# Predicting Communication Intention in Social Networks

Charalampos Chelmis
*Department of Computer Science*
*University of Southern California*
*Los Angeles, CA, USA*
*chelmis@usc.edu*

Viktor K. Prasanna
*Ming Hsieh Department of Electrical Engineering*
*University of Southern California*
*Los Angeles, CA, USA*
*prasanna@usc.edu*

***Abstract*—In social networks, where users send messages to each other, the issue of what triggers communication between unrelated users arises: does communication between previously unrelated users depend on friend-of-a-friend type of relationships, common interests, or other factors? In this work, we study the problem of predicting directed communication intention between two users. Link prediction is similar to communication intention in that it uses network structure for prediction. However, these two problems exhibit fundamental differences that originate from their focus. Link prediction uses evidence to predict network structure evolution, whereas our focal point is directed communication initiation between users who are previously not structurally connected. To address this problem, we employ topological evidence in conjunction to transactional information in order to predict communication intention. It is not intuitive whether methods that work well for link prediction would work well in this case. In fact, we show in this work that network or content evidence, when considered separately, are not sufficiently accurate predictors. Our novel approach, which jointly considers local structural properties of users in a social network, in conjunction with their generated content, captures numerous interactions, direct and indirect, social and contextual, which have up to date been considered independently. We performed an empirical study to evaluate our method using an extracted network of directed @-messages sent between users of a corporate microblogging service, which resembles Twitter. We find that our method outperforms state of the art techniques for link prediction. Our findings have implications for a wide range of social web applications, such as contextual expert recommendation for Q&A, new friendship relationships creation, and targeted content delivery.**

***Keywords*-social networking analysis; corporate microblogging; communication intention; social factors; prediction;**

## I. Introduction

Link prediction refers to the problem of predicting the existence of a link between two entities in an entity relationship graph, by calculating entity similarity based on entity attributes [1] and the graph structure [2]–[4]. Social link prediction in particular has gained a lot of attention, since it can assist users into discovering and making new friends, improving their overall user experience. On the other hand, using link prediction, companies exploit social networking sites for monetization, by selectively expanding their targeted user base and by triggering targeted advertising campaigns.

Link prediction in social networks is a challenging problem, as social networking data is inherently noisy and heterogeneous. Overall, social networking users provide scarce information about their interests in their profiles, which are often incomplete and obsolete. Further, user-generated published content mostly comprises of free, unstructured text, which often does not adhere to grammatical and syntactical rules, contains slag terms and abbreviations, and is often of restricted length (e.g. 140 characters in Twitter). Hence, social networking content includes useful information for social link prediction, but this information is not well structured, and is often misleading or ambiguous. For this reason, most social link prediction approaches calculate graph-based proximity scores [2], [3], asserting that the "closer" two nodes are in the social graph, the more likely they are to become linked in the future. Liben-Nowell et al. [3] showed that the Adamic/Adar metric performs best in scientific coauthorship networks, while Backstrom et al. [5] introduced a supervised link prediction approach based on supervised random walks.

While social networking data present challenges in social link prediction, they also exhibit a wealth of information to be used for that cause. Social networking data, often have some sort of "context" associated with them, including user provided annotations (e.g. description, hashtags) and system information (e.g. upload time). "Individual features might be noisy or unreliable but collectively they provide revealing information" [6] about users. Schifanella et al. [1] showed that strong correlations exist between annotations and social proximity, and used semantic similarity between user annotations as statistical predictors of friendship links.

Network proximity metrics measure the likelihood of an interaction between two users, regardless of the existence of a path between them. Proximity metrics used in prior work include neighborhood based methods and methods based on the ensemble of all paths [3]. The greater the number of paths connecting two users, the closer they are considered to be in the network. However, information spread in social networks depends, not only on the underlying network structure, but also on the information itself, and the nature of the process by which nodes interact [7]. In social media, users broadcast information to all their neighbors (i.e. one-

to-many interactions) or specific groups (e.g. groups of friends in Facebook or circles in Google+), whereas public status updates in social networks may be distributed to every single user (i.e. broadcast). One-to-one interactions exist in the form of directed messages between users (e.g. @reply messages in Twitter).

We propose a technique that takes into account both structural properties of the social network and user interactions (both explicit and implicit) in order to predict initiation of communication between nodes. We compare our approach to state of the art methods used in linked prediction on a corporate microblogging service, which resembles Twitter. Our problem differs from social link prediction. Instead of trying to give an answer to the classification problem of whether an edge between users $u$ and $v$ exists or not, we are trying to understand the factors behind the intension of user $u$ to sent a direct message to user $v$. An edge represents a conversation between users rather than friendship and its directionality matters; it depends on the user who starts the conversation and is asymmetric ($e_1 = (u, v) \neq e_2 = (v, u)$). Further, each edge is created under specific context (e.g. a message sent under a specific topic in group $g_1$, using a set of hashtags $S_{t_1}$). It is not intuitive whether methods that work well for link prediction would work as well in this setting.

In this paper we make the following contributions:

- We define the communication intention prediction problem for social networks.
- We propose a new framework for proximity calculation in directed graphs, which jointly considers local structural properties of the nodes along with direct and indirect, semantically enriched interactions between them.
- We evaluate our approach on a @replies network, inferred from a corporate microblogging service, demonstrating that communication intention prediction can be accurately performed.

The paper is organized as follows. We first discuss relevant prior work. Next, we introduce our methodology for representing users within a social network, and we describe our algorithm for computing semantic similarity between users. We then discuss in detail the results of a quantitative experimental study, designed to discover which factors trigger communication between unrelated users, and measure how the importance of such factors changes with respect to different users.

## II. Related Work

The problem of link prediction for social networks has been well studied. Liben-Nowell et al. [3] explored several similarity metrics for social link prediction. Markines et al. [8] systematically analyzed semantic similarity measures based on folksonomies. Instead, we are using an augmented social graph, into which many different object types coexist simultaneously. Schifanella et al. [1] utilized vocabulary overlap between users as indicator of user connectivity in Flickr. We are extending this hypothesis to other aspects of users activity, semantically enriching them, and considering them in conjunction to local network structure.

Jacovi et al. [9] recommended "interesting" people using prior evidence of user interests in terms of following and tagging. Chen et al. [10] recommended new friends, based either on social or content similarity. Their approach required at least one prior interaction between users and employed simple keyword matching schemes. Instead, we do not assume prior relationships between users. Further, we are trying to predict communication, which does not entail friendship (and vice versa), and to understand the factors that trigger such communication, based on semantically enriched content and semantic similarity, in conjunction to network proximity.

Different approaches for social link prediction based on random walks include [5], [11], [12]. Such techniques mainly focus on the social network structure and are computationally expensive. Fire et al. [13] addressed the computational cost, using efficient topological features. Probabilistic inferencing approaches include [14], [15], but also exhibit high computational cost. Machine learning approaches include [16]–[18]. All such approaches however do not exploit the full extend of information available in a social network. Sadilek et al. [19] used network structure, content, and user location to infer social ties. Their probabilistic model relies on a dense representation of all possible, symmetric friendships, and requires training. Instead, we are focusing on asymmetric, directed communication. Our approach does not require training to "learn" probability distributions for every node pair, but can dynamically keep track of changes and recompute pairwise similarities incrementally, when necessary.

Perhaps the most close work to ours is [20]. Their focus, however, is on information diffusion around topics for given time slice, having past communication evidence. Sousa et al. [21] investigated factors that motivate users to reply to other users in Twitter, whereas our focal point is to determine what motivates users to initiate a conversation with others.

## III. Communication Intention Prediction in Microblogging Sites

Twitter is a social networking site which allows users to share their status updates as well as interact with others by sending short text messages. Users can follow other people and have followers themselves. The following relationship in this case may not be reciprocal. Users can "retweet" posts, make use of #hashtags, and directly address a message to another user ('@' followed by a username). Such messages are referred to as @replies. In this setting, users interact either directly or indirectly. Users are explicitly connected when there is a "following" relationship (social link) between them. Users are implicitly connected through indirect

activities, such as common use of #hashtags, retweets, and @replies. We can infer a directed network from @replies messages, representing users' interaction flow [21].

In this paper we aim to understand what motivates communication between users. In particular, we seek to answer the following question: "What makes people sent @reply messages to *strangers*?". Our hypothesis is that information on the @reply network can help predict users' intention to communicate in microblogging services. To this end, we consider a directed social graph $G = (V, E)$, where each node $u \in V$ represents a user, and an edge $e = (u, v) \in E$ exists if and only if user $u$ has sent at least one @reply message to user $v$. Each edge may have a weight $w_{uv}$ attached to it, such that $w_{uv}$ equals the frequency of replies sent from user $u$ to user $v$. We have chosen this intuitive definition for edges so as to represent the "transer" of content from user $u$ to user $v$ when user $u$ sends user $v$ a @reply message. An undirected edge $e_{uv}$ between users $u$ and $v$ if either user sent a message to the other would not capture the semantics of directed communication, which may or may not be reciprocal.

We formulate the communication intention prediction problem as follows:

**Communication Intention Prediction Problem Definition.** *Given $G' = (V', E')$, a subgraph of $G$ consisting of all nodes in $G$ ($V' \equiv V$) and a subset of edges in $G$ ($E' \subset E$), output a ranked list $L$ of edges (links), not present in $G'$, that are predicted to appear in $G$, such that $E' \cup L \equiv E$.*

## IV. User Representation

Social networks provide a variety of contextual features, that are dependent on the type of the resource (e.g. photos may be geotagged). We consider a representation of microblogs using each feature according to its type, adapting the definition of event from [22]. Using this definition, each microblog is characterized by a set of attributes, both textual and non-textual, some of which are unique for each post, while others may be missing or have multiple instances. For example, each post has exactly one date attached to it. Other features, such as hashtags, may be arbitrarily many.

**Textual Features:** These features include raw textual content (bag-of-words), as well as user provided hashtags and group participation. We formally represent raw textual content using tf.idf weight vectors [23] and then utilize the cosine similarity metric to compute similarity between such vectors. We clean the text by performing stemming and basic stop-word elimination. The use of cosine similarity lacks semantics and ignores semantic associations between terms with similar meaning but poor lexical similarity. Hashtags are meant to be a selective set of highly descriptive keywords of the content of microblogs. Groups are indicative of communities of interest. Stemming and/or stop-word removal, and cosine similarity do not seem appropriate for them, hence we are calculating semantic similarity for these facets.

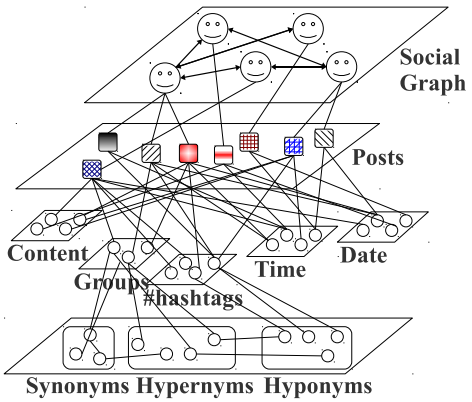



Figure 1. Augmented social graph.

**Temporal Features:** These features regard the date and time a post was made. We represent date values as the number of minutes elapsed since the Unix epoch.

A user can be modeled as a union of her connections and her content, based on the features described above. Using this user model, we form an augmented, directed social graph, presented in Figure 1. In the rest of this section we describe in detail our approach, which consists of calculating users proximity through aggregation over their microblogs similarity and similarity with respect to their network neighborhood.

### A. Semantic Similarity of Textual Features

To compute semantic similarity between hashtags (similarly for groups), we utilize WordNet - a lexical database for English [24]. The WordNet toolkit permits search of relevant concepts in terms of conceptual, semantic and lexical relations: *a*) Synonyms: terms that denote the same concept (e.g. "car" - "automobile"); *b*) Hypernyms: more general concepts (e.g. "furniture" is a hypernym of "bed"); and *c*) Hyponyms: more specific concepts (e.g. "bed" is a hyponym of "furniture").

*Semantic Similarity of Concepts*: Given two concepts $a$ and $b$, let $S_a$ denote a set of terms (specified below) that describe $a$, and $S_b$ a set of terms that describe $b$. The similarity $s(a, b)$ between $a$ and $b$ is then defined as the Jaccard index:

$$s(a, b) = s(S_a, S_b) = \frac{|S_a \bigcap S_b|}{|S_a \bigcup S_b|}, \quad (1)$$

where $|\cdot|$ is set cardinality, $\bigcup$ is set union, and $\bigcap$ denotes set intersection. It holds that $0 \leq s(a, b) \leq 1$, and $s(a, b) = 1$ if $S_a$ and $S_b$ are identical, and $s(a, b) = 0$ if they do not share any terms at all. We use this similarity measure, which has been found to be a good trade-off between simplicity and performance [25], to calculate similarity between textual concepts $a$ and $b$. The system returns all synonym concepts of $a$, denoted with $S_a$, as well as all synonym

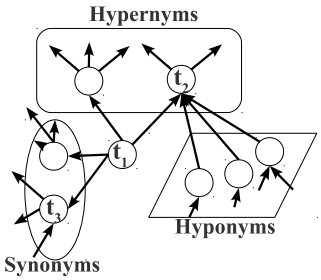



Figure 2. Example of hashtag hierarchy.

concepts of $b$, denoted as $S_b$. We define the *synonym-based* similarity between concepts $a$ and $b$ as $s_s(a,b) = s(a,b) = s(S_a, S_b)$. Similarly, we define *hypernym-based* similarity as $s_h(a,b) = s(a,b) = s(H_a, H_b)$ and *hyponym-based* similarity as $s_{hp}(a,b) = s(a,b) = s(Hp_a, Hp_b)$.

The semantic similarity between two hashtags (same for groups) $a$ and $b$ can then be computed as the weighted sum of the measures described above. This metric however does not discriminate between cases where hashtags belong to the same subtree as shown in Figure 2 (e.g. $t_2$ is a hypernym of $t_1$). To resolve this issue, we compute similarity between the union of annotations for each hashtag. To account for lexical similarity between hashtags $a$ and $b$, we consider their Levenshtein similarity. We use the $\max$ operator to select the highest similarity, either semantic or lexical. Formally, we define the semantic similarity between two hashtags $a$ and $b$ ($s_g$ for groups) as follows:

$$\begin{aligned} s_{tg}(a,b) \;=\; & \max\{LevenshteinSimilarity(a,b), \\ & w_s s_s(a,b) + w_h s_h(a,b) + w_{hp} s_{hp}(a,b), \\ & s(S_a \cup H_a \cup Hp_a, S_b \cup H_b \cup Hp_b)\}, \end{aligned} \tag{2}$$

where $w_s = w_h = w_{hp} = 1/3$. We chose symmetric weights, since we did not find any particular reason to weight differently the similarity contribution of synonyms, hypernyms, and hyponyms.

*Textual Similarity:* We compute textual similarity $s_{tx}(x,y)$ between two bag-of-words $x$ and $y$, represented as $tf.idf$ weight vectors, using cosine similarity [23].

### B. Date Similarity

We compute similarity between dates $d_1$ and $d_2$ as follows:

$$s_d(d_1,d_2) = \begin{cases} 0 & \text{if } |d_1 - d_2| \geq T_d \\ 1 - \frac{|d_1 - d_2|}{T_d} & \text{otherwise} \end{cases}, \tag{3}$$

where $T_d = 365$. In other words, if $d_1$ and $d_2$ are more than one year apart, we define their similarity as zero. Otherwise, we define their similarity as one minus their difference in days.

### C. Time Similarity

We compute similarity between time instances $t_1$ and $t_2$ as follows:

$$s_t(t_1,t_2) = \begin{cases} 0 & \text{if } |t_1 - t_2| \geq T_t \\ 1 - \frac{|t_1 - t_2|}{T_t} & \text{otherwise} \end{cases}, \tag{4}$$

where $T_t = 86400$. In other words, if $t_1$ and $t_2$ are more than one day apart, we define their similarity as zero. Otherwise, we define their similarity as one minus their difference in seconds.

Overall, we compute similarity between timestamps $x$ and $y$ as $s_{df}(x,y) = w_d s_d(x_d, y_d) + w_t s_t(x_t, y_t)$, where $s_d(.,.)$ is calculated using Equation 3, $s_t(.,.)$ is calculated using Equation 4, and $w_d + w_t = 1$. Different $T_d$ and $T_t$ values may yield optimal results for different datasets. We leave users the ability to set these thresholds according to their respective needs.

### D. Feature Set Similarity

We use a variation of Hausdorff point set distance measure [26] to calculate similarity between two sets of features $A : \{a_1, a_2, ..., a_m\}$ (e.g. $X$'s hashtags) and $B : \{b_1, b_2, ..., b_n\}$ (e.g. set of hashtags associated with post $Y$), as follows:

$$S_H(A,B) = \frac{1}{|A|} \sum_{k=1}^{|A|} \max_i \{sim(a_k, b_i)\}, \tag{5}$$

where $sim(a_k, b_i)$ is any similarity measure on any two set elements $a_k$ and $b_i$. This is the average of the maximum similarity of features in set $A$ with respect to features in set $B$ [22]. Like the original Hausdorff distance metric, this similarity measure is asymmetric with respect to the sets: $S_H(A,B) \neq S_H(B,A)$.

### E. Content Proximity

To compute similarity between two posts $p_1$ and $p_2$ we compute the similarity between each of their attributes respectively. Combining all similarity measures described above in a weighted sum, we get the similarity between two posts $p_1$ and $p_2$ as follows:

$$\begin{aligned} S(p_1,p_2) \;=\; & w_g s_g(p_{1_g}, p_{2_g}) + w_{tg} S_H(p_{1_{tg}}, p_{2_{tg}}) + \\ & w_{tx} s_{tx}(p_1,p_2) + w_{df} s_{df}(p_1,p_2), \end{aligned} \tag{6}$$

where $w_{df} + w_g + w_{tg} + w_{tx} = 1$. In our experiments we consider numerous weighting schemes and report our observations.

The similarity measure between two users $u$ and $v$ with respect to their microbloggs can then be computed using the modified Hausdorff distance as follows:

$$S_C(u,v) = \frac{1}{|u_p|} \sum_{k=1}^{|u_p|} \max_i \{S(u_{p_k}, v_{p_i})\}, \tag{7}$$

where $u_p$ denotes the set of user $u$'s microbloggs. Our content proximity metric is easily extensible to other types

of resources, such as documents, videos etc., that have contextual features attached to them.

### F. Network Proximity

To compute similarity between two users with respect to their network proximity we considered numerous proximity methods proposed in the literature. For simplicity and reduced complexity, we chose to use a modification of the common neighbors metric. We define network proximity between users $u$ and $v$ as:

$$S_N(u,v) = s(\Gamma_u, \Gamma_v) = \frac{|\Gamma_u \cap \Gamma_v|}{|\Gamma_u|}, \tag{8}$$

where $\Gamma_u$ denotes the set of $u$'s neighbors. Normalizing by $|\Gamma_u|$, $S_N$ becomes asymmetric with respect to users. This way closeness is calculated on the premises of the percentage of common neighbors instead of the absolute number, with higher percentage indicating greater intersection of common interests.

### G. User Similarity

We calculate user similarity as a weighted function of content and network proximity. We define similarity between two users $u$ and $v$ as:

$$S(u,v) = \lambda S_C(u,v) + (1-\lambda) S_N(u,v), \tag{9}$$

where $\lambda$ controls the tradeoff between content and network proximity.

For our prediction problem, we first construct the augmented social graph $G(V,E)$. Given a user $u$, we compute user similarity in a top down fashion for all facets, for all $u$'s posts with respect to all other users in the network that do not belong in the set of user $u$'s direct contacts, using Equations 7-9. To reduce complexity, we can restrain similarity calculation to users being up to distance $d$ from user $u$, instead of considering the complete user corpus.

## V. Data set

Our dataset is a complete snapshot (June 2010 - August 2011) of a corporate micro-blogging service, which resembles Twitter, consisting of 4,213 unique users, who have posted 16,438 messages in total, distributed over 8,139 threads and 88 groups. Out of all messages, 8,174 are broadcast (e.g. status updates) and 8,264 are @replies. The number of unique hastags is 637. The corporate micro-blogging site does not impose any restrictions on the way people interact or who they chose to follow, much similar to Twitter. We inferred a network based on the @replies messages, represented as a directed graph $G = (V,E)$, where each node $u \in V$ represents a user, and each edge $e = (u,v) \in E$ exists if and only if user $u$ has sent at least one @reply message sent to user $v$. Of the 4,213 total users, 582 belong in the largest connected component, contributing 11,684 messages and sharing 3,773 edges. We did not impose restrictions on the minimum number of neighbors or messages per user, resulting in a variety of users' activity ranging from a minimum of one message per user to over 400 messages maximum.

## VI. Data Characteristics

The existence of edge $e_{ij}$ does not guarantee that the reciprocal edge $e_{ji}$ also exists. Hence, the relationship is not symmetric. If user A sends a message to user B, the edge $e_{AB}$ is created, but not vise versa. We call user B the "follower" of user A. If B also replies to A, then they become each other's "mutual followers". Figure 3 shows the scatter plot of the number of followees versus the number of followers. The points are scattered around the diagonal, indicating equal numbers of followers and followees. The cumulative distribution of the out-degree to in-degree ratio exhibits high correlation between in-degree and out-degree, which can be explained as a result of symmetric links being created due to the tendency of users to reply back when they receive a message from other users.

We further examine the probability distributions of the number of messages $n_m$ and the number of replies $n_r$ per user, the distribution of the number of groups $n_g$ to which a post belongs and the probability of finding a user with a number $n_t$ of distinct hashtags in his vocabulary, as well as the total number $t$ of hashtag assignments per user (a hashtag used twice is counted twice) and the total number $g$ of group related messages per user (the number of messages sent to a group). More precisely, if $f_u(t)$ is the frequency of hashtag $t$ being used by user $u$, then the total number of hashtag assignments of user $u$ is given by: $t_u = \sum_t f_u(t)$. Similarly, if $f_u(g)$ is the number of times user $u$ has sent a message to group $g$ (either privately to another group member or broadcast to the group), then the total number of group messages of user $u$ is given by: $g_u = \sum_g f_u(g)$. All activities show behavior consistent with power law networks; the majority of users show small activity patterns with few nodes being significantly more active. All distributions are broad, indicating that the activity patterns of users are highly heterogeneous.

### A. Correlations Between Features and the Network

We now examine correlations between user activities and the structure of the @replies network. Specifically, we investigate if there is a connection between the number of neighbors a user has and the activity patterns of such user. We characterize the average activity of users with $k$ neighbors (we consider in-degree and out-degree separately) in the @replies network using the following quantities: (i) the average number of messages $n_m$ of users with $k$ neighbors, (ii) the average number of replies $n_r$ of users with $k$ neighbors, (iii) the average number of distinct groups $n_g$ (similarly for total number of group messages) of users with $k$ neighbors, (iv) the average number of distinct hashtags

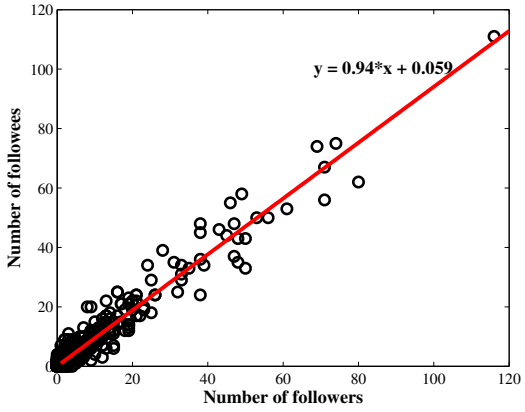



Figure 3. Scatter plot of the number of followers and the number of followees.

$n_t$ (similarly for total hashtag assignments) of users with $k$ neighbors. For example,

$$n_m(k) = \frac{1}{|u : k_u = k|} \sum_{u:k_u=k} n_m(u). \qquad (10)$$

Figure 4 shows the probability distributions of such quantities. All activities have an increasing trend for increasing values of $k$ (both for in-degree and out-degree). Large fluctuations can be observed for large values of $k$ due to the fewer highly connected users over whom the averages are performed. Notably, the average number of messages and replies are very well correlated to the number of neighbors, as is the average number of distinct groups. The average number of (distinct) hashtag assignments exhibits more heterogeneity than the other measures, but overall the trend is increasing with increasing values of $k$. Users with many contacts but using very few hashtags and sending very few group messages can however be observed.

### B. Homophily as a Function of Network Proximity

We examine user similarity in terms of hashtag usage, with respect to their distance in the @replies network. We argued earlier that users of the corporate micro-blogging service mostly hashtag their own content. This observation along with the personal character of hashtaging make us anticipate that there will be no global hashtag vocabulary across users, commonly found in social bookmarking sites [27], [28], or if such a vocabulary exists, it will be extremely incoherent. To test the existence of a globally shared vocabulary, we selected pairs of users at random and measured the number of their shared hashtags, which, on average is $\approx 1.001$.

Even though random pairs of users don't have common hashtags, adjacent users in social networks tend to share common interests, a property known as homophily [29], [30] or assortative mixing [31]. We measure user homophily

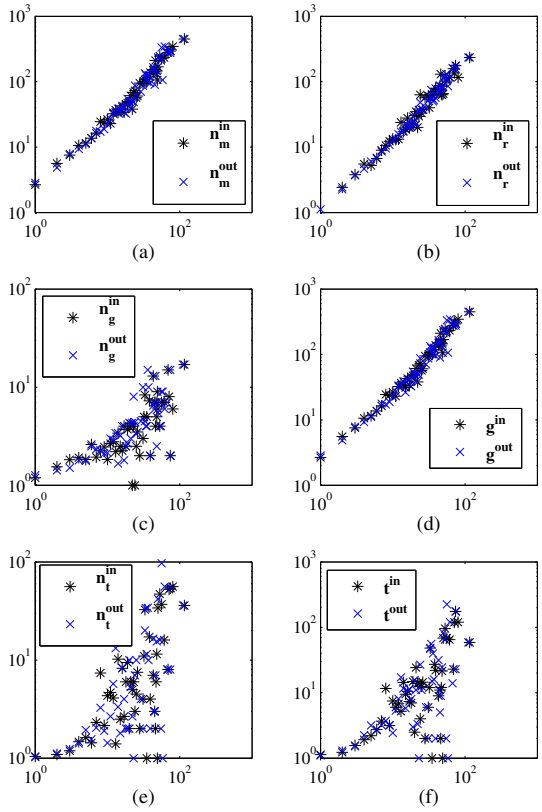



Figure 4. From left to right and top to bottom, average number of (a) messages $n_m$, (b) replies $n_r$, (c) distinct groups $n_g$ and (d) groups $g$, (e) distinct hashtags $n_t$ and (f) total hashtag assignments $t$ of users having $k$ neighbors in the @replies network.

with respect to hashtags as a function of users' distance in the @replies network. We regard hashtag assignments of user $u$ as a feature vector, whose elements correspond to hashtags and whose entries correspond to frequencies of hashtag usage for user $u$. Hence, the normalized similarity between two users $u$ and $v$ with respect to their hashtag vectors, $\sigma_{hashtags}(u, v)$ can be computed as follows:

$$\sigma_{hashtags}(u, v) = \frac{\sum_t f_u(t) f_v(t)}{\sqrt{\sum_t f_u(t)^2 \sum_t f_v(t)^2}}, \qquad (11)$$

$\sigma_{hashtags}(u, v)$ is equal to 0 if users $u$ and $v$ have no hashtags in common, and 1 if they have used exactly the same hashtags. We further define the normalized similarity between two users $u$ and $v$ with respect to their distinct hashtag usage, $\sigma_{U_{hashtags}}(u, v) = \frac{\sum_t \delta_u^t \delta_v^t}{\sqrt{n_t(u) n_t(v)}}$, where $n_t(u)$ is the total number of distinct hashtags of user $u$ and $\delta_u^t = 1$ if user $u$ has used hashtag $t$ at least once, and 0 otherwise.

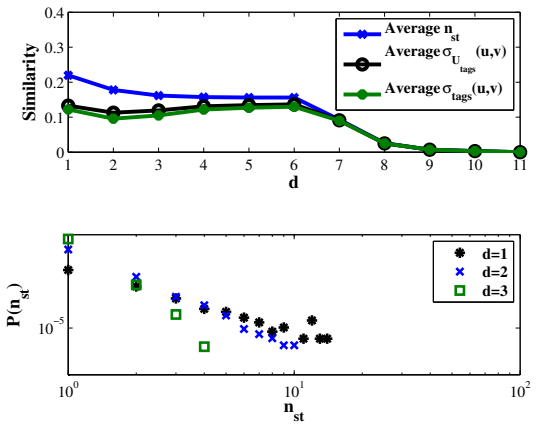



Figure 5. Top: Average number of shared hashtags $n_{st}$, $\sigma_{hashtags}(u,v)$, and $\sigma_{U_{hashtags}}(u,v)$ of two users as a function of their distance $d$ in the network. Bottom: Probability distribution of the number of shared hashtags $n_{st}$ of two users being at distance $d$ on the network, for $d = 1, 2, 3$.

To compute averages of the aforementioned similarities we performed an exhaustive investigation of the @replies network up to distance equal to the network diameter. Figure 5 demonstrates the dependency of user similarity on distance, by showing the average number of shared hashtags and the corresponding average cosine similarities of two users as a function of $d$. The average number of shared hashtags remains almost constant for $d \leq 6$, after which point it drops rapidly[1]. High hashtag alignment is observed between neighbors for greater distance than traditional online social networks [1], due to the fact that inside the corporate microblogging site users exhibit more focused interests aligned with their discipline, day to day responsibilities and ongoing projects.

## VII. Evaluation

We demonstrate the effectiveness of our framework by removing some of the edges in the @replies network and recommending the links based on the pruned graph. We use four-fold cross validation by randomly dividing the set of edges into four partitions, use one partition for prediction, and retain the links in the other partitions. We randomly sample 100 users and recommend the top-$k$ links for each user. We use precision, recall and mean reciprocal rank (MRR) for reporting accuracy. We measure precision at $k$ as: $P@k = \frac{1}{|S|}\sum_{p\in S}\frac{N_k(p)}{k}$, where $S$ is the set of sampled users and $N_k(p)$ is the number of truly linked persons in the top-$k$ list of user $p$. Similarly, we measure recall at $k$ as $R@k = \frac{1}{|S|}\sum_{p\in S}\frac{|F_p \cap R_p|}{|F_p|}$, where $F_p$ denotes the truly linked user set of person $p$ and $R_p$ denotes the set of recommended users of person p ($|R_p| = k$). Finally, we measure MRR at $k$ as $MRR@k = \frac{1}{|S|}\sum_{p\in S}\frac{1}{rank_p}$, where $rank_p$ denotes the rank of the first correctly recommended link of user $p$.

[1]We got similar results when we examined users homophily with respect to groups as a function of their distance in the @replies network

Table I
Weighting schemes.

| Metric | $w_{tx}$ | $w_{tg}$ | $w_g$ | $w_{df}$ | $w_d$ | $w_t$ |
|---|---|---|---|---|---|---|
| SS_Uniform | 0.25 | 0.25 | 0.25 | 0.25 | 0.5 | 0.5 |
| SS_Tags | 0.0 | 1.0 | 0.0 | 0.0 | 0.5 | 0.5 |
| SS_Groups | 0.0 | 0.0 | 1.0 | 0.0 | 0.5 | 0.5 |
| SS_Text | 1.0 | 0.0 | 0.0 | 0.0 | 0.5 | 0.5 |
| SS_Time_1 | 0.0 | 0.0 | 0.0 | 1.0 | 0.8 | 0.2 |
| SS_Time_2 | 0.0 | 0.0 | 0.0 | 1.0 | 0.5 | 0.5 |
| SS_Time_3 | 0.0 | 0.0 | 0.0 | 1.0 | 0.2 | 0.8 |
| SS_Mix_1 | 0.2 | 0.3 | 0.4 | 0.1 | 0.8 | 0.2 |
| SS_Mix_2 | 0.2 | 0.5 | 0.2 | 0.1 | 0.8 | 0.2 |
| SS_Mix_3 | 0.3 | 0.45 | 0.1 | 0.15 | 0.8 | 0.2 |

We compare our approach against four baseline approaches described below:

- **Random Selection:** Randomly select a pair of users and create a link between them.
- **Shortest Distance:** Create a link between user $u$ and the user $v$ closest to him (length of shortest path).
- **Common Neighbors:** $\sigma(u,v) = |N(u)\bigcap N(v)|$, where $N(u)$ is the set of neighbors of user $u$ in the social network.
- **Shared Vocabulary:** Following [32] and our analysis on user homophily as a function of network proximity, we regard the vocabulary of a user as a feature vector whose elements correspond to hashtags and whose entries are the hashtag frequencies for that specific user's vocabulary. To compare the hashtag feature vectors of two users, we use the standard cosine similarity defined in Equation 11.

We use SS_Uniform to denote our method using a uniform weighting scheme. We experiment with multiple weighting schemes, resulting in numerous variations of our approach, shown in Table I.

### A. Methods Comparison

Here we compare the accuracy of our conversation initiation prediction scheme (SS_Uniform) against the baselines. Tables II and III list average precision, recall and MRR as calculated over our four-fold cross validation experiment for 100 randomly chosen users. We indicate the best performing baseline, against which we compute percentage lift, i.e. the % improvement that our method achieves over the best performing baseline.

Random selection performs the worst as expected, since the more users and edges in the graph the tougher it becomes to recommend correct links by random selection. Shortest Distance also performs poorly, while Common Neighbors perform slightly better. Even though we do not report results for the Adamic/Adar and Katz metrics, they perform as bad as common neighbors. This indicates that graph structure

Table II
PREDICTION PRECISION ACHIEVED BY DIFFERENT METRICS.

| Metric | $P@1$ | $P@5$ | $P@10$ | $P@20$ | $P@50$ |
|---|---|---|---|---|---|
| Random | 0.0070 | 0.0036 | 0.0057 | 0.0059 | 0.0048 |
| Shortest Distance | 0.0716 | 0.0716 | **0.0643** | **0.0492** | 0.0303 |
| Common Neighbors | **0.1050** | **0.0768** | 0.0637 | 0.0487 | **0.0303** |
| Shared Vocabulary | 0.0327 | 0.0318 | 0.0247 | 0.0193 | 0.0141 |
| SS_Uniform,$\lambda = 0.8$ | 0.162 | 0.109 | 0.089 | 0.066 | 0.039 |
| Precision Lift % | **54.29** | **41.93** | **38.41** | **34.14** | **28.71** |

Table III
PREDICTION RECALL AND MRR ACHIEVED BY DIFFERENT METRICS.

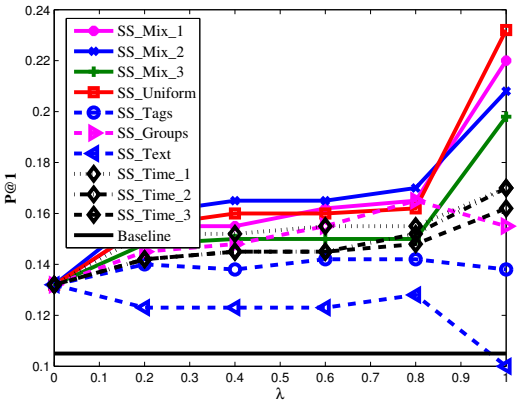
| Metric | $R@1$ | $R@10$ | $MRR@1$ | $MRR@10$ |
|---|---|---|---|---|
| Random | 0.0020 | 0.0021 | 0.0067 | 0.0016 |
| Shortest Distance | 0.0269 | **0.0204** | 0.0716 | 0.0156 |
| Common Neighbors | **0.0321** | 0.0198 | **0.1050** | **0.0177** |
| Shared Vocabulary | 0.0062 | 0.0069 | 0.0327 | 0.0068 |
| SS_Uniform,$\lambda = 0.8$ | 0.162 | 0.283 | 0.162 | 0.25 |
| Lift % | **404.67** | **1287.25** | **54.28** | **1312.43** |

has some predictive power but is insufficient by itself to perform well. Intuitively, close proximity due to short @replies path does not necessarily entail a direct @reply message to be sent. Shared Vocabulary is comparable in accuracy to network-based metrics, even though it performs worse overall, most probably due to the high hashtag alignment which is observed between neighbors for distance $d \leq 6$ in this dataset. Hence, vocabulary commonality alone is not indicative of communication intension either, but could potentially prove complementary to structural features. Nonetheless, all approaches exhibit a drop in accuracy as a function of $k$, which can be explained by the average degree of nodes in our dataset. It is impossible to get more correct results in the top-$k$ list once the maximum value of correct neighbors is reached.

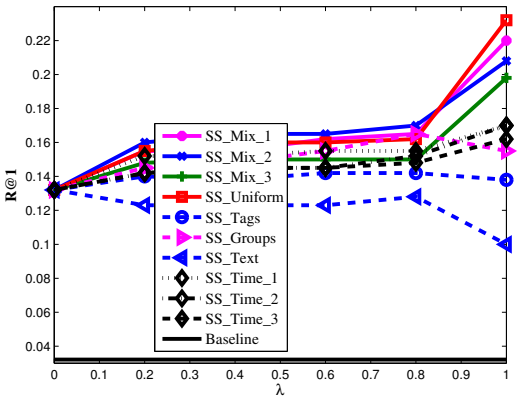
Our approach outperforms the baselines with respect to all accuracy metrics, often by a considerable margin, by mediating local structural characteristics with content and rich semantics about content's metadata. Although average precision achieved by our approach appears low, ranging from 16.2% for $k = 1$ to 3.9% for $k = 50$, it is at least an order of magnitude higher than precision achieved by baselines. Similarly, our approach performs better in terms of recall and MRR, where it achieves the best improvement over the best performing baseline. Note that for $k = 50$, the recall of our approach is **50%**. Precision values follow a heavy-tailed distribution, indicating a strong difficulty in making accurate predictions for some users, while achieving very high precision (100% or close to 100%) for others.

### B. Weight Scheme Selection

There are two types of parameters in our approach: $\lambda$, and six weighting factors ($w_{tx}$, $w_{tg}$, $w_g$, $w_{df}$, $w_d$, $w_t$) each controlling the significance of different facets into the proximity calculation. Different datasets may lead to




(a) Precision.



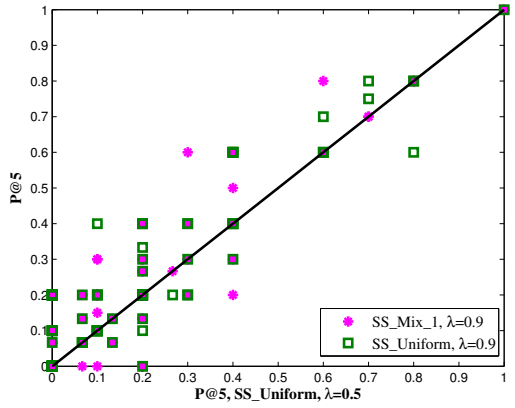

(b) Recall.

Figure 6. (a) Precision @1 and (b) Recall @1 as a function of λ..

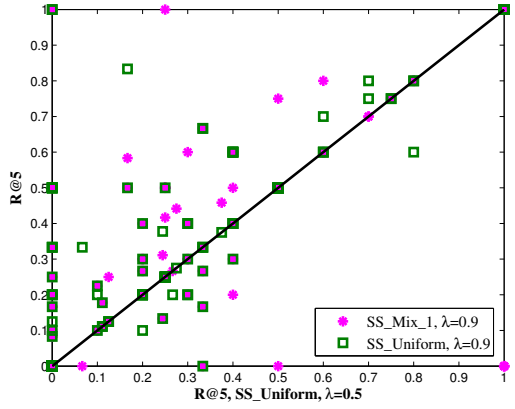
different optimal values for these parameters. We obtain the best values for our dataset by performing a grid search over ranges of values for these parameters and measuring accuracy on the validation set for each of the configuration settings. Table I lists some of the weighting schemes we experimented with.

*1) Effect of Parameter λ:* Parameter $\lambda$ controls the trade-off between structural proximity and content similarity. The higher its value the more significance is given to content similarity. A value of 0 only considers network proximity, whereas a value of 1 only considers content similarity. Figure 6 demonstrates the effect of parameter $\lambda$ in Precision and Recall @1, achieved by the weighting schemes presented in Table I.

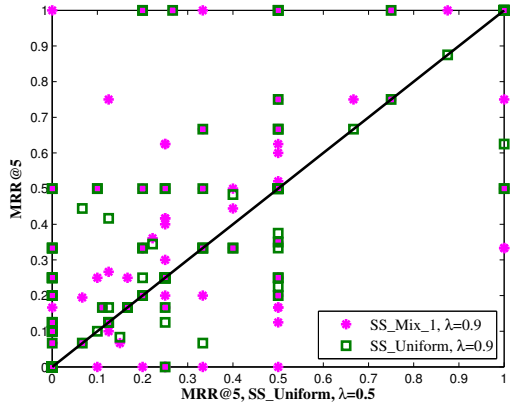
Schemes which consider only one type of content facet (i.e. SS_Uniform, SS_Tags, SS_Groups, SS_Text, and the three SS_Time schemes) perform better than the baseline, since they still combine network and content proximity scores to make a good prediction. Interestingly, time schemes perform better than schemes considering hastags or




(a) Precision.

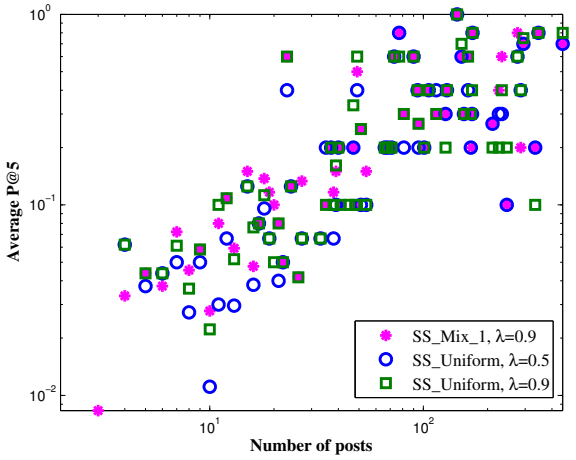



(b) Recall.

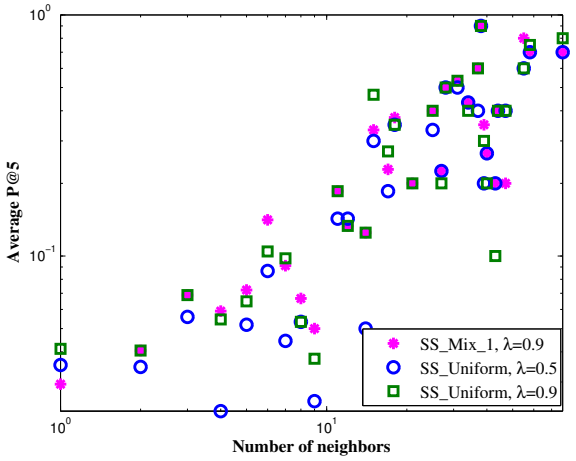



(c) MRR.

Figure 7. Impact of weighting scheme on accuracy (measured @5).




(a) Precision as a function of content.




(b) Precision as a function of network structure.

Figure 8. Average precision (measured@ 5) of users having $k$ (a) posts or (b) neighbors in the @replies network.

text alone, but have inferior performance than SS_Groups. This indicates that timing between replies is essential in this dataset, an outcome which can be explained as a result of the corporate environment, which requires prompt answers.

Among the mix schemes, SS_Mix_3 performs worse, probably due to the discounted weighting of the group facet. SS_Mix_2, which gives more emphasis on the hashtags and treats equally the textual and group facets is the best performing weighting scheme, apart from when $\lambda = 1$. In this case, the uniform scheme outperforms the rest. Nonetheless, all weighting schemes considerably outperform the baseline. The $\lambda$ value that provides the best tradeoff between content and network appears to be $\lambda = 0.8$ (when also considering $k \in \{5, 10, 20, 50\}$).

*2) Effect of Weighting Scheme:* Figure 6 hints on how weighting schemes affect accuracy overall. Figure 7 demonstrates the effect of weighting schemes on accuracy per user. Here, we compare accuracy@5 of three schemes, however, our observations hold for all of our schemes, for all top-$k$ results. In most cases, both $SS\_Mix\_1$, $\lambda = 0.9$ and $SS\_Uniform$, $\lambda = 0.9$ significantly outperform

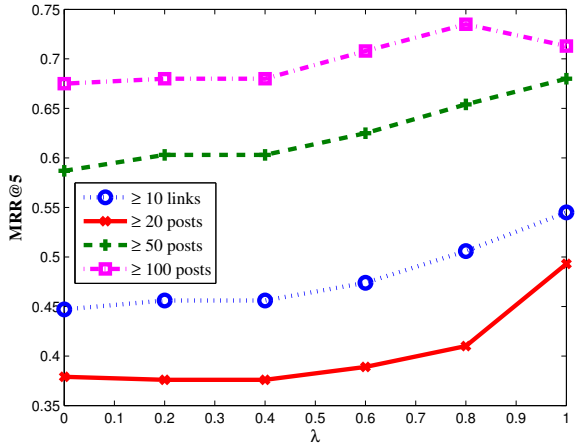



Figure 9. MRR (measured @ 5) as a function of $\lambda$. Different plots impose structural or content availability restrictions. All measurements refer to weighting scheme $SS\_Mix\_1$.

$SS\_Uniform$, $\lambda = 0.5$ with respect to precision, recall and MRR. However, in few cases $SS\_Uniform$, $\lambda = 0.5$ performs better than the other two weighting schemes. Moreover, different weighting schemes perform better for different users (e.g. $SS\_Mix\_1$, $\lambda = 0.9$ and $SS\_Uniform$, $\lambda = 0.9$ achieve different accuracy values for same users). This can be explained as a result of different criteria being important for different users (i.e. content versus time, or network proximity). Hence, personalization is imperative in order to achieve better accuracy overall.

*3) Content Availability and Structural Proximity:* Figure 8a shows average precision as a function of the number of available posts. Figure 8b shows how average precision depends on network structure. To test the effect of these two factors we performed an experiment where we imposed either structural or content restrictions. Structural restrictions implicitly impose some content restrictions since an edge represents at least one @reply message. This is not a 1-to-1 mapping, since many @replies between the same pair of users refer to the same edge. The number of messages does not necessarily reflect number of @replies since many posts may be broadcast instead of direct messages.

Figure 9 shows MRR as a function of $\lambda$ for different restrictions. Intuitively, the greater the number of posts (or the number of neighbors) available for a user, the greater the statistical evidence, resulting in more accurate predictions. In fact, by restricting users to have $\geq 50$ posts, we achieve on average (over all $k \in \{1, 5, 10, 20, 50\}$) 58.7%-68% MRR (32.7%-37.2% precision and 34.1%-38% recall). Considering users with $\geq 100$ posts, we achieve on average 67.5%-73.5% MRR (38.9%-45.2% precision and 38.9%-45.3% recall).

## VIII. Conclusion

We introduced the problem of communication intention prediction in social networks. We addressed this problem using a novel framework that exploits both local structural characteristics and semantically enriched, user generated content. We tested the effectiveness of our approach on an extracted directed network, inferred from directed @reply messages sent between users of a corporate microblogging service. We showed that the more statistical evidence available per user, the better accuracy we can achieve. Based on our findings, our methodology shows great potential to help users identify "interesting" people to initiate conversations with, collaboratively solve problems, or simply create new friends. Although we didn't explore temporal effects on our prediction problem, we found evidence that personalized weighting schemes can greatly improve overall accuracy. We leave this as future work, along with experimentation on larger datasets from Twitter and Facebook.

## Acknowledgment

This work is supported by Chevron Corp. under the joint project, Center for Interactive Smart Oilfield Technologies (CiSoft), at the University of Southern California.

## References

[1] R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer, "Folks in folksonomies: social link prediction from shared metadata," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, ser. WSDM '10. ACM, 2010, pp. 271–280.

[2] L. L and T. Zhou, "Link prediction in complex networks: A survey," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 6, pp. 1150–1170, 2011.

[3] D. Liben-Nowell and J. Kleinberg, "The link prediction problem for social networks," in *Proceedings of the Twelfth International Conference on Information and Knowledge Management*, ser. CIKM '03. New York, NY, USA: ACM, 2003, pp. 556–559.

[4] Y. Koren, S. C. North, and C. Volinsky, "Measuring and extracting proximity graphs in networks," *ACM Trans. Knowl. Discov. Data*, vol. 1, December 2007.

[5] L. Backstrom and J. Leskovec, "Supervised random walks: predicting and recommending links in social networks," in *Proceedings of the fourth ACM international conference on Web search and data mining*, ser. WSDM '11. New York, NY, USA: ACM, 2011, pp. 635–644.

[6] H. Becker, M. Naaman, and L. Gravano, "Learning similarity metrics for event identification in social media," in *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, ser. WSDM '10. New York, NY, USA: ACM, 2010, pp. 291–300.

[7] R. Ghosh, K. Lerman, T. Surachawala, K. Voevodski, and S.-H. Teng, "Non-conservative diffusion and its application to social network analysis," *CoRR*, vol. abs/1102.4639, 2011.

[8] B. Markines, C. Cattuto, F. Menczer, D. Benz, A. Hotho, and S. Gerd, "Evaluating similarity measures for emergent semantics of social tagging," in *Proceedings of the 18th international conference on World wide web*, ser. WWW '09. New York, NY, USA: ACM, 2009, pp. 641–650.

[9] M. Jacovi, I. Guy, I. Ronen, A. Perer, E. Uziel, and M. Maslenko, "Digital traces of interest: Deriving interest relationships from social media interactions." in *ECSCW'11*, 2011, pp. 21–40.

[10] J. Chen, W. Geyer, C. Dugan, M. Muller, and I. Guy, "Make new friends, but keep the old: recommending people on social networking sites," in *Proceedings of the 27th international conference on Human factors in computing systems*, ser. CHI '09. New York, NY, USA: ACM, 2009, pp. 201–210.

[11] M. Chen, J. Liu, and X. Tang, "Clustering via random walk hitting time on directed graphs," in *Proceedings of the 23rd national conference on Artificial intelligence - Volume 2*. AAAI Press, 2008, pp. 616–621.

[12] Y. Zhijun, M. Gupta, T. Weninger, and J. Han, "A unified framework for link recommendation using random walks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2010 International Conference on*, aug. 2010, pp. 152–159.

[13] M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici, "Link prediction in social networks using computationally efficient topological features," in *Proceedings of the IEEE Third International Conference on Social Computing (SocialCom)*, October 2011.

[14] H. Kashima and N. Abe, "A parameterized probabilistic model of network evolution for supervised link prediction," in *Proceedings of the Sixth International Conference on Data Mining*, ser. ICDM '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 340–349.

[15] J. Kunegis and A. Lommatzsch, "Learning spectral graph transformations for link prediction," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 561–568.

[16] M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki, "Link prediction using supervised learning," in *In Proceedings of SDM 06 workshop on Link Analysis, Counterterrorism and Security*, 2006.

[17] A. Popescul and L. H. Ungar, "Statistical relational learning for link prediction," in *IJCAI03 Workshop on Learning Statistical Models from Relational Data*, 2003.

[18] T. Feyessa, M. Bikdash, and G. Lebby, "Node-pair feature extraction for link prediction," in *Proceedings of the IEEE Third International Conference on Social Computing (SocialCom)*, October 2011.

[19] A. Sadilek, H. Kautz, and J. P. Bigham, "Finding your friends and following them to where you are," in *Proceedings of the fifth ACM International Conference on Web Search and Data Mining*, ser. WSDM '12. ACM, 2012.

[20] M. De Choudhury, H. Sundaram, A. John, and D. D. Seligmann, "Contextual prediction of communication flow in social networks," in *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, ser. WI '07. Washington, DC, USA: IEEE Computer Society, 2007, pp. 57–65.

[21] D. Sousa, L. Sarmento, and E. Mendes Rodrigues, "Characterization of the twitter @replies network: are user ties social or topical?" in *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, ser. SMUC '10. New York, NY, USA: ACM, 2010, pp. 63–70.

[22] B. Shevade, H. Sundaram, and L. Xie, "Modeling personal and social network context for event annotation in images," in *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries*, ser. JCDL '07. New York, NY, USA: ACM, 2007, pp. 127–134.

[23] G. Kumaran and J. Allan, "Text classification and named entities for new event detection," in *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, ser. SIGIR '04. New York, NY, USA: ACM, 2004, pp. 297–304.

[24] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to wordnet: An on-line lexical database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, dec 1990.

[25] M. D. Lee, B. Pincombe, and M. Welsh, *An Empirical Evaluation of Models of Text Document Similarity*. Mahwah, NJ: Erlbaum, 2005, pp. 1254–1259.

[26] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, sep 1993.

[27] C. Marlow, M. Naaman, D. Boyd, and M. Davis, "Ht06, tagging paper, taxonomy, flickr, academic article, to read," in *Proceedings of the seventeenth conference on Hypertext and hypermedia*, ser. HYPERTEXT '06. New York, NY, USA: ACM, 2006, pp. 31–40.

[28] P. Mika, "Ontologies are us: A unified model of social networks and semantics," *Semantic Web*, vol. 5, pp. 5–15, March 2007.

[29] J. Weng, E.-P. Lim, J. Jiang, and Q. He, "Twitterrank: finding topic-sensitive influential twitterers," in *Proceedings of the third ACM international conference on Web search and data mining*, ser. WSDM '10. New York, NY, USA: ACM, 2010, pp. 261–270.

[30] M. McPherson, L. S. Lovin, and J. M. Cook, "Birds of a feather: Homophily in social networks," *Annual Review of Sociology*, vol. 27, no. 1, pp. 415–444, 2001.

[31] M. E. J. Newman, "Mixing patterns in networks," *Phys. Rev. E*, vol. 67, p. 026126, Feb 2003.

[32] C. Cattuto, D. Benz, A. Hotho, and S. Gerd, "Semantic grounding of tag relatedness in social bookmarking systems," in *Proceedings of the 7th International Conference on The Semantic Web*, ser. ISWC '08. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 615–631.