From Individual Behavior to Influence Networks: A Case Study on Twitter *

Arlei Silva, Hérico Valiati, Sara Guimarães, Wagner Meira Jr.
Universidade Federal de Minas Gerais
Computer Science Department
Belo Horizonte, Brazil
{arlei,herico,sara,meira}@dcc.ufmg.br

ABSTRACT

Understanding social influence and its related phenomena is a major challenge in the study of the human collective behavior. In the recent years, the availability of internet-based communication and interactivity data has enabled studies on social influence at an unprecedented scale and time resolution. In this work, we study how individual behavior data may provide knowledge regarding influence relationships in a social network. We define what we call the influence network discovery problem, which consists of identifying influence relationships based on user behavior across time. Our objective is the design of accurate models that are able to exploit different types of behavior in order to discover how people influence each other. Several strategies for influence network discovery are proposed and discussed. Moreover, we present a case study on the application of such strategies using a follower-followee network and user activity data from Twitter, which is a popular microblogging and social networking service. We consider that a follower-followee interaction defines a potential influence relationship between users and the act of posting a tweet, a URL or a hashtag represents an individual behavior on Twitter. The results show that, while tweets may be used effectively in the discovery of influence relationships, hashtags and URLs do not lead to good performance in such task. Moreover, strategies that consider the time when an individual behavior is observed outperform those that do not and by combining such information with the popularity of the behaviors, even better results may be achieved.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Information Systems Applications—*Miscellaneous*

Keywords

Social Influence, Social Network Analysis, Link Prediction, Twitter

1. INTRODUCTION

*This work was partially supported by CNPq, CAPES, FINEP, FAPEMIG, and InWeb.

Social influence has been argued to play an important role in our society. The basic principle of social influence is that people can affect each other through social ties. Therefore, someone's decisions are often driven by the opinion or influence of people from his/her social circle. Previous work has found that 2/3 of the economy of the United States is supported by the "word-of-mouth" dissemination process [9]. Moreover, from voting to purchasing decisions, from information propagation to innovation diffusion, social influence has been an intriguing matter and important research topic in social sciences, psychology, and, more recently, computer science. Understanding social influence and its related phenomena is a major challenge in the study of the human collective behavior. From a computer science perspective, social influence is a key aspect in the design of effective recommendation systems, viral marketing/advertising strategies, and information diffusion mechanisms.

An important challenge in the study of influence and information diffusion in social networks is the lack of data at a large enough scale. Most of the social network datasets available for research contain just static topological information (i.e., persons and relationships) and do not contain key information for analyzing influence and information diffusion. Further, social influence analysis requires temporal information that indicates, for instance, eventual association of persons to information items. As a consequence of such scarcity, a significant part of the existing models and analysis of social influence are based on synthetic data, which is frequently based on epidemiological models. Nevertheless, previous work has shown that such assumption may not be appropriate[12]. However, this scenario changed in the recent years, as a consequence of the availability of internet-based communication and interactivity data, resulting in studies on social influence on an unprecedented scale and time resolution. Blogs, news media websites, viral marketing campaigns, photo and video sharing services, and online social networks in general have provided rich datasets that supported several interesting findings regarding social influence and information diffusion in real scenarios.

In this work, we make use of a large scale dataset in order to study how individual behavior data can provide knowledge regarding influence relationships in a social network. Table 1 illustrates what we call individual behavior data. For a set of individuals ($u1$-$u5$) we define a set of possible behaviors ($b1$-$b4$). Moreover, we also define a set of timestamps t_1-t_4. Each individual may express some of the defined behaviors in the time interval considered. The individual $u1$ acts as $b1$ at time t_1, for example. Examples of behaviors in real settings include purchasing a given product, expressing an opinion, or posting a comment in a blog.

user	behavior	time
$u1$	$b1$	t_1
$u1$	$b2$	t_2
$u2$	$b1$	t_2
$u2$	$b2$	t_2
$u2$	$b3$	t_2
$u1$	$b3$	t_3
$u3$	$b2$	t_3
$u3$	$b1$	t_3
$u3$	$b3$	t_3
$u4$	$b3$	t_3
$u1$	$b4$	t_4
$u5$	$b1$	t_4
$u5$	$b4$	t_4

Table 1: Example of individual behavior data

This paper studies how influence relationships can be discovered based on such individual behavior information in social networks. Based on the well-accepted hypothesis that social influence affects individual behavior significantly, we track what people do over time in order to learn implicit influence ties among them. Our objective is the design of accurate models that are able to exploit different types of behavior in order to discover how people influence each other.

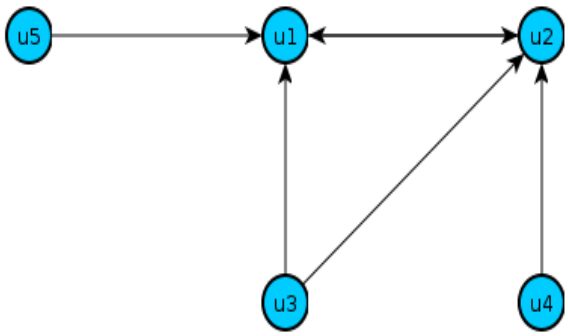


Figure 1: Example of an influence network

An influence relationship defines whether the behavior of a given individual is affected by the behavior of another one. These relationships integrate what we call an influence network, which may be represented by a directed graph $G(V, E)$ where each vertex $v \in V$ corresponds to an individual and each edge $(v_i, v_j) \in E$ represents an influence relationship from v_j to v_i. Figure 1 shows an example of an influence network involving the individuals in the illustrative individual behavior dataset from Table 1. In such network, individuals $u5$, $u3$, and $u2$ are influenced by $u1$. The relationship between $u1$ and $u2$ is of mutual influence.

This work focuses on the problem of inferring how individuals influence each other, as shown in the example from Figure 1, using individual behavior data, such as in the dataset from Table 1. Formally, given two individuals u_i and u_j, lets consider an inference model that considers how many times u_j presented the same behavior of u_i in a subsequent instant of time in order to evaluate whether the behaviors of u_i might influence the behaviors of u_j. The individual $u3$, for example, repeated the behavior of $u1$ twice ($b1$ and $b2$). On the other hand, $u5$ never behaved like $u3$. Therefore, such model would infer that $u3$ is influenced by $u1$ and $u5$ is not influenced by $u3$. Along this paper, we study models for the discovery of influence relationships based on individual behavior information.

We present a case study using influence network and user behavior data from *Twitter*. Twitter is a popular microblogging and social networking service. In Twitter, a user represents an individual and influence is expressed, in our approach, through the relationships associated with the act of posting a particular content, which can be a tweet, a URL, or a hashtag. Our analysis of such scenario reached some interesting conclusions that include:

- Tweets are discriminative in the discovery of influence relationships. However, the use of URLs and hashtags does not achieve good performance in such task.
- The order in which the individuals perform a given behavior is a relevant information of influence network discovery.
- Inter-posting time is more useful than co-occurrence information in the identification of influence edges. However, when the order in which the behaviors are expressed is known, strategies that consider co-occurrence are more effective than those that consider only inter-activity time information.

The remaining of this paper is organized as follows. Section 2 discusses some topics related to the the problem of discovering influence relationships from user behavior data. In Section 3, we characterize important aspects of the dataset used in this work. In Section 4, we present several strategies for identifying influence relationships based on user behavior along time. Such strategies are evaluated in a case study using data from Twitter in Section 5. The work finishes with the conclusions and future work in Section 6.

2. RELATED WORK

In this section, we discuss related work on influence and information propagation in social networks, and link prediction, which are research topics related to this work.

2.1 Influence and Information Propagation in Social Networks

Social influence and information propagation have become popular research topics in Computer Science in the recent years. We divide the existing work on influence and information propagation into three main groups: (1) understanding influence, (2) characterizing influence and information propagation in real networks, and (3) detecting influentials.

Understanding influence: Few people would argue that social influence does not exist or does not affect our society. However, there is no consensus regarding how influence occurs and may be measured. A significative effort has been made on distinguishing social influence from other phenomena, such as homophily and confounding [1, 2, 14]. Moreover, while most of the research community has supported the idea of the existence of a minority of influentials with the power to determine the decisions of social networks, others have argued that the role played by influentials should be studied more carefully [4, 26].

Characterization: Previous work has characterized social influence and information propagation (a.k.a., information diffusion)

in several real-life scenarios such as blogs [18], viral marketing campaigns [16, 24, 12], knowledge-sharing sites [23, 12], photo-sharing services [7], and Twitter [6]. Despite the absence of a standard methodology for measuring social influence and information propagation, such studies have discovered interesting empirical evidences regarding how individuals affect each other and how information spreads in social networks.

Detecting influentials and maximizing influence: Selecting a subset of nodes that are able to maximize the influence in a social network [13, 8] is an optimization problem that has attracted the interest of the research community in the recent years. A related problem is the selection of nodes in order to detect outbreaks in networks [17]. In [5] and [27], the authors propose techniques for ranking users in terms of their influence on Twitter.

2.2 Link Prediction

Link prediction is a relational learning problem presented in [19]. It consists of predicting links that will be added to the network in the future, based on a current snapshot of the network. The basic idea is to compute simple proximity metrics, such as distance and number of common friends, for pairs of nodes in order to predict new links. More recent approaches [25, 20, 3] proposed machine learning techniques for link prediction. Such approaches assume the availability of part of the network in order to predict new links.

The problem of discovering influence relationships based on user behavior data can be seen as a particular case of the classical link prediction task. However, different from most of the link prediction methods we do not assume the availability of any initial information regarding the structure of the influence network.

In [15], the authors introduce the *cold start link prediction problem*, which is a version of the link prediction problem when the network is totally missing. The idea is to consider information about the nodes in order to predict the network structure. Such problem is very close to the influence network discovery problem studied in this work. However, while they consider friendship relationships, which are undirected, we consider directed influence relationships.

The problem of inferring networks of diffusion and influence was proposed by [10]. The authors present an approximate algorithm for the identification of the optimal network that explains individual transmissions over the set of nodes. This work has a different focus, which is understanding how different types of individual activity across time may be used as evidences in order to discover influence relationships. Moreover, while we present a case study using data from Twitter, they have applied their technique to infer how information flows through media sites and blogs.

In [11], the authors propose models for computing influence probabilities based on individual activity and social network data. Such influence probabilities measure how the actions of a user will be influenced by the actions of its neighbohrs. They applied the proposed models to a social graph from Flickr, achieving good results in the prediction whether and when individuals perform a given action. This paper considers a different problem in which there is no social network data available. Moreover, instead of computing influence probabilities, our problem consists of identifying influence relationships in a social network.

3. TWITTER DATA CHARACTERIZATION

#tweets	2,956,941
#users	318,627
#URLs	508,538
#hashtags	430,084
#edges	269,217,548

Table 2: Dataset description

This paper studies how individual behavior data can be employed in the discovery of influence relationships in social networks. We use a dataset crawled from Twitter[1]. The dataset is a sample of tweets related to the Brazilian Soccer Championship, which is one of the biggest sport events in Brazil.

Twitter is a social network and microblogging website. In the recent years, Twitter has been extensively used in the study of social influence and information propagation in social networks. The next sections give an overview of the individual behavior and the influence network data employed in this work.

3.1 Individual Behavior

We call individual behavior data any information that represents particular individual activities performed across time. Examples of individual behavior include buying a given product, voting for a particular candidate, or practicing a given sport activity. Table 1 shows an illustrative example of individual behavior data where individuals $u1$-$u5$ can act as $b1$-$b4$. The availability of internet-based communication and interaction data in the recent years enables the analysis of rich individual behavior data in large scale. Examples of online individual behavior include buying a particular product in an e-market (e.g., Amazon, eBay), watching or uploading a specific video through video-sharing website (e.g., Youtube, Vimeo), posting or commenting about a particular topic on a blog website (e.g., Blogspot, Blogger), and exchanging a message regarding a particular subject in an online social network (e.g., Facebook, MySpace).

This work uses individual behavior data crawled from Twitter. We consider three types of content generated by users: tweets, URLs, and hashtags. We distinguish expanded from non-expanded URLs in our study, since compressing URLs is a common practice in order to generate short messages in Twitter. Table 2 shows important statistics about our dataset. It contains a signficative amount of users, tweets, hashtags, URLs, and edges. Figure 2 presents the popularity, activity, and inter-posting time distributions of tweets, non-expanded URLs, expanded URLs, and hashtags in our dataset.

The popularity distributions of the contents in our dataset (Figures 2a, 2b, 2c, and 2d) are heavy-tailed, as expected. The shape of the distributions is characteristic of power-law distributions. Most of the content shared reach few users and only a small fraction of the tweets, URLs and hashtags get popular. The top popularity tweet, non-expanded URL, expanded URL, and hashtag, have reached 4,757, 3,906, 14,996, and 8,268 users, respectively. Along this paper we investigate the impact of the frequency of a given behavior over its power in predicting influence relationships.

The activity distributions (Figures 2e, 2f, 2g, and 2h) are also heavy-tailed. In general, most of the users post few tweets, URLs and hashtags. On the other hand few users have posted more than 1,000

[1] http://twitter.com

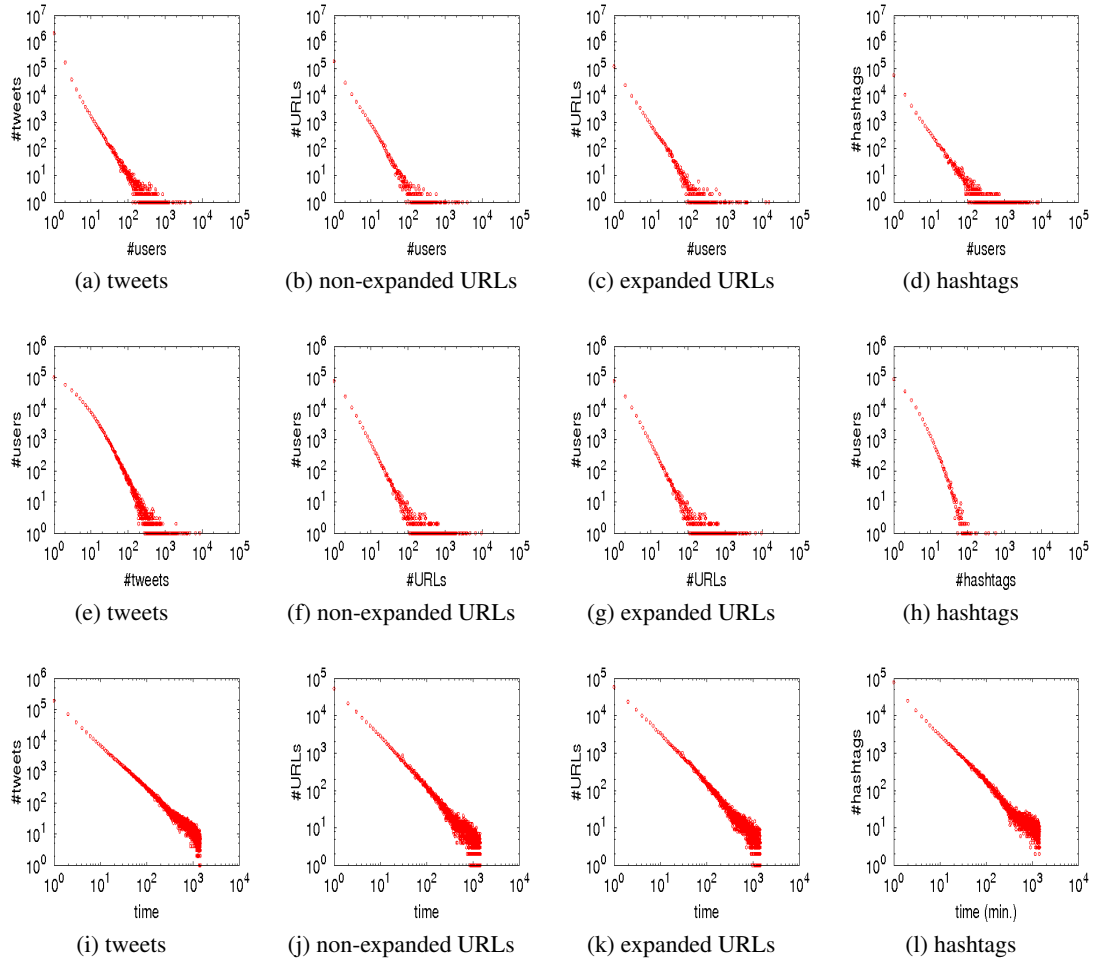


Figure 2: Popularity, activity, and inter-posting time distributions of tweets, URLs and hashtags in Twitter

tweets, URLs, and hashtags in our dataset. Since this paper focuses on the problem of discovering influence relationships based on user activity, we are specially interested in active users, which may produce significative evidences regarding the influence network.

Figures 2i, 2j, 2k, and 2l show the inter-posting time distributions for tweets, non-expanded URLs, expanded URLs, and hashtags, respectively. Inter-posting time also presents heavy-tailed distributions, what means that sequential posts of a given tweet, URL or hashtag are usually submitted in a short period of time. Moreover, few tweets, URLs and hashtags present long inter-posting time. In Section 4, we propose strategies to consider the inter-activity time in the discovery of influence relationships.

Along this section, we have presented the individual behavior data used in this work, which is a set of tweets, URLs and hashtags from Twitter. In the next section, we describe the influence network data that is based on follower-followee interactions in Twitter.

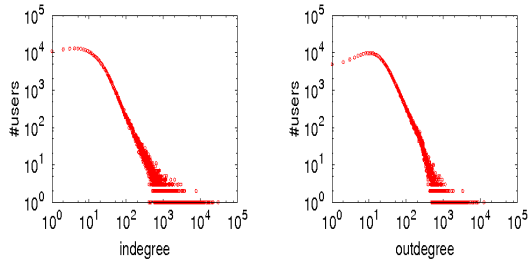
3.2 Influence Network

This section characterizes the influence network employed in this work, which is composed by follower-followee interactions in Twitter. If a user u_i follows a user u_j, u_i will receive all the posts of u_j on his or her Twitter page. Therefore, a follower is potentially influenced by his or her followee. Twitter is an interesting mechanism for the spread of information between its users, enabling online influence relationships among them.

In this paper, we employ a follower-followee network, composed of 318,627 users and 269,217,548 edges, crawled from Twitter. The users in the network are those who have posted, at least, one tweet from the user behavior dataset described in the last section. Figure 3 show the indegree and outdegree distributions from our network.

The indegree of a user in our influence network corresponds to the number of followers he/she has. Figure 3a shows that the indegree may be characterized by a power-law distribution. Most of the users have few followers and only few of them are followed by many. Power-law degree distributions occur in several real-life networks and Twitter was not expected to be different [21]. The outdegree of a user represents the number of followees it has. Similarly to the indegree distribution, the outdegree distribution also present a slope characteristic of power-law distributions. The top indegree and outdegree users in our dataset have a degree of 30,146 and 12,549, respectively.

Since follower-followee relationships are asymetric, it is interesting to check the reciprocity rate of such relationships in Twitter.



(a) indegree (b) outdegree

Figure 3: Degree distributions

The reciprocity rate is the fraction of edges (u_i, u_j) for which there is a reciprocal edge (u_j, u_i). We found a reciprocity rate of 41% in our dataset. While some of the strategies for influence network discovery studied in this paper do not take the order in which the behaviors occur, generating the same scores for reciprocal edges, we also study how temporal information may be useful in the identification of the direction of influence relationships.

In the next section we present the influence network discovery problem, which consists of identifying influence relationships based on individual behaviors across time.

4. INFLUENCE NETWORK DISCOVERY

In this section we study the influence network discovery problem. First, in Section 4.1,we give a formal definition for the problem of discovering an influence network from individual behavior data. Section 4.2 defines a general framework for influence network discovery. Such framework may apply different scoring functions, which are described in Section 4.3

4.1 Problem Formulation

The influence network discovery problem can be seen as a particular case of the traditional link prediction problem. The objective is to identify influence relationships based on individual behavior across time. The set of influence relationships composes what we call an influence network.

DEFINITION 1. **Influence Network Discovery Problem:** Given a set of individuals $\mathcal{I}$, a set of possible behaviors $\mathcal{B}$, and a function $\mathcal{T} : B \times \mathcal{I} \rightarrow T$ defined as follows:

$$\mathcal{T}(u_i, b_j) = \begin{cases} 0, & \text{if } u_i \text{ has never behaved as } b_j \\ t, & \text{if } u_i \text{ behaved as } b_j \text{ in } t(t > 0) \end{cases}$$

The problem consists of discovering a set of influence relationships $\mathcal{E}$, such that $\mathcal{E} = \{(u_i, u_j) | u_i \in \mathcal{I} \wedge u_j \in \mathcal{I}\}$.

We may evaluate a solution for the influence network discovery problem by comparing the influence network discovered against ground truth information about influence relationships. Based on the problem definition presented in this section, the next section defines a general framework for influence network discovery.

4.2 General Framework

Algorithm 1 is a high-level description of a general framework for influence network discovery. The framework receives the set of individuals $\mathcal{I}$, the set of behaviors $\mathcal{B}$, and a scoring function ϕ, which gives a score for an influence edge (u_i, u_j) based on a particular behavior. The output π is a function that gives, for which pair of users (u_i, u_j), where $u_i \neq u_j$, a value proportional to the probability of (u_i, u_j) to be an influence edge.

A scoring function ϕ receives as parameters two individuals u_i and u_j, such that $u_i \in \mathcal{I}$ and $u_j \in \mathcal{I}$, a behavior $b \in \mathcal{B}$, and the function $\mathcal{T}$. Based on such information, it returns a score $r \in \mathbb{R}$, which is a measure of how the adoption of b by u_i and u_j gives evidence concerning the existence of the influence edge (u_i, u_j). In the next section, we define several scoring functions for influence network discovery.

Algorithm 1: Framework for influence network discovery

```
Input  : I, B, T, φ
Output: π
1 for u_i ∈ I do
2     for u_j ∈ I do
3         π(u_i, u_j) = 0;
4 for b ∈ B do
5     I_b ← {u_i ∈ I | T(u_i, b) > 0};
6     for u_i ∈ I_b do
7         for u_j ∈ I_b do
8             if u_i ≠ u_j then
9                 π(u_i, u_j) = π(u_i, u_j) + φ(u_i, u_j, b, T);
```

The framework first sets the score of each edge to 0 (lines 1-3) and then applies the scoring function ϕ to each pair of users (u_i, u_j), for each behavior $b \in \mathcal{B}$. The score of a pair (u_i, u_j) is given by the following expression:

$$\pi(u_i, u_j) = \sum_{b \in \mathcal{B}} \phi(u_i, u_j, b, \mathcal{T})$$

4.3 Scoring Functions

This section presents scoring functions for influence network discovery. As described in the last section, a scoring function ϕ receives two individuals, u_i and u_2, a behavior b, and the function $\mathcal{T}$, and returns a score $r \in \mathbb{R}$. A scoring function may be given as input to the general framework described by Algorithm 1 in order to compute scores for candidate influence relationships based on user behavior data. We propose 11 scoring functions that consider three types of information: (1) co-occurrence, (2) popularity, and (3) inter-activity time. Table 3 shows the scoring functions employed in this work, which will be detailed in the following sections. Each function is described in terms of an expression and a condition that must hold for a given pair of vertices to receive the score given by the expression.

4.3.1 Co-occurrence functions

Co-occurrence scoring functions compute the score of a candidate edge (u_i, u_j) as the number of common behaviors shared by u_i and u_j. The basic idea is that the higher the number of behaviors shared by individuals, the more likely is an influence relationship between them. We define two co-occurrence scoring functions: the directed co-occurrence scoring function (*DC*) and the undirected co-occurrence scoring function (*UC*).

Name	Expression	Condition
Directed co-occurrence function (DC)	1	$\mathcal{T}(u_i, b) \wedge \mathcal{T}(u_j, b) \wedge \Delta > 0$
Undirected co-occurrence function (UC)	1	$\mathcal{T}(u_i, b) \wedge \mathcal{T}(u_j, b)$
Directed co-occurrence function with linear frequency decay (DCLIF)	$\sigma(b)^{-1}$	$\mathcal{T}(u_i, b) \wedge \mathcal{T}(u_j, b) \wedge \Delta > 0$
Directed co-occurrence function with logarithmic frequency decay (DCLOF)	$(log(\sigma(b)))^{-1}$	$\mathcal{T}(u_i, b) \wedge \mathcal{T}(u_j, b) \wedge \Delta > 0$
Directed co-occurrence function with exponential frequency decay (DCEF)	$e^{-\sigma(b)}$	$\mathcal{T}(u_i, b) \wedge \mathcal{T}(u_j, b) \wedge \Delta > 0$
Undirected co-occurrence function with linear frequency decay (UCLIF)	$\sigma(b)^{-1}$	$\mathcal{T}(u_i, b) \wedge \mathcal{T}(u_j, b)$
Undirected co-occurrence function with logarithmic frequency decay (UCLOF)	$(log(\sigma(b)))^{-1}$	$\mathcal{T}(u_i, b) \wedge \mathcal{T}(u_j, b)$
Undirected co-occurrence function with exponential frequency decay (UCEF)	$e^{-\sigma(b)}$	$\mathcal{T}(u_i, b) \wedge \mathcal{T}(u_j, b)$
Directed co-occurrence function with linear time decay (DCLIT)	Δ^{-1}	$\mathcal{T}(u_i, b) \wedge \mathcal{T}(u_j, b) \wedge \Delta > 0$
Directed co-occurrence function with logarithmic time decay (DCLOT)	$log(\Delta)^{-1}$	$\mathcal{T}(u_i, b) \wedge \mathcal{T}(u_j, b) \wedge \Delta > 0$
Directed co-occurrence function with exponential frequency decay (DCET)	$e^{-\Delta}$	$\mathcal{T}(u_i, b) \wedge \mathcal{T}(u_j, b) \wedge \Delta > 0$

Table 3: Scoring functions

The directed co-occurrence scoring function scores edges according to the order in which individuals presented the behaviors. If two users u_i and u_j have shown the same behavior, but u_i did it before u_j, then only the edge (u_j, u_i) is scored based on b. We use the symbol Δ to represent the time interval in between the individual behaviors of u_i and u_j (i.e., $\Delta = \mathcal{T}(u_i, b) - \mathcal{T}(u_j, b)$). Differently from its directed version, the undirected co-occurrence scoring function computes the number of co-occurrences ignoring which user has shown the behavior first.

4.3.2 Popularity decay functions

The popularity decay functions extend the co-occurrence functions, described in last section, by considering a decay based on the popularity of the behaviors. Since less popular behaviors are expected to be more informative for the discovery of influence relationships, we expect that such approach may be an improvement over the co-occurrence models. We represent the popularity (or frequency) of a behavior b by $\sigma(b)$.

We propose six scoring functions that consider the co-occurrence of behaviors with a popularity decay in the discovery of the influence network. Such functions can be classified into directed (*DCLIF, DCLOF,* and *DCEF*) or undirected (*UCLIF, UCLOF,* and *UCEF*), depending on whether they consider the order in which the individuals expressed a given behavior, and according to the decay function applied, which can be linear (*DCLIF* and *UCLIF*), logarithmic (*DCLOF* and *UCLOF*), or exponential (*DCEF* and *UCEF*).

4.3.3 Inter-activity time decay functions

The last group of scoring functions employed in this work extend the co-occurrence function presented in Section 4.3.1 by considering the inter-activity time. The intuition is that the closer in time two individuals presented a given behavior, the more likely it is an evidence of an influence relationship. Therefore, differently from the directed functions that take into account which individual has shown the behavior first, the inter-activity time decay functions consider also the length of the time interval in which the individuals presented a given behavior.

Since inter-activity decay functions are based on the time when the individuals have shown a given behavior, they are naturally directed. However, the decay of such functions can vary according to the function employed. The directed co-occurrence function with linear time decay (*DCLIT*), directed co-occurrence function with logarithmic time decay (*DCLOT*), and directed co-occurrence function with exponential time decay (*DCET*), which are detailed in Table 3 apply a linear, logarithm, and exponential decay functions, respectively.

The next section applies the strategies for influence network discovery presented along this section in a case study using individual behavior and influence network data from Twitter.

5. EXPERIMENTAL RESULTS

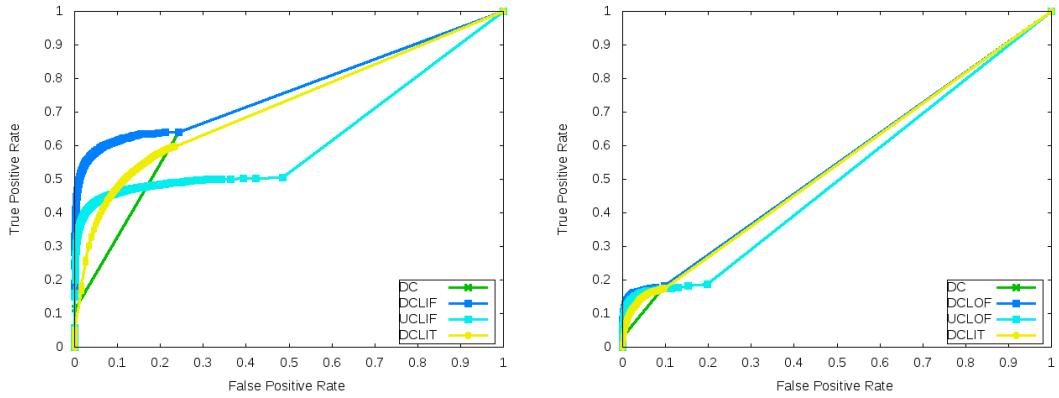
In this section, we present a case study on the influence network discovery problem using a real dataset from Twitter. In Section 3, we have characterized important properties of such dataset, which consists of a set of user activities (posting a tweet, a URL, or a hashtag) and a network defined by follower-followee interactions. We consider user activities as individual behavior data and the follower-followee interactions as influence relationships.

We evaluate the quality of a solution for the influence network discovery problem using the ROC (Receiver Operating Characteristic) analysis [22]. A ROC curve is a plot of the true positive rate (TPR = TP/(TP+FN)) versus the false positive rate (FPR = FP/(FP+TN)), where TP is the number of true positives, FN is the number of false negatives, FP is the number of false positives, and TN is the number of true negatives. For a given influence network $G(V, E)$ and a solution E', a true positive is an edge $e \in E \cap E'$, a false negative is an edge e, such that $e \in E$ and $e \notin E'$, a false positive is an edge e, such that $e \in E'$ and $e \notin E$, and a true negative is an edge e, such that $e \notin E'$ and $e \notin E$. Based on a ROC curve, we compute the area under curve (AUC) as a measure of effectivity. In our evaluation, we consider edges (v_i, v_j) for which there is, at least, one common behavior (posting a tweet, a URL, or a hashtag) shared by v_i and v_j, since other edges could not be predicted using any scoring function described in Section 4.

Table 4 shows the value of the AUC for the scoring functions defined in Section 4 and for the user activities described in Section 3. Posting a tweet is the most effective behavior in the discovery of influence relationships, achieving values of AUC from 0.54 to 0.77, depending on the scoring function applied. Expanded and non-expanded URLs have shown not to be effective in the discovery of influence relationships, achieving AUC values close to 0.50. Non-expanded URLs present slightly better results than the expanded ones, what was expected, since URLs propagated through an influence edge may maintain their original compressed form, evidencing a particular propagation. Moreover, the results show that

Function	Tweets	Non-expanded URLs	Expanded URLs	Hashtags
DC	0.71	0.54	0.53	0.39
UC	0.54	0.50	0.47	0.28
DCLIF	**0.77**	**0.55**	**0.54**	0.35
DCLOF	**0.77**	**0.55**	**0.54**	0.45
DCEF	**0.77**	**0.55**	**0.54**	**0.55**
UCLIF	0.62	0.51	0.49	0.31
UCLOF	0.62	0.51	0.49	0.30
UCEF	0.62	0.51	0.49	**0.52**
DCLIT	0.72	0.54	**0.54**	0.37
DCLOT	0.72	0.54	0.53	0.38
DCET	0.70	0.54	0.53	0.46

Table 4: Area under curve for different scoring functions and user activities



(a) tweets (b) non-expanded URLs

(c) expanded URLs (d) hashtags

Figure 4: ROC curves for the most effective scoring functions of each group of functions

hashtags are ineffective in the influence network discovery problem. Only two scoring functions (*DCEF* and *UCEF*) were able to achieve AUC values higher than 0.50 using hashtag data.

In terms of the scoring functions defined in this work, the results from Table 4 support some interesting conclusions. The directed scoring functions always outperform their non-directed versions, in spite of the reciprocity rate of 41% in the influence network. Moreover, inter-posting time has shown to be more effective than co-occurrence information. However, functions that consider co-occurrence and the order in which the activities are performed lead to the best results. Specific decay functions (logarithmic, linear, or exponential) have shown to be specially useful for hashtag data. In particular, the good results achieved by exponential frequency decay functions (*UCEF* and *DCEF*) when compared to the other functions evidences that very popular hashtags (see Figure 2d) carry little knowledge about influence relationships on Twitter, what was expected, since their popularity results in low discriminative power.

Figure 4 shows the ROC curves for the top scoring functions of

each group of functions and for each user activity. Scoring functions are divided into 4 groups: co-occurrence functions (*DC* and *UC*), directed popularity decay functions (*DCLIF, DCLOF,* and *DCEF*), undirected popularity decay functions (*UCLIF, UCLOF,* and *UCEF*), and inter-activity time decay functions (*DCLIT, DCLOT,* and *DCET*).

Figure 4a shows the ROC curves for the top scoring functions of each group and considering tweets. We can notice that the *DCLIF* function presents the best results in general. The *UCLIF* function achieves good performance for the top influence edges, but its not able to achieve similar results when the number of edges considered is increased. An opposite behavior is presented by the *DC* function, which presents poor effectivity for top edges but performs similarly to the *DCLIF* function for a large fraction of influence edges.

The ROC curves for the top scoring functions for the non-expanded URLs are shown in Figure 4b. We can notice that non-expanded URLs cover a very small fraction of the edges from the influence network, what explains the poor performance presented in Table 4. Similar conclusions can be drawn from the ROC curves for the top scoring functions for the expanded URLs (Figure 4c) and hashtags (Figure 4d). Figures 2e, 2f, 2g, and 2h have shown that users are more active in posting tweets than URLs and hashtags. In general, we found that URLs and hashtags are not discriminative in the identification of influence relationships.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we have studied how individual behavior data may be applied in the identification of influence relationships in social networks. We formalized such problem as the influence network discovery problem and presented several strategies for solving the proposed problem. The basic idea is to score potential influence edges based on co-occurrences of particular behaviors for pairs of individuals. Along Section 4, we presented several scoring functions for influence network discovery.

In order to evaluate the proposed strategies for influence network discovery, we applied them to the identification of follower-followee interactions on Twitter based on the content (tweets, URLs, and hashtags) generated by users across time. The results show that tweets may be very useful in the discovery of influence relationships. However, URLs and hashtags do not achieve such good performance. Moreover, it is important to consider the order in which two individuals present the same behavior and the popularity of such behaviors in the influence network discovery problem.

As future work, we will study techniques for extending the set of influence relationships discovered. Basically, potential influence edges identified based on individual behavior may be applied as probabilistic topological information in the generation of new edges. A similar approach have been proposed for the link prediction problem in [15]. Moreover, it would be interesting to apply the influence relationships discovered in order to predict individual behavior, such as in [11].

7. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD*, 2008.

[2] S. Aral, L. Muchnik, and A. Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *PNAS*, 2009.

[3] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *WSDM*, 2011.

[4] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *WSDM*, 2011.

[5] C. Bigonha, T. N. C. Cardoso, M. M. Moro, V. A. F. Almeida, and M. A. Gonçalves. Detecting Evangelists and Detractors on Twitter. In *WebMedia*, 2010.

[6] M. Cha, H. Haddadi, F. Benevenuto, and P. K. Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.

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

[8] W. Chen, Y. Wang, and S. Yang. Efficient influence maximization in social networks. In *KDD*, 2009.

[9] R. Dye. Buzz on Buzz. *Harvard Business Review*, 2000.

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

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

[12] J. L. Iribarren and E. Moro. Information diffusion epidemics in social networks. Jun 2007.

[13] D. Kempe, J. Kleinberg, and E. Tardos. Maximizing the spread of influence through a social network. In *KDD*, 2003.

[14] T. La Fond and J. Neville. Randomization tests for distinguishing social influence and homophily effects. In *WWW*, 2010.

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

[16] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 2007.

[17] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *KDD*, 2007.

[18] J. Leskovec, M. Mcglohon, C. Faloutsos, N. Glance, and M. Hurst. Patterns of cascading behavior in large blog graphs. In *SDM*, 2007.

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

[20] R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *KDD*, 2010.

[21] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 2003.

[22] F. J. Provost, T. Fawcett, and R. Kohavi. The case against accuracy estimation for comparing induction algorithms. In *ICML*, 1998.

[23] M. Richardson and P. Domingos. Mining knowledge-sharing sites for viral marketing. In *KDD*, 2002.

[24] M. R. Subramani and B. Rajagopalan. Knowledge-sharing and influence in online social networks via viral marketing. *Commun. ACM*, 2003.

[25] B. Taskar, M. fai Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *NIPS*, 2004.

[26] D. Watts and P. Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, 2007.

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