A Visibility-based Model for Link Prediction in Social Media

Linhong Zhu and Kristina Lerman
USC Information Sciences Institute
4676 Admiralty Way
Marina Del Rey, CA 90292
{linhong,lerman}@isi.edu

Abstract

A core task of social network analysis is to predict the formation of new social links. In the context of social media, link prediction serves as the foundation for forecasting the evolution of the follower graph and predicting interactions and the flow of information between users. Previous link prediction methods have generally represented the social network as a graph and leveraged topological and semantic measures of similarity between two nodes to estimate the probability of a new link between them. In this work, we suggest another link creation mechanism for social media that is based on the ease of discovering the new node. Specifically, a user v creates a link to another user u after seeing u's name on his or her screen; in other words, visibility of a user (name) is a necessary condition for new link formation. We propose a model for link prediction, which estimates the probability a user will see another user's name, and use this model to predict new links. We estimate a set of parameters in the proposed model using Maximum-Likelihood and Minorize-Maximize methods. Empirical results show that the proposed model can more accurately predict both follow and co-mention links than alternative state-of-the-art methods. Our work suggests that the effort required to discover a new social contact is negatively correlated with link formation, and the easier it is to discover a user, the higher the likelihood a link will be created.

1 Introduction

The link prediction task attempts to infer what new connections may form in a network in the future given its current state. Link prediction is the basis of many social recommendation tasks, including friend suggestion, product recommendation and advertisement matching. Due to its wide applicability, this problem has received much attention from the research community [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11]. Much of the focus of ongoing research has been on analyzing structural and content information in the network to identify features which are predictive of new links. A popular class of approaches measures structural proximity between nodes and uses this as a heuristic for predicting new links [2, 12, 13]. A variant of this approach that combined structural information with node- and edge-level attributes was used to predict new friendship links in an online social network [14, 6]. In contrast, we propose a simple model of link formation in social media that accounts for the mechanisms of information discovery. Before a social media user can create a new link, e.g., follow another user, she must first discover that user. Outside of explicit system recommendations, a user may discover another by seeing that user's name mentioned by one of the people she follows. *We suggest that name visibility is the primary mechanism for link formation in social media.* This, in turn, allows us to build more effective models for link prediction.

Specifically, we claim that the attention a user receives on Twitter is positively related to how many new social links will form. A link to user u forms when someone follows u or when someone mentions u in a post for the first time. Attention is related to visibility of the user name, i.e., how likely other users are to see it in their stream. For example, on Twitter, I may see a u's screen name in my stream because one of the people I follow retweeted u's post, and decide to follow u directly [9]. Alternatively, I may see u's screen name because one of the people I follow mentioned u in a post, perhaps as a result of a conversation that person had with u, and I may decide to follow u myself. Both of these link formation events can be explained by name visibility. Of course, there are other ways I may discover a new user to follow, e.g., through Twitter's own "who to follow" recommendations. However, since we cannot detect when and what recommendations Twitter makes, we do not include this source of name visibility in our model.

In this paper we present a mathematical model of name visibility and use it to predict new social links. Although we develop the model in the context of Twitter, it is more general and can be applied to other social media platforms with different user interfaces. Our model is based on psychological research that shows that people do not allocate their attention uniformly. When presented with a list of items, people attend more to items that are higher on the list than to lower ones. This cognitive bias, known as position bias, affects many facets of behavior, including the answers people select in response to multiple-choice questions [15, 16], where on the screen they look [17, 18], and search results they choose to click on [19]. Additionally, as a result of position bias, items at the top of a user's social media stream have higher visibility — and are more likely to be viewed — than items in lower positions [20, 21]. This extends to names mentioned in the posts as well: screen names in items at the top of a user's stream have higher visibility than names appearing in lower positions. Thus, to compute a name's visibility to a user, we need to know its position in the user's stream at the time of her visit. Unfortunately, this information is not always available. Instead, we estimate the average visibility from the available data. Specifically, the visibility of some screen name depends on the time it appears in the user's steam, and the number of new screen names the user receives between the time of the given name's appearance and the time the user visits the stream. Given this information, we calculate the probability the user views the name, and use this quantity to predict links.

In the proposed mathematical model, several user-specific parameters are estimated from data. To reduce computation and improve the generality of the model, we adopt out-of-sample extension technique [22], which maps the user-specific parameters to the functions of population-specific hyper-parameters. A minorize–maximize (MM [23] [24]) method is utilized to estimate the hyper-parameters by solving the constrained likelihood maximization problem. With the parameter estimation approach, we conduct experimental evaluations of the performance of prediction tasks by comparing with a set of state-of-the-art methods.

Our work generalizes previous research on predicting new links in social media within a unified framework of name visibility. Visibility is a heuristic for the effort required to discover a user in social media. The more visible some user's name is (because he or she is mentioned in many posts, or her own messages are retweeted by others), the easier it is to discover that user and the greater the likelihood that user will acquire new followers or will be mentioned by others. We can leverage our knowledge about the visibility of names to improve the accuracy of link prediction.

2 Related Work

There exist a variety of techniques for link prediction, ranging from feature-based supervised classification to matrix factorization and probabilistic graphical models. In the following, we summarize a set of representative works from each type of approaches, where more details can be found in a recent survey paper [25].

Much of the work on this problem is based on structural proximity. The survey paper [2] has studied a set of both local proximity measures and global proximity measures. The local proximity measures, which includes Preferential Attachment [1], Adamic-Adar [26], triadic closure [27, 28, 29] examines the neighborhood similarity of two nodes. The global proximity measure [30, 31, 32, 33, 34], takes into account indirect connections between network nodes. For instance, Katz score [30] counts paths between two nodes, and a set of other works [32, 34] developed measures of proximity between nodes based on random walks. Some existing works also model the link prediction problem as a supervised classification task, where each data point corresponds to a pair of vertices in the social network graph. In existing research on link prediction, majority of the features are a set of structural proximity measures which are extracted from the graph topology such as [7, 35, 36, 37].

While structural proximity measures the similarity of nodes in terms of graph topology, matrix factorization-based approaches utilize the similarity of nodes in the latent space for link predication. The general idea about these approaches [6, 10, 11] is to map the nodes into a certain latent space by factorizing the input graph matrix and/or content matrix and then reconstruct links based on the similarity of nodes in the latent space.

Probabilistic models for link prediction [3, 4, 5, 8] have branched in two directions. In one class of models, the main idea is to obtain a posterior probability that denotes the chance of co-occurrence of node pairs we are interested in [4, 12]. The second class of models focuses on modeling the evolution of the network. For example, Huang et al. [8] propose a Parameterized Social Activity Model using continuous-time stochastic process to predict aggregate social activities.

In the social context, links are formed not only based on node similarity (e.g., homophily), but also due to the information flow and influence [38, 13]. Considering this scenario, previous works have examined how information diffusion affects follow link formation [9]. The basic idea behind this work is that information diffusion via retweets causes new follow links that create a shortcut to the information source. In this work, we suggest that in social media, name visibility through users activity, is an important way to create "follow" and "mention" links. Our visibility model is more general in that it explains both the creation of shortcuts in information diffusion and new co-mentions using the same mechanism of name visibility.

Our model is similar to the stochastic model used by [20] to predict user retweets of posts. That model used post visibility to predict whether it will be retweeted, while we use name visibility to predict new links.

3 Data and Empirical Observations

In this work, we use a collection of Twitter data related to November 2012 California ballot initiatives. To collect this data, we identified a set of advocates, which denote Twitter accounts that explicitly promoted a position on one of the ballot initiatives, and added them to the list of monitored accounts. The position could be for the initiative, against it, or neutral. The statistics of advocates before the start of data collection are reported in Table 1, where $N_{\texttt{followers}}$, $N_{\texttt{friends}}$ and $N_{\texttt{statuses}}$ denote the numbers of followers, friends and posts made by the user respectively. By "friends" we denote accounts being followed by a particular user, while "followers" are the accounts following her. We monitored each advocate's follower lists every ten seconds by querying Twitter. The query time can be used as approximate link formation time for new followers obtained between two query intervals. By doing this, we obtain a *dynamic follow graph* which is used to learn and predict new followers for each advocate.

Besides monitoring the follow links, we also monitored user activity (i.e., posts or tweets made by accounts). Given the set of advocates and their followers, we collected their entire tweet history, as well as tweets which mentioned their screen names, during the monitoring period. This gives us a

Table 1: Statistics of a target set of users (advocates) on Twitter.

	$N_{\texttt{followers}}$	$N_{\texttt{friends}}$	$N_{\texttt{statuses}}$
mean	17589.6	1429.0	5935.1
standard dev.	118272.2	3567.8	12872.3
max	1065377	25162	101876

dynamic mention graph with time stamps of mentions that can be used for co-mentioning prediction. Note that we do not distinguish between mentions in original posts from mentions in retweets.

In the following, we report results of empirical analysis to provide more insights about name visibility, user attention and social links. Unfortunately, it is difficult to directly measure how much attention user pays to an advocate. Instead, we determine attention indirectly, by measuring whether user v has received the screen name of an advocate u in her stream, e.g., via a mention '@u', when v does not follow u. Our intuition is that exposure and attention are correlated: the more exposure u receives, the more attention she receives.

Specifically, the screen name of u, can be exposed to a non-follower v through one of the following mechanisms:

- if a friend of v posts or reposts any message which contains "@u" (Case 1), or
- if any user (not a friend of v) mentions u and v together in the same message (Case 2)

In general, social media friends "virtually" introduce people to new social contacts. Whenever any of v's friends posts or reposts a message that mentions u, then u's screen name will appear in v's stream (as '@u' embedded in the text of the message), giving v a chance to learn about u and create a new link. There are other mechanisms of exposure. User v may also learn about u through system recommendation, or an email received from Twitter, but we ignore these mechanisms since we do not have any data to measure they effects.

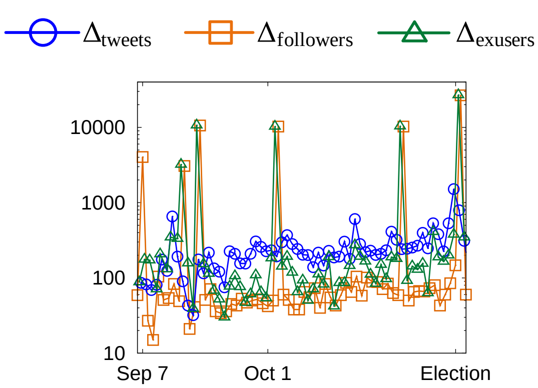


Figure 1: Time-series correlation among number of new exposed non-followers, number of new followers, and number of new tweets mentioning @CArightToKnow.

The following empirical observations inform our visibility model.

Table 2: Testing results of mutual information and transfer entropy among number of new followers ($\Delta_{\texttt{followers}}$), number of new tweets ($\Delta_{\texttt{tweets}}$) and number of new exposed non-followers $\Delta_{\texttt{exusers}}$.

X	Y	$I(X:Y)$	TE($X \to Y$)	TE($Y \to X$)
$\Delta_{\texttt{followers}}$	$\Delta_{\texttt{tweets}}$	0.108	0.004	0.083
$\Delta_{\texttt{followers}}$	$\Delta_{\texttt{exusers}}$	0.005	0.046	0.249
$\Delta_{\texttt{tweets}}$	$\Delta_{\texttt{exusers}}$	0.149	-0.047	-0.014

3.1 Effect of Exposure

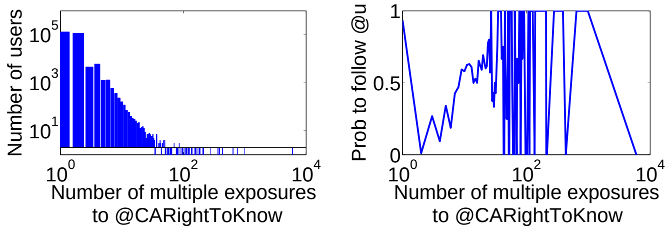
Observation 1 *The more exposure an advocate u receives, via others mentioning u in their posts or retweeting u's posts, the more followers u will attract.*

We first illustrate with a special case: @CARightToKnow, who was an active advocate for one of the ballot initiatives. In Figure 1 we plot the daily numbers of new followers ($\Delta_{\texttt{followers}}$), new tweets mentioning @CARightToKnow ($\Delta_{\texttt{tweets}}$) and new non-followers exposed to @CARightToKnow ($\Delta_{\texttt{exusers}}$). There is usually a peak in the number of new exposed non-followers right before a peak in the number of new followers, especially before election, which suggests a causal relationship between visibility of a name and new followers.

Now we statistically examine the casuality among activities ($\Delta_{\texttt{tweets}}$), number of new exposed non-followers ($\Delta_{\texttt{exusers}}$) and number of new followers ($\Delta_{\texttt{followers}}$) using mutual information and transfer entropy [39, 40]. The mutual information of two time-series variables X and Y, is defined as $I(X:Y) = H(X)$ - $H(X \mid Y)$; while transfer entropy from X to Y, is defined as $TE(X \to Y) = H(Y^t \mid Y^{t-1})$ - $H(Y^t \mid Y^{t-1}, X^{t-1})$. The higher value of mutual information of two variables X and Y denotes stronger mutual dependency between variables while a higher value of transfer entropy from X to Y denotes X exerts certain influence on Y.

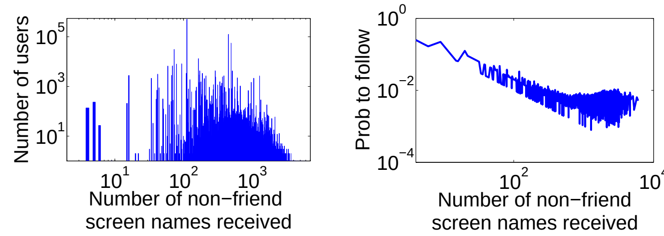
For each advocate, we construct three time-series vectors $\Delta_{\texttt{followers}}$, $\Delta_{\texttt{tweets}}$, and $\Delta_{\texttt{exusers}}$), and then report the averaged testing results of all the advocates in Table 2. The result shows that follow link formation is correlated with user's mentioning activity with mutual information 0.108. In addition, there is a possible causality from user activity to follow link formation since transfer entropy from $\Delta_{\texttt{tweets}}$ to $\Delta_{\texttt{followers}}$ is much larger than that from $\Delta_{\texttt{followers}}$ to $\Delta_{\texttt{tweets}}$. Though user activity is correlated with visibility with mutual information 0.149, the negative transfer entropies of both sides between $\Delta_{\texttt{tweets}}$ and $\Delta_{\texttt{exusers}}$ indicate that there is little causality between user activity and visibility. To the contrary, though the correlation between visibility and follow link formation is low with mutual information 0.005, the transfer entropy from $\Delta_{\texttt{exusers}}$ to $\Delta_{\texttt{followers}}$ is up to 0.249 and much larger than that from $\Delta_{\texttt{followers}}$ to $\Delta_{\texttt{exusers}}$, which suggests a possible causal relationship between visibility and follower link formation.

However, it is still difficult to evaluate whether these values (e.g., transfer entropy = 0.249) are significant enough to suggest a causal relationship from name visibility to the formation of follow links. Consequently, we conduct another way of assessing significance of the causal relationship from $\Delta_{\texttt{exusers}}$ and $\Delta_{\texttt{followers}}$. Given the aim of distinguishing



(a) (b)

Figure 2: 2(a): The distribution of number of multiple exposures to @CARightToKnow. 2(b): Probability of v starts following u vs number of times v was exposed to u (number of multiple exposures).



(a) (b)

Figure 3: 3(a): The distribution of number of non-friend screen names received. 3(b): Probability of v starts following any @u versus number of names received by v.

between the null hypothesis H_0 of no causality and the hypothesis H_1 of causality:

$$\begin{aligned} H_0: &\quad TE(\Delta_{\texttt{exusers}} \rightarrow \Delta_{\texttt{followers}}) = 0 \\ H_1: &\quad TE(\Delta_{\texttt{exusers}} \rightarrow \Delta_{\texttt{followers}}) > 0 \end{aligned} \tag{1}$$

we create the distribution of H_0 by reshuffling $\Delta_{\texttt{exusers}}$, while keeping the order of $\Delta_{\texttt{followers}}$ intact. We then conduct one-tail t-test to validate whether the value of the causality measure obtained on the shuffled data is smaller than the value of causality measure of original data. We accept the H_1 hypothesis of causality with p-value$= 8.995 \times 10^{-6}$ ($\ll 0.01$).

We conduct the same hypothesis test for user activity and the formation of follow links, and accept the H_1: $\mathrm{TE}(\Delta_{\texttt{tweets}} \rightarrow \Delta_{\texttt{followers}}) > 0$ with p-value$= 4.833 \times 10^{-6}$ ($\ll 0.01$)

Based on this observation, we build our link predictor for a user based upon the visibility of his/her name to others and others' activities.

3.2 Effect of Multiple Exposures

Observation 2 *For a specific advocate u, the majority of non-followers who are exposed to u receive u's name in their streams only a few times. In addition, exposing the same user multiple times does not necessarily increase the probability a new link will form.*

Figure 2(a) plots the distribution of non-followers who were exposed to @CARightToKnow. The plot shows a heavy-tailed distribution. About 100K users received @CARightToKnow once over the data collection period. In contrast, 27 users received @CARightToKnow more than 100 times. Figure 2(b) examines the relationship between how often a user v received @CARightToKnow and the average probability of v followed back @CARightToKnow. Many users directly followed back @CARightToKnow after a single exposure. However, for a large population (130K users) between 2-20 exposures to @CARightToknow decreased the follow probability. The follow probability rose again as the number of exposures increased above 50.

3.3 Effect of the Number of Names in the Stream

Observation 3 *In our data set, the majority of users received 50–200 screen names in their streams, with a few users receiving many more or fewer names. In addition, the attention a name receives from a user depends on the number of names in the user's stream.*

Figure 3(a) shows that more than 500K users received around 112 screen names in their streams, with only 139 users receiving more than 3000 screen names and 399 users receiving fewer than 10 screen names. Figure 3(b) plots the average probability v starts following another user vs the number of non-friend screen names she received in her stream. When v receives more screen names, she is less likely to follow any specific name.

Observations 2 and 3 suggest that attention received by u is not simply determined by the number of exposures u's name had, but also depends on other factors, such as the number of other names non-followers receive. All these factors affect the visibility of a name, i.e., the likelihood that a non-follower will see the name in his or her stream.

In summary, we have presented some observations derived from real data collected from Twitter. Based on these observations, we believe that the more attention users receive from non-followers, by having their screen names appear in non-followers streams, the more links will be created. Friends act as a vector for link formation, by facilitating "virtual introductions" via @mentions or retweets. Without such exposure, it is difficult for users to attract new followers. These observations suggest that visibility of a user name drives the attention the user receives, and that we can exploit this connection for the purpose of link prediction. In the next section, we explain the computational framework for estimating name visibility and using it in a predictive model.

4 The Name Visibility Model

Due to the randomness of social activities, we seek to utilize a stochastic model to describe the process through which a user v from a population pays attention to another user u who is visible to user v. In contrast to prior research that focused on the visibility of a post [20] or a news story [41], we examine the effects of name visibility. The visibility of a name in turn causes other users in a population to pay

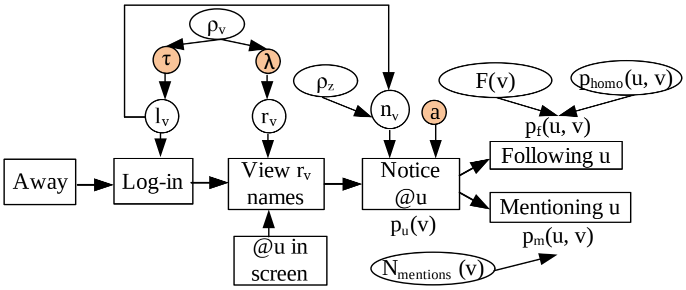


Figure 4: State diagram of a user v's behavior regarding to another user u who is in the screen of v. The ovals represent parameters which are measured, while circles represent parameters to be estimated from data. To simplify the model, only the filled circles which represent the hyper-parameters are to be finally estimated from data.

Table 3: Notations and explanations

Notations	Explanations
n_v	average number of screen names received by v between two log-in interval
r_v	average number of viewed names by v
l_v	average log-in frequency of v
ρ_v	activity of v
z	a friend of v
$p_u(v)$	maximal probability name u is visible to v
$p_{\texttt{homo}}(u,v)$	homophily between users u and v
$F^+(v)$	probability v follows similar users
$F^-(v)$	probability v follows dissimilar users
$p_f(u,v)$	probability user v follows u
$p_m(u,v)$	probability user v mentions u
$N_{\texttt{mentiones}}(v)$	number of users v has mentioned so far
a, τ, λ	hyper-parameters

attention and respond in terms of new "follow" or "@mention" relationships.

Figure 4 shows the process through which a user v from a population views another user u (e.g., an advocate), notices u and then responds to u. For ease of presentation, we summarize notations used in Figure 4 in Table 3. When a user v visits a social media site to view her stream, she may view a limited portion of her stream, hence seeing only a limited number of screen names. Whether an advocate u, who appears in v's stream, will be visible to v, depends on the total number of screen names n_v, the log-in frequency of v and as well as the average number of names v views per log-in. Once v notices u's name (with probability $p_u(v)$), v might then create a new "follow" or "mention" link to u.

Below, we first discuss the computation of the probability that u's name is visible to v ($p_u(v)$), as well as link formation probability $p_f(u,v)$ and $p_m(u,v)$.

4.1 Name Visibility

Let X denote the position of u's name in v's stream at the time v logs-in to the social media site to view her stream. Since Twitter lays the stream out in a chronological order, the position of u's name corresponds to its arrival order in v's stream. Suppose that n_v candidate names arrived in v's stream since v's last log-in, and that v views only r_v names. Let S_{n_v,r_v} denote the event of v noticing u after viewing r_v

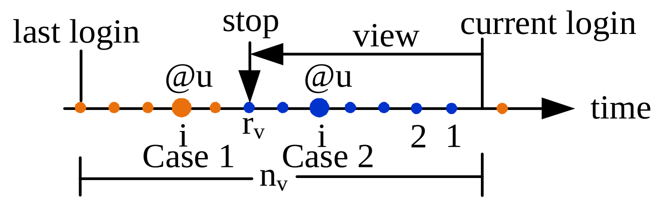


Figure 5: An illustrative example about how a user u becomes visible to v. Each dot represents a candidate screen name, while large dots denote the target user u and blue dots represent a list of viewed users. The screen names are chronologically ordered by their arrival time.

of the n_v candidate names in the stream. First we show that X is uniformly distributed on $\{1, 2, \cdots, n_v\}$. This is easy to prove, since the candidates arrive in a random order. Next, we consider the likelihood of noticing u given that u's name is in i^{th} position.

We compute the conditional probability u's name is visible to v, where u is in i^{th} position. There are two cases to consider, as illustrated in Figure 5. The figure shows screen names that have arrived in v's stream, chronologically ordered by their arrival time. Upon logging in, v views names in reverse chronological order, starting with the most recent names at the top of the stream. In Case 1, u's name arrived before the last of the r_v names viewed by v; therefore, the event S_{n_v,r_v} fails. In Case 2, the arrival order of u is $i \le r_v$, then the event will succeed with certain probability. Intuitively, we suggest that:

- The success probability (i.e., visibility) is positively correlated with the frequency v visits the stream (activity rate).
- The visibility of a name decays with its position i in the stream.

The former intuition suggests that the visibility of names is related to the total effort the users are willing to exert on examining their stream. More active users visit their stream more frequently or inspect larger portions of it; hence, they notice more names. The latter intuition, a manifestation of position bias, asserts that users are more likely to view recent names near the top of their stream rather than older names.

Thus, for $n_v \in N^+$, we model the probability that u is visible to v when v views r_v names ($0 \le r_v \le n_v$) as:

$$p(S_{n_v,r_v} \mid X = i) = \begin{cases} 0 & \text{if } i > r \\ l_v e^{-ai} & \text{if } i \le r_v. \end{cases} \tag{2}$$

where a is a parameter and l_v denotes the log-in frequency of the user v. Note that here we utilize the exponentially decaying function e^{-ai}, but we argue that any other decaying functions such as power-law, inverse gaussian, etc., can be used depending on the specific scenario.

We compute the visibility of u when v views r_v names from among n_v candidates:

$$
\begin{aligned}
p(S_{n_v,r_v}) =& \sum_{i=1}^{r_v} p(X=i)p(S_{n_v,r_v} \mid X=i) \\
&+ \sum_{i=r_v+1}^{n_v} p(S_{n_v,r_v} \mid X=i)p(X=i) \\
=& \frac{l_v e^{-a}(1-e^{-ar_v})}{n_v(1-e^{-a})}
\end{aligned} \tag{3}
$$

subject to $0 \le r_v \le n_v$

For simplicity, we use $p_u(v)$ to denote the maximum visibility probability of u to v.

4.2 Link Formation Probability

Once user v views u's name, she may respond to it. In this study, we focus on two kinds of responses: "following" and "co-mentioning", i.e., when v follows u or mentions u in her post.

We only have the complete profiles of followers for each advocate in our data set. Therefore, we limited the set of user u to the set of advocates and then calculated the probability that a user v follows u after seeing u's screen name. The proposed model characterizes user following response $R_u(v)$ as a combination of two: the homophily $p_{\texttt{homo}}(u,v)$ between two users u and v, and the response probability of v in terms of following other advocates $F^+(v)$ and $F^-(v)$:

$$R_u(v) = p_{\texttt{homo}}(u,v)F^+(v) + (1 - p_{\texttt{homo}}(u,v))F^-(v)$$

Given an advocate u and another user v, we estimate the homophily score between these two as the average stance similarity on the ballot initiatives in our political data set. Specifically, let $\overrightarrow{a_u}$ denote a vector of stances (for, against, or neutral) where $\overrightarrow{a_u(i)}$ represents the stance of user u on the i^{th} initiative, which can be computed using supervised learning approach proposed by Smith et al. [42] or unsupervised approach proposed by Zhu et al. [43]. We then define the $p_{\texttt{homo}}(u,v)$ as:

$$p_{\texttt{homo}}(u,v) = \cos \texttt{sim}(\overrightarrow{a_u}, \overrightarrow{a_v})$$

The second factor determining a follow response is the probability that a user v follows a same-stance/different-stances advocate after seeing his screen name, $F^+(v)/F^-(v)$. We approximate $F^+(v)$ and $F^-(v)$ as:

$$
\begin{aligned}
F^+(v) =& \frac{N_{\texttt{friends}\cap\texttt{stances}}(v)}{N_{\texttt{stances}}(v)} \\
F^-(v) =& \frac{N_{\texttt{friends}\cap\texttt{diffstances}}(v)}{N_{\texttt{diffstances}}(v)}
\end{aligned}
$$

where $N_{\texttt{friends}\cap\texttt{stances}}(v)$ denotes number of same-stance advocates v has followed, $N_{\texttt{friends}\cap\texttt{diffstances}}(v)$ denotes number of different-stances advocates v has followed, $N_{\texttt{stances}}(v)$ denotes number of advocates who hold same position with v, and $N_{\texttt{diffstances}}(v)$ denotes number of advocates who hold different positions with v.

Combining the above factors, the proposed model gives the probability of a user v follows an advocate u as:

$$p_f(u,v) = p_u(v)R_u(v) \tag{4}$$

For "co-mentioning", we have the complete activities (i.e., posts) of all the users and we then define the probability that a user v mentions another user u after a user z mentions both u and v as:

$$p_m(u,v) = p_u(v)N_{\texttt{mentions}}(v)/n \tag{5}$$

where $N_{\texttt{mentions}}(v)$ denotes number of users v has mentioned so far, and n is total number of users in the entire population.

4.3 Parameter Estimation

Unfortunately, none of the parameters (log-in frequency l_v of v, average number of screen names between two log-in intervals n_v, and average number of viewed names r_v) are available in our data. Moreover, these parameters are highly variable, and we do not want to estimate them for each individual. Here we adopt the Out-of-Sample Extensions [22], to transfer the parameter estimation for each individual to hyper-parameter estimation for the entire population. The idea is similar to the mean-field theory in physics and probability theory, where the interactions to any individual are replaced with an average or effective interaction.

Both log-in frequency l_v and the number of viewed names r_v are positively correlated with user's activity rate ρ_v. Thus, we introduce two hyper-parameters τ and λ, to capture the relations between l_v and r_v and user's activity ρ_v. Specifically, we use the simple linear model and define l_v and r_v as:

$$l_v = \tau\rho_v;\ r_v = \lambda\rho_v \tag{6}$$

The replacement of average number of screen names with hyper-parameters is a more complex process. Intuitively, number of screen names received by v is related to the activity of v's friends. If v follows more people, and they are actively mentioning or re-tweeting others, then v might receive more tweets as well as screen names. On the other hand, since n_v denotes the average number of screen names received between log-ins, then n_v is inversely proportional to frequency of log-in l_v. The above discussion motivates us to define n_v as:

$$n_v = \frac{\sum_{z\in\texttt{friends}(v)} \rho_z}{\tau\rho_v} \tag{7}$$

The parameter dependencies are summarized in Figure 4 and the final parameters are $\theta = \{\tau, \lambda, a\}$. We estimate them by maximum likelihood: the choices of model parameters that maximize the probability of the observed responses from a set of users according to the model. By plugging Eq. 6 and Eq. 7 into Eq. 4, the Log-likelihood function can be rewritten as:

$$
\begin{aligned}
\mathcal{L}(\theta) =& \sum_{(u,v)\in E} \log \frac{A(u,v)\tau^2(1-e^{-a\lambda\rho_v})}{e^a - 1} \\
&+ \sum_{(u,v)\notin E} \log(1 - \frac{A(u,v)\tau^2(1-e^{-a\lambda\rho_v})}{e^a - 1}) \\
&\text{subject to } \forall v, 0 \le \lambda\rho_v \le \frac{\sum_{z\in\texttt{friends}(v)} \rho_z}{\tau\rho_v} \\
&\text{where } A(u,v) = \frac{R_u(v)\rho_v^2}{\sum_{z\in\texttt{friends}(v)} \rho_z}, u \text{ is an} \\
&\text{advocate, and } v \text{ is a user who was exposed to } u
\end{aligned} \tag{8}
$$

Equation 8 is an inequality constraint likelihood function, which makes it intractable to directly take the derivatives. Thus, we use a minorize–maximize (MM) algorithm [23] [24] to estimate the parameters. This algorithm, of which EM is a special case, finds a lower-bound function and then maximizes it, as follows. First, using the concavity[1] of $\ln(1-x)$, we have:

$$\begin{aligned}&\sum_{(u,v)\notin E} \ln(1 - A(u,v)\frac{\tau^2(1-e^{-a\lambda\rho_v})}{(e^a-1)})\\ \geq &\sum_{(u,v)\notin E}[\ln(1-A(u,v)B(\widetilde{\theta},v)) \\ +&\frac{A(u,v)}{1-A(u,v)B(\widetilde{\theta},v)}(B(\widetilde{\theta},v)-B(\theta,v))]\end{aligned} \tag{9}$$

where

$$B(\widetilde{\theta},v) = \frac{\widetilde{\tau}^2(1-e^{-\widetilde{a}\widetilde{\lambda}\rho_v})}{(e^{\widetilde{a}}-1)},$$
$$B(\theta,v) = \frac{\tau^2(1-e^{-a\lambda\rho_v})}{(e^a-1)},$$

and $\widetilde{\theta}$ denote values of parameters obtained from previous iteration.

In each iteration, we can consider $B(\widetilde{\theta},v)$ and $A(u,v)$ as constants which are independent of parameters, thus maximizing the $L(\theta)$ is equal to maximizing the following lower bound function:

$$\begin{aligned}&Q(\theta \mid \widetilde{\theta}) = \sum_{(u,v)\in E}\ln B(\theta,v) - \sum_{(u,v)\notin E}\frac{A(u,v)B(\theta,v)}{1-A(u,v)B(\widetilde{\theta},v)}\\ &\text{subject to } \forall v, 0 \leq \lambda\rho_v \leq \frac{\sum_{z\in\texttt{friends}(v)}\rho_z}{\tau\rho_v}\end{aligned} \tag{10}$$

We use block coordinate ascent to solve Eq. 10, which leads to the following updating rules:

$$\tau = \min(C(v), \sqrt{\frac{|E|\widetilde{\tau}^2}{\sum_{(u,v)\notin E}\frac{A(u,v)B(\widetilde{\theta},v)}{1-A(u,v)B(\widetilde{\theta},v)}}}) \tag{11}$$

$$\lambda = \max(0, \min(D(v), \frac{\widetilde{\lambda}\sum_{(u,v)\in E}\rho_v E(\widetilde{\theta},v)}{M(\widetilde{\theta},v)})) \tag{12}$$

$$a = \ln\frac{F(\widetilde{\theta},v)-G(\widetilde{\theta},v)-H(\widetilde{\theta},v)}{F(\widetilde{\theta},v)-G(\widetilde{\theta},v)-H(\widetilde{\theta},v)-|E|} \tag{13}$$

where

$$C(\widetilde{\theta},v) = \min_v(\frac{\sum_z\rho_z}{\widetilde{\theta}\rho_v^2}),$$
$$D(\widetilde{\theta},v) = \min_v(\frac{\sum_z\rho_z}{\widetilde{\tau}\rho_v^2}),$$
$$E(\widetilde{\theta},v) = e^{-\widetilde{a}\widetilde{\lambda}\rho_v},$$
$$M(\widetilde{\theta},v) = \sum_{(u,v)\notin E}\frac{A(u,v)\rho_v E(\widetilde{\theta},v)\widetilde{\tau}^2}{(1-A(u,v)B(\widetilde{\theta},v))(e^{\widetilde{a}}-1)},$$

[1]A continuously differentiable function of one variable is convex on an interval if and only if the function lies above all of its tangents: $f(x) \geq f(y) + f'(y)(x-y)$ for all x and y in the interval.

Algorithm 1 A MM algorithm to learn parameters in visibility model

Input: A dynamic network G=(V, E)
Output: $\theta = \{a, \tau, \lambda\}$

```
Randomly initialize a > 0, τ > 0, λ > 0
repeat
    for each advocate u and each exposed user v
        compute A(u,v), B(θ̃,v), C(θ̃,v), D(θ̃,v)
        E(θ̃,v), M(θ̃,v), F(θ̃,v), G(θ̃,v), H(θ̃,v)
    Calculate τ using Eq. 11
    Calculate λ using Eq. 12
    Calculate a using Eq. 13
until Convergence
```

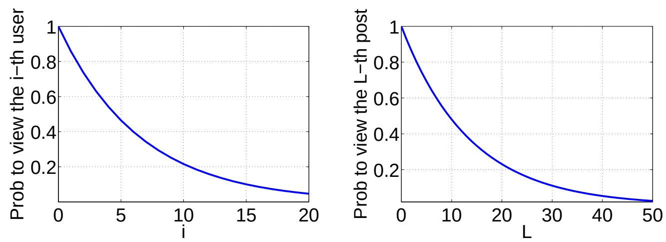


Figure 6: Study of position bias: How probability of viewing a screen name and the corresponding probability to view a post are decaying with positions.

$$F(\widetilde{\theta},v) = \sum_{(u,v)\in E}\frac{\widetilde{\lambda}\rho_v E(\widetilde{\theta},v)}{1-E(\widetilde{\theta},v)},$$
$$G(\widetilde{\theta},v) = \sum_{(u,v)\notin E}\frac{\widetilde{\lambda}\widetilde{\tau}^2\rho_v A(u,v)E(\widetilde{\theta},v)(1-E(\widetilde{\theta},v))}{1-A(u,v)B(\widetilde{\theta},v)},$$
$$H(\widetilde{\theta},v) = \sum_{(u,v)\notin E}\frac{A(u,v)\widetilde{\tau}^2}{2(1-A(u,v)B(\widetilde{\theta},v))(e^{\widetilde{a}}-1)^2}.$$

We summarize the proposed algorithm in Algorithm 1.

The parameter estimation in "co-mention" link prediction is similar to the above process.

4.4 Empirical Study of Parameters

With all available data, we estimate values of parameters using the MM algorithm, which leads to a=0.15326, θ=0.128, τ=3.83. Using the estimated value a, the probability a user v views a screen name u at position i is plotted in the left side of Figure 6. We notice that the viewing probability decays much faster than the law of surfing [20], which models the relation between probability to view a post and its position. This is understandable since not every post contains a screen name @u. Empirically, on average 47.8% of posts in our dataset contain a screen name. Thus, if we map the probability of viewing a screen name in the i-th position to the probability of viewing a post in the L-th ($L = 2.09i$) position, we get the plot shown on the right of Figure 6. The plotted relation between position of posts and probability to view posts is similar to relation modeling the law of surfing [20].

Next, we empirically investigate the values of the average number of viewed names r_v and average number of screen

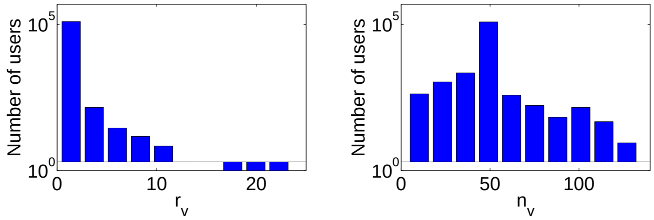


Figure 7: Distributions of average number of viewed names r_v and average number of screen names n_v.

names per log-in n_v. We do not analyze the values of l_v since from Eq. 6, we get that l_v is linearly proportional to r_v. We still choose @CARightToKnow as a target advocate, and then plot the distribution of n_v and r_v among its exposed users. The results are shown in Figure 7. The distribution of r_v follows a power-law, which indicates that the majority of users only view few names with some exceptions who view more than 10 names. While r_v follows a power-law distribution, the distribution of n_v is more like a skewed normal distribution with mean 51.8264.

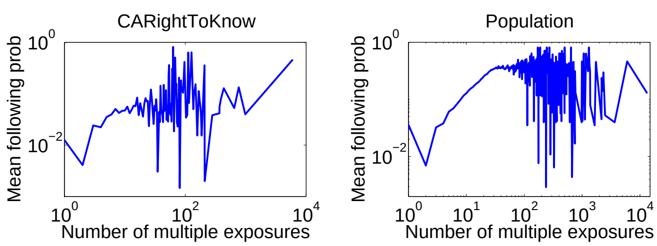


Figure 8: The relation between number of multiple exposures and estimated probability of following. These two figure plot the relation between the frequency of multiple exposures and average value of estimated probability of following for exposed users with a give exposure frequency of @CARightToKnow and the entire population, respectively.

In Figure 2(b), we plotted the relation between the number of times of a user was exposed to an advocate and probability he or she followed the advocate. Now we want to validate whether the estimated following probability using Eq. 4 and estimated parameters still preserves similar relation to the frequency of multiple exposures. In Figure 8 we plot the frequency of multiple exposures of @CARightToKnow to exposed users and the estimated following probability. We find that the estimated following probability using the proposed stochastic visibility model preserves the empirically observed relationship, even though exposure frequency is neither a parameter from observation nor a to-be-estimated parameter in the model. In other words, the empirical study supports the proposed stochastic visibility model and shows that its parameters are accurately estimated.

5 Experiments

In this section, we evaluate the stochastic visibility model on the link prediction task.

5.1 Evaluation Metrics and Baselines

In this experimental study, we employ two widely used metrics to evaluate the link prediction results: the Area Under ROC Curve (AUC) and mean average precision (MAP). ROC curve is created by plotting the fraction of true positives out of the total actual positives (true positive rate) vs. the fraction of false positives out of the total actual negatives (false positive rate), at various threshold settings. Thus, AUC score evaluates the overall ranking yielded by the algorithms with a larger AUC indicating better link prediction performance.

Suppose there are m missing links for a user, then the average precision for this user at n is defined as:

$$ap@n = \frac{\sum_{k=1}^{n} p(k)}{\min(m, n)}$$

where $p(k)$ the precision at cut-off k in the item list. The mean average precision for N users at position n is the average of the average precision of each user, i.e.,

$$MAP@n = \frac{\sum_{i=1}^{N} ap@n_i}{N}$$

The MAP metric focuses on the top-n predicted (or recommended) results. The larger precision means the higher predictive accuracy.

We compare the proposed approach, *stochastic visibility* model (**SV**), with a set of baselines:

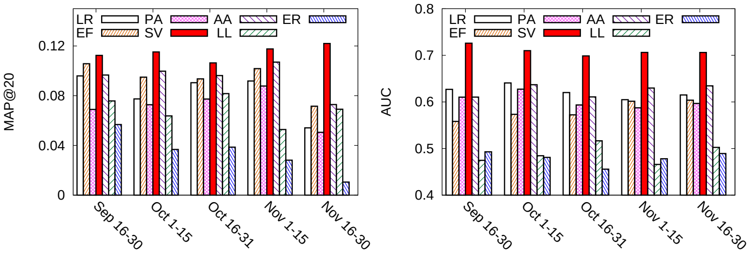
- *Linear Regression (**LR**)*: We extract a set of features from observations such as number of followers and friends, number of status updates, mentions and mentioned counts for both u and v, and then train linear regression model from existing links.
- *Edge Rank (**ER**)* [2]: Edgerank can be considered as a personalized PageRank (or a rooted PageRank) over a graph. It is defined as the stationary probability of v in a random walk that returns to u with probability α each step, moving to a random neighbor with probability 1-α. Both ER and Regression (either logistic regression or linear regression LR), are very effective in friends/follower prediction such as Facebook Link prediction challenge[2].
- *Preferential Attachment (**PA**)* [1]: This model assumes that probability of a new link depends on the number of existing links.

  $$p_f(u, v) = N_{followers}(u) \times N_{friends}(v)$$

  We use preferential attachment model because it is widely used to model link formation in networks.
- *Adamic-Adar (**AA**)* [26]: Probability of a link in this model depends on the number of common neighbors, attenuated by the log of the degree of a neighbor. Here, degree $deg(z)$ is the sum of the in- and out-degree of the node (friends and followers respectively).

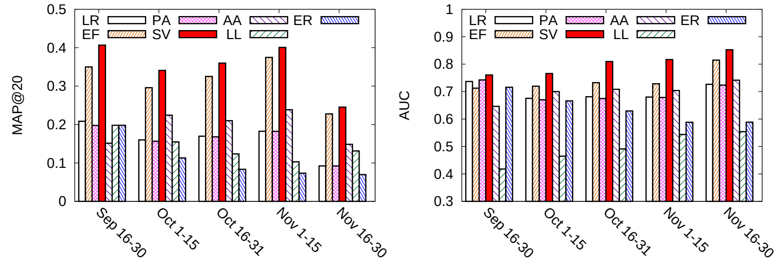
  $$p_f(u, v) = \sum_{z \in (friends(v) \cap followers(u))} \frac{1}{\log(deg(z))}$$

[2]http://blog.echen.me/2012/07/31/edge-prediction-in-a-social-graph-my-solution-to-facebooks-user-recommendation-contest-on-kaggle/



(a) (b)

Figure 9: Comparison of different algorithms on the task of predicting follow links using (a) MAP and (b) AUC scores.



(a) (b)

Figure 10: Comparison of different algorithms on the task of predicting co-mentions using (a) MAP and (b) AUC scores.

AA has been verified as a most competitive baseline for static link prediction by various works such as [2, 13].

- *Latest Activity (**LL**)* [5]: Rank links in descending order of exposure time, with most recent exposure first. LL is a simple but effective baseline for temporal link prediction [12].
- *Exposure Frequency (**EF**)*: We compute the number of times an advocate u is exposed to user v, and then rank pairs based on the exposure frequency.

Note that the baselines are defined for follow link prediction, but they can be easily adapted to co-mention prediction with the help of mention graph.

5.2 Follow Link Prediction Results

Figure 9 show results of seven algorithms on the task of predicting new follow links. Note that x-axis denotes the testing time interval and all the available data before testing time are for training. The proposed approach SV (solid red color) outperforms all other baselines in terms of both MAP and AUC scores. The second best performer is either EF or AA for MAP metric and LR or AA for AUC metric. PA, LL, and ER perform much worse than the others but there is no consistent loser among these three.

5.3 Co-mention Prediction Results

The results of co-mentioning prediction are shown in Figure 10. Similar to follow link prediction, the proposed approach SV achieves consistently better results than all other baselines. While EF performs worse than SV, it has much better performance than other baselines in terms of MAP. On average, the prediction precision of SV and EF can be twice as high as the others. In terms of AUC score, EF obtains second-best performance four out of five times. Since EF can be considered as a simple model of visibility using exposure frequency only, the success of visibility-based approach indicates that visibility leads to creation of new links especially co-mentioning links.

6 Conclusion

Recent research has shown that the ease of discovering an item, such as a tweet [20] or a scientific paper [44], determine the likelihood of response, i.e., a retweet or a citation respectively. In this paper we have shown that the same applies to social link formation in social media: the easier it is to discover a new user, the greater the likelihood of a new social tie forming, either as a new follow relationship or a new mention. We capture the ease of discovery via the concept of dynamic visibility [21], so that the more visible a user is, the easier she is to discover, the more links to her will form.

We have presented an intuitive, yet effective, model that uses visibility of a name to predict new links. The model accounts for visibility of another user's name to a given user based on the number of messages containing new names a user receives in her social media stream and the frequency with which she visits the stream to view new messages. The greater the number of names a user has to inspect in her stream to find a particular name, the less likely she is to view it and to form a new link to that user. We have developed an algorithm to estimate a set of hyper-parameters proposed in this model and conducted empirical studies to verify the effectiveness of the proposed approach. We showed that the proposed model can better predict new follow and co-mention links that alternative link prediction algorithms.

In the future, we will use the visibility model to explore the evolution of social networks, and how sharing and resharing of information in social media shapes the follower graph.

Acknowledgement

This material is based upon work supported in part by the Air Force Office of Scientific Research under Contract Nos. FA9550-10-1-0569, by the Air Force Research Laboratories under contract FA8750-12-2-0186, and by DARPA under Contract No. W911NF-12-1-0034.

References

[1] M. E. J. Newman, "The structure and function of complex networks," *SIAM REVIEW*, vol. 45, pp. 167–256, 2003.

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

[3] H. Kashima and N. Abe, "A parameterized probabilistic model of network evolution for supervised link prediction," in *ICDM*, 2006, pp. 340–349.

[4] C. Wang, V. Satuluri, and S. Parthasarathy, "Local probabilistic models for link prediction," in *ICDM*, 2007, pp. 322–331.

[5] T. Tylenda, R. Angelova, and S. Bedathur, "Towards time-aware link prediction in evolving social networks," in *SNA-KDD*, 2009, pp. 9:1–9:10.

[6] S. Gao, L. Denoyer, and P. Gallinari, "Temporal link prediction by integrating content and structure information," in *CIKM*, 2011, pp. 1169–1174.

[7] W. Cukierski, B. Hamner, and B. Yang, "Graph-based features for supervised link prediction," in *IJCNN*, 2011, pp. 1237–1244.

[8] S. Huang, M. Chen, B. Luo, and D. Lee, "Predicting aggregate social activities using continuous-time stochastic process," in *CIKM*, 2012, pp. 982–991.

[9] L. Weng, J. Ratkiewicz, N. Perra, B. Gonçalves, C. Castillo, F. Bonchi, R. Schifanella, F. Menczer, and A. Flammini, "The role of information diffusion in the evolution of social networks," in *SIGKDD*, 2013, pp. 356–364.

[10] J. Ye, H. Cheng, Z. Zhu, and M. Chen, "Predicting positive and negative links in signed social networks by transfer learning," in *WWW*, 2013, pp. 1477–1488.

[11] G.-J. Qi, C. C. Aggarwal, and T. Huang, "Link prediction across networks by biased cross-network sampling," in *ICDE*, 2013.

[12] P. Sarkar, D. Chakrabarti, and M. I. Jordan, "Nonparametric link prediction in dynamic networks," in *ICML*, 2012.

[13] K. Narang, K. Lerman, and P. Kumaraguru, "Network flows and the link prediction problem," in *SNA-KDD*, 2013.

[14] L. Backstrom and J. Leskovec, "Supervised random walks: Predicting and recommending links in social networks," in *WSDM*, 2011.

[15] S. L. Payne, *The Art of Asking Questions*. Princeton University Press, 1951.

[16] N. J. Blunch, "Position bias in multiple-choice questions," *Journal of Marketing Research*, vol. 21, no. 2, pp. pp. 216–220, 1984.

[17] G. Buscher, E. Cutrell, and M. R. Morris, "What do you see when you're surfing?: using eye tracking to predict salient regions of web pages," in *SIGCHI*, 2009.

[18] S. Counts and K. Fisher, "Taking it all in? visual attention in microblog consumption," in *ICWSM*, 2011.

[19] N. Craswell, O. Zoeter, M. Taylor, and B. Ramsey, "An experimental comparison of click position-bias models," in *WSDM*, 2008.

[20] T. Hogg, K. Lerman, and L. M. Smith, "Stochastic models predict user behavior in social media," in *SocialCom*, 2013.

[21] N. Hodas and K. Lerman, "How limited visibility and divided attention constrain social contagion," in *SocialCom*, 2012.

[22] Y. Bengio, J.-F. Paiement, and P. Vincent, "Out-of-sample extensions for lle, isomap, mds, eigenmaps, and spectral clustering," in *NIPS*. MIT Press, 2003, pp. 177–184.

[23] H. Zhou and Y. Zhang, "EM vs MM: A Case Study," *Comput. Stat. Data Anal.*, vol. 56, no. 12, pp. 3909–3920, Dec. 2012.

[24] D. R. Hunter, K. Lange, D. O. Biomathematics, and H. Genetics, "A tutorial on mm algorithms," *Amer. Statist*, pp. 30–37, 2004.

[25] M. A. Hasan and M. J. Zaki, "A survey of link prediction in social networks," in *Social Network Data Analytics*, 2011, pp. 243–275.

[26] L. A. Adamic and E. Adar, "Friends and neighbors on the web," *SOCIAL NETWORKS*, vol. 25, pp. 211–230, 2001.

[27] S. Golder and S. Yardi, "Structural predictors of tie formation in twitter: Transitivity and mutuality," in *SocialCom*, Aug. 2010, pp. 88–95.

[28] D. M. Romero and J. M. Kleinberg, "The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter," in *ICWSM*, 2010.

[29] T. Lou, J. Tang, J. Hopcroft, Z. Fang, and X. Ding, "Learning to predict reciprocity and triadic closure in social networks," *ACM TKDD*, vol. 7, no. 2, pp. 5:1–5:25, aug 2013.

[30] L. Katz, "A new status index derived from sociometric analysis," *Psychometrika*, vol. 18, no. 1, pp. 39–43, March 1953.

[31] G. Jeh and J. Widom, "Simrank: a measure of structural-context similarity," in *SIGKDD*, 2002, pp. 538–543.

[32] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks*, vol. 30, no. 1–7, pp. 107–117, 1998.

[33] P. Symeonidis, E. Tiakas, and Y. Manolopoulos, "Transitive node similarity for link prediction in social networks with positive and negative links," in *RecSys*, 2010, pp. 183–190.

[34] R.-H. Li, J. X. Yu, and J. Liu, "Link prediction: the power of maximal entropy random walk," in *CIKM*, 2011, pp. 1147–1156.

[35] D. Yin, L. Hong, and B. D. Davison, "Structural link analysis and prediction in microblogs," in *CIKM*. ACM, 2011, pp. 1163–1168.

[36] S. A. Golder and A. Marwick, "A structural approach to contact recommendations in online social networks," SSM Workshop.

[37] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici, "Link prediction in social networks using computationally efficient topological features," in *SocialCom/PASSAT*, 2011, pp. 73–80.

[38] K. Bischoff, "We love rock 'n' roll: Analyzing and predicting friendship links in last.fm," in *WebSci*, 2012, pp. 47–56.

[39] T. Schreiber, "Measuring information transfer," *Physical review letters*, vol. 85, no. 2, pp. 461–464, 2000.

[40] G. Ver Steeg and A. Galstyan, "Information transfer in social media," in *WWW*, 2012, pp. 509–518.

[41] T. Hogg and K. Lerman, "Social dynamics of digg," *EPJ Data Science*, vol. 1, no. 5, June 2012.

[42] L. M. Smith, L. Zhu, K. Lerman, and Z. Kozareva, "The role of social media in the discussion of controversial topics," in *SocialCom*, 2013.

[43] L. Zhu, A. Galstyan, J. Cheng, and K. Lerman, "Tripartite graph clustering for dynamic sentiment analysis on social media," in *ACM SIGMOD/PODS*, 2014.

[44] D. Wang, C. Song, and A.-L. L. Barabási, "Quantifying long-term scientific impact." *Science (New York, N.Y.)*, vol. 342, no. 6154, pp. 127–132, Oct. 2013.