Bridging Offline and Online Social Graph Dynamics

Manuel Gomez-Rodriguez
Stanford University
MPI for Intelligent Systems
manuelgr@stanford.edu

Monica Rogati
LinkedIn
monica@rogati.com

ABSTRACT

The online and offline worlds are converging. Location-based services, ubiquitous mobile devices and on-the-go social network accessibility are blurring the distinction between in-person activities and their virtual counterpart. An important effect of this convergence is the rapid and powerful impact of offline events (meetings, conferences) on the evolution and temporal dynamics of the *online* connectivity between members of social and professional networks. However, these effects have been largely unexplored, in part due to the lack of datasets that provide a reliable mapping between events attendees and their online identity and connections.

We are seeking to bridge this gap by using data from LinkedIn, a popular business-related social networking site with more than 120 million members and 10,000 real world events. We find that offline events may induce connectivity changes in the online network – there is a dramatic increase in the number of connections between event attendees shortly after the date of the event. Building on these insights, we describe a non-supervised framework that exploits connectivity changes temporally correlated to real world events to successfully infer more than 40% of specific event attendees. Finally, we revisit the link prediction problem by including user contributed information about offline events to achieve higher link prediction performance.

Categories and Subject Descriptors: H.2.8 [**Database Management**]: Database Applications – *Data mining*
General Terms: Algorithms; Experimentation.
Keywords: Social networks, real world events, temporal dynamics, link prediction.

1. INTRODUCTION

In recent years, there has been an increasing effort and significant progress in understanding the global structure and evolution of social networks [3, 5, 8, 16, 17, 22]. However, the mechanism and motivation underlying individual edge creation is still under-explored [4, 16]. In many circumstances, we may be unable to understand the evolution and dynamics underlying a social network by limiting our inputs to node features, edge features and the topological structure of the network.

In the context of social and professional networks, external factors such as social gatherings and professional conferences trigger new connections between people (nodes) in the network and are key to understanding its evolution. Understanding these mechanisms and their motivation is important - not only for its intrinsic value, but for its potential to improve link prediction algorithms, detecting offline events (meetings, conferences, parties, etc.) that caused the connection, or finding attendees with common interests that facilitate both edge creation and the above-mentioned events. In particular, external events allow us (i) to predict *when* the connection between two people will be created (*i.e.*, it is more likely to happen just before or after an event in which both attend), and (ii) to predict connections between people that are distant in terms of network distance, geography or both.

Moreover, since online networks are usually very large and sparse [22], sampling has become a challenging task. For example, even highly connected LinkedIn members (>1000 connections) are connected to less than 0.001% of all the LinkedIn members. In this context, external factors provide an efficient and valuable meaningful sampling of a real network that goes beyond first and second degree connections or co-membership to a specific discussion group or online community.

Present work. We study how real world professional events and social gatherings relate to the temporal dynamics and evolution of a professional network. We show that the number of new connections among attendees to events increases significantly in a short time window just after the dates of the events. Building on this empirical insight, we first describe how to infer attendees to an event from changes in the connectivity of a social network. Later on, we revisit the link prediction problem to account for real world events, achieving a higher performance.

We use data from LinkedIn, an online professional network with more than 120 million members. In addition to the social graph, defined by the professional connections among LinkedIn members, we record a public list of attendees, often incomplete, for the largest 10,000 real world public events that created a page on `events.linkedin.com`. The lists are often incomplete or partial since we only account for members that publicly RSVP'ed to an event using `events.linkedin.com`. This dataset gives us a comprehensive direct mapping between a subset of the attendees to events and members of a social network.

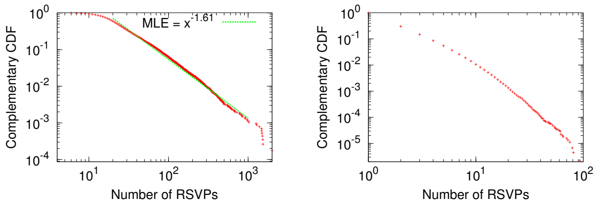
Related work. Our work builds upon several lines of research that lie in two main categories: link prediction and *event detection*.

The link prediction problem in networks has raised much interest in recent years. The problem has been posed as an unsupervised [18] as well as a supervised [24, 30] machine learning problem. Many approaches to link prediction are based on proximity measures on the network topology [1, 19] and community detection [11, 14]. However, there have been substantial developments that extend the feature space beyond proximity measures. Link prediction based on the combination of node features (*i.e.*, user profile information), edge features (*i.e.*, interaction information) and network structure has been shown to improve performance [4]. Recently, geographical proximity among nodes has been also considered in the context of link prediction [25, 28, 31]. Finally, even temporal traces of diffusion that allow for network inference methods can also be viewed as new features for predicting links in networks [12, 13, 21, 29]. Link prediction has been evaluated in a broad range of networks: coauthorship networks [4, 14] (arXiv e-print archive, PubMed), social networks [4, 20] (Facebook, Twitter), mobile phone networks [10, 31], and location-based social networks [10, 28] (Gowalla, Brightkite).

Event detection has been an active research area, focused on identifying events in social media. Traditionally, text mining and clustering techniques have been applied to document feature representations of news articles in blogs and news media sites in order to find events of interest [2, 15, 32]. More recently, approaches to event detection that exploit the underlying structure and temporal dynamics of social networks have been suggested, especially for Twitter data [6, 7, 9, 23, 26, 27]. However, previous studies aim to detect news events or natural disasters, but not *events* in the sense of meetups and social or professional gatherings, their attendees and their impact on the online connectivity between attendees to these events.

To the best of our knowledge, the influence that real world events have on the evolution of the connectivity between members of a social network is a novel research problem that has not been studied in previous work. We believe this work is a starting point towards better understanding interactions among members of a social network, inferring attendees of real world events and helping to improve link prediction. The main contribution of our work is twofold. First, we find that real world events are temporally correlated to connectivity changes in networks and we show that these changes alone allow us to infer attendees to such events. Second, we modify well known methods for link prediction to account for real world events, achieving a higher performance.

The remainder of the paper is organized as follows. In section 2, we explore how and to which extent real world events may influence the dynamics and evolution of a professional network. Section 3 shows how to infer attendees to real world events from changes in the connectivity of a professional network. Section 4 describes how to use real world events for improving non supervised link prediction algorithms. Finally, conclusions and future work are discussed in section 5.



(a) RSVPs per event (b) RSVPs per member

Figure 1: Both the number of RSVPs per event (Panel (a)) and the RSVPs per member (Panel (b)) are heavy-tailed distributions, as many other natural processes. About 75% of the real world events have between 10 and 50 RSVPs, and about 70% of the members report attendance to a single event.

2. EVENT DYNAMICS

In this section, we describe the data that allows us to link real world events to social network dynamics. We start by computing general statistics about events and attendees. We then show empirical evidence that real world events are temporally correlated to an increase of the connectivity rate in the social network that the attendees belong to.

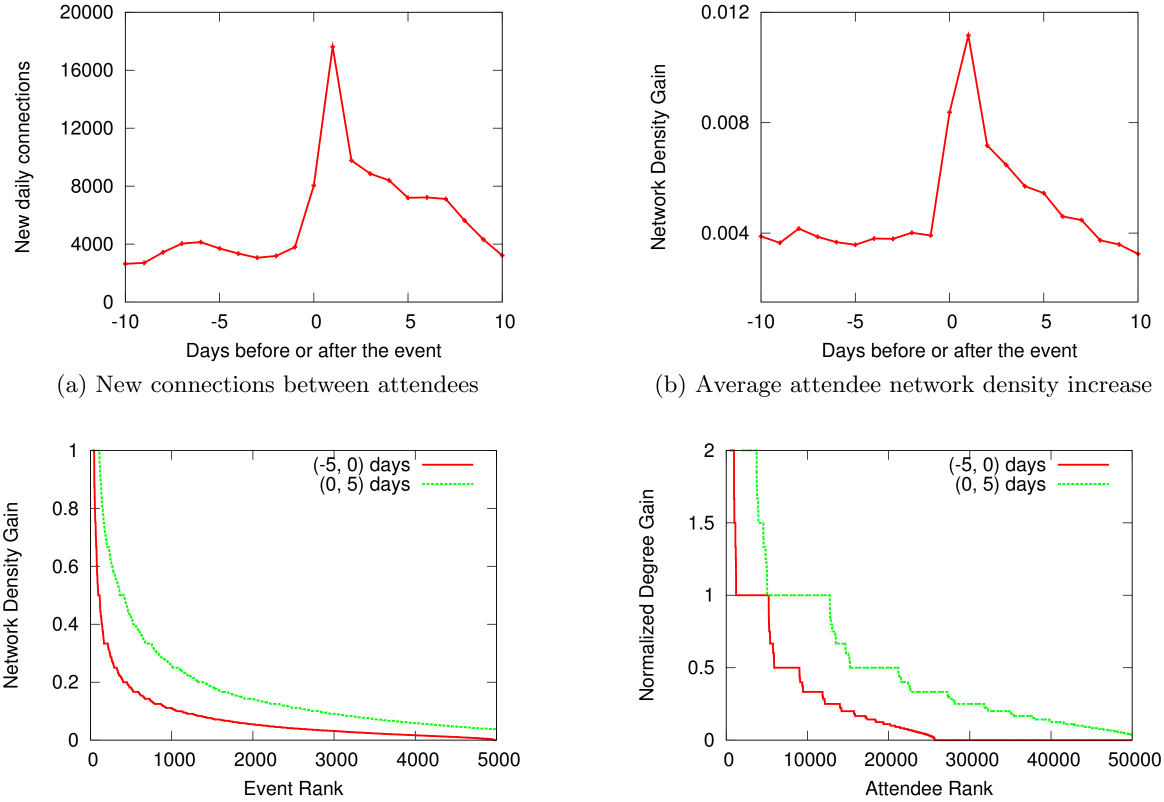
Data. We use data from a popular business-related social networking site, LinkedIn, with more than 120 million members that is mainly used for professional networking. In addition to the professional connections among LinkedIn members that define the social graph of the site, we record the dates and lists, often incomplete, of LinkedIn members that attended more than 10,000 real world events that have a public webpage at `events.linkedin.com`. The lists are often incomplete or partial because we only account for members that RSVP'ed to an event using `events.linkedin.com`, but the actual complete list of attendees is hidden and may be larger.

First, we compute the distribution for the number of attendees per event that RSVP'ed and for the number of events that a member RSVP'ed[1]. Figure 1(a) shows the complementary cumulative distribution (Complementary CDF) for the number of attendees per event that RSVP'ed. We observe a heavy-tailed distribution, as many other natural processes – more than 90% of the real world events have more than 10 RSVPs but only 15% of real world events have more than 50 RSVPs. That means, 75% of the real world events in `events.linkedin.com` have between 10 and 50 RSVPs. Figure 1(b) shows the complementary CCDF for the number of events that a member RSVP'ed. Again, we observe a heavy-tailed distribution, with 70% of the members reporting attendance to a single real world event.

Now, we continue by computing several quantities that allow us to study how real world events temporally correlate to an increase of new professional connections between attendees to those events.

Connections, density and events. We record the dates when events take place, the connections between attendees of such events that RSVP'ed using `events.linkedin.com` and the day in which those attendees become connected. Figure 2(a) shows the absolute daily number of new con-

[1] We have considered only LinkedIn members that RSVP'ed to an event on `events.linkedin.com` at least once.



(a) New connections between attendees

(b) Average attendee network density increase

(c) Network density gain per event

(d) Normalized degree gain per attendee

Figure 2: Daily connections and network density. In the figures, we observe that there is a higher connectivity rate (and network density increase) between attendees (that RSVP'ed) on and up to 10 days after the dates in which events take place. This occurs in average (Panels (a), (b)), across events (Panel (c)) and across attendees (Panel (d)).

nections. There are several interesting patterns. First, we find a sharp increase in the daily number of new connections during and up to 10 days after the events. Second, 10 days after the event, the daily number of new connections declines and it is even lower than before the event. This empirical insights are also supported by the average daily density gain over the subgraphs induced by real world events on the full social graph[2], as shown in Figure 2(b). We define density of a subgraph G_e as:

$$D(G_e) = 2|E_e|/(|V_e| \cdot (|V_e| - 1)),$$

where V_e and E_e are the set of nodes and connections in G_e, and we define density gain of the social graph of an event at day t as $\left(D(G_e^{t+1}) - D(G_e^t)\right)/D(G_e^t)$, where $D(G_e^{t+1})$ and $D(G_e^t)$ are the subgraph densities at days $t+1$ and t respectively.

We have observed an average higher connectivity rate and density increase on and up to 10 days after the dates in which events take place. However, does this density increase occur consistently across the full spectrum of real world events with a website in LinkedIn? As Figure 2(c) shows, it does occur across events. This figure shows the density gain for the subgraph induced by each event in the 5 days before the event, $\left(D(G_e^{t_e}) - D(G_e^{t_e-5})\right)/D(G_e^{t_e-5})$, and the 5 days after the event, $\left(D(G_e^{t_e+5}) - D(G_e^{t_e})\right)/D(G_e^{t_e})$. For each time window, the events are sorted by decreasing density gain. We observe that across the full range of events, there is a greater density increase (gain) during the 5 days after the date of the event than during the 5 days before. This supports the empirical findings that we discussed in average in the paragraph above.

Now, we break down events by attendees, and compute the normalized degree gain per attendee for the 5 day time window before the event and the 5 day time window after the event. We define normalized degree of an attendee as the number of connections of the attendee to other attendees divided by the total number of attendees to the event minus one. Figure 2(d) shows the normalized degree gain for the attendees of all events. For each time window, the attendees are sorted by decreasing normalized degree gain. In this case, we observe that only half the attendees increases significantly their normalized degree by connecting to other attendees during the 5 days before the event, but there is an increase in normalized degree across the full range of attendees during the 5 days that follow the the date of each event.

[2] The subgraph G_e induced by a real world event e on a social graph G is composed of all nodes from G that RSVP'ed to the event e using `events.linkedin.com` and the connections among them.

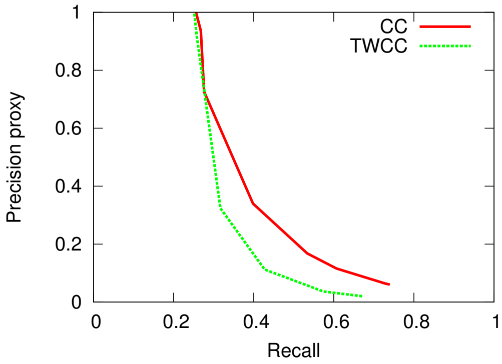


Figure 3: Tradeoff between average recall (correctly identified attendees) and precision proxy (ratio between the list of attendees that RSVP'ed and the size of the inferred set of attendees) for connection counting (CC) and temporally weighted connection counting (TWCC). We achieve this tradeoff by tuning the time window $[t_e - w_{min}, t_e + w_{max}]$.

Although we have found a connectivity increase between attendees to real world events shortly after the dates of the events, this does not imply, strictly speaking, causality – we cannot claim that the connectivity increase is a direct cause of the event. However, in practice, we can still exploit this coincidence to infer attendees to events or predict links between attendees.

3. INFERRING ATTENDEES

In this section, we describe and evaluate two simple methods that perform surprisingly well inferring attendees to events by simply exploiting this network connectivity increase. Both methods allow for a tradeoff between recall and precision by parameter tuning.

3.1 Algorithms

Given a undirected network $G = (V, E)$ and a real world event e, we define the set of nodes that attended a real world event e as $A_e \subseteq V$, the set of nodes that RSVP'ed to the event e as $S_e \subseteq A_e$, and the set of nodes that attended the event e but did not RSVP'ed as $I_e \subseteq A_e$. We assume that nodes that RSVP'ed typically attend the event and therefore $I_e \cup S_e \approx A_e$ and $I_e \cap S_e = \emptyset$. In many cases I_e is unknown and our goal is to find the nodes that belong to I_e given the seed set S_e, for every real world event e. We now describe two simple methods to achieve this goal: connection counting and temporally weighted connection counting.

Connection counting (CC). We build the set of inferred attendees $\hat{I}$ by considering all nodes in G that have n or more than n new connections to nodes in S_e in a time window $[t_e - w_{min}, t_e + w_{max}]$, where t_e is the (starting) date of the event e,

$$\hat{I} = \{i \in V \backslash S_e : |j \in S_e, -w_{min} \leq (t_{i,j} - t_e) \leq w_{max}| \geq n\},$$

where $t_{i,j}$ is the time in which nodes i and j become connected. We achieve a tradeoff between recall and precision by tuning w_{min}, w_{max} and n. For simplicity, in the remainder of the paper, we work with symmetric time windows around the (starting) date of the event (*i.e.*, $w = w_{min} = w_{max}$); however, this does not restrict our ability to choose different values for w_{min} and w_{max}.

Temporally weighted connection counting (TWCC). We build the set of inferred attendees $\hat{I}$ by considering nodes in G such that the temporally weighted sum of their connections to nodes in S_e exceeds a threshold ε,

$$\hat{I} = \{i \in V \backslash S_e : \sum_{j \in S_e} e^{-\alpha \cdot |t_{i,j} - t_e|} \geq \varepsilon\},$$

where $t_{i,j}$ is the time in which nodes i and j become connected and α is a decay factor that accounts for the evolution of the connectivity increase with respect to the date of the event, shown in Figure 2(b). We achieve a tradeoff between recall and precision by tuning α and ε.

3.2 Experimental evaluation

To evaluate the performance of both connection counting and temporally weighted connection counting, we would like to study the tradeoff between precision and recall in average across all 10,000 real world events.

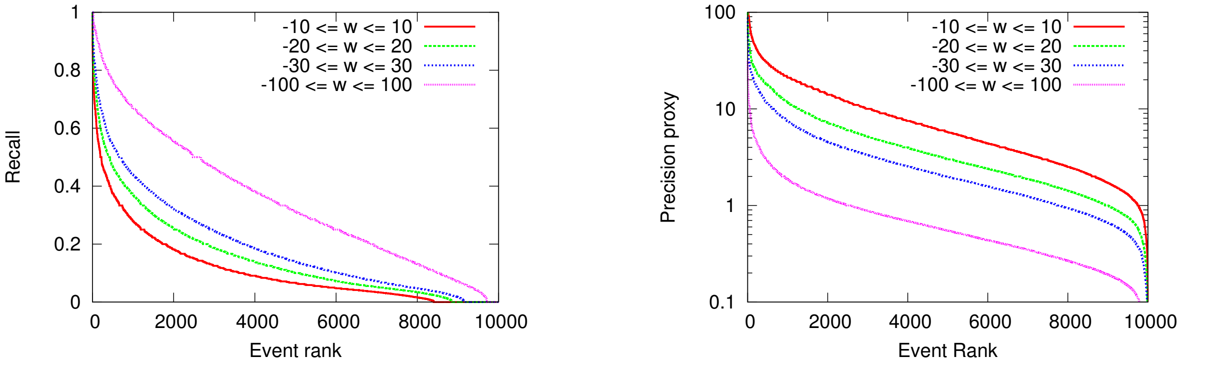
If a complete list of attendees (ground truth) for a real world event is available, precision is the fraction of nodes in the inferred set of attendees, $\hat{I}$, present in the complete list of attendees of the event that did not RSVP'ed (*i.e.*, $|I_e \cap \hat{I}_e|/|\hat{I}_e|$) and recall is the fraction of nodes in the list of attendees of the event that did not RSVP'ed, I_e, that are present in the inferred set of attendees $\hat{I}_e$ (*i.e.*, $|I_e \cap \hat{I}_e|/|I_e|$).

Unfortunately, in general, we do not have access to a complete list of attendees or ground truth for each real world event but only to an incomplete list of people that RSVP'ed through `events.linkedin.com`. However, in addition to estimate recall (*i.e.*, correctly identified attendees) using crossvalidation, we are able to identify and measure a *precision proxy*.

To estimate the recall for an event, we perform leave-one-out crossvalidation (LOOCV) for every member in the list of attendees that RSVP'ed, S_e. In particular, we solve $|S_e|$ inference problems, one for each member $i \in S_e$. For each inference problem, we create the sets $S'_e = S_e \backslash i$ and $I'_e = \{i\}$, and infer I'_e from S'_e. We then compute the recall for each of these inference problems and estimate the total recall computing the average.

We cannot estimate the precision for an event given only a list of attendees that RSVP'ed through `events.linkedin.com`, S_e. Instead, we compute a precision proxy as follows. For each event e, we let CC and TWCC include members $i \in S_e$ in the inferred set $\hat{I}_e$. We then compute the ratio between the list of attendees that RSVP'ed, S_e, and the size of the inferred set of attendees, $\hat{I}_e$, *i.e.*, $|S_e|/|\hat{I}_e|$ for each method (and event). This ratio can be relatively low for events in which not many attendees RSVP but the size of the event is actually high. In some cases, a very small value may also indicate a lack of precision.

We now start by evaluating the tradeoff between recall and precision proxy in average across all the events for both connection counting and temporally weighted connection counting. Later, we show that the method that achieves the best tradeoff between recall and the precision proxy works well across all 10,000 events, not only in average. Finally, we evaluate both methods in terms of precision and recall in a case study of a particular event in which both a list of attendees at LinkedIn and complete list of attendees at the



(a) Recall for connection counting with n = 2

(b) Precision proxy for connection counting with n = 2

Figure 4: Recall (correctly identified attendees) and precision proxy (ratio between the list of attendees that RSVP'ed and the size of the inferred set of attendees). Panel (a) and (b) show that the tradeoff between recall and precision proxy shown for connection counting in Figure 3 is consistent across events. Longer time windows result on higher recall and smaller precision proxy. Note that if we have an estimate of the total number of attendees to an event $|A_e|$, we can tune w and n to achieve $|S_e \cup \hat{I}_e| = |A_e|$

official website of the event are available.

Tradeoff between recall and precision proxy. We achieve a tradeoff between recall and precision proxy by tuning the time window $[t_e - w, t_e + w]$ in connection counting and the threshold ε in temporally weighted connection counting. Figure 3 shows the average recall vs precision proxy for both methods. There are several interesting observations. First, if we assume that the list of RSVPs is complete (*i.e.*, all attendees have RSVP'ed and thus $S_e = A_e$), we correctly identify approximately 27% of the attendees that RSVP'ed with both CC and TWCC. Second, if we allow for an event size three times larger than the list of RSVPs, we manage to identify 40% of the RSVPs with CC and 32% with TWCC. Finally, allowing for an event size ten times larger than the RSVP list leads to discover more than 60% of the attendees that RSVP'ed with CC but only 44% with TWCC. We observe that for any fixed value of the precision proxy CC achieves always a higher recall value than TWCC.

Recall across events. We now pay attention to the individual recall across all the 10,000 real world events for connection counting (CC). Figure 4(a) shows the recall for different time windows $[t_e - w, t_e + w]$. For $w = 100$ (*i.e.*, the time window spanning three months before and after the date of the event), we achieve an average recall as high as 40% across 10,000 real world events and a recall higher than 40% for more than 50% of the events and higher than 60% for more than 20% of the events.

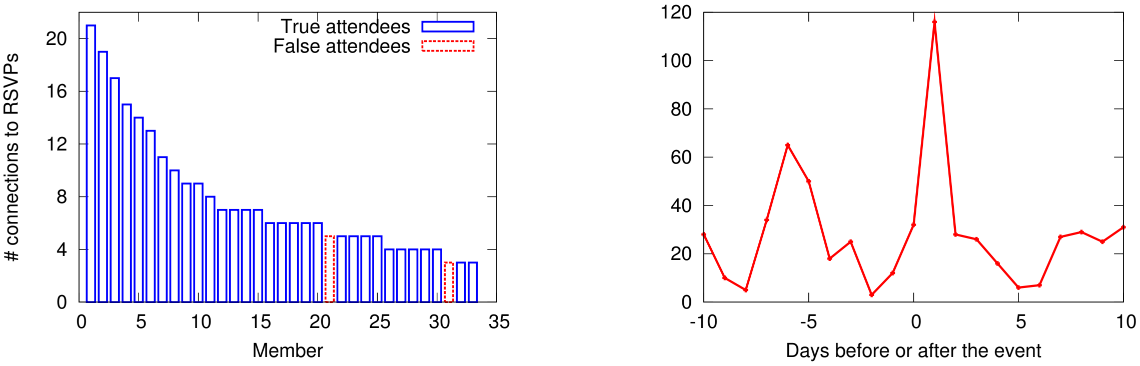
Proxy to precision across events. In Figure 4(b), we observe the ratio between the list of attendees that RSVP'ed, S_e, and the size of the inferred set of attendees, $\hat{I}_e$, that may include members $i \in S_e$, across all 10,000 real world events for connection counting. It is difficult to judge the performance because the real number of attendees per event is unknown, and we only have access to the list of attendees that RSVP'ed using `events.linkedin.com`. Note that if we have an estimate of the total number of attendees to an event $|A_e|$, we can either tune the parameters of connection counting to achieve $|S_e \cup \hat{I}_e| = |A_e|$.

Example and case study: precision and recall for a professional event[3]. Although true event attendee lists are rarely made public, we have examined how our techniques perform in one case where such information is known. In spite of its small size, the event helps us exemplify our techniques and grounds our precision proxy. The official website of the event that we have chosen contains links to the LinkedIn profiles of each attendee that has a LinkedIn account and therefore, we have a reliable mapping between both the list of attendees at LinkedIn and the complete list of attendees at the official website.

All 21 people that RSVP'ed are also listed as attendees in the official website of the event. However, there are a total of 63 attendees with LinkedIn account listed in the official website (out of 67 attendees), *i.e.*, there are 42 LinkedIn members that attended the event that did not RSVP'ed. Figure 5(b) shows the daily connections to members that RSVP'ed. As in section 2 for all events in average, we also observe a peak in new connections just before and after the event, and later on a decline in the number of new connections. Figure 5(a) shows the number of connections to members that RSVP'ed in a time window spanning 10 days before and after the event for every member in the inferred set of attendees returned by connection counting. More than 75% of the inferred attendees created 5 or more connections to members that RSVP'ed. Our aim is to infer the set of 42 members that did not RSVP using only the connectivity increase in the social graph.

Using connection counting with a time window spanning 20 days before and after the event and a threshold of 2 connections, the recall on the set of 42 members that did not RSVP is 71.4%. Importantly, connection counting returns only 2 LinkedIn members that are not listed in the official website nor RSVP'ed, *i.e.*, if we assume that only people

[3] Drupal executives meeting in Brussels, 8-10 October, 2010. Official event website: `http://cxo.drupaldays.org`. LinkedIn event website: `http://www.linkedin.com/osview/canvas?_ch_page_id=1\&_ch_panel_id=1\&_ch_app_id=7083120\&_applicationId=2000\&_ownerId=0\&appParams=\{"go_to":"events/421548","referrer":"public"\}`.



(a) Connections to RSVPs on LinkedIn per inferred attendee

(b) Daily connections to RSVPs vs time

Figure 5: Drupal executives meeting event. In Panel (a), we observe that more than 75% of the inferred attendees created 5 or more connections to members that RSVP'ed through LinkedIn in the time window spanning 20 days before and after the event. In Panel (b) we find a sharp increase on the number of new connections just before and after the event and later on a decline in the number of new connections.

in the list of attendees in the official website attended the event, the precision of our method is 95.5%. Moreover, if we perform leave-one-out crossvalidation (LOOCV) on the set of 21 members that RSVP'ed, the recall on the 21-member set is 100%.

Using temporally weighted connection counting with $\alpha = 0.1$ and $\varepsilon = 1$, the recall on the set of 42 members that did not RSVP is 83.3%, increasing 11.9% with respect to connection counting. However, the precision is 81.4%, slightly worse than connection counting (-14.1%). As in connection counting, the recall on the 21-member set of members that RSVP'ed using leave-one-out crossvalidation is 100%.

In both connection counting and temporally weighted connection counting we can achieve a tradeoff between precision and recall by tuning the parameters w, n (in CC) and α and ε (in TWCC).

4. INFERRING CONNECTIONS

In this section, we modify two well-known quantities on the graph topology, which have been used successfully for link prediction [1, 18], to leverage from the lists of attendees that RSVP'ed to real world events. We then evaluate the modified quantities in a non-supervised setting for link prediction and show that including information about events enables us to achieve a higher performance.

4.1 Algorithms

Given a undirected network $G = (V, E)$ and a real world event e with (starting) date t_e, we define the set of nodes that RSVP'ed to the event e through `events.linkedin.com` as $S_e \subseteq V$. Our aim is to predict new connections in dates close to t_e in which at least one of the peers belongs to S_e.

Baseline methods. We first recall two baseline methods based on ranking measures on the graph topology that have been shown to achieve a relatively good performance in the link prediction problem in social networks: normalized common neighbors and Adamic-Adar. Normalized common neighbors (CN) between two nodes i and j is defined as the number of connections that nodes i and j have in common normalized by the product of the connections of each node,

$$CN(i,j) = \sqrt{\frac{|\Gamma(i) \cap \Gamma(j)|}{|\Gamma(i)||\Gamma(j)|}}, \qquad (1)$$

where $\Gamma(n) = \{m \in V : (m,n) \in E\}$. Adamic-Adar (AA) modifies common neighbors by weighting each neighbor by her degree instead of simply counting,

$$AA(i,j) = \sum_{n \in \Gamma(i) \cap \Gamma(j)} \frac{1}{\log |\Gamma(n)|}. \qquad (2)$$

Both quantities (baselines) do not take into account whether two nodes attended the same real-world event. However, the probability that two nodes become connected in a social network increases if they get to know each other in person in a real-world event.

Event-based methods. The rationale for an event-based link prediction approach is better understood after having a close look at Figure 6. The figure shows the new daily connections between attendees that RSVP'ed to an event vs the total number of new daily connections created by these attendees to any node in the network. Importantly, we observe that an attendee to an event tends to create almost one order of magnitude more connections to attendees of the same event in dates closer to the event than in other days far from the date of the event. We then introduce two simple methods based on normalized common neighbors and Adamic-Adar that given a list of RSVP's to a real-world event achieve a greater performance on the link prediction task for dates close to the date of the event.

Normalized common attendees (CA$_e$) between two nodes i and j given an event e is defined as the number of connections to attendees of the event e that nodes i and j have in common normalized by the product of the connections of each node that are attendees to the event,

$$CA_e(i,j) = \sqrt{\frac{|\Gamma_e(i) \cap \Gamma_e(j)|}{|\Gamma_e(i)||\Gamma_e(j)|}}, \qquad (3)$$

where $\Gamma_e(n) = \{m \in S_e : (m,n) \in E\}$. In this case, we assume that two nodes are more likely to get to know each

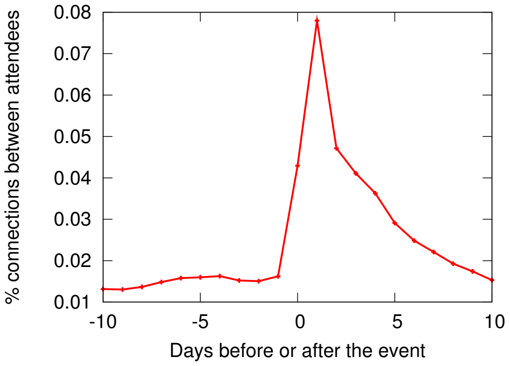


Figure 6: New daily connections between attendees of an event vs the total number of new daily connections created by attendees of the event to any node in the network. An attendee to an event tends to create almost one order of magnitude more connections to attendees of the same event in dates closer to the event than in other days far from the date of the event

other and become connected in the social network if they have common connections that attended an event e. Finally, event-based Adamic-Adar (AA$_e$) simply modifies common attendees in the same way that Adamic-Adar modifies common neighbors, penalizing nodes with high degree,

$$AA_e(i,j) = \sum_{n \in \Gamma_e(i) \cap \Gamma_e(j)} \frac{1}{\log |\Gamma(n)|}, \tag{4}$$

where $\Gamma_e(n)$ and $\Gamma(n)$ are defined as above. Common attendees and event-based Adamic-Adar use both the list of attendees to an event and the network topology.

4.2 Experimental evaluation

We evaluate our baseline and event-based methods as follows. For each attendee to an event, we consider (i) her second degree connections up to w_{min} days before the day of the event and (ii) the other attendees to the event to build the list P_e of potential connections that may be created by attendees to an event e during the time window $(t_e - w_{min}, t_e + w_{max})$. Then, we generate for each method a list of top-k most *likely* connections per event $\hat{L}_{e,k} \subseteq P_e$. Finally, sweeping over k values allows us to obtain different points in the precision recall curve.

Recall and precision. For each event, we compute the precision and recall of the baseline and event-based methods on the connections L_e that the attendees create during the time window $(t_e - w_{min}, t_e + w_{max})$. We define precision as the fraction of connections in the list of top-k most *likely* connections $\hat{L}_{e,k}$ present in the list of connections L_e (*i.e.*, $|L_e \cap \hat{L}_{e,k}|/|\hat{L}_{e,k}|$) and recall as the fraction of connections in the list of connections L_e present in the list of top-k most *likely* connections $\hat{L}_{e,k}$ (*i.e.*, $|L_e \cap \hat{L}_{e,k}|/|L_e|$).

For our experiments we set $w_{min} = w_{max} = 10$ days, *i.e.*, we try to find the connections created in a 20-day time window centered on the (starting) date of each event. First, we filter out events with less than 10 attendees, since we have observed that attendees to such small events are typically heavily connected between them and events do not provide additional information, and events with more than 50 attendees for computational reasons, since they were only 15% of the total number of events. Then, we generate two sets of events: a set of 500 events with the smallest number of connections between attendees up to $t_e - w_{min}$ and a set of 100 random events. For each event, the set of potential connections that we rank are (i) connections between each attendee and her second degree connections (second degree connections up to $t_e - w_{min}$ days before the (starting) date of the event) and (ii) connections between attendees.

The average set size of potential connections per event is 730,330 connections for the 500-event set and 1,518,700 connections for the 100-event set, while the average set of true connections contains only 534 connections and 453 connections respectively, *i.e.*, the probability of choosing a true connection at random is at most $7 \cdot 10^{-4}$ for the 500-event set and $3 \cdot 10^{-4}$ for the 100-event set. For each of the methods we ranked in total more than 500 million potential connections.

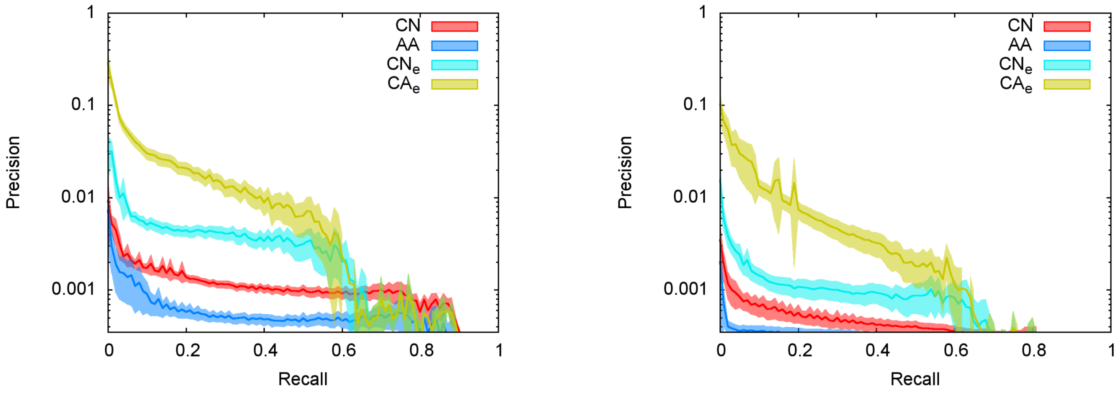
Figure 7 shows the average precision vs recall curves across events with 1.96· standard error $(\sigma/\sqrt{N})$ bands, which result of sweeping over k on the lists of top-k most *likely* connections in the event-based and baseline methods for both set of events. In both event sets, AA$_e$ outperforms both baselines in terms of precision for more than an order of magnitude for recall values up to 50%. For example, for a 10% recall, AA$_e$ achieves a precision of approximately 4% in the 500-event set and 1.5% in the 100-event set while CA$_e$ precision is 0.5% in the 500-event set and 0.25% in the 100-event set. The precision for both CN and AA goes down to a value below 0.2% in the 500-event set and below 0.06% in the 100-event set. Due to the heavily unbalance dataset that the algorithms need to deal with, they output solutions with relatively low precision value. If we compare the performance between both sets of real world events, we observe that event-based methods gives a greater additional mileage in the 500-event set with the smallest number of connections between attendees up to $t_e - w_{min}$ than in the 100-event random set. A possible explanation behind this difference in performance is that a small number of connections among attendees w_{min} days before an event makes inferring connections using only the network topology more difficult.

Here, we aim to give empirical evidence that real world events improve performance by modifying two simple methods – including real world events information in more sophisticated link prediction methods may help to increase precision.

Performance vs. event size. We now perform a stratified analysis of the performance with respect to event size. Figure 8 plots the average area under curve (AUC) values on the precision recall curves against event size. We observe that the performance does not change significantly with respect to the event size and event-based methods outperform the baselines across the full range of event sizes.

5. CONCLUSIONS

We have given empirical evidence that real world events shape the temporal dynamics of a social network. Real-world events may facilitate connections between attendees in an on-line social network. We conclude this after studying a business-related social network, LinkedIn, with more than 115 million members and 10,000 real-world events. To the best of our knowledge, our work tries to bridge, for the



(a) 500-event set with lowest number of connections before $w_{min} = 10$

(b) 100-event random set

Figure 7: Precision vs recall for the link prediction task. Event-based methods (AA$_e$ and CA$_e$) perform better than the baselines (CN and AA) in terms of precision across almost the full range of recall values. For recall values below 50%, AA$_e$ achieves precision values an order of magnitude higher than both baselines CN and AA.

first time, the gap between off-line and on-line social graph dynamics.

We exploit the bridge between off-line and on-line dynamics in two research problems: attendee inference and link prediction. First, we show that simple methods that account for event-induced connectivity changes in a social network can be fruitfully applied to uncover attendees to real-world events. We are able to successfully infer more than 40% of specific event attendees using only event-induced connectivity changes. Second, we modify well-known non supervised link prediction methods to account for the event-induced network dynamics and we show that these modifications lead to a significant improvement. In particular, we achieve precision values more than an order of magnitude higher than traditional methods that do not account for event-induced network dynamics.

There are several research directions to build on and benefit from our framework. First, we have studied how real-world events shape a business-related social network but it is an open question if similar patterns occur in a non business-related social network (*e.g.*, Facebook, Twitter, etc.), and to study the (possible) differences between them. Second, since sampling of social networks is becoming increasingly challenging due to the network sizes, real-world events could be used as an efficient meaningful sampling mechanism that would go beyond first and second degree connections or community-based sampling. Third, temporal records of real world events and connections in a social network are often noisy, inaccurate or unobserved. In such cases, it is necessary to develop and apply inference and reconstructing algorithms for the temporal data.

Finally, we have shown that it is possible to infer attendees to an event based on the network dynamics. But, is it possible to go a step further and infer an event itself and its attendees based on the network dynamics? Would it be feasible to perform real-time detection of events from data streams of online network activity? Would geotemporal traces left by nodes in a social network give an additional mileage in detecting real-world events?

6. REFERENCES

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

[2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 37–45. ACM, 1998.

[3] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD '06: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 44–54. ACM, 2006.

[4] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *WSDM '11: Proceedings of the fourth ACM International Conference on Web Search and Data Mining*, pages 635–644. ACM, 2011.

[5] A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509, 1999.

[6] H. Becker, M. Naaman, and L. Gravano. Learning similarity metrics for event identification in social media. In *WSDM '10: Proceedings of the third ACM International Conference on Web search and Data Mining*, pages 291–300. ACM, 2010.

[7] H. Becker, M. Naaman, and L. Gravano. Beyond trending topics: Real-world event identification on twitter. In *AAAI '11: Proceedings of the 25th AAAI Conference on Artificial Intelligence*, 2011.

[8] A. Blum, T. Chan, and M. Rwebangira. A random-surfer web-graph model. In *Proceedings of the eighth Workshop on Algorithm Engineering and Experiments and the third Workshop on Analytic Algorithmics and Combinatorics*, volume 123, page 238. Society for Industrial Mathematics, 2006.

[9] C. Castillo, M. Mendoza, and B. Poblete. Information credibility on twitter. In *WWW '10: Proceedings of the 20th international conference on World wide web*,

pages 675–684. ACM, 2011.

[10] E. Cho, S. Myers, and J. Leskovec. Friendship and mobility: User movement in location-based social networks. In *KDD '11: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 2011.

[11] A. Clauset, C. Moore, and M. Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

[12] M. Gomez-Rodriguez, D. Balduzzi, and B. Schölkopf. Uncovering the Temporal Dynamics of Diffusion Networks. In *ICML '11: Proceedings of the 28th International Conference on Machine Learning*, 2011.

[13] M. Gomez-Rodriguez, J. Leskovec, and A. Krause. Inferring Networks of Diffusion and Influence. In *KDD '10: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, pages 1019–1028, 2010.

[14] K. Henderson and T. Eliassi-Rad. Applying latent dirichlet allocation to group discovery in large graphs. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1456–1461. ACM, 2009.

[15] J. Kleinberg. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery*, 7(4):373–397, 2003.

[16] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD '08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 462–470. ACM, 2008.

[17] J. Leskovec, J. Kleinberg, and C. Faloutsos. Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, page 187, 2005.

[18] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American Society for Information Science and Technology*, 58(7):1019–1031, 2007.

[19] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33):11623, 2005.

[20] B. Meeder, B. Karrer, A. Sayedi, R. Ravi, C. Borgs, and J. Chayes. We know who you followed last summer: inferring social link creation times in twitter. In *WWW '11: Proceedings of the 20th International Conference on World wide web*, pages 517–526. ACM, 2011.

[21] S. Myers and J. Leskovec. On the Convexity of Latent Social Network Inference. In *NIPS '10: Advances in Neural Information Processing Systems*, 2010.

[22] M. Newman. The structure and function of complex networks. *SIAM review*, pages 167–256, 2003.

[23] A. Popescu and M. Pennacchiotti. Detecting controversial events from twitter. In *KDD '10: Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, pages 1873–1876. ACM, 2010.

[24] A. Popescul and L. Ungar. Statistical relational learning for link prediction. In *IJCAI '03: Workshop on Learning Statistical Models from Relational Data*, volume 149, page 172. Citeseer, 2003.

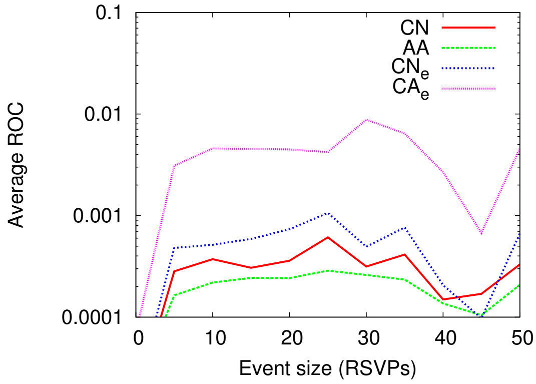


Figure 8: Average AUC values vs event size. The performance does not change significantly with respect to the event size (number of RSVPs to the event) and CA$_e$ outperforms CN and AA for one order of magnitude

[25] M. Rivera, S. Soderstrom, and B. Uzzi. Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *Annual Review of Sociology*, 36:91–115, 2010.

[26] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: real-time event detection by social sensors. In *WWW '10: Proceedings of the 19th International Conference on World Wide Web*, pages 851–860. ACM, 2010.

[27] J. Sankaranarayanan, H. Samet, B. Teitler, M. Lieberman, and J. Sperling. Twitterstand: news in tweets. In *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 42–51. ACM, 2009.

[28] S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *KDD '11: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 2011.

[29] T. Snowsill, N. Fyson, T. de Bie, and N. Cristianini. Refining causality: who copied from whom? In *KDD '11: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 2011.

[30] B. Taskar, M. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *NIPS '03: Advances in Neural Information Processing Systems*, 2003.

[31] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. Barabási. Human mobility, social ties, and link prediction. In *KDD '11: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, 2011.

[32] Y. Yang, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. In *SIGIR '98: Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 28–36. ACM, 1998.