On the Predictability of Human Contacts: Influence Factors and the Strength of Stronger Ties

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

***Abstract*—While the analysis of online social networks is a prominent research topic, offline real-world networks are still not covered extensively. However, their analysis can provide important insights into human behavior. In this paper, we analyze influence factors for link prediction in human contact networks. Specifically, we consider the prediction of new links, and extend it to the analysis of recurring links. Furthermore, we consider the impact of stronger ties for the prediction. The results and insights of the analysis are a first step onto predictability applications for human contact networks.**

I. Introduction

With the growing amount of social data, ubiquitous systems, and mobile social media applications transcending everyday life, the analysis of social networks is receiving increased attention. This especially relates to the dynamics and the creation of links between the networks' subjects [1], e. g., concerning their mobility [32], [8] and dynamic behavior [33], [31]. While there is a large body of research concerning *online* social networks, e. g., [21], [27], [14], [23], [18], [36], [19], [20], important aspects of *offline* social networking still remains largely unexplored. The analysis of such networks can potentially provide more direct answers to fundamental questions, e. g., how do personal links get established, is it possible to correlate this with roles, how does the intensity of personal communication evolve?

In this paper, we aim at providing first insights for answering such questions. We focus on real-world offline networks of *human contacts*, that is, *face-to-face* conversations between persons. In contrast to virtual networks, these contacts were acquired using a ubiquitous RFID-based system that allows us to collect face-to-face contacts. Thus, we can observe and analyze social interaction at a very detailed level, including the specific event sequences and durations.

We consider link prediction [21], [27], [14], [23], [18], [36], [17], [16] in the context of networks of human contacts. We aim to predict *new* contacts based on network properties, as an adaptation of methods for online social networks. In addition, we extend the analysis in two important directions: First, we consider the length of the contacts in more detail, and analyze the impact of longer conversations. Second, we consider the prediction of future *recurring* contacts, i. e., renewed contacts between specific actors. For these, we analyze influence factors and patterns for establishing such contacts, and also consider their specific *durations* in a fine-grained dynamic analysis. Essentially, this leads to the analysis of the impact of *stronger ties* for new and recurring contacts.

For the analysis, we apply real-world data collected at the LWA 2010 conference in Kassel, Germany, and the Hypertext 2011 conference in Eindhoven, The Netherlands, using the Conferator [3] system. The results of the analysis indicate that stronger ties have a strong influence on the contact behavior and the prediction performance. We show, that there are clear influence patterns of the contact durations. Furthermore, considering the contact durations in the ranking of the predicted contacts significantly improves the performance. This can be generalized for both conferences.

Our contribution is four-fold and can be summarized as follows:

1) Concerning link prediction, we analyze the problem of predicting links in real-world *human contact* networks, focusing on *new* links.
2) We extend the basic link prediction problem setting for predicting *recurring* links, considering different event windows, e.g., day one vs. the subsequent days.
3) We consider (and adapt) different state-of-the-art network proximity measures for the prediction.
4) Finally, we analyze the influence of stronger ties for the prediction and show its impact using real-world data of two conferences.

The context of our work is established by the social conferencing application Conferator [3] implemented using the Ubicon system.[1] It provides ubiquitous access to conference information and allows conference participants to manage their contacts at the conference and to personalize their conference program. Using the system, conference participants can recall their individual contacts after the conference, e. g., as virtual business cards. In addition, recommendations for contacting *interesting* persons is provided. The system utilizes active RFID technology from the Sociopatterns project[2] which allows us to analyze the collected contact (proximity) data between the participants as a proxy for their face-to-face conversations.

The rest of this paper is structured as follows: In Section II we discuss related work. Section III describes the RFID hardware setting and the collected datasets. After that, we present the used network proximity measures and link prediction techniques in Section IV. We discuss the results in Section V. Finally, Section VI concludes with a summary and interesting options for future work.

[1] http://www.ubicon.eu
[2] http://www.sociopatterns.org

II. Related Work

Below, we discuss related work concerning the analysis of human contact behavior, and its connection to link prediction.

A. Analysis of Human Contacts

Contacts patterns in social networks, and their underlying mechanisms, e.g., homophily [25] are a classic topic of social network analysis. However, the analysis of offline social networks, focusing on human contacts, has been largely neglected. In this context, Eagle et al. [12] and Zhoe et al. [14] presented an analysis of proximity information collected by devices based on Bluetooth communication, similar to Xu et al. [35], who also related this to online social networks. However, in all these experiments it was not possible to detect reliable face-to-face contacts. In our experiments, we use a new generation of active RFID tags (proximity tags). The technical innovation of these tags is the possibility to detect the proximity of other tags, which allows us to recognize face-to-face contacts at a high detailed level including specific points in time and their durations.

One of the first experiments using proximity tags was conducted by Cattuto and colleagues in [2] at the ESWC 2009 conference. Here, the authors presented a novel application that combines online and offline data from the conference attendees. In [11], Cattuto and colleagues compared the attendees' contact patterns with their research seniority and their activity in social web platforms. They also extended their analysis to healthcare environments [9] and schools [29]. However, no analysis or application towards link prediction has been performed using the approaches discussed above. We discuss this important aspect in more detail below.

B. Link Prediction

The prediction of new links between nodes in a social network is a challenging task. A first comprehensive fundamental analysis was done by Liben-Nowell and Kleinberg in [21]. Here, the authors defined the link prediction problem and studied link prediction approaches based on proximity measures of nodes in a co-authorship network of physics. Wang et al. examined the impact of human mobility on link prediction in [32]. In [27] Murata and Moriyasu presented weighted variants of the network proximity measures *Adamic-Adar*, *Common Neighbors* and *Preferential Attachment*. The authors applied the weighted measures to networks of question-answer bulletin boards systems and showed that these measures outperform existing measures. In [23] the authors present an approach to analyzing the role of weak ties in social networks.

The fundamental difference between our work and existing literature is that to the best of our knowledge we present the first link prediction analysis of a human face-to-face contact network. In addition, we extend our studies to the predictability of strong ties and recurring links. Furthermore, we present new insights into the communication behavior of participants during a conference. We expect, that these results can help to improve the quality of link prediction in social networks in the future.

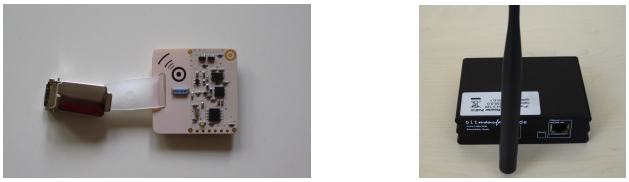


Fig. 1. Proximity Tag (left) and RFID Reader (right)

III. Face-to-Face Contact Data

In the following section, we first describe the applied active RFID technology for collecting the data and constructing the network of human contacts. After that, we define our problem setting for link detection, and describe, how we model the underlying networks. Next, we present the two datasets collected at the LWA 2010,[3], and the Hypertext 2011[4] conferences, and provide initial characteristic statistics.

A. RFID Setup

For our experiments we asked each conference participant to wear an active RFID tag (see Figure 1). These so called proximity tags are developed by the SocioPatterns project. One decisive factor of these tags is the possibility to detect other proximity tags within a range of up to 1.5 meters which allows us to identify and analyze human face-to-face contacts. Each RFID tag sends signals to RFID readers that are placed at fixed positions in the conference area. The RFID readers (see Figure 1) forward these signals to a central server, where all signals are stored into a database. Each signal contains the ID of the transmitting tag and the IDs of all RFID tags in its proximity. At both conferences we also offered a visualization of each participant's localitation. To determine the location of each participant, we used techniques described in [28] and [26]. For more information about the proximity tags we refer to Barrat et al. [10] and the OpenBeacon website.[5]

B. Problem Statement

We model the social network as an undirected multi-graph $G = (V, E)$, where V is the set of participants and an edge $e = (x, y) \in E$ with weight $w(e)$ represents a face-to-face contact between two participants x and y with contact duration $w(e)$. Additionally, each edge $e \in E$ is labeled with the start time $t(e)$ of the conversation. As in [30], 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 a signal from each other for more than 60 seconds.

Let t_s be the starting time of the conference, t_e its end time, and $t \in [t_s, t_e]$. We consider all conversations during $[t_s, t]$ as training data and conversations during $(t, t_e]$ as test data for the prediction task (For the sake of simplicity, we assume that there is no conversation taking place at time t, which holds in particular whenever t is set during the night).

[3] http://www.kde.cs.uni-kassel.de/conf/lwa10/
[4] http://www.ht2011.org/
[5] http://www.openbeacon.org

More precisely, for a given $t \in [t_s, t_e]$, we define

$$E^{\leq t} = \{e \in E \mid t(e) \leq t\} \ .$$

For our analysis, we use the graph $G^{\leq t} = (V, E^{\leq t})$ to train the prediction models. The test set is defined as follows: In analogy to $E^{\leq t}$, we define

$$E^{>t} = \{e \in E \mid t(e) > t\} \ .$$

Let V_{core} be the set of participants who have at least one contact during the training interval and at least one contact during the test interval. By restricting $E^{>t}$ to those edges where both vertices are contained in V_{core}, we obtain

$$E^{>t}_{\text{core}} = E^{>t} \cap V_{\text{core}} \times V_{\text{core}}$$

and finally $G^{>t} = (V_{\text{core}}, E^{>t}_{\text{core}})$.

The different link prediction tasks that we consider within this paper are to predict:

1) New links only (as in [21]), i. e., all links in $E^{>t}_{\text{core}} \setminus E^{\leq t}$.
2) Recurring links, i. e., all links in $E^{>t}_{\text{core}} \cap E^{\leq t}$.

Note that — following the approach in [21] — the training set $G^{\leq t}$ contains all vertices of G, while the test set $G^{>t}$ contains only those vertices that are present in the core.

C. RFID Data

For the LWA 2010 and Hypertext (HT) 2011 conferences we used the first day of the conference as training data. Hence, we aim to predict new and recurring conversations of day two and three. Table I gives a detailed description of the collected datasets. In Figure 2, we observe the typical distribution of all face-to-face contacts for both conferences. Confirming previous findings, e.g. in [15], [4], [24], most of the contacts take less than one minute and the contact durations of both conferences show a long-tailed distribution. In addition, Figure 2 shows, that the number of contacts at LWA 2010 was significantly higher than at HT 2011. The diameter and average path length of G is similar to the results presented in [15], [4].

IV. Network Proximity Measures

In this section, we discuss the proximity measures (see Table II) used in our analysis for the prediction tasks: In [21] Liben-Nowell and Kleinberg analyzed several network proximity measures. In their analysis they showed that the network proximity measures *Common Neighbors*, and *Adamic Adar* [1] perform best. The measure *Common Neighbors* is based on the assumption that it is more likely that two nodes are connected if these two nodes have many neighbors in common. *Adamic Adar* is similar to *Common Neighbors*, but here the *Common Neighbors* are weighted with respect to their degree. *Preferential Attachment* is based on the assumption, that the probability [8] of a new node being connected to node x is proportional to the degree of x. Zhou et al. [36] presented a new measure called *Resource Allocation*. This measure is similar to *Adamic Adar*, but in [36] the authors show that it performs better (in most cases) than *Common Neighbors*

TABLE I
General statistics for the collected datasets. Here d is the diameter, *APL* the average path length and *LCN* the largest clique number.

	Hypertext 2011	**LWA 2010**
#days	3	3
$\lvert V\rvert$	62	77
$\lvert E\rvert$	640	1004
Avg.Deg.(G)	41.3	52.16
APL (G)	1.7	1.7
LCN (G)	14	16
d (G)	3	3
$\lvert V_{\text{core}}\rvert$	49	57
$\lvert E^{\leq t}\rvert$	481	426
$E^{>t}_{\text{core}} \setminus E^{\leq t}$	132	394
$E^{>t}_{\text{core}} \cap E^{\leq t}$	134	242
Avg.Deg.$(G(\leq t))$	32.1	27.04
APL $(G(\leq t))$	1.84	1.9
LCN $(G(\leq t))$	13	9
d $(G(\leq t))$	4	4
Avg.Deg.$(G(> t))$	21.7	44.6
APL $(G(> t))$	1.99	1.64
LCN $(G(> t))$	8	13
d $(G(> t))$	4	3

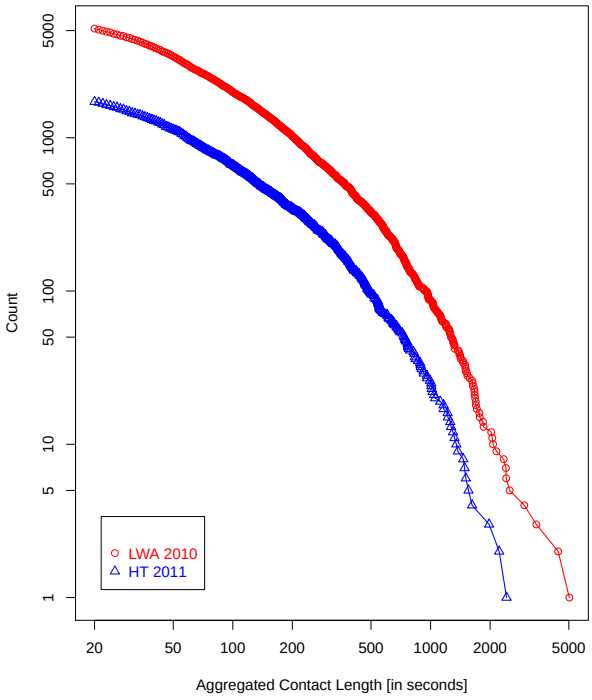


Fig. 2. Cumulated contact length distribution of all face-to-face contacts of the LWA 2010 and the HT 2011 conference, respectively: The x-axis displays the minimum length of a contact in seconds, the y-axis the number of contacts having at least this contact length, respectively. The axes are scaled logarithmically.

TABLE II
OVERVIEW OF NETWORK PROXIMITY MEASURES.

Measure	Unweighted	Weighted
Common Neighbors	$CN(x,y) = \lvert N(x) \cap N(y) \rvert$	$WCN(x,y) = \sum_{z \in N(x) \cap N(y)} w(x,z) + w(y,z)$
Adamic-Adar	$AA(x,y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{\log \lvert N(z) \rvert}$	$WAA(x,y) = \sum_{z \in N(x) \cap N(y)} \frac{w(x,z)+w(y,z)}{\log (\sum_{z' \in N(z)} w(z',z))}$
Jaccard's Coefficient	$J(x,y) = \frac{\lvert N(x) \cap N(y) \rvert}{\lvert N(x) \cup N(y) \rvert}$	$WJ(x,y) = \frac{\sum_{z \in N(x) \cap N(y)} w(x,z)+w(y,z)}{\sum_{x' \in N(x)} w(x,x') + \sum_{y' \in N(y)} w(y,y')}$
Resource Allocation	$RA(x,y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{\lvert N(z) \rvert}$	$WRA(x,y) = \sum_{z \in N(x) \cap N(y)} \frac{w(x,z)+w(y,z)}{\sum_{z' \in N(z)} w(z',z)}$
Pref. Attachment	$PA(x,y) = \lvert N(x) \rvert \cdot \lvert N(y) \rvert$	$WPA(x,y) = \sum_{x' \in N(x)} w(x,x') \cdot \sum_{y' \in N(y)} w(y,y')$

and *Adamic Adar*. We also apply *Preferential Attachement* and *Jaccard's Coefficient* for link prediction. In [27] Murata and Moriyasu presented weighted variants of *Adamic Adar*, *Preferential Attachement* and *Common Neighbors*. A weighted variant of *Resource Allocation* is presented in [23].

All these proximity measures are defined using the assumption that two nodes which are close to each other in the graph have a higher probability of becoming connected in the future. In this paper, we also analyze the predictive power of the weighted variants compared to the original "unweighted" proximity network measures in the context of human contact networks. Furthermore, we extend the group of weighted network proximity measures to a weighted version of *Jaccard's Coefficient*. For all network proximity measures we need the definition of the neighborhood for a node x. The set of neighbors $N(x)$ for node x is defined as

$$N(x) = \{y | y \in V, (x,y) \in E\}$$

Table II provides a detailed overview of the used "unweighted" and weighted proximity measures. Previous work on weighted link prediction used the contact count of two persons for the weight of the link between them. Contact count means, for example, the number of telephone contacts or collaborations between the respective pair of actors. In this work, we use contact count (number of contacts on the first day) as well as contact duration (sum of all contact durations on the first day) for weighting a link between two persons.

V. LINK PREDICTION

In this section, we study the link prediction problem on human contact networks. As already done in literature [32], [21], [27], we analyze the predictability of several network proximity measures (see Table II). To the best of the authors' knowledge, this is the first time that link prediction is analyzed in the context of human (face-to-face) contact networks. In contrast to previous work, we also extend our studies to the prediction of stronger links and recurring links.

A. Human Communication Statistics and Basic Analysis

Knowledge about human communication behaviour is important to improve the prediction of future links. We therefore present some new insights into the communication behaviour of participants during a conference. In Figure 3, we analyze the average contact length distribution with the longest, second longest, . . ., tenth longest contact. In average, each participant talks to his longest contact for more than one third of his total communication time. This fraction decreases rapidly (from 38 percent for the longest contact), when we consider the fraction of the second longest contact. Here, the fraction of the contact length compared to the overall contact length is approximately 17 to 19 percent. Interestingly, both barplots look quite similar at both conference datasets. This might indicate, that this is the typical behavior in a conference setting.

Will particpants who had a contact at the first day of the conference talk to each other again on the second or third day? For answering this question and for understanding its mechanisms it is important to consider the contact length from the first day of the conference. In Figure 5, we observe the clear trend, that a contact is more likely to be renewed the longer the contact on the first day. In Figure 4, we plot the distribution of all contacts for the second and third day, depending on the contact length of the first day. We observe, that a longer contact is more likely, the longer the contact on the first day. An interesting further question is then to find typical features to predict renewed contacts and their length.

B. Role-based Influence Factors

In the following, we analyze the impact of a number of (external) role-based factors for the link-prediction problem, relating to properties of the people collaborating in the contact network. Specifically, we focus on the prediction of new contacts and recurring links.

We use pattern mining for identifying *characteristic patterns* [5] describing subgroups with a high share of *new contacts*. The applied technique is subgroup discovery, e.g., [6],

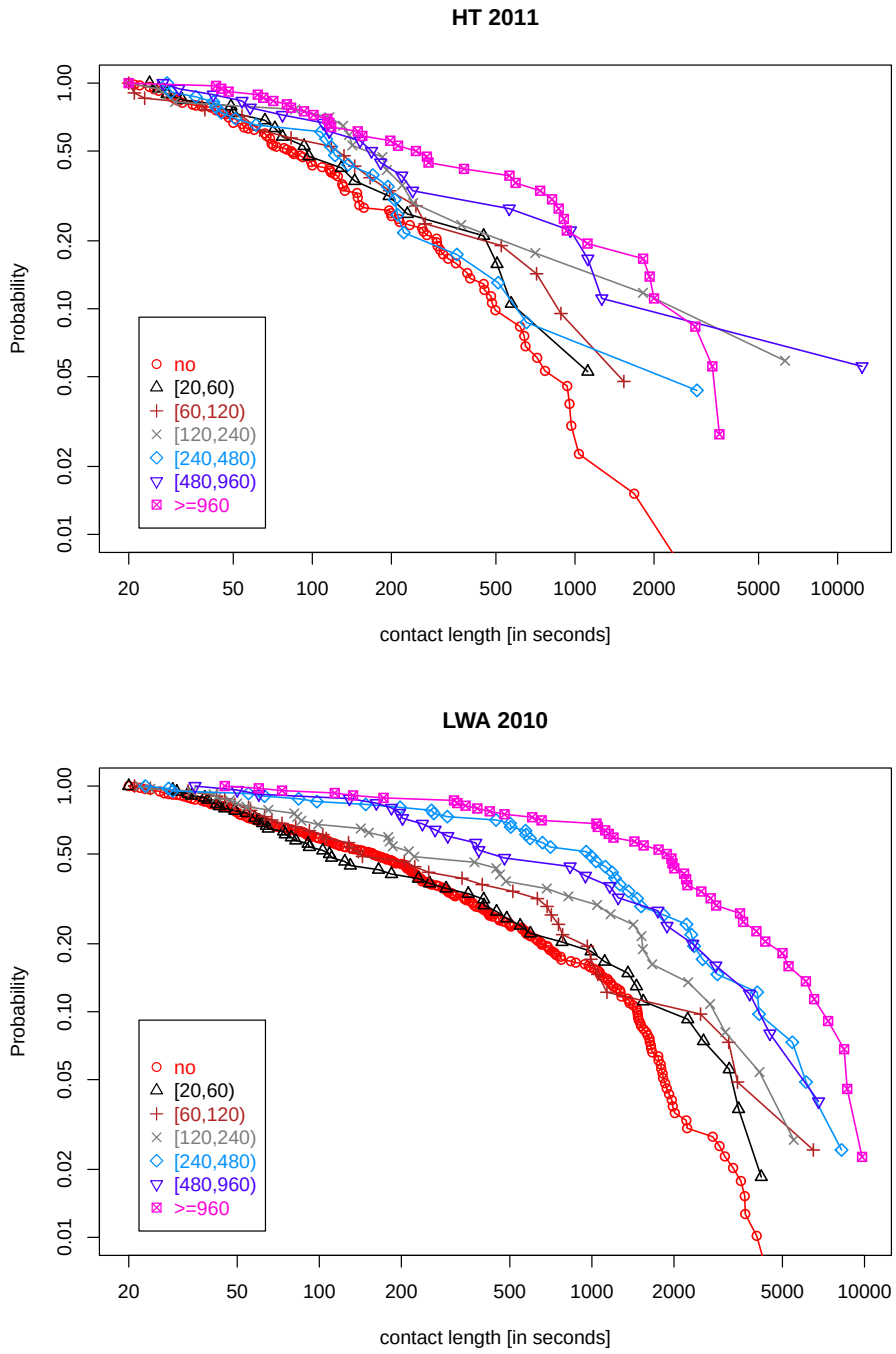
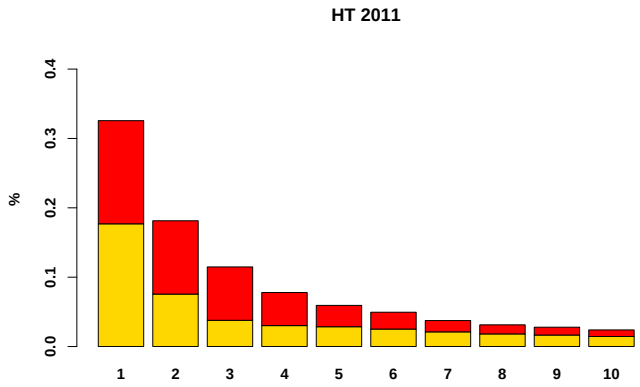
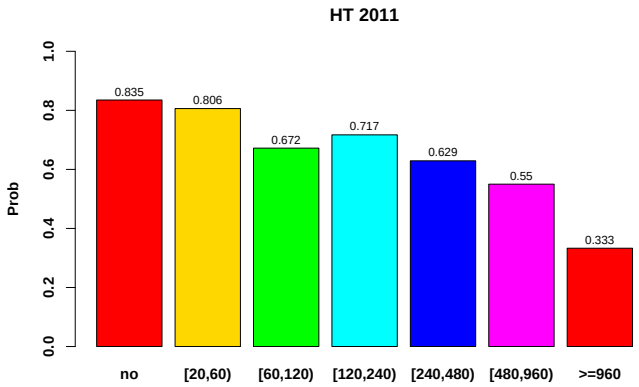
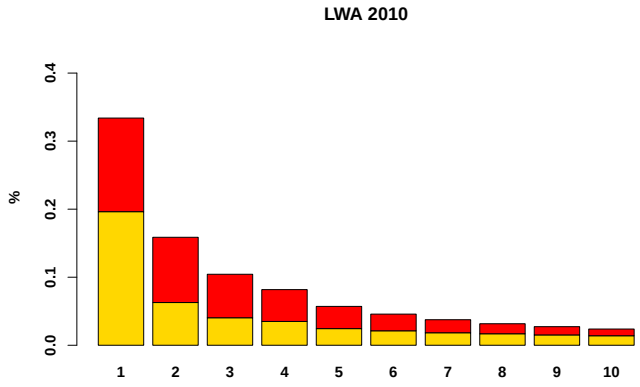


Fig. 4. Impact of the contact duration between two participants on the first day, and the contact length of a recurring contact at the second and third day, for the HT 2011 and the LWA 2010 conference, respectively. The red line labeled with 'no' (circle symbol) in the LWA 2010 plot, for example, shows the distribution of all contacts between participants at day two and three, which had no contact at the first day. The line labeled with $[60, 120)$ (cross symbol) shows the distribution of all contacts between participants at day two and three, which had a contact with contact duration between 60 and 120 seconds at the first day of the conference.

[7], [34]: Basically, we aim at discovering subgroups of participants described by combinations of factors, e.g., *session chair AND strong affiliation* that show a high share of a certain target property, an increased mean of new contacts compared to the default share. Intuitively, we identify conjunctions of attribute values describing subsets of a dataset that maximize a given property, e.g., an increased mean of an attribute in the subset compared to the whole dataset. In the patterns described below, this *target attribute* is given by the *mean contact count* of new contacts.







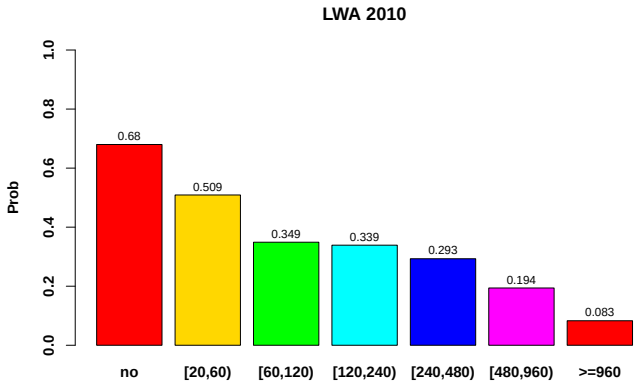


Fig. 3. Average fraction of contact duration to each participant's longest, second longest, . . ., tenth longest contact, for the HT 2011 and the LWA 2010 conference, respectively. The lower "sub-bar" contained in each bar shows the standard deviation. The x-axis represents the $i-$th longest contact; the y-axis shows the fraction of contact duration to the $i-th$ longest contact. Here, for example, the left bar (labeled with 1) in the HT 2011 barplot means, that in average each participant talks to his longest contact for approximately 38 percent of his total contact duration.

TABLE III
OVERWIEW ABOUT THE NUMBER OF PARTICPANTS FOR DIFFERENT TIME THRESHOLDS USED IN FIGURE 5

	LWA 2010		**HT 2011**	
	$\sum$ Contacts	$\#no$ Contacts	$\sum$ Contacts	$\#no$ Contacts
no	1230	836	798	666
$[20, 60)$	110	56	98	79
$[60, 120)$	63	22	64	43
$[120, 240)$	56	19	60	43
$[240, 480)$	58	17	62	39
$[480, 960)$	31	6	40	22
≥ 960	48	4	54	18

Fig. 5. Impact of contact duration between two participants on the first day on the contact between these two particpants for the remaining days (day two and three), for the HT 2011 and the LWA 2010 conference, respectively. The left bar labeled with 'no' in the HT 2011 barplot means, for example, that two participants who had no contact at the first day of the conference had no contact until the end of the conference in 83.5%. The bar labeled with $[20, 60)$ in the HT 2011 barplot means that two participants who had a contact with a duration between 20 and 60 seconds on the first day had no contact on the second and third day in 80.6% of all cases. In Table III we present the detailed numbers for these figures. The column $\sum$ Contacts represents the number of contacts for the specified type of contact. Here for example (for HT 2011) the row 'no' means, that there are 1230 pairs of particpants who had no contact at the first day. 836 of these pairs had no contact at the second and third day, either. The row $[20, 60]$ means that there are 110 participants, who had a contact with contact duration between 20 and 60 seconds. 56 of these had no further contact on the second and third day.

We focused on different subgroup structures, i. e., partitionings, induced by academic status, affiliation with the Hypertext conference series, and affiliation with one of the four conference tracks. In Table IV, we present some statistics about the different subgroups. We classify participants as highly affiliated with the Hypertext conference series if they presented a paper more than three times at Hypertext conferences in different years. The affiliation of a participant is low when he or she has never presented a paper or presented a paper at Hypertext 2011 for the first time. All other participants are classified with a medium affiliation.

The tables show the lift of the pattern assessing the ratio of the mean of new contacts covered by the pattern and the fraction of the whole dataset, the size of the pattern extension (number of described participants), and the description itself. The first line in Table V, for example, shows that being a session chair with a strong affiliation to HT 2011 increases the mean number of new contacts by 58%.

TABLE IV
PARTITIONS OF THE SET OF PARTICIPANTS INTO SUBGROUPS ACCORDING TO ACADEMIC STATUS AND AFFILIATION WITH HT 2011.

Academic Status	
Professor	14
PhD-candidate	34
PhD	20
Other	7

Affiliation with HT	
high	12
medium	17
low	46

Below, we exemplarily show interesting patterns with respect to the Hypertext 2011 conference. We collected conference and participants roles and analyzed their correlation with the emergence of new contacts. As shown in Table V, as expected we observe an influence of being a session chair at the conference; this is even increased for participants with stronger affiliation to the conference, i.e., if participants are more experienced and also have more publications at previous conferences. As expected, we observe that *presenters* encounter a lot of new contacts. Also, the academic status of *Professor* increases the contact count. This is also confirmed by the LWA 2010 data.

TABLE V
EXEMPLARY TOP 5 ROLE INFLUENCE PATTERNS FOR THE HYPERTEXT 2011 CONFERENCE MEASURING THE INCREASE IN NEW CONTACTS.

#	Lift	Mean	Size	Description
1	1.58	8.50	6	session chair AND strong affiliation
2	1.55	8.36	11	professor
3	1.35	7.25	8	session chair
4	1.31	7.08	12	strong affiliation
5	1.08	5.81	16	presenter

In addition to new contacts, we also analyzed recurring contacts and their contact durations. Table VI shows exemplary patterns for the Hypertext 2011 conference. While we observe, that people with a low affiliation, i. e., participants that are new to the conference are still very active after the first day, an interesting finding for Hypertext is, that being a session chair and being a professor increases the mean duration of contacts by 10% while the single factors alone inhibit the duration (-13% and -18%, respectively). For the LWA 2010 we found a slightly different pattern; the organizers were still very active (increase by 34%), but the professors scored as expected (increase by 17%).

TABLE VI
EXEMPLARY ROLE INFLUENCE PATTERNS FOR THE HYPERTEXT 2011 CONFERENCE MEASURING THE MEAN OF RECURRING CONTACTS.

#	Lift	Mean	Size	Description
1	2.10	5944.17	6	PhD AND low affiliation
2	1.52	4297.15	26	low affiliation
3	1.09	3089.00	6	session chair AND professor
4	1.08	3038.67	21	PhD candidate
5	1.06	3003.93	14	PhD
6	0.87	2461.25	8	session chair
7	0.82	2326.18	11	professor

C. Evaluation Method

For the evaluation of link prediction measures, often the precision of the top n predicted links is used [21], where n is the number of positive links (i.e. the number of new or renewed links on day 2 and 3). In this work, we measure the accuracy by the area under a receiver operating characteristic (AUC) [13]. In short, receiver operating characteristic (ROC) graphs plot the true positive rate on the y-axis and the false positive rate on the x-axis, concerning the set of predictions (ranking). The advantage of AUC is that it considers the whole ranking. In the context of link prediction AUC has already been used, e. g., in [22]. For the prediction of new or recurring 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.

D. Prediction of New Links

In this subsection, we evaluate the quality of several link prediction measures (see Table II) to predict new links, i. e., all links in $E^{>t}_{\text{core}} \setminus E^{\leq t}$. In Table VII, we present the predictor scores of the original network proximity measure as well as the weighted variants of these measures. In the following, we index the network measure with dur when we use the contact duration as the weight of the link; we index the network measure with cc when we use the contact count as the weight of the link between two participants. Table VII suggests, that the network structure helps to improve the prediction accuracy, because all measures outperform the random predictor (the AUC value of a random predictor is 0.5). This also means that in a human contact network the network topology contains useful information for the prediction of new links. This result is not suprising, since it confirms the results of [21] and [32]. Here the authors analyzed the predictive power of proximity network measures in a co-authorship network and a mobile phone caller network. For the HT 2011 and LWA 2010 datasets the weighted variants of *Resource Allocation* and *Preferential Attachment* performed best.

In Table VII we further compared the AUC values of the original and the two versions (contact duration and contact count) of the weighted proximity measures: We observe that the weighted variants always achieve better results than the unweighted versions. However, there is no clear winner between measures weighted with contact count and those measures weighted with contact duration. Figure 6 shows the development of the AUC values for the original and weighted versions of the *Common Neighbors* and *Resource Allocation* network proximity measure, when we focus more and more on longer conversations. This means that we do not take into account conversations with contact length lower than a time threshold t (value on the x-axis) and examine only the ranking positions of conversations greater than the time threshold t. In Figure 6 we see an interesting development. On both datasets, the one for LWA 2010 and the one for the HT 2011, longer conversations tend to be placed higher in the ranking than shorter conversations.

TABLE VII
BASELINE RESULTS

	HT 2011	LWA 2010
	AUC	AUC
N	0.6224	0.6397
WN_{dur}	0.6473	0.6556
WN_{cc}	0.6493	0.6500
J	0.6171	0.6131
WJ_{dur}	0.6491	0.6431
WJ_{cc}	0.6428	0.6348
AA	0.6264	0.6398
WAA_{dur}	0.6496	0.6548
WAA_{cc}	0.6520	0.6496
RA	0.6265	0.6368
WRA_{dur}	0.6536	0.6425
WRA_{cc}	0.6527	0.6400
PA	0.6010	0.6503
WPA_{dur}	0.6425	0.6596
WPA_{cc}	0.6479	0.6514

TABLE VIII
NUMBER OF NEW CONTACTS ON THE SECOND AND THIRD DAY.

	Hypertext 2011	LWA 2010
$\geq$ 20 sec.	132	394
$\geq$ 60 sec.	82	275
$\geq$ 120 sec.	52	212
$\geq$ 240 sec.	30	148
$\geq$ 480 sec.	14	98
$\geq$ 960 sec.	4	63
$\geq$ 1920 sec.	1	17
$\geq$ 3840 sec.	0	4

E. Prediction of Recurring Links

In this subsection, we analyze the predictability of recurring links, i.e all links in $E^{>t}_{\text{core}} \cap E^{\leq t}$. For the evaluation of the predictability we apply the different proximity network measures described in Table II, similar to the evaluation in Section V-D.

The "advantage" of recurring links in contrast to new links is, that there is already a contact on the first day of the conference. Therefore, we use and analyze the total contact length (CL) of the first day as predictor score. As a quality measure for the resulting rankings, we use again the AUC measure for comparison (see Section V-C).

As shown in Figure 7, we observe that the contact duration outperforms all analyzed network proximity measures. Only if we look at all recurring contacts, i. e., with no time threshold, the network proximity measures present better results than the contact length predictor. Most of the network proximity measures achieve relatively low AUC values for the prediction of recurring links, even for high time thresholds. We hypothesize, that the measures need to be adapted in order to account for the new link structure, e.g., being extended to include the path structure of the network.

In Table IX we analyze, if there is a correlation between the score of the proximity measures and the contact duration.

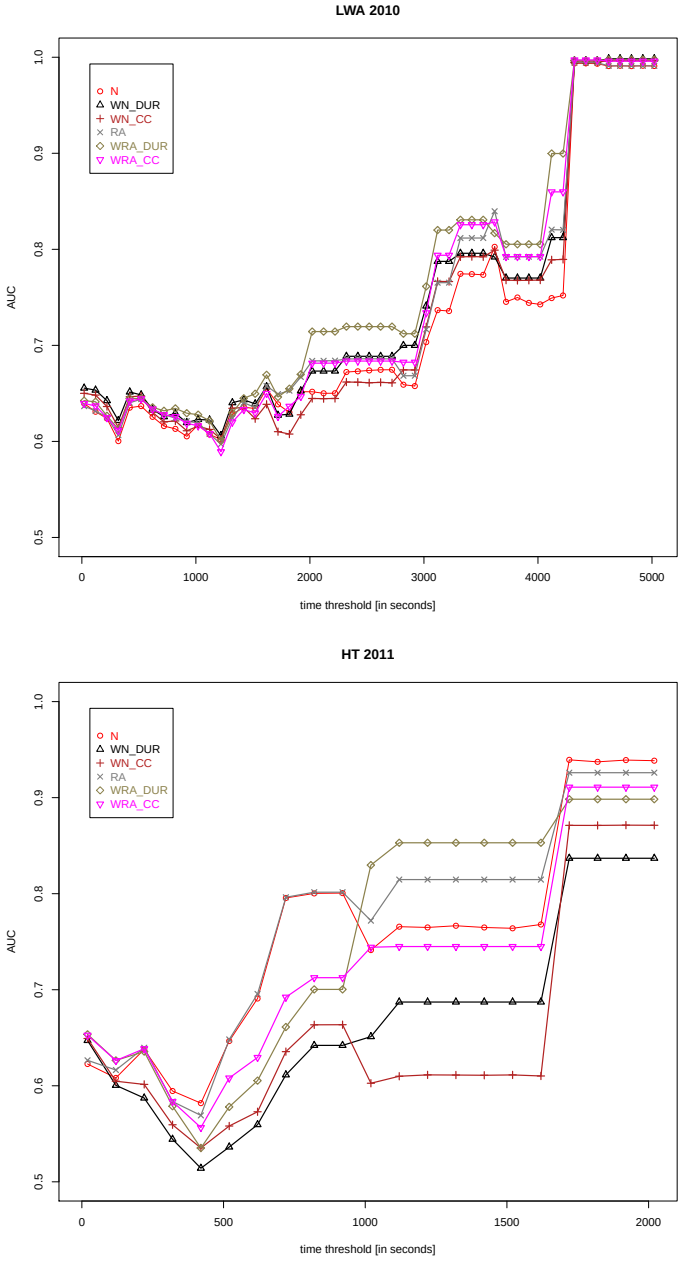
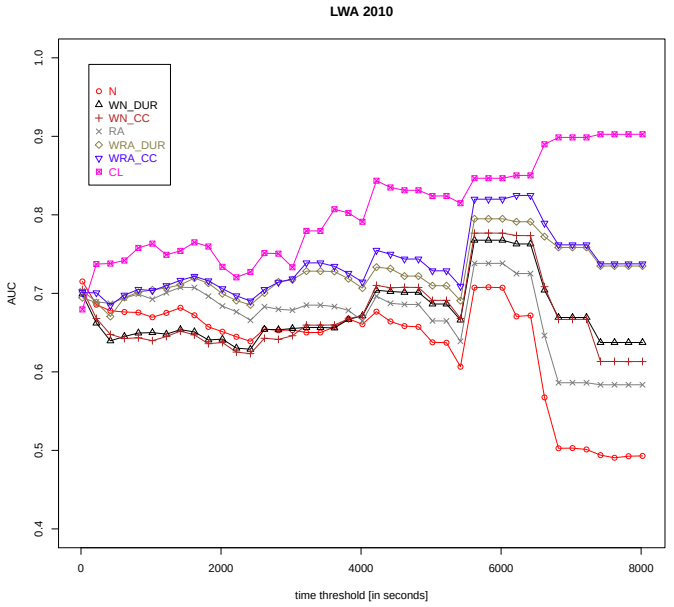


Fig. 6. Threshold-based analysis of contact length and AUC-values for the prediction of new links, focussing on the ranking positions of longer conversations, for the LWA 2010 and the HT 2011 conference, respectively. The x-axis represents the minimum contact duration and the y-axis shows the AUC value for the prediction of new links with a contact duration at least this contact length. The detailed number of new contacts for different time thresholds we present in Table VIII.

For each possible contact pair $x, y \in V_{\text{core}}$, we determine the corresponding predictor score. We apply the spearman correlation coefficient for measuring the correlation. The results of this analysis indicate, that there is in fact a correlation between the contact length and the predictor score. We also observe, that the correlation for the weighted proximity measures is higher than the correlation for the original ones.



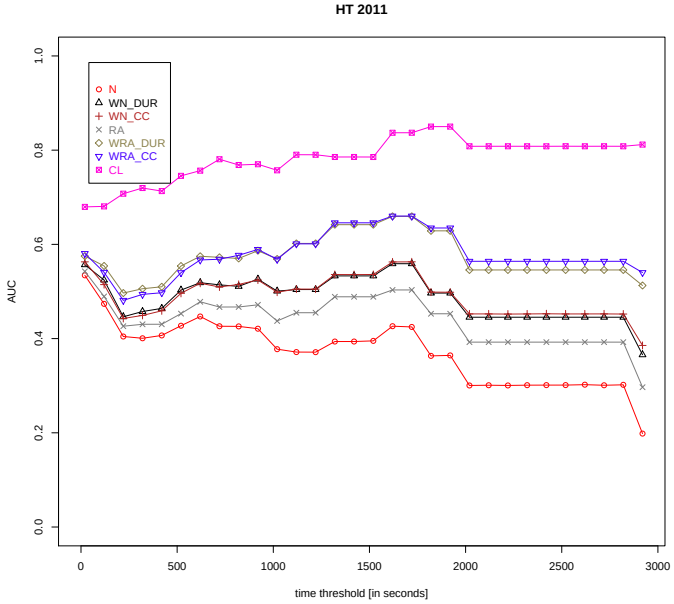


Fig. 7. Threshold-based analysis of contact length and AUC-values for the prediction of recurring links, for the LWA 2010 and the HT 2011 conference, respectively. We compare the ranking obtained using the value of the respective network proximity measure, and the ranking created due to the contact length, of the first day. The x-axis represents the minimum contact duration of a contact and the y-axis shows the AUC value for the prediction of new links with a contact duration of least this contact length. The CL labeled lines are the results for the prediction of recurring links using the contact length of the first day for the predictor score.

Figure 8 shows (exemplarily for the *Weighted Resource Allocation*) that longer contacts have in average indeed a higher WRA_{Dur} value than shorter contacts.

VI. Conclusions

In this paper, we considered the predictablity of human face-to-face contacts and presented an analysis of influence

TABLE IX
Correlation between contact length (duration) and different predictor scores.

	HT 2011		LWA 2010	
	cor	p-val	cor	p-val
N	0.2079	≤ 0.001	0.3694	≤ 0.001
WN_{dur}	0.2390	≤ 0.001	0.3840	≤ 0.001
WN_{cc}	0.2421	≤ 0.001	0.3797	≤ 0.001
J	0.2104	≤ 0.001	0.3293	≤ 0.001
WJ_{dur}	0.2376	≤ 0.001	0.3718	≤ 0.001
WJ_{cc}	0.2375	≤ 0.001	0.3668	≤ 0.001
AA	0.2144	≤ 0.001	0.3728	≤ 0.001
WAA_{dur}	0.2423	≤ 0.001	0.3853	≤ 0.001
WAA_{cc}	0.2471	≤ 0.001	0.3841	≤ 0.001
RA	0.2213	≤ 0.001	0.3742	≤ 0.001
WRA_{dur}	0.2589	≤ 0.001	0.3800	≤ 0.001
WRA_{cc}	0.2566	≤ 0.001	0.3800	≤ 0.001
PA	0.1850	≤ 0.001	0.3510	≤ 0.001
WPA_{dur}	0.2371	≤ 0.001	0.3936	≤ 0.001
WPA_{cc}	0.2436	≤ 0.001	0.3853	≤ 0.001

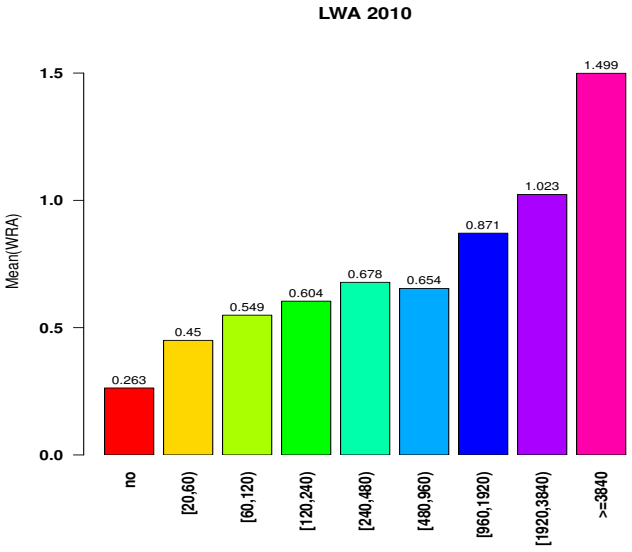


Fig. 8. Exemplary analysis for the LWA 2010 conference: For different time intervals (x-axis) the mean score (y-axis) of the *Weighted Resource Allocation* (WRA_{Dur}) is measured over all possible contact pairs $x, y \in V_{\text{core}}$, where the total contact durations belong to the respective time intervals.

factors for link prediction in such human contact networks. Additionally, we considered and analyzed the strength of stronger ties within the network. Specifically, we considered the standard problem of predicting new links, and extended it to the analysis of recurring links.

Furthermore, we considered (and adapteded) different network proximity measures for the prediction, and took descriptive properties of human participants into account. We also considered the impact of stronger ties for the prediction, i.e., considering longer contact durations. To the best of the authors' knowledge, this is the first time that such techniques and analyses have been performed concerning face-to-face

contacts in human contact networks. The analysis' results provided interesting insights especially concerning the impact of the contact durations and the strength of such stronger ties. These insights are a first step onto predictability applications for human contact networks.

For future work, we aim to embed the indicators, patterns, and influence factors into more advanced prediction models in the context of human contact networks. Furthermore, we plan to extend the analysis towards more dynamic approaches including movement and location-based events for improving the prediction further. Another interesting option is the inclusion of extended temporal aspects into prediction models.

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 utilized active RFID technology which was developed within the SocioPatterns project, whose generous support we kindly acknowledge. Our particular thanks go the SocioPatterns team, especially to Ciro Cattuto, who enabled access to the Sociopatterns technology, and who supported us with valuable information concerning the setup of the RFID technology.

REFERENCES

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

[2] H. Alani, M. Szomszor, C. Cattuto, W. V. den Broeck, G. Correndo, and A. Barrat. Live Social Semantics. In *International Semantic Web Conference*, pages 698–714, 2009.

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

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

[5] M. Atzmueller, F. Lemmerich, B. Krause, and A. Hotho. Who are the Spammers? Understandable Local Patterns for Concept Description. In *Proc. 7th Conference on Computer Methods and Systems*, 2009.

[6] M. Atzmueller and F. Puppe. SD-Map - A Fast Algorithm for Exhaustive Subgroup Discovery. In *Proc. 10th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD 2006)*, pages 6–17, Heidelberg, Germany, 2006. Springer.

[7] M. Atzmueller, F. Puppe, and H.-P. Buscher. Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In *Proc. 19th Intl. Joint Conference on Artificial Intelligence (IJCAI-05)*, pages 647–652, Edinburgh, Scotland, 2005.

[8] A.-L. Barabasi. Linked the New Science of Networks, 2002.

[9] A. Barrat, C. Cattuto, V. Colizza, L. Isella, C. Rizzo, A. E. Tozzi, and W. V. den Broeck. Wearable Sensor Networks for Measuring Face-to-Face Contact Patterns in Healthcare Settings. In *eHealth*, pages 192–195, 2010.

[10] A. Barrat, C. Cattuto, V. Colizza, J.-F. Pinton, W. V. den Broeck, and A. Vespignani. High Resolution Dynamical Mapping of Social Interactions with Active RFID. *CoRR*, abs/0811.4170, 2008.

[11] A. Barrat, C. Cattuto, M. Szomszor, W. V. den Broeck, and H. Alani. Social dynamics in conferences: Analyses of data from the live social semantics application. In *International Semantic Web Conference (2)*, pages 17–33, 2010.

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

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

[14] P. Hui, A. Chaintreau, J. Scott, R. Gass, J. Crowcroft, and C. Diot. Pocket Switched Networks and Human Mobility in Conference Environments. In *Proc. ACM SIGCOMM workshop on Delay-tolerant networking*, pages 244–251, New York, NY, USA, 2005. ACM.

[15] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. V. den Broeck. What's in a Crowd? Analysis of Face-to-Face Behavioral Networks. *CoRR*, 1006.1260, 2010.

[16] K. Jahanbakhsh, V. King, and G. C. Shoja. Predicting Missing Contacts in Mobile Social Networks. In *Proc. 12th IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks, WOWMOM 2011*, pages 1–9, 2011.

[17] K. Jahanbakhsh, G. C. Shoja, and V. King. Human Contact Prediction Using Contact Graph Inference. In *Proc. IEEE/ACM Intl. Conference on Green Computing and Communications and Intl. Conference on Cyber, Physical and Social Computing*, pages 813–818, Los Alamitos, CA, USA, 2010. IEEE Computer Society.

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

[19] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Community Structure in Large Networks: Natural Cluster Sizes and the Absence of Large Well-Defined Clusters, 2008.

[20] J. Leskovec, K. J. Lang, and M. W. Mahoney. Empirical Comparison of Algorithms for Network Community Detection, 2010.

[21] D. Liben-Nowell and J. M. Kleinberg. The Link Prediction Problem for Social Networks. In *CIKM*, pages 556–559, 2003.

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

[23] L. Lü and T. Zhou. Link Prediction in Weighted Networks: The Role of Weak Ties. *EPL (Europhysics Letters)*, 89:18001, 2010.

[24] B.-E. Macek, C. Scholz, M. Atzmueller, and G. Stumme. Anatomy of a Conference. In *Proc. 23rd ACM Conference on Hypertext and Social Media*, pages 245–254, New York, NY, USA, 2012. ACM Press.

[25] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annu. Rev. Sociol.*, 27(1):415–444, 2001.

[26] M. Meriac, A. Fiedler, A. Hohendorf, J. Reinhardt, M. Starostik, and J. Mohnke. Localization Techniques for a Mobile Museum Information System. In *Proceedings of WCI*, 2007.

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

[28] C. Scholz, S. Doerfel, M. Atzmueller, A. Hotho, and G. Stumme. Resource-Aware On-Line RFID Localization Using Proximity Data. In *Proc. ECML/PKDD 2011*, pages 129–144, 2011.

[29] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotto, W. V. den Broeck, C. Régis, B. Lina, and P. Vanhems. High-Resolution Measurements of Face-to-Face Contact Patterns in a Primary School. *CoRR*, abs/1109.1015, 2011.

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

[31] C. Tantipathananandh and T. Y. Berger-Wolf. Constant-Factor Approximation Algorithms for Identifying Dynamic Communities. In J. F. E. IV, F. Fogelman-Soulié, P. A. Flach, and M. Zaki, editors, *KDD*, pages 827–836. ACM, 2009.

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

[33] D. J. Watts and S. H. Strogatz. Collective Dynamics of /'small-world/' Networks. *Nature*, 393(6684):440–442, june 1998.

[34] S. Wrobel. An Algorithm for Multi-Relational Discovery of Subgroups. In *Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97)*, pages 78–87, Heidelberg, Germany, 1997. Springer.

[35] B. Xu, A. Chin, H. Wang, L. Chang, K. Zhang, F. Yin, H. Wang, and L. Zhang. Physical Proximity and Online User Behavior in an Indoor Mobile Social Networking Application. In *Proc. 4th IEEE Intl. Conf. on Cyber, Physical and Social Computing (CPSCom 2011)*, 2011.

[36] T. Zhou, L. Lu, and Y.-C. Zhang. Predicting Missing Links via Local Information. *The European Physical Journal B - Condensed Matter and Complex Systems*, 71:623–630, 2009.