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.

We study these effects by using data from LinkedIn, a popular business-related social networking site. 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 method 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

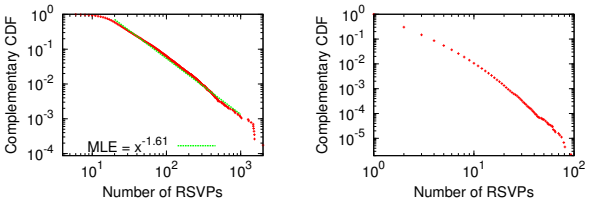
In recent years, there has been an increasing effort and significant progress in understanding the global structure and evolution of social networks [1, 3, 5]. However, the mechanism and motivation underlying individual edge creation is still under-explored [2, 4]. In many circumstances, we may be unable 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.

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.

2. EVENT DYNAMICS

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



(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.

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.

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

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 and the 5 days after the event. 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 before.

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 each event.

3. INFERRING ATTENDEES

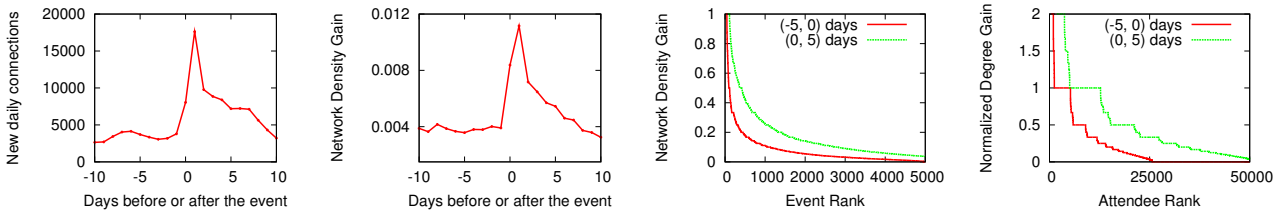
Algorithm. Given a undirected network $G = (V, E)$ and a real world event e, we define the set of nodes that attended a real world event e as $A_e \subseteq V$, the set of nodes that RSVP'ed to the event e as $S_e \subseteq A_e$, and the set of nodes that attended the event e but did not RSVP'ed as $I_e \subseteq A_e$. 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 a simple method to achieve this goal.

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, $\hat{I} = \{i \in V \backslash S_e : |j \in S_e, -w_{min} \leq (t_{i,j} - t_e) \leq w_{max}| \geq n\}$, where $t_{i,j}$ is the time in which nodes i and j become connected and t_e is the (starting) date of the event e. 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; however, this does not restrict our ability to choose different values for w_{min} and w_{max}.

Experimental evaluation. To evaluate the performance of our method, we would like to study the tradeoff between precision and recall in average across all 10,000 real world events.

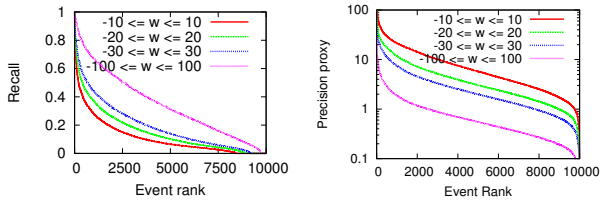
If a complete list of attendees (ground truth) for a real world event is available, precision is the fraction of nodes in the inferred set of attendees, $\hat{I}$, present in the complete list of attendees of the event that did not RSVP'ed (*i.e.*, $|I_e \cap \hat{I}_e|/|\hat{I}_e|$) and recall is the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

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



(a) New connections (b) Network density increase (c) Network density gain (d) Normalized degree gain

Figure 2: Daily connections and network density. 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.



(a) Recall (b) Precision proxy

Figure 3: Recall and precision proxy ($n = 2$) for the attendee prediction task.

not estimate the precision for an event given only a list of attendees that RSVP'ed through LinkedIn, S_e. Instead, we compute a precision proxy as follows. For each event e, we let our method 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.

Figure 3(a) shows the recall across events for different time windows $[t_e - w, t_e + w]$. For $w = 100$, 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. In Figure 3(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. 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 LinkedIn.

Example and case study[2]. 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.

[2]Drupal executives meeting in Brussels, 8-10 October, 2010. Official event website: `http://cxo.drupaldays.org`.

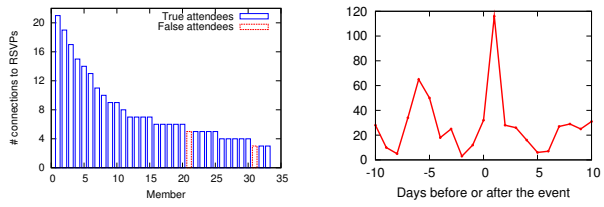
All 21 people that RSVP'ed are also listed as attendees in the official website of the event. However, there are a total of 63 attendees with LinkedIn account listed in the official website (out of 67 attendees), *i.e.*, there are 42 LinkedIn members that attended the event that did not RSVP'ed. Figure 4(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 4(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 our method. More than 75% of the inferred attendees created 5 or more connections to members that RSVP'ed. Using our method for inferring attendees 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%. The method returns only 2 LinkedIn members that are not listed in the official website nor RSVP'ed, *i.e.*, if we assume that only people in the list of attendees in the official website attended the event, the precision of our method is 95.5%.

4. INFERRING CONNECTIONS

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 LinkedIn 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.

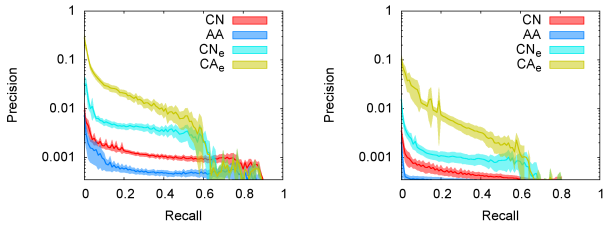
We first recall two baseline methods based on ranking measures on the graph topology that have been shown to achieve a relatively good performance in the link prediction problem in social networks: normalized common neighbors and Adamic-Adar. Normalized common neighbors (CN) between two nodes i and j 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. Adamic-Adar (AA) modifies common neighbors by weighting each neighbor by her degree instead of simply counting.

The rationale for an event-based link prediction approach relies on the observation 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. Nor-



(a) Total connections to RSVPs (b) Daily connections to RSVPs

Figure 4: Case study: Drupal executives meeting event.



(a) 500-event set (b) 100-event random set

Figure 5: Precision vs recall for the link prediction task.

malized 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. In this case, we assume that two nodes are more likely to get to know each 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.

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 $\hat{L}_{e,k} \subseteq P_e$. Sweeping over k values allows us to obtain different points in the precision recall curve.

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 $(t_e - w_{min}, t_e + w_{max})$. 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 up to $t_e - w_{min}$ days before the (starting) date of the event and (ii) connections between attendees.

Figure 5 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.

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. We exploit the bridge between off-line and on-line dynamics in two research problems: attendee inference and link prediction. First, we show that a simple method 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.

6. REFERENCES

[1] L. Backstrom, D. Huttenlocher, J. Kleinberg, and X. Lan. Group formation in large social networks: membership, growth, and evolution. In *KDD*, 2006.

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

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

[4] J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. Microscopic evolution of social networks. In *KDD*, 2008.

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