant information about our datasets. The TW-CARS, TW-SOCCER, and TW-ELECTIONS datasets are collections of tweets containing terms associated to cars, the Brazilian Soccer Championship, and the presidential elections in the United States, respectively. We crawled these tweets using the Twitter streaming API. The TW-LARGE dataset was obtained from previous work [32] and is estimated to contain about 20-30% of all public tweets published during the collection period. The MEME dataset is the complete set of phrases from MemeTracker. It is important to point out that the methods studied in this experimental section, including ProfileRank, make no use of textual information.

Although ProfileRank makes no use of a social network, some of our evaluations employ this information. Therefore, we obtained two follower networks from Twitter, one for TW-SOCCER and another for TW-LARGE, which are shown in Table 2. In particular, the TW-LARGE network was obtained from a previous work [17].

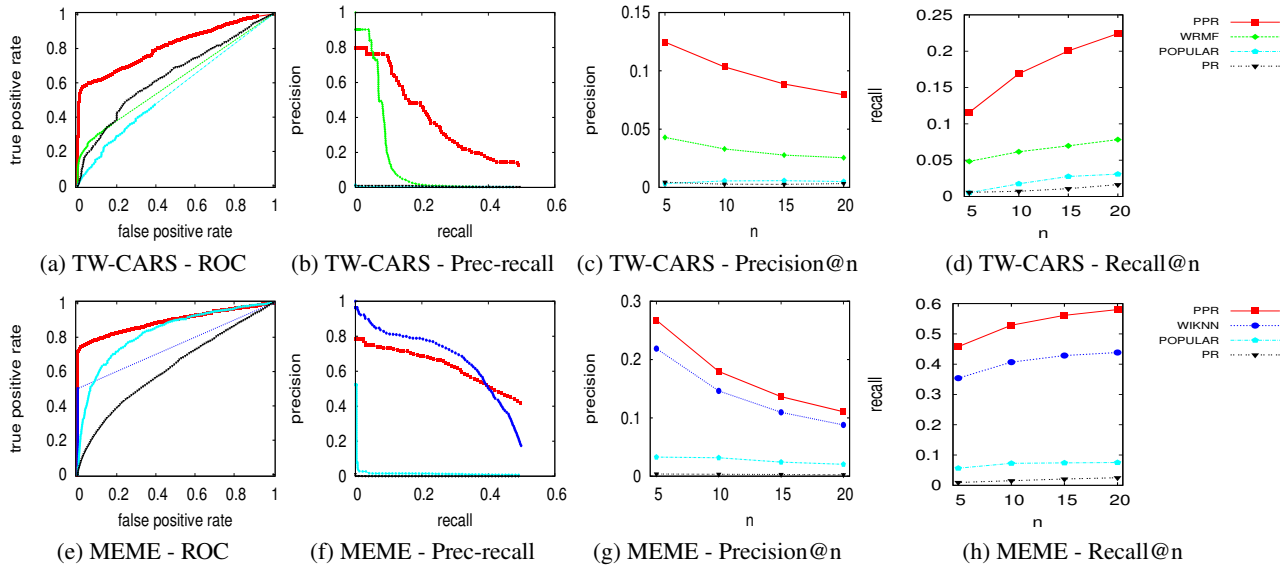
5.2 Content Recommendation

We evaluate the content recommendations given by the personalized (PPR) and the global (PR) versions of ProfileRank using the TW-CARS and MEME datasets. Due to lack of space, the results for the other datasets were omitted. For PPR, we compute content relevance scores for each user using training data and show how these scores are positively correlated with the likelihood that the user will propagate (e.g., retweet) the content in the test data. In the case of PR, we evaluate the same correlation using global, instead of personalized, relevance scores. When considering the global version (PR), the relevance of each piece of content is the overall sum of its relevance to all users. The evaluation metrics applied are the ROC analysis, precision-recall, precision@n, and recall@n. For each user in the test set, we rank the contents according to the relevance scores learned from training and recommend those at the top. A hit occurs whenever a method predicts one of the propagations (e.g., a retweet) in the test data.

We consider a comprehensive set of collaborative filtering techniques implemented by the *MyMediaLite*[2] recommender system library as baselines. The basic idea of these techniques is to rely on the interest similarity between users to recommend new items. Two given users u_1 and u_2 are considered similar if they share common interests (e.g., post the same content) and then the model identifies potential items consumed by u_1 to be recommended to

[1] http://code.google.com/p/profilerank/

[2] http://www.ismll.uni-hildesheim.de/mymedialite/



(a) TW-CARS - ROC (b) TW-CARS - Prec-recall (c) TW-CARS - Precision@n (d) TW-CARS - Recall@n

(e) MEME - ROC (f) MEME - Prec-recall (g) MEME - Precision@n (h) MEME - Recall@n

Figure 2: Content recommendation. Evaluation of Personalized ProfileRank (PPR), a method that recommends the most popular content (POPULAR), global ProfileRank (PR) and the best collaborative filtering methods (WRMF and WIKNN) in terms of ROC, Precision-recall, Precision@n, and Recall@n using TW-CARS and MEME. PPR outperforms the baseline strategies (see Table 3 for more results).

u_2, and vice-versa. Though we applied all the item recommendation techniques available in the library, we will report only the results obtained by the most effective ones, which are WRMF (matrix factorization), WBPRMF (Optimization), WIKNN (K-nearest neighbor based on items), and WUKNN (K-nearest neighbor based on users), and also the technique that recommends the most popular content, which we call POPULAR. We set all the recommendation techniques with their default parameters, including ProfileRank.

For the TW-CARS dataset, we restricted our evaluation to those users who posted at least 5 times and those tweets with at least 1 retweet – users and tweets were removed recursively until these thresholds were met. For each tweet, we split its occurrences (tweet and retweets) into training and test sets, each one containing 50% of the tweet occurrences, considering the occurrence time. Users that did not appear in the training set were removed from the test set. Since we make no use of the following network – but use only information diffusion data – there is no leak of information from the training into the test set. The resulting number of tweets and users in the training dataset were 8,793, and 2,456, respectively. Moreover, the test dataset contains 3,491 tweets and 536 users.

Figures 2a, 2b, 2c, and 2d show the ROC, precision-recall, precision@n, and recall@n curves obtained by Personalized ProfileRank (PPR) and WRMF, which obtained the best results among the techniques evaluated, for the TW-CARS dataset. We also show the results achieved by the global ProfileRank (PR) and the POPULAR technique. Table 3a shows the evaluation of the performance of seven content recommendation techniques on TW-CARS. PPR achieves better results than the baseline methods in terms of all the evaluation metrics considered. As expected, the global version of ProfileRank (PR) does not achieve as good results.

Due to the large scale of the MEME dataset and the cost of generating recommendation models for each user, we decided to limit our analysis to a smaller version of the MEME dataset containing the memes from 08/01/2008 to 08/07/2008. We also restricted the set of users and memes to those with at least 5 memes and 2 occurrences, respectively. For each meme, we split its occurrences into training and test sets, each one containing 50% of the meme occurrences. Users that did not appear in the training set were removed from the test set. The resulting number of memes and users in the training dataset were 208,595, and 66,410, respectively. Moreover, the test dataset contains 26,952 memes and 11,127 users.

Method	AUC	BEP	P@5	P@20	R@5	R@20
PPR	**0.81**	**0.28**	**0.12**	**0.08**	**0.12**	**0.22**
PR	0.64	0.01	0.01	0.01	0.01	0.02
WRMF	0.61	0.11	0.04	0.03	0.05	0.08
WBPRMF	0.58	0.08	0.02	0.01	0.03	0.04
WIKNN	0.57	0.13	0.05	0.03	0.05	0.09
WUKNN	0.57	0.13	0.05	0.03	0.05	0.09
POPULAR	0.55	0.01	0.01	0.01	0.01	0.03

(a) TW-CARS

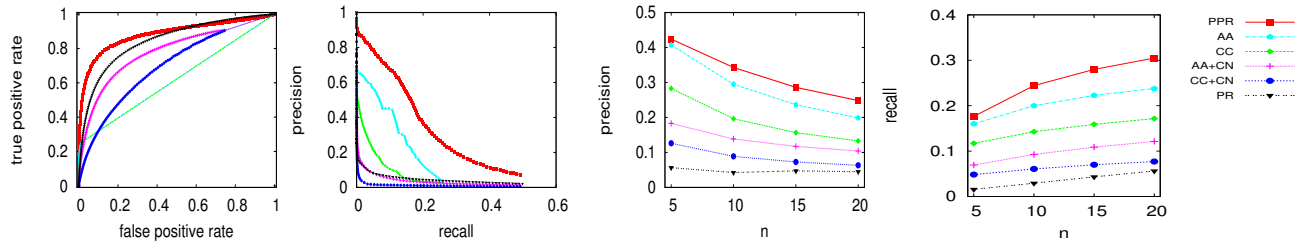
Method	AUC	BEP	P@5	P@20	R@5	R@20
PPR	**0.89**	**0.46**	**0.27**	**0.11**	**0.46**	**0.58**
WIKNN	0.75	0.43	0.22	0.09	0.35	0.44
WUKNN	0.75	0.38	0.21	0.09	0.35	0.44
WBRMF	0.71	0.09	0.04	0.02	0.07	0.13
WRMF	0.71	0.05	0.01	0.01	0.01	0.01
POPULAR	0.65	0.01	0.01	0.01	0.01	0.02
PR	0.62	0.01	0.01	0.01	0.01	0.02

(b) MEME

Table 3: Evaluation of the content recommendation methods in terms of Area Under the ROC Curve (AUC), Precision-recall Breakeven Point (BEP), Precision at n (P@n), and Recall at n (R@n) using the TW-CARS and MEME datasets. Personalized ProfileRank (PPR) outperforms all the baseline techniques.

Content recommendation results of the best methods (PPR and WIKNN) and also for POPULAR and PR on the MEME dataset are shown in Figures 2e, 2f, 2g, and 2h. Results for these and other methods are presented in Table 3b. Again, PPR outperforms the baseline techniques in terms of all evaluation metrics considered.

The results presented in this section show that the Personalized ProfileRank (PPR) provides effective content recommendations. In fact, there is strong evidence that ProfileRank handles the sparsity of information diffusion data better than the collaborative filtering approaches. According to the precision-recall curves (Figures 2b



(a) TW-SOCCER - ROC (b) TW-SOCCER - Prec-recall (c) TW-SOCCER - Precision@n (d) TW-SOCCER - Recall@n

Figure 3: User recommendation. Evaluation of Personalized ProfileRank (PPR), Adamic-Adar (AA), Common content (CC), Adamic-Adar with Common Neighbors (AA+CN), Common Content with Common Neighbors (CC+CN), and global ProfileRank (PR) in terms of ROC, Precision-recall, Precision@n, and Recall@n using TW-SOCCER. PPR outperforms the baseline techniques (see Table 4 for more results).

and 2f), collaborative filtering techniques outperform PPR for top global recommendations, which correspond to recommendations of popular content to very active users. Nevertheless, since ProfileRank is based on diffusion data, and not user similarity, it is able to recommend content with higher overall accuracy, even for less active users and not so popular content.

5.3 User Recommendation

In this section, we evaluate the performance of ProfileRank (PR and PPR) in user recommendation using TW-SOCCER. Again, due to lack of space, we omitted the other datasets. We compute personalized (PPR) and global (PR) influence scores for pairs of users based on information diffusion data and evaluate how these scores are positively correlated with the likelihood of the existence of a follower interaction between them. For each user in the test set, we rank the other users according to the influence scores learned from training and recommend those at the top. A hit occurs whenever a method predicts a follower edge from the network. Again, for the global version (PR), the influence of each user is the overall sum of her influence on all user. We compare ProfileRank against a set of cold start link prediction techniques [18]. In particular, we implemented two strategies that compute the likelihood of an influence edge as a function of the number of tweets shared (CC) and also as the Adamic-Adar score (AA) between pairs of users [1]. Moreover, we also extend these simple strategies using a probabilistic common neighbors measure, as done by a previous work [18]. We call CC+CN and AA+CN, the link prediction strategies that combine the common neighbors measure with the number of tweets shared (CC) and the Adamic-Adar score (AA), respectively. Though cold start link prediction techniques do not consider time information (i.e., the order in which the content appears), we found that this information does not improve their performance.

We restricted our evaluation to those users who posted at least 10 times and to those tweets with at least 1 retweet. The resulting set of 101,688 tweets from 14,844 users was given as training data to the user recommendation methods. Again, there is no information leak from training into the testing set, since we use the follower network only and solely for evaluating the results. Figures 3a, 3b, 3c, and 3d show the ROC, precision-recall, precision@n, and recall@n curves obtained by all the user recommendation techniques for the TW-SOCCER dataset. Table 4 shows more detailed results for the user recommendation task. PPR achieves better results than the baseline methods w.r.t. the evaluation metrics. Similarly to what we found in content recommendation, the global version of ProfileRank achieves worse results, showing that global influence is not a good predictor for influence edges between users.

Content diffusion information is highly associated to the underlying influence network in Twitter. However, while it is known that many following relationships on Twitter do not lead to retweets [5], ProfileRank is able to identify only those relationships that induce information diffusion on the network.

Method	AUC	BEP	P@5	P@20	R@5	R@20
PPR	**0.88**	**0.25**	**0.42**	**0.25**	**0.18**	**0.30**
PR	0.84	0.07	0.06	0.05	0.02	0.06
AA+CN	0.78	0.06	0.18	0.10	0.07	0.12
CC+CN	0.70	0.02	0.13	0.06	0.05	0.08
AA	0.62	0.17	0.41	0.20	0.16	0.24
CC	0.61	0.10	0.28	0.13	0.12	0.17

Table 4: Evaluation of the user recommendation methods in terms of Area Under the ROC Curve (AUC), Precision-recall Breakeven Point (BEP), Precision at n (P@n), and Recall at n (R@n) using the TW-SOCCER dataset. Personalized ProfileRank (PPR) outperforms the baseline techniques.

5.4 Relevant Content and Influential Users

As described in Section 4.2, ProfileRank computes both personalized and global user influence and content relevance based on a circular definition. Due to the absence of ground truth information on globally relevant and influential users, recommendation tasks are an interesting alternative scenario for the quantitative evaluation of ProfileRank. This section presents a qualitative analysis of the results given by ProfileRank using the TW-ELECTIONS and TW-LARGE datasets. We also compare ProfileRank with other methods for measuring user influence and content relevance.

Tables 5a and 5b show the top relevant tweets and influential users from TW-ELECTIONS. Most of the relevant tweets are from two members (Liam Payne and Josh Devine) of a pop band called One Direction that make reference to Barack Obama. These tweets have attracted more interest from the Twitter community than posts from the candidates themselves. Moreover, two tweets about Obama's support for gay marriage were identified as relevant by ProfileRank. Besides the presidential candidates, top influential users include two profiles associated to Obama's campaign, three profiles that share comedy content about the elections, and the One Direction leading singer's (Liam Payne) profile.

The results from TW-LARGE dataset are shown in Tables 5c and 5d. Top relevant tweets are viral marketing campaigns. Many people are interested in posting this kind of content in order to be eligible for receiving the offered gift. The top relevant tweet, for instance, corresponds to a promotional tweet created by Spoofcard. Top influentials are users that produce online social engagement, such as news media websites, artists, and comedians. It is interesting to notice that only one of the top 10 users (joinred) authored one of the top 5 tweets.

content	description
@BarackObama hi mr Obama have you got up all night yet?	Message from Liam Payne to Barack Obama
That was one of the strangest days ever will smith taylor swift justin bieber michelle obama wow what it going on with my life!!	Liam Payne about the 2012 Kid's Choice Award
Obama, congratulations on being the first sitting President to support marriage equality. Feels like the future, and not the past. #NoFear	Lady Gaga about Obama's support for gay marriage
"Same-sex couples should be able to get married."– President Obama	Obama about his support for gay marriage
Summertime with @NiallOfficial and @BarackObama! http://t.co/KNnWnfz7	Josh Devine about a picture including a Obama's statue

(a) TW-ELECTIONS - Relevant content

user	description
BarackObama	US President and Democrat candidate
Obama2012	Obama's campaign
UberFacts	Comedy facts
BorowitzReport	Comedy news
StephenAtHome	Comedian
truthteam2012	Obama's campaign
Real_Liam_Payne	Pop singer
MittRomney	Republican candidate
thinkprogress	Political blog
realDonaldTrump	Businessman

(b) TW-ELECTIONS - Influential users

content	description
Send one tweet and get a free #SpoofCard to spoof your Caller ID, change your voice & record calls	Advertising
Regalamos 1000 dominios .com.mx a 1000 de nuestros followers, mas detalles en http://is.gd/2OplD	Advertising
District Lines is giving away 30 free shirts, contest ends midnight 8/27, ENTER HERE http://district	Advertising
It's World AIDS Day. Turn Twitter (RED) - literally! Use #red or #laceupsavelives & turn tweets th...	Campaign
we got some google wave invites... you need one? RT this !! #googlewave #wave	Sharing

(c) TW-LARGE - Relevant content

user	description
mashable	News website and blog
lilduval	Comedian
smashingmag	Magazine
justinbieber	Pop singer
johncmayer	Pop Singer
iamdiddy	Rapper
joinred	Initiative against HIV
shitmydadsays	Comedian
paulocoelho	Writer
myfabolouslife	Rapper

(d) TW-LARGE - Influential users

Table 5: Top relevant content and influential users discovered from TW-ELECTIONS and TW-LARGE using ProfileRank.

	ProfileRank	PageRank	#propag.	#followers
ProfileRank	-	n/a	0.89	n/a
PageRank	0.28	-	n/a	n/a
#propag.	0.81	0.30	-	n/a
#followers	0.29	0.81	0.32	-

(a) User metrics

	ProfileRank	#content propag.	PageRank	#user propag.	#followers
ProfileRank	-	0.36	n/a	0.42	n/a
#content propag.	0.22	-	n/a	0.44	n/a
PageRank	0.26	-0.02	-	n/a	n/a
#user propag.	0.27	0.11	0.42	-	n/a
#followers	0.25	-0.01	0.83	0.45	-

(b) Content metrics

Table 6: Pairwise ranking correlations among ProfileRank, PageRank, number of propagations for users and content and number of followers. Results for TW-LARGE are on the lower triangular matrix (in blue) and results for MEME are on the upper triangular matrix (in red). PageRank and number of followers are not applicable (n/a) to MEME due to the lack of an explicit network.

Using the TW-LARGE and MEME datasets, we compare ProfileRank with the following metrics: (1) user's PageRank in the Twitter following network, (2) user's number of followers, (3) number of propagations of a user's content, (4) number of propagations of a content. A propagation corresponds to a retweet in TW-LARGE and a new occurrence of a phrase in MEME. A user's content propagations are aggregated as this user's propagations. Kendall-τ rank correlation coefficient, which varies from -1 to 1, was used to assess the level of agreement between the scores given by ProfileRank and the other measures. Tables 6a and 6b show the pairwise correlation between the metrics for user influence and content relevance, respectively. Some measures are not applicable (n/a) to MEME because, different from Twitter, it has no explicit network.

Correlation results show that user influence computed by ProfileRank is strongly correlated with the users' number of propagations. On the other hand, content relevance is more correlated with the number of user's than content propagations, as a consequence of the impact of the author over his content. It is interesting to notice that both the Common Content (CC) and the POPULAR strategy for user and content recommendation, respectively, which apply propagation information, are outperformed by ProfileRank. Neither content relevance nor user influence scores agree with the user' PageRank or number of followers, which are based on the Twitter network.

Our results show that information diffusion supports effective user influence and content relevance assessments. The proposed model can be further extended in order to incorporate specific types of content information (e.g., text, images). Regarding the social networks that support the information diffusion, we have shown that by tracing content itself, we can identify influence relationships in these networks accurately. Moreover, social interactions that are not associated to influence do not provide evidence for user influence and content relevance in information diffusion platforms.

5.5 Running Time Evaluation

In this section, we evaluate the running time of ProfileRank. As discussed in Section 4.2, influence and relevance scores can be computed as the PageRank of the users and content, respectively, in the information diffusion graph. However, because ProfileRank computes these scores using two matrices, user-content and content-user, instead of a single large matrix that combines both users and content, it outperforms the strategy based on PageRank significantly (3 times faster on average), as shown in Table 7.

Dataset	ProfileRank	PageRank
TW-CARS	3.85	10.04
TW-SOCCER	39.32	133.55
TW-ELECTIONS	5.28	9.20
TW-LARGE	17.74	59.33
MEME	1.23	3.86

Table 7: Running time (in seconds).

6. CONCLUSIONS

In this paper, we introduced ProfileRank, an information diffusion model that measures content relevance and user influence based on random walks over a user-content bipartite graph. The basic principle exploited by ProfileRank is that relevant content is created and propagated by influential users and influential users create relevant content. This principle supports the formulation of an algorithm that resembles traditional ranking algorithms for hyperlinked environments, such as PageRank [22] and HITS [16].

We have evaluated ProfileRank as a content and user recommender system using data from Twitter and MemeTracker. The results have shown that ProfileRank is able to make accurate content and user recommendations, outperforming all the baselines considered in the experiments. Furthermore, we have applied ProfileRank in the identification of relevant content and influential users from Twitter and MemeTracker. Based on an analysis of the results, we showed that relevant content and influential users discovered by ProfileRank provide valuable knowledge in the analysis of information diffusion. We also investigated the level of agreement between ProfileRank scores and other metrics, showing that user influence given by ProfileRank is strongly correlated with the user's number of propagations. On the other hand, ProfileRank evaluations for both content and users are not correlated with metrics based on network data. Finally, we showed that ProfileRank is computationally efficient, outperforming an alternate strategy that computes relevance and influence scores as the PageRank of vertices in an information diffusion graph.

This work opens several promising directions for future research. Given the parallel between ProfileRank and PageRank, it would be also interesting to combine ProfileRank with a filtering approach in the development of a search engine for Twitter. In addition, we want to incorporate the temporal dynamics of content propagation into ProfileRank in order to provide updated user influence and content relevance measures. Another research direction would be incorporating textual and network information, in order to better measure each user's influence. Finally, due to the large scale and dynamic nature of information diffusion data, the use of MapReduce and fast update approaches [28] would enable the application of ProfileRank to terabyte-scale data.

7. REFERENCES

[1] L. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[2] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*, volume 82. Addison-Wesley New York, 1999.

[3] S. Baluja, R. Seth, D. Sivakumar, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video suggestion and discovery for youtube: taking random walks through the view graph. In *WWW*, 2008.

[4] M. Bianchini, M. Gori, and F. Scarselli. Inside pagerank. *ACM Trans. Internet Technol.*, 5(1):92–128, 2005.

[5] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi. Measuring User Influence in Twitter: The Million Follower Fallacy. In *ICWSM*, 2010.

[6] M. Cha, A. Mislove, and K. P. Gummadi. A measurement-driven analysis of information propagation in the flickr social network. In *WWW*, 2009.

[7] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. Chi. Short and tweet: experiments on recommending content from information streams. In *CHI*, 2010.

[8] M. De Choudhury, S. Counts, and M. Czerwinski. Identifying relevant social media content: leveraging information diversity and user cognition. In *HT*, 2011.

[9] M. Franceschet. Pagerank: standing on the shoulders of giants. *Commun. ACM*, 54(6):92–101, June 2011.

[10] N. Friedkin. *A structural theory of social influence*, volume 13. Cambridge University Press, 2006.

[11] M. Gomez Rodriguez, J. Leskovec, and A. Krause. Inferring networks of diffusion and influence. In *SIGKDD*, 2010.

[12] A. Goyal, F. Bonchi, and L. V. Lakshmanan. Learning influence probabilities in social networks. In *WSDM*, 2010.

[13] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *WWW*, 2004.

[14] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *RecSys*, 2010.

[15] Y. Kim and K. Shim. Twitobi: A recommendation system for twitter using probabilistic modeling. *ICDM*, 2011.

[16] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. In *SODA*, 1998.

[17] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *WWW*, 2010.

[18] V. Leroy, B. B. Cambazoglu, and F. Bonchi. Cold start link prediction. In *SIGKDD*, 2010.

[19] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1):5, 2007.

[20] J. Leskovec, L. Backstrom, and J. Kleinberg. Meme-tracking and the dynamics of the news cycle. In *SIGKDD*, 2009.

[21] D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM*, 2003.

[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, 1999.

[23] F. Ricci, L. Rokach, and B. Shapira. Introduction to recommender systems handbook. *Recommender Systems Handbook*, 2011.

[24] D. M. Romero, W. Galuba, S. Asur, and B. A. Huberman. Influence and passivity in social media. In *PKDD*, 2011.

[25] A. Silva, H. Valiati, S. Guimarães, and W. Meira Jr. From individual behavior to influence networks: A case study on twitter. In *Webmedia*, 2011.

[26] J. Tang, J. Sun, C. Wang, and Z. Yang. Social influence analysis in large-scale networks. In *KDD*, 2009.

[27] H. Tong, C. Faloutsos, and J.-Y. Pan. Fast random walk with restart and its applications. In *ICDM*, 2006.

[28] H. Tong, S. Papadimitriou, P. S. Yu, and C. Faloutsos. Proximity tracking on time-evolving bipartite graphs. In *SDM*, 2008.

[29] D. Watts. A twenty-first century science. *Nature*, 445(7127), 2007.

[30] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, 2010.

[31] V. V. Williams. Multiplying matrices faster than coppersmith-winograd. In *STOC*, 2012.

[32] J. Yang and J. Leskovec. Patterns of temporal variation in online media. In *WSDM*, 2011.