New Insights and Methods for Predicting Face-to-Face Contacts

Christoph Scholz and **Martin Atzmueller** and **Gerd Stumme**
Knowledge and Data Engineering Group, University of Kassel, Germany
{scholz, atzmueller, stumme}@cs.uni-kassel.de

Alain Barrat
Centre de Physique Théorique
CNRS UMR 6207, Marseille, France
alain.barrat@cpt.univ-mrs.fr

Ciro Cattuto
Data Science Laboratory
ISI Foundation, Torino, Italy
ciro.cattuto@isi.it

Abstract

The prediction of new links in social networks is a challenging task. In this paper, we focus on predicting links in networks of face-to-face spatial proximity by using information from online social networks, such as co-authorship networks in DBLP, and a number of node level attributes.

First, we analyze influence factors for the link prediction task. Then, we propose a novel method that combines information from different networks and node level attributes for the prediction task: We introduce an unsupervised link prediction method based on rooted random walks, and show that it outperforms state-of-the-art unsupervised link prediction methods. We present an evaluation using three real-world datasets. Furthermore, we discuss the impact of our results and of the insights we glean in the field of link prediction and human contact behavior.

1 Introduction

With the rise of location-based services and mobile social networks, there is increasing interest in the analysis of networks of physical proximity and behavioral interaction, e.g., networks that involve spatial relations like co-location or face-to-face proximity. In this context, physical devices, e.g., mobile phones or RFID devices, can help to link relations in the digital domain to relations in physical space, and vice-versa, so that interactions are becoming more interdependent. Specifically, the prediction of links is a challenging task concerning both online and offline social networks.

In this paper, we focus on the prediction of links of human face-to-face proximity in physical space. Our application context is given by the Conferator social conferencing application (Atzmueller et al. 2011) built on top of the RFID-based proximity sensing system developed by the SocioPatterns collaboration (http://www.sociopatterns.org).

For the link prediction task we use both offline and online data: As our main data source we consider the contact networks of the participants of three different conferences. In addition, we analyze the co-location relations of participants (as a proxy for their encounters), the co-authorship network of DBLP, and information about the textual content of the papers of each participant.

Using the different offline and online data, we extend previous work of (Scholz, Atzmueller, and Stumme 2012), and are able to merge the DBLP online network with the human contact network. Furthermore, we enrich them using content information about the nodes of the respective networks. For this hybrid dataset, we present our novel unsupervised link prediction method.

In our application setting we focus on unsupervised methods for two reasons. First, human contact networks are rather sparse. Second, in an unsupervised setting we can avoid the cold-start problem (which is also enhanced by the relative sparsity of the data) which was also a practical requirement for the deployment of the proposed method in the Conferator system. Good link predictions are very important in social network platforms, because friendship recommenders are often responsible for a significant number of new links.

The contribution of this work is as follows:

1. We analyze influence factors for the link prediction task in human contact networks considering both close-range proximity relations, co-authorship information and node-specific content information.

2. We present a novel unsupervised link prediction method based on rooted random walks, integrating sets of networks for the prediction task and show that it outperforms state-of-the-art unsupervised link prediction methods.

3. Using the presented link prediction method, we show that the predictability of physical face-to-face proximity can be further improved using information from online social networks.

The rest of this paper is structured as follows: Section 2 discusses related work. Section 3 describes the RFID hardware setting and presents a detailed overview of the collected datasets. Section 4 analyzes specific influence factors of human contact behavior. After that, we present our novel unsupervised link prediction method. Section 5 reports a detailed evaluation using three real-world datasets. Finally, Section 6 summarizes our results and discusses future work.

2 Related Work

In this section, we discuss related work, and start by giving an overview of the analysis of human contact behavior. After that, we focus on specialized link prediction techniques.

Human Contact Behavior

The analysis of human contact patterns and their underlying structure is an interesting and challenging task in social network analysis. In this context, (Eagle, Pentland, and Lazer 2009) and (Hui et al. 2005) presented an analysis using proximity information collected by bluetooth devices as a proxy for human proximity. However, given the range of interaction of bluetooth devices, the detected proximity does not necessarily correspond to face-to-face contacts.

The SocioPatterns collaboration developed an infrastructure that detects close-range and face-to-face proximity (1-1.5 meters) of individuals wearing proximity tags with a temporal resolution of 20 seconds (Cattuto et al. 2010). They presented an application (at the ESWC 2009 conference) that combines online and offline data from conference attendees (Alani et al. 2009). (Zuo et al. 2012) also study the influence between offline and online properties using a mobile social application in the context of academic conferences. (Barrat et al. 2010) compared the attendees' contact patterns with their research seniority, their co-authorship and their activity in social web platforms.

The SocioPatterns sensing infrastructure was also deployed in other environments in order to study the dynamics of human contacts, such as healthcare environments (Isella et al. 2011a), schools (Stehlé et al. 2011) and museums (Isella et al. 2011b). (Macek et al. 2012) analyzes the interactions and dynamics of the behavior of participants at conferences, and also the connection between research interests, roles and academic jobs of conference attendees.

Link Prediction

In the field of link prediction in social networks a first detailed and comprehensive analysis was done by (Liben-Nowell and Kleinberg 2003): The authors defined the link prediction problem and analyzed the predictability of unsupervised methods that use only proximity information of nodes in the social network graph. (Murata and Moriyasu 2007) presented and analyzed weighted variants of the network proximity measures *Adamic-Adar*, *Common Neighbors* and *Preferential Attachment*. (Lü and Zhou 2010) presented an approach to analyze the role of weak ties in social networks. (Zhuang et al. 2012b) presented a method using active learning to infer social ties. (Wang et al. 2011) examined the impact of human mobility on link prediction.

(Lichtenwalter, Lussier, and Chawla 2010) introduce a novel unsupervised method, i.e., a restricted variant of rooted PageRank, and a new supervised method (Lichtenwalter and Chawla 2012) for link prediction. Concerning the general prediction technique of combining different sources, (Backstrom and Leskovec 2011), for example, introduced a supervised method, that combines different node level attributes. The method for predicting new links used in (Backstrom and Leskovec 2011) is based on supervised random walks.

Most of these works analyzed the predictability of new links in online social networks like co-authorship in DBLP or arXiv.org. The prediction of new links in real-world social contacts has been largely neglected. (Zhuang et al. 2012a) presented prediction techniques using location-based proximity as a proxy for face-to-face encounters and online social networks. In contrast, (Scholz, Atzmueller, and Stumme 2012) conducted a first analysis concerning the predictability of new links in real face-to-face contact networks. We extend this prior work by enriching the feature space with data from both the real and the online world.

The fundamental difference between our work and existing literature is that we analyze the relation between offline and online data for link prediction, and determine various influence factors in this context. To the best of the authors' knowledge, this is the first time that the connection between human contact networks and online information is analyzed for predicting face-to-face contacts. Furthermore, we also propose a novel hybrid algorithm for link prediction that incorporates a set of networks for improving the algorithmic performance.

3 Face-To-Face Contact Data

In this section, we describe the active RFID technology used for collecting the contact networks in conferences. We then define the link prediction problem. Next, we give an overview of the real world datasets collected at the LWA 2010, the Hypertext 2011, and the LWA 2012 conferences. We describe the contact networks, encounter data, co-authorship (DBLP) data, and the node-specific content information given by a set of manuscript data of the conference participants.

RFID Setup

At the LWA 2010 and Hypertext 2011 conferences we asked participants to wear active RFID devices that can sense and log the close-range face-to-face proximity of individuals wearing them. This allows us to map out time-resolved networks of face-to-face contacts among the conference attendees. In the following, we will refer to these active RFID tags as *proximity tags*.

A proximity tag sends out two types of radio packets: Proximity-sensing signals and tracking signals. Proximity radio packets are emitted at very low power and their exchange between two devices is used as a proxy for the close-range proximity of the individuals wearing them. Packet exchange is only possible when the devices are in close enough contact to each other (1-1.5 meters). The human body acts as an RF shield at the carrier frequency used for communication (Cattuto et al. 2010).

As in (Szomszor et al. 2010), we record a face-to-face contact when the length of a contact is at least 20 seconds. A contact ends when the concerning proximity tags do not detect each other for more than 60 seconds.

The proximity tags also send out tracking signals, at different power levels, that are received by antennas of RFID readers installed at fixed positions in the conference environment. These tracking signals are used to relay proxim-

ity information to a central server and also to provide approximate (room-level) positioning of conference participants (Scholz et al. 2011). This allows us to monitor encounters, e.g., the number of times a pair of participants is assigned to the same set of nearest readers. All the packets emitted by a proximity tag contain a unique numeric identifier of the tag, as well the identifiers of the detected nearby devices. For more information about the proximity sensing technology, we refer the reader to the website of SocioPatterns (http://www.sociopatterns.org).

Problem Statement

Let t be a point in time during the conference. For the prediction task, we consider all conversations starting before t as training data and conversations starting later as test data. The training data is thus the undirected graph $G^{\leq t} = (V^{\leq t}, E^{\leq t})$, where $V^{\leq t}$ is the set of all participants who had at least one face-to-face contact with some other participant before t. Two participants $u, v \in V^{\leq t}$ are connected by an edge $e := (u, v)$ in $E^{\leq t}$ if they had at least one face-to-face contact before t; the weight $w(e)$ is the sum of the durations of all their face-to-face contacts before t.

Let V_{core} be the set of participants who had at least one contact during the training interval and at least one contact during the test interval. As test data, we consider the graph $G^{>t} = (V_{\text{core}}, E^{>t}_{\text{core}})$: Two participants $u, v \in V_{\text{core}}$ are connected by an edge $e = (u, v)$ in $E^{>t}_{\text{core}}$ if they had at least one face-to-face contact after t. Following (Liben-Nowell and Kleinberg 2003), our aim is now to predict, for each pair of users who had no face-to-face-contact before t, i. e., for each $(u, v) \in (V^{\leq t} \times V^{\leq t}) \setminus E^{\leq t}$, whether $(u, v) \in E^{>t}_{\text{core}}$ holds or not. We compute a predictor score for each pair $(u, v) \in (V \times V) \setminus E^{\leq t}$. In an application, one would then set a threshold and predict all pairs with a predictor-score above the threshold. For evaluation purposes, however, we will follow the standard approach of determining the AUC value (see Section 5) directly based on the predictor scores. During the evaluation, we will also analyze if longer face-to-face contacts are easier to predict. Therefore, we also consider $G^{>t}$ as a weighted graph, where $w(u, v)$ is the sum of the durations of all face-to-face contacts of the participants u and v after t.

Datasets

For the link prediction task we combine face-to-face contact data with online co-authorship data, and node-specific information in order to achieve a better predictability of new links in the physical face-to-face contact network. Below, we provide a detailed overview on the collected RFID datasets, the co-author and full-text data.

RFID-Data At the conferences LWA 2010, Hypertext 2011, and LWA 2012, we collected three networks of face-to-face proximity using the SocioPatterns wearable sensor infrastructure described in Section 3. A link in the face-to-face proximity network indicates physical proximity between two conference participants; each link can be weighted by the cumulated duration of all contacts between the linked individuals.

	LWA 2010	HT 2011	LWA 2012
#days	3	3	3
$\lvert V \rvert$	77	68	42
$\lvert E \rvert$	1004	698	478
Avg.Deg.(G)	26.07	20.53	22.76
APL (G)	1.7	1.76	1.45
d (G)	3	4	3
AACD	797	529	1023
$\lvert V_{\text{core}} \rvert$	57	49	32
$\lvert E^{\leq t} \rvert$	426	481	263
$E^{>t}_{\text{core}} \setminus E^{\leq t}$	394	132	134

Table 1: General statistics for the collected datasets. Here d is the diameter, AACD the average aggregated contact-duration (in seconds) and *APL* the average path length. The last two lines represent the size of training and test datasets used for the evaluation.

Table 1 reports a detailed overview of the collected face-to-face proximity datasets. The distributions of the contact lengths of all aggregated face-to-face contacts between conference participants are heavy-tailed (see Figure 1), as observed in many other contexts (Cattuto et al. 2010; Isella et al. 2011b). More than 50% of all aggregated face-to-face contacts last less than 200 seconds and the average contact-duration is less than one minute, but very long contacts are also observed. The diameter, average degree and average path-length of G are similar to the results presented in (Isella et al. 2011b; Atzmueller et al. 2012).

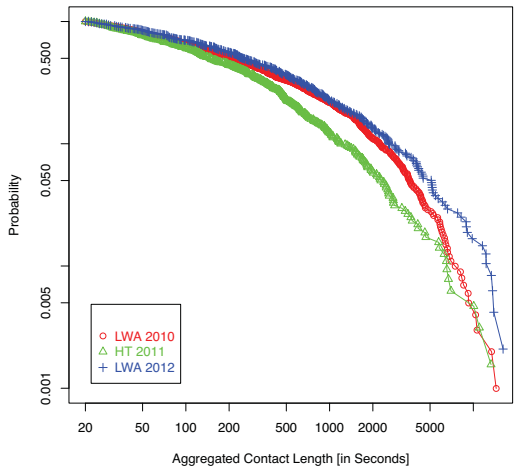


Figure 1: Distributions of the aggregated contact-durations for all pairs of individuals for the three conferences (LWA 2010, HT 2011, LWA 2012). The x-axis displays the duration of an aggregated face-to-face contact in seconds, the y-axis the probability of observing a link between two individuals having at least this duration. Both axes are scaled logarithmically.

In addition, we recorded (for each 20 second time slot)

for each conference participant the closest, second-closest, and third-closest RFID reader. We define the encounter value $enc_i(x, y)$ between two participants x and y as the number of time slots in which participants x and y are assigned to the same set of $i = 1, 2, 3$ closest readers. We can then create the according encounter-networks $encounter_i$, for which an edge between node x and node y (denoting the respective participants) is weighted by $enc_i(x, y), i = 1, 2, 3$. We only create an edge between nodes x and y, if the respective weight is greater than zero.

DBLP Data For our link prediction task we combine online data collected from the co-authorship network DBLP with face-to-face proximity data collected during the three conferences. Accordingly, we collected the co-authorship network from DBLP for the years 2010, 2011, and 2012.

In Table 2 we summarize the number of nodes and edges of the obtained DBLP-networks for the different years. In order to provide an overview of the connections in the network, Table 3 reports the distribution of shortest path distances between pairs of conference participants. A shortest path distance of 1 indicates co-authorship. Furthermore, a shortest path distance of ∞ indicates that either no path exists between the conference participants x and y, or no DBLP entry for at least one of the participants x or y exists. The large number of pairs with distance ∞ at LWA 2010 (compared to HT 2011 and LWA 2012) is due to the fact that some student assistants (with no publications yet) joined the conference participants.

	2010	**2011**	**2012**
$\|V\|$	962.318	1.062.140	1.112.337
$\|E\|$	7.056.220	8.053.134	8.606.320

Table 2: Statistics on the co-authorship network DBLP collected for the years 2010, 2011 and 2012.

Manuscript Data In addition to the co-authorship network of DBLP we analyzed the content of all papers from all conference participants since 2006 that are listed in DBLP. For each participant, we created a bag-of-words representation, as a *paper profile*: We considered the union of the papers of the respective participant for constructing the bag-of-words. As preprocessing, we applied the Porter Stemmer algorithm and removed a number of stop words. In Table 4, we give a short overview of the manuscript-data of all conference participants. Figure 2 displays the cumulative distribution of the number of papers crawled for each participant for the three conferences.

4 Link Prediction

The goal of link prediction scenarios is the prediction of new links, i. e., all links in $E^{>t}_{\text{core}} \setminus E^{\leq t}$. In this section we first describe the applied evaluation method. Then, we analyze influence factors for link prediction considering co-authorship, paper similarity, and encounter. Finally we present the proposed novel unsupervised *Hybrid Rooted PageRank* link prediction method.

	LWA 2010	**HT 2011**	**LWA 2012**
1	49 (0.898)	18 (0.889)	16 (0.94)
2	151 (0.543)	53 (0.283)	30 (0.767)
3	360 (0.425)	182 (0.267)	112 (0.571)
4	341 (0.304)	427 (0.286)	142 (0.507)
5	121 (0.179)	311 (0.402)	47 (0.744)
6	13 (0.25)	83 (0.385)	4 (1)
7	0	7 (0.43)	0
∞	1891 (0.31)	810 (0.344)	510 (0.519)

Table 3: Statistics on the shortest-path distances between pairs of conference participants in DBLP network. The first line (shortest path distance 1) gives the number of co-authors for each conference, the second line gives the number of authors with distance two in the DBLP network. In brackets we present the fraction of pairwise particpants (for the corresponding shortest path-distance) who had a face-to-face contact. At LWA 2010, for example, there are 49 pairs of participants with distance one and 151 pairs of participants with distance two in the DBLP co-authorship network. Concerning the co-authors, 89.8% of all co-authors had at least a short face-to-face contact at the LWA 2010.

	LWA 2010	**HT 2011**	**LWA 2012**
$\|P\|$	1187	608	470

Table 4: Numbers of papers published since 2006 by the participants, resp., according to DBLP.

Evaluation Method

For the evaluation of link prediction measures, the precision of the top n predicted links is often used (Liben-Nowell and Kleinberg 2003), where n is the number of positive events (i.e., the real number of observed new links, in our case in the second and third days of a conference, taking the first day as training set.

In this work, we measure the accuracy by using the area under a receiver operating characteristic (AUC), e. g., (Hanley and McNeil 1982), i.e., the area under the ROC plot with the true positive rate on the y-axis and the false positive rate on the x-axis. The advantage of AUC is that it considers the whole ranking instead of focusing on only the top-n positions.

For link prediction, AUC has already been used, e. g., in (Lichtenwalter, Lussier, and Chawla 2010). For the prediction of new links each network proximity measure (predictor) outputs a ranked list in decreasing order of confidence. Since we know the real contacts of the second and third day we can evaluate the AUC value for each proximity measure.

Influence Factors for Link Prediction

The prediction of new links in face-to-face contact networks is a challenging and difficult problem. Knowledge on possible influence-factors of human communication behavior is therefore an essential asset. In the following, we provide some new insights into influence factors for human communication behavior at a conference.

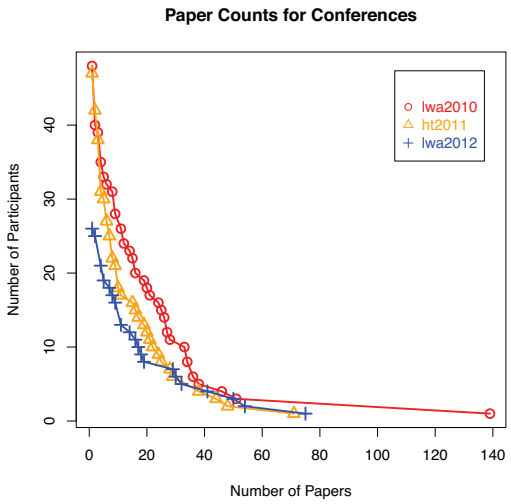


Figure 2: Distribution of the number of papers for each participant analyzed at the three conferences. The x-axis displays the minimum number of papers, the y-axis the number of participants having at least this paper count.

1. We present connections between the "digital world" and the real world networks, by analyzing the relationship between the distance in the co-authorship network DBLP and the aggregated face-to-face contact length of participants. We mention here that we use the aggregated contact over the whole conference in this analysis.
2. We analyze, whether pairs of participants are more frequently in face-to-face contact when their papers' bag-of-words model has a higher cosine similarity.
3. Finally we evaluate the relation between participants' encounter value and participants' face-to-face proximity, in order to assess the influence of frequent locational encounters on face-to-face contacts.

Co-Authorship Figure 3 shows the aggregated face-to-face contact length distributions for pairs of participants having different shortest path distances in the DBLP co-author network. The figure indicates that co-authorship has a significant influence on the contact durations. Pairs of participants at distance one and two have a higher probability for a longer duration of their aggregated face-to-face proximity contact than pairs of participants with a shortest path distance greater than three. In Table 3 we see (as expected) that most of the co-authors have at least a short face-to-face contact during the conference.

As expected, we observe that co-authors have the highest probability for a face-to-face contact. For the LWA 2010 dataset, we even observe the trend that participants have a higher probability for a face-to-face contact when they are close in the DBLP-network, while this can only be observed for smaller distances in the HT 2011 and LWA 2012 datasets.

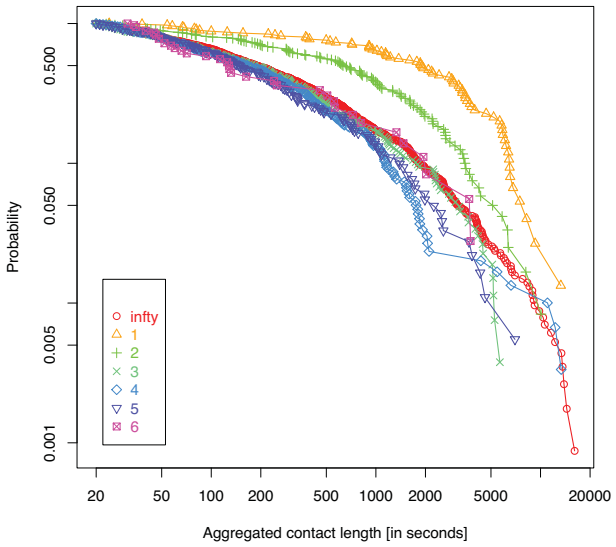


Figure 3: Aggregated contact length distributions between pairs of participants according to the respective different shortest path-distances in the DBLP network.

Paper Profile Similarity As described in Section 3, we analyzed the content of the papers of conference participants considering all papers since 2006 that were listed in DBLP, and created a bag-of-words representation for each of them. We weighted terms with a simple *term frequency* (TF) weighting, such that a term is weighted by the number of times it occurs in the bag-of-words.

Table 5 shows the influence of the paper-similarity profile, especially for longer relative contact thresholds. For the contacts above a normalized 10% level (with respect to the longest contact) we observe a large increase in AUC that monotonically dominates the filtered contacts.

Encounter We also used the encounters between all conference participants as an indicator for a face-to-face contact. As defined in Section 3 the encounter value of two participants x and y is computed as the number of time slots in which participants x and y are assigned to the same set of $i = 1, 2, 3$ closest readers. The basic time interval we choose is the minimum face-to-face contact length of 20 seconds.

Table 6 shows that the encounter value on the LWA2010 and LWA2012 dataset is good proxy for a face-to-face proximity. Like the paper-similarity the encounter-value is better suited to predict longer face-to-face contacts. Surprisingly, we obtain very good results even with just one RFID reader.

The Hybrid Rooted PageRank Method

In this section, we present a new unsupervised prediction method, *Hybrid Rooted PageRank*, that combines the information of different networks. The proposed algorithm is an extension of the *rooted PageRank* algorithm (Liben-Nowell and Kleinberg 2003).

rclt	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
AUC (LWA 2010)	0.559	0.622	0.628	0.645	0.668	0.681	0.691	0.687	0.677	0.670	0.664
Size	880	350	242	171	133	111	85	68	64	54	46
AUC (HT 2011)	0.500	0.513	0.549	0.564	0.574	0.565	0.567	0.549	0.584	0.610	0.622
Size	612	257	182	148	115	94	76	66	55	48	40
AUC (LWA 2012)	0.530	0.521	0.553	0.581	0.573	0.563	0.559	0.544	0.531	0.562	0.559
Size	338	167	111	86	66	53	43	33	31	27	24

Table 5: Statistics on users' paper cosine similarity and area under curve for different relative contact length thresholds (rclt). The relative contact length thresholds means that each participant's contact length is normalized by the maximal contact length of a participant).

rclt	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
AUC (LWA 2010, 3 Reader)	0.745	0.802	0.820	0.850	0.846	0.832	0.847	0.848	0.856	0.851	0.886
AUC (LWA 2010, 2 Reader)	0.752	0.804	0.825	0.855	0.855	0.847	0.859	0.864	0.879	0.878	0.878
AUC (LWA 2010, 1 Reader)	0.745	0.802	0.820	0.850	0.846	0.832	0.847	0.849	0.856	0.851	0.848
Size	2008	697	468	323	252	196	154	120	105	88	77
AUC (HT 2011, 3 Reader)	0.653	0.644	0.653	0.662	0.677	0.687	0.701	0.689	0.683	0.700	0.716
AUC (HT 2011, 2 Reader)	0.674	0.663	0.676	0.686	0700	0.700	0.719	0.709	0.701	0.703	0.712
AUC (HT 2011, 1 Reader)	0.677	0.667	0.684	0.689	0.697	0.714	0.727	0.723	0.716	0.721	0.739
Size	1280	518	355	265	194	150	122	101	83	74	63
AUC (LWA2012, 3 Reader)	0.682	0.780	0.791	0.808	0.822	0.848	0.857	0.850	0.836	0.832	0.826
AUC (LWA2012, 2 Reader)	0.695	0.797	0.816	0.832	0842	0.870	0.880	0.871	0.856	0.855	0.844
AUC (LWA2012,1 Reader)	0.705	0.794	0.815	0.823	0.837	0.855	0.870	0.866	0.861	0.858	0.847
Size	956	370	239	195	140	105	88	70	59	49	42

Table 6: Statistics on user's encounter value and area under curve for different relative contact length thresholds (rclt). The encounter value of each participant p is also normalized by the maximal encounter value of participant p.

The *Hybrid Rooted PageRank* computes the stationary distribution of nodes under the random walk described in Algorithm 1. In each step the walk selects a network with respect to a given probability distribution. A link in this network is then selected using link-weights as transition probabilities. When no link exists in the chosen network (i.e., the node is isolated), the algorithm jumps back to the root node. In this way, one can integrate different networks for the link prediction.

The *Hybrid Rooted PageRank* is executed for each conference participant (i. e., the conference participant is the root node r). We then build and evaluate the ranking of the predicted links between the root and the other nodes using the values of the stationary distributions of the *Hybrid Rooted PageRank*. The major advantage of this approach is given by the observation that different networks can complement each other in the prediction task using the proposed hybrid algorithm. In addition, the network probabilities allow us to evaluate the influence of each individual network. Note that the *Hybrid Rooted PageRank* is exactly the *rooted PageRank* in the case when there is only one network.

Input : Networks $N = \{N_1, \ldots, N_n\}$,
Network-Probabilities $P = \{p_1, \ldots, p_n\}$,
Probability α, Root node r
Output: Stationary distribution weight of node v under the following random walk:

With probability α jump to root node r.
With probability $1 - \alpha$:
  Choose Network $N_i \in N$ with respect to probability distribution P.
  if *There exist no outgoing edges* **then**
    Jump to root node r
  else From the current node c jump to a neighbor n selected with a probability $\frac{w(c,n)}{\sum_{c \to d} w(c,d)}$, i. e., proportional to the weight $w(c, n)$ of the edge (c, n).

Algorithm 1: *Hybrid Rooted Random Walk*

5 Evaluation

In this section, we evaluate the proposed novel unsupervised link prediction method and analyze how different networks interact, when new links are created. We start with a discussion of baseline predictors, before we present and discuss the results.

Baseline Predictors

In this section we discuss baseline predictors for unsupervised link prediction. Liben and Kleinberg (Liben-Nowell and Kleinberg 2003) proposed and analyzed several unsupervised link prediction methods. Most of these methods generate predictor scores based on the nodes' neighborhood (e.g. *Adamic Adar*, *Resource Allocation*, *Common Neighbors*, *Jaccard's coefficient*, ...) or path information (e.g. *Katz*, *rooted PageRank*, and more). The authors show that the network proximity measures *Adamic Adar* (Adamic and Adar 2003), *Common Neighbors* and Katz (Katz 1953) perform very well. In (Katz 1953) the authors also show that *unweighted Katz* is more effective than the weighted variant.

The measure *Common Neighbors* is based on the assumption that it is more likely that two nodes become connected if they have many neighbors in common. *Adamic Adar* is similar to *Common Neighbors*, but here the common neighbors are weighted with respect to their degree. *Jaccard's coefficient* divides the number of common neighbors by the number of total neighbors.

The proximity measure *Preferential Attachment* is simply the product of the degrees of the corresponding nodes. Zhou et al. (Zhou, Lu, and Zhang 2009) present and analyze a new measure called *Resource Allocation*. This measure is similar to *Adamic Adar*, but in (Zhou, Lu, and Zhang 2009) the authors show that in most cases it performs better than *Adamic Adar*.

In (Lü and Zhou 2010; Murata and Moriyasu 2007; Scholz, Atzmueller, and Stumme 2012) the authors presented and analyzed weighted variants for most of these predictor scores. All these proximity measures are based on the assumption that two nodes have a higher probability of becoming connected in the future when they are close to each other in the network.

(Scholz, Atzmueller, and Stumme 2012) show that *weighted Resource Allocation* and *weighted Jaccard's coefficient* perform well for the prediction of new links in networks of face-to-face proximity. In our experiments we consider *weighted Resource Allocation*, *weighted Jaccard's coefficient* and *rooted PageRank* as baseline predictors in our experiments: We compare the results of the *Hybrid Rooted PageRank* to those predicted by these baseline predictors. For a pair of nodes (x, y) *weighted Resource Allocation* is defined as (Lü and Zhou 2010):

$$WRA(x,y) = \sum_{z \in N(x) \cap N(y)} \frac{w(x,z) + w(y,z)}{\sum_{z' \in N(z)} w(z',z)},$$

where $N(x) = \{y | y \in V, (x,y) \in E\}$ is the neighborhood of a node x. The proximity measure *weighted Jaccard's coefficient* is defined as (Scholz, Atzmueller, and Stumme 2012):

$$WJC(x,y) = \frac{\sum_{z \in N(x) \cap N(y)} w(x,z) + w(y,z)}{\sum_{x' \in N(x)} w(x,x') + \sum_{y' \in N(y)} w(y,y')}.$$

A description of *rooted PageRank* algorithm can be found in the section on the *Hybrid Rooted PageRank* method.

Results and Discussion

For our experiments we considered three real-world datasets collected at the LWA 2010, HT 2011 and LWA 2012 conferences. We compare the predictor-results of the *Hybrid Rooted PageRank* algorithm described in Algorithm 1 with those predicted by *weighted Resource Allocation*, *weighted Jaccard's coefficient* and *rooted PageRank*.

In addition, we analyze and compare the prediction performance of different networks: Face-to-face proximity contact network, paper-based similarity network, DBLP co-authorship network and the encounter network. We created the paper-similarity network analogously to the encounter network, weighting each link by the cosine similarity between the respective bag-of-words profiles.

In our analysis, we especially focus on the combination of two networks: The DBLP co-authorship network and the face-to-face proximity contact network. As return parameter (i.e., parameter α) for the *Hybrid Rooted PageRank* and *rooted PageRank* we used $\alpha = 0.15$; we tried different parameters, and finally selected $\alpha = 0.15$ in accordance with (Liben-Nowell and Kleinberg 2003), where this value also yields satisfactory performance.

We also use a parameter $\beta \in [0,1]$ that gives the probability to choose the DBLP co-authorship network at each step of the random walk ($1 - \beta$ is the probability to choose the face-to-face proximity contact network). Finally, we use an increasing time threshold x that filters out all new links with weight lower than x. This means, given time threshold x, that

$$NewLinks(x) = \{e \in \{E_{\text{core}}^{>t} \setminus E^{\leq t}\} | w(e) \geq x\}$$

is the set of new links we want to predict. Table 7 provides an overview of the number of new links for the different time thresholds. In our experiments, we also want to analyze the predictability of stronger links. For this purpose, we need to make the contact length distributions of different participants comparable. Therefore, we normalize each face-to-face contact length of participant $v \in V$ by the maximal contact length of participant v. The result is that the contact graph becomes directed (from the normalization step it follows that for an edge $(u,v) \in E$ it is not necessarily valid that $w(u,v) = w(v,u)$). Analogously, we also normalize the encounter values in the different networks for the respective participants.

Figure 4 shows the results for the HT2011 dataset, focussing only on participants with an edge in the co-authorship network DBLP. This means that we focus on conference participants with at least one co-authorship. We plot the results for *Hybrid Rooted PageRank* with parameters $\beta \in \{0, 0.5, 1\}$. Choosing $\beta = 0$ corresponds to the *rooted PageRank* with a random walk on the face-to-face proximity contact network. For $\beta = 1$ the random walk is only executed on the co-authorship network DBLP. The results of Figure 4 show that the interaction of the DBLP and face-to-face proximity network performs better for predicting new links than each network separately. In addition, we

observe that the *Hybrid Rooted PageRank* performs better on stronger links than on weaker links. Here, we note that a time threshold of $x \in [0, 1]$ means that we filter out the x percent of the shortest face-to-face contacts of each conference participant.

In our prediction task we also want to predict new links between participants who have no link in the DBLP co-author network. In Figure 5 we use the *Hybrid Rooted PageRank* on different networks and compare the result to those of the baseline predictors. In this figure, Contact+DBLP means that we use the *Hybrid Rooted PageRank* with $\beta = 0.5$ on the combination of the Face-to-Face Contact and DBLP-network. We observe, that the *Hybrid Rooted PageRank* outperforms unsupervised baseline methods for time thresholds greater than 0.1 on all datasets. Considering the low number of co-authorships at each conference, this is a surprising result. On the LWA 2010 dataset the network proximity measure *weighted Jaccard's coefficient* has a compareable performance for time thresholds smaller than 0.4. Furthermore it is interesting to see, that the prediction on the paper similarity network works a bit better than the prediction on the DBLP co-author network. Additionally the face-to-face proximity network performs better than the encounter network. This confirms the intuition that proximity information is more effective to predict new links. However, the advantage of the encounter network is that it contains information about talks attended together at the conference.

Focusing on all new contacts (time threshold is 0), the network proximity measure *Resource Allocation* shows the best results. Here we argue, that links in the DBLP-network are better suited to predict new links that lead to longer face-to-face proximity. This statement is also supported by Figure 4. Here we see, that focussing on all links the DBLP-network performs very weakly. For longer relative thresholds the DBLP-network performs as well as the face-to-face contact network.

Figure 6 reports the influence of the DBLP co-authorship network and face-to-face contact network on the prediction of new links. Here, we predict the AUC-values on each each dataset for each $\beta \in \{0, 0.1, 0.5, 0.9, 1\}$ using the *Hybrid Rooted PageRank* algorithm. We see that on the LWA2010 dataset $\beta = 0.9$ performs best (that means, in each step of the random walk the probability to choose the DBLP network is 0.9, and therefore the probability to choose the face-to-face contact network 0.1). On the HT2011 dataset $\beta = 0.5$ performs best.

Overall, we observe that for all tested β-values we get better prediction results on all datasets, if we combine face-to-face proximity data with DBLP data. In addition we notice that a higher weight of the DBLP-network does not necessarily lead to better prediction results.

6 Conclusions and Future Work

In this paper, we considered the predictability of human face-to-face proximity and presented a novel unsupervised link prediction technique that combines different networks in a hybrid approach. We showed that the proposed approach outperforms state-of-the-art unsupervised link prediction methods. Furthermore we demonstrated that the predictability of human face-to-face contact networks could be improved by using information from online social networks. In addition we studied several factors influencing the face-to-face proximity of individuals at conference gatherings.

One open research question is the estimation of optimal weights for the different networks. For a better parameter estimation we aim to extend our analysis to more datasets. An alternative approach is the automatic parameter estimation of weights for the different networks. Here our goal is to develop an adequate weighting scheme using the network topology of the training data. However it is unclear whether the given network structure is appropriate for this task.

Like most past link prediction studies this work also focused on an unsupervised approach. Another open research question is whether a supervised approach is preferable in our scenario. In future work, we will analyze if a supervised approach could help to further improve the prediction quality using features of online and offline networks.

	LWA 2010	HT 2011	LWA 2012
0	788	264	268
0.1	261	123	103
0.2	163	83	65
0.3	106	68	50
0.4	85	52	37
0.5	62	45	28
0.6	44	39	23
0.7	35	36	19
0.8	28	28	17
0.9	22	27	15
1	20	23	10

Table 7: Statistics on the number of new links for various normalized time thresholds (in the first column).

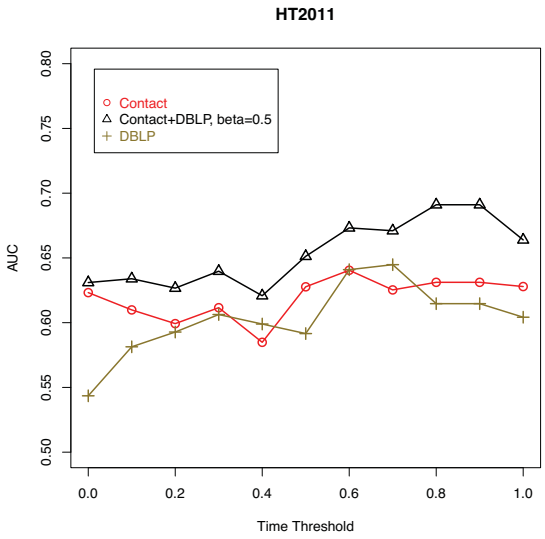
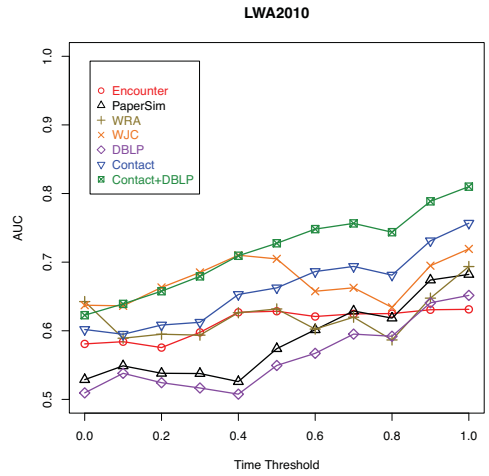
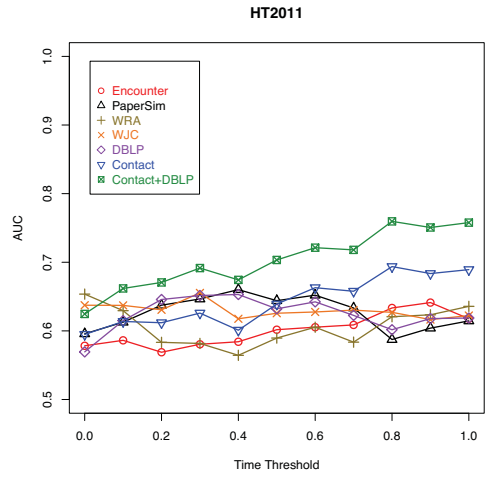


Figure 4: AUC-values of the contact prediction restricted to participants with a DBLP-Account. Here we run the *Hybrid Rooted PageRank* with parameters $\beta \in \{0, 0.5, 1\}$.





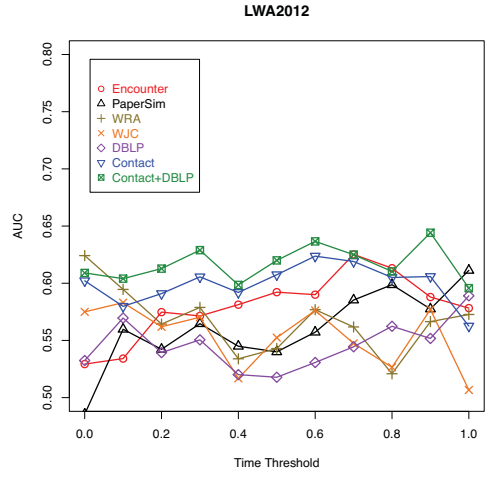
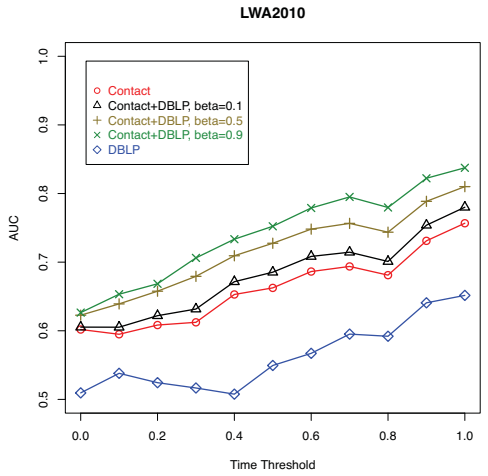
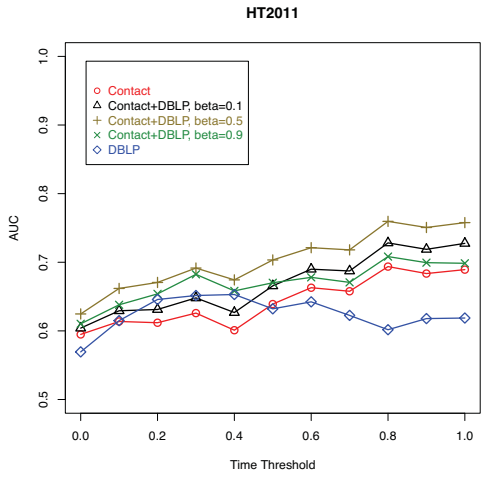


Figure 5: AUC-values of the contact predictions with *Hybrid Rooted PageRank* for different networks. WRA means *weighted Resource Allocation*, WJC means *weighted Jaccard's coefficient* on the contact network.





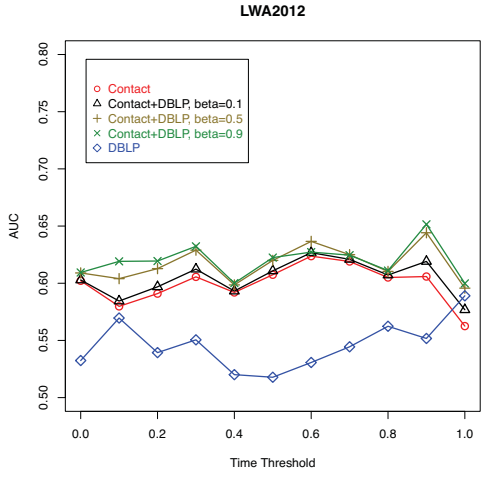


Figure 6: AUC-values of the contact prediction with *Hybrid Rooted PageRank* for different β-parameters. We choose $\beta \in \{0, 0.1, 0.5, 0.9, 1\}$.

Acknowledgements

This work has been supported by the VENUS research cluster at the interdisciplinary Research Center for Information System Design (ITeG) at Kassel University.

We thank the SocioPatterns collaboration for providing privileged access to the SocioPatterns sensing platform that was used in collecting the contact data.

References

Adamic, L. A., and Adar, E. 2003. Friends and Neighbors on the Web. *Social Networks* 25(3):211–230.

Alani, H.; Szomszor, M.; Cattuto, C.; den Broeck, W. V.; Correndo, G.; and Barrat, A. 2009. Live Social Semantics. In *Intl. Semantic Web Conf.*, 698–714.

Atzmueller, M.; Benz, D.; Doerfel, S.; Hotho, A.; Jäschke, R.; Macek, B. E.; Mitzlaff, F.; Scholz, C.; and Stumme, G. 2011. Enhancing Social Interactions at Conferences. *it+ti* 3:1–6.

Atzmueller, M.; Doerfel, S.; Hotho, A.; Mitzlaff, F.; and Stumme, G. 2012. Face-to-Face Contacts at a Conference: Dynamics of Communities and Roles. In *Modeling and Mining Ubiquitous Social Media*, volume 7472 of *LNAI*. Springer.

Backstrom, L., and Leskovec, J. 2011. Supervised Random Walks: Predicting and Recommending Links in Social Networks. In *WSDM*, 635–644.

Barrat, A.; Cattuto, C.; Szomszor, M.; den Broeck, W. V.; and Alani, H. 2010. Social Dynamics in Conferences: Analyses of Data from the Live Social Semantics Application. In *Intl. Semantic Web Conf.*, 17–33.

Cattuto, C.; Van den Broeck, W.; Barrat, A.; Colizza, V.; Pinton, J.-F.; and Vespignani, A. 2010. Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. *PLoS ONE* 5(7):e11596.

Eagle, N.; Pentland, A.; and Lazer, D. 2009. From the Cover: Inferring Friendship Network Structure by using Mobile Phone Data. *Proceedings of The National Academy of Sciences* 106:15274–15278.

Hanley, J. A., and McNeil, B. J. 1982. The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve. *Radiology* 143(1):29–36.

Hui, P.; Chaintreau, A.; Scott, J.; Gass, R.; Crowcroft, J.; and Diot, C. 2005. Pocket Switched Networks and Human Mobility in Conference Environments. In *Proc. ACM SIGCOMM Workshop on Delay-tolerant Networking*, WDTN '05, 244–251. New York, NY, USA: ACM.

Isella, L.; Romano, M.; Barrat, A.; Cattuto, C.; Colizza, V.; Van den Broeck, W.; Gesualdo, F.; Pandolfi, E.; Ravà, L.; Rizzo, C.; and Tozzi, A. 2011a. Close Encounters in a Pediatric Ward: Measuring Face-to-Face Proximity and Mixing Patterns with Wearable Sensors. *PLoS ONE* 6:e17144.

Isella, L.; Stehlé, J.; Barrat, A.; Cattuto, C.; Pinton, J.-F.; and Broeck, W. V. D. 2011b. What's in a Crowd? Analysis of Face-to-Face Behavioral Networks. *Journal of Theoretical Biology* 271:166–180.

Katz, L. 1953. A New Status Index Derived from Sociometric Analysis. *Psychometrika* 18(1):39–43.

Liben-Nowell, D., and Kleinberg, J. M. 2003. The Link Prediction Problem for Social Networks. In *CIKM*.

Lichtenwalter, R., and Chawla, N. V. 2012. Vertex Collocation Profiles: Subgraph Counting for Link Analysis and Prediction. In *WWW*, 1019–1028.

Lichtenwalter, R.; Lussier, J. T.; and Chawla, N. V. 2010. New Perspectives and Methods in Link Prediction. In *KDD*, 243–252.

Lü, L., and Zhou, T. 2010. Link Prediction in Weighted Networks: The Role of Weak Ties. *EPL* 89:18001.

Macek, B. E.; Scholz, C.; Atzmueller, M.; and Stumme, G. 2012. Anatomy of a Conference. In *23rd ACM Conference on Hypertext and Social Media, HT '12*, 245–254. Milwaukee, WI, USA, June 25-28, 2012: ACM.

Murata, T., and Moriyasu, S. 2007. Link Prediction of Social Networks Based on Weighted Proximity Measures. In *Web Intelligence*, 85–88.

Scholz, C.; Atzmueller, M.; and Stumme, G. 2012. On the Predictability of Human Contacts: Influence Factors and the Strength of Stronger Ties. In *SocialCom 2012*. Boston, MA, USA: IEEE Computer Society.

Scholz, C.; Doerfel, S.; Atzmueller, M.; Hotho, A.; and Stumme, G. 2011. Resource-Aware On-Line RFID Localization Using Proximity Data. In *Proceedings ECML/PKDD)*, volume 6913 of *LNCS*, 129–144. Berlin: Springer.

Stehlé, J.; Voirin, N.; Barrat, A.; Cattuto, C.; Isella, L.; Pinton, J.-F.; Quaggiotto, M.; Van den Broeck, W.; Régis, C.; Lina, B.; and Vanhems, P. 2011. High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. *PLoS ONE* 6(8):e23176.

Szomszor, M.; Cattuto, C.; den Broeck, W. V.; Barrat, A.; and Alani, H. 2010. Semantics, Sensors, and the Social Web: The Live Social Semantics Experiments. In *ESWC (2)*, 196–210.

Wang, D.; Pedreschi, D.; Song, C.; Giannotti, F.; and Barabási, A.-L. 2011. Human Mobility, Social Ties, and Link Prediction. In *KDD*, 1100–1108.

Zhou, T.; Lu, L.; and Zhang, Y.-C. 2009. Predicting Missing Links via Local Information. *Europ. Phys. Journal B - Condensed Matter and Complex Systems* 71:623–630.

Zhuang, H.; Chin, A.; Wu, S.; Wang, W.; Wang, X.; and Tang, J. 2012a. Inferring Geographic Coincidence in Ephemeral Social Networks. In *ECML/PKDD*. Springer.

Zhuang, H.; Tang, J.; Tang, W.; Lou, T.; Chin, A.; and Wang, X. 2012b. Actively Learning to Infer Social Ties. *Data Min. Knowl. Discov.* 25(2):270–297.

Zuo, X.; Chin, A.; Fan, X.; Xu, B.; Hong, D.; Wang, Y.; and Wang, X. 2012. Connecting People at a Conference: A Study of Influence between Offline and Online Using a Mobile Social Application. In *GreenCom*, 277–284. Los Alamitos, CA, USA: IEEE Computer Society.