On the Predictability of Recurring Links in Networks of Face-to-Face Proximity

Christoph Scholz Martin Atzmueller Gerd Stumme

Knowledge and Data Engineering Group, University of Kassel
Wilhelmshöher Allee 73, D-34121 Kassel, Germany
{scholz, atzmueller, stumme}@cs.uni-kassel.de

ABSTRACT

This paper focuses on the predictability of *recurring links*: These links are generated repeatedly in a network for different forms of social ties, e. g., by face-to-face interactions in offline social networks. In particular, we analyse the predictability of recurring links in networks of face-to-face proximity using several path-based measures, and compare these to network-proximity measures based on the nodes' neighbourhood. Furthermore, we show that the *current tie strength* is a good predictor for this link prediction task. In addition we show that the removal of weak ties improves the predictability for most of the considered network proximity measures. For our analysis we utilize three real-world datasets collected at different scientific conferences using the Conferator (`http://www.conferator.org`) system.

1. INTRODUCTION

Link prediction is a prominent research topic in offline and online social networks in order to understand structural mechanisms of link creation and its dynamics, e. g., for supporting applications such as recommendation systems. Specifically, we consider the prediction of *recurring links*: These are generated repeatedly in a network, i. e., if a tie between actors is formed multiple times. A prominent case of recurring links are face-to-face interactions. In this paper, we analyse the predictability of recurring links in such networks of face-to-face proximity comparing several path-based measures to standard network proximity measures as a reference. Furthermore, we analyse the impact of *strong* and *weak* ties for the prediction.

Our contribution is summarised as follows:

1. We compare neighbourhood-based and path-based network proximity measures for link prediction in networks of face-to-face proximity, focusing on recurring links. Moreover, we compare the performance of all measures to the *current tie strength* predictor.
2. Furthermore, we analyse the impact of stronger ties for the prediction, in a threshold-based analysis for both neighbourhood-based and path-based methods.
3. We also analyse the role of weak ties, and show that they weaken the performance of the predictors.

We analyse three real-world datasets collected at different scientific conferences using the social conference guidance system Conferator.

The rest of this paper is structured as follows: Section 2 discusses related work. After that, Section 3 describes the applied RFID hardware setting, that we used to collect our datasets. Furthermore, we give a detailed overview on the collected real-world datasets at the LWA 2010, HT 2011 and LWA 2012 conferences. Section 4 discusses the applied measures for link prediction. In Section 5, we present our results and discuss these in the context of the originating academic conferences. Finally, we conclude with a summary and discuss future work in Section 6.

2. RELATED WORK

A first comprehensive analysis of link prediction using unsupervised methods was done by Liben-Nowell and Kleinberg in [11]. Murata and Moriyasu [14] analysed weighted variants of different network proximity measures. Lichtenwalter et al. [12] presented a new unsupervised (a restricted variant of rooted PageRank) and a new supervised method for the prediction of new links. Backstrom and Leskovec introduced in [2] a supervised method, based on supervised random walks, for the prediction of new links.

However, most of these approaches analysed the predictability of new links in online social networks like Facebook or DBLP. The prediction of links in offline social networks has been largely neglected. For reliably detecting face-to-face proximity, a new generation of active RFID tags has been developed by the SocioPatterns collaboration, cf. [4], which we also applied for the data analysed in this work. In [15], we presented a first analysis concerning the predictability of new and recurring links in real world face-to-face contact networks. In [7], we showed that the predictability of new links can be further improved by data from online networks, proposing a new unsupervised link prediction method that combines the information of different networks. In [17] Tsugawa and Ohsaki also analysed the quality of unsupervised methods in the context of link prediction in face-to-face proximity networks; they compared the predictability of links in face-to-face contact networks and other types of social networks, supporting our earlier work in [15].

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2579235.

	LWA 2010	HT 2011	LWA 2012
#days	3	3	3
$\|V\|$	77	68	42
$\|E\|$	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

Table 1: Collected datasets. d is the diameter, AACD the average aggregated contact-duration (in seconds) and *APL* the average path length.

3. FACE-TO-FACE CONTACT DATA

In this section, we summarize the framework used for collecting face-to-face contact networks, before we briefly describe the collected datasets.

At the conferences LWA 2010, Hypertext (HT) 2011, and LWA 2012, we collected networks of face-to-face proximity. Each link in the network indicates physical proximity and can be weighted by the cumulated duration of all face-to-face proximity contacts between the linked persons.

For the three conferences we asked all participants to wear the active RFID devices described above, which 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 [6]. As in [16], we record a face-to-face contact when the length of a contact is at least 20 seconds. A contact ends when the proximity tags do not detect each other for more than 60 seconds. 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 the SocioPatterns project (http://www.sociopatterns.org).

Table 1 provides a summary on the characteristics of the three collected face-to-face proximity datasets. As already observed in many other contexts [6, 9, 13] the distributions of all aggregated face-to-face contacts lengths between conference participants are heavy-tailed. The diameter, average degree and average path length of G are similar to the results presented in [1, 9]. For more details on the applied datasets, we refer to, e. g., [7].

4. NETWORK PROXIMITY MEASURES

In this section, we discuss neighbourhood-based and path-based measures used in our analysis for the prediction tasks. Focussing on unsupervised methods, most of the predictor scores are based on either nodes' neighbourhoods or path information. All of these proximity measures are based on the assumption that two nodes have a higher probability to become connected, if these two nodes are close in the graph.

We model the social network as an undirected weighted multi-graph $G = (V, E, w)$, where V is the set of participants and an edge $(u, v) \in E$ represents a face-to-face contact between two participants u and v, and where the weight $w(u, v)$ of edge (u, v) is the sum of the durations of all face-to-face contacts between participants u and v.

Neighbourhood-based Network Proximity Measures.

In Table 2, we provide a detailed overview of the used unweighted and weighted proximity measures. The measure *Common Neighbours* is based on the assumption that it is more likely that two nodes are connected if these two nodes have many neighbours in common. *Adamic Adar* and *Resource Allocation* are similar to *Common Neighbours*, but here the *Common Neighbours* are weighted with respect to their degree. Considering *Jaccard's Coefficient* it is more likely that two nodes are connected, if these two nodes share a high fraction of their respective neighbourhood. *Preferential Attachment* is based on the assumption, that the probability [3] of a new node being connected to node x is proportional to the degree of x. We define the neighbourhood for a node x, i. e., the set of neighbours $N(x)$, as

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

Path-based Network Proximity Measures.

The *rooted PageRank* [11] predictor is an adaption of the PageRank algorithm [5] for the link prediction task. The *rooted PageRank (RPR)* predictor score between participants x and y is defined by the stationary probability distribution of participant y under the following random walk [11]:

- With probability α, jump to x.
- With probability $1 - \alpha$, jump to a random neighbour of the current node.

For the *weighted rooted PageRank (WRPR)* predictor, the random walk selects from the current node c a random neighbour n of node c with probability $\frac{w(c,n)}{\sum_{c \to d} w(c,d)}$, where $w(c, d)$ is the weight of the edge (c, d).

The *Katz* [10] predictor is defined as

$$Katz(x, y) = \sum_{l=1}^{\infty} \beta^l \cdot |\text{path}_{x,y}^l|,$$

where $\text{path}_{x,y}^l$ is, for $x, y \in V$, the set of paths from x to y with length l. We note that $\beta \in [0, 1]$ is a damping factor that weights short paths higher/lower in the summation.

5. ANALYSIS

In this section, we analyse the predictability of recurring links in face-to-face contact networks. We especially compare the predictability of path-based and neighbourhood-based network proximity measures. Furthermore, we focus on the prediction of stronger recurring links and analyse the importance of stronger links for this link prediction task. In addition we analyse the role of weak ties for the prediction task. We start with a definition of the research problem.

5.1 Problem Statement

Let t be a point in time during the conference. For the prediction task, we define all face-to-face contacts starting before t as training data and face-to-face contacts starting later as test data. The training data is then the undirected graph

Table 2: Overview of network proximity measures based on the nodes' neighbourhood.

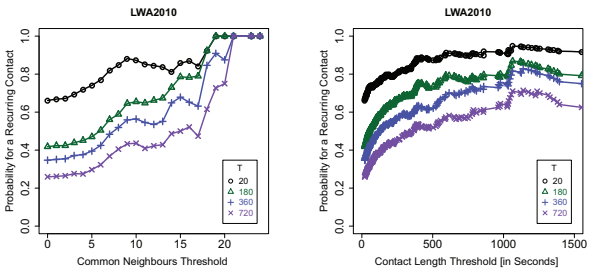
Measure	Unweighted	Weighted
Common Neighbours	$CN(x,y) = \lvert N(x) \cap N(y) \rvert$	$WCN(x,y) = \sum_{z \in N(x) \cap N(y)} w(x,z) + w(y,z)$
Adamic-Adar	$AA(x,y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{\log \lvert N(z) \rvert}$	$WAA(x,y) = \sum_{z \in N(x) \cap N(y)} \frac{w(z,x)+w(z,y)}{\log (\sum_{z'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

$G^{\leq t} = (V^{\leq t}, E^{\leq t})$, 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 $(u,v) \in E^{\leq t}$, if they had at least one face-to-face contact before t. The weight $w_{\leq t}(u,v)$ 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. We consider the graph $G^{>t} = (V_{\text{core}}, E^{>t}_{\text{core}})$ as test data: Two participants $u, v \in V_{\text{core}}$ are connected by an edge $(u,v) \in E^{>t}_{\text{core}}$ if u and v had at least one face-to-face contact after t.

Then, the prediction task, which we focus on in this paper, is to predict *recurring links*,i. e., *all links in* $E^{\geq t}_{core} \cap E^{<t}$. In order to do this, we compute a predictor score for each pair $(u,v) \in (V_{core} \times V_{core}) \cap E^{\leq t}$. 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 [8] directly based on the predictor scores. During the evaluation, we will also analyse if longer face-to-face contacts are easier to predict. Therefore we also consider $G^{>t}$ as a weighted graph, where $w_{>t}(u,v)$ is the sum of the durations of all face-to-face contacts of the participants u and v after t.

5.2 Influence Factors for the Prediction of Recurring Links

In the recurring link prediction problem we want to predict whether a link between two participants u and v will recur or not. Unlike the *new link* prediction problem [11,15], in the recurring link prediction problem we can also use the information about the already existing tie strength of the corresponding participants u and v. We then analyse the influence of the number of common neighbours and the already existing tie strength on the recurrence of a link. In Figure 1, we plot the probability for a recurring link with tie strength T as a function of common neighbours and as function of the already existing tie strength. Given a face-to-face contact between two participants at the first day of the conference, we compute in this analysis whether a contact (with minimum contact duration T) recurs or not on the second or third day of the conference, depending on the number of common neighbours and existing tie strength of the first day. We observe, that the probability increases almost linearly the higher the number of common neighbours and the higher the already existing tie strengths are.



(a) (b)

Figure 1: Probability for a recurring link with strength T as a function of common neighbours and tie strength. In both figures the y-axis shows the probability for a recurring link (with tie strength T), given at least: a) a specific number of common neighbours or b) a specific tie strength. The respective thresholds are defined by the x-axis.

5.3 Predictability of Recurring Links in Face-to-Face Proximity Networks

In this section we evaluate and compare the quality of network-based and path-based network proximity measures to predict recurring links. Furthermore, we use the *current tie strength* between two participants as predictor. The *current tie strength* between participants u and v is defined as $w_{\leq t}(u,v)$, where here t is the end of the first day of the conference. Then we use these predictor scores to analyse its prediction quality with respect to whether a link will recur or not towards the end of the conference. In Figure 2, we plot the AUC-values for all network proximity measures measures and the *current tie strength*. First, we observe

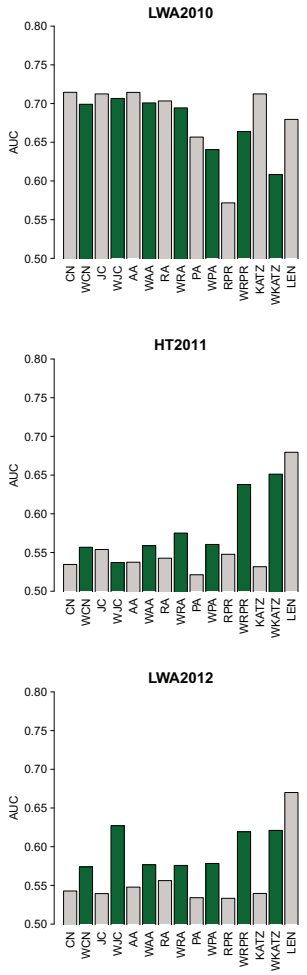


Figure 2: AUC values/network proximity measures. LEN here indicates the *current tie strength* as predictor.

that the network structure helps to improve the prediction accuracy, because all predictors outperform the random predictor. Here, we note that the AUC-value of a random predictor is 0.5. These results are not too surprising, since this has already been shown for the new link prediction problem [11, 15]. Furthermore, we notice that the first day's tie strength performs very well as predictor on all datasets. With respect to the HT 2011 and LWA 2012 dataset we see that path-based network proximity measures perform better than measures based on the nodes' neighbourhood. However this result does not hold on the LWA 2010 dataset.

In Figure 3, we focus more and more on longer face-to-face contacts for the link prediction task. This means that we only consider face-to-face contacts longer than a given time threshold T. In Figure 3, this time threshold T is defined by the x-axis. Considering longer contacts, we observe that weighted path-based measures clearly outperform network proximity measures based on the nodes' neighbourhood. Furthermore, the weighted variants of the path-based measures perform much better than the unweighted variants. In addition, we notice that (also for longer contacts) using the first day's tie strength as predictor performs very well on all datasets. Except for the HT 2011 dataset, this predictor performs best. This is surprising, because we expected that the path-based measures would significantly outperform all other measures. Apparently, the combination of information of the node's neighbourhood with the first day's tie strength is boosting the performance. Considering the neighbourhood-based measures, we see that the unweighted *Preferential Attachment* predictor performs very weak on all datasets.

5.4 The Role of Weak Ties for the Prediction of Recurring Links

We also analyse the role of weak ties for our prediction scenario. Exemplarily we focus here on the prediction of stronger links with a time threshold of 15 minutes, but the results are very similar for other time thresholds. For the analysis, we compute the AUC value for several network proximity measures, using the face-to-face contact networks, where all links have been removed that fall below a given time threshold T. In Figure 4, this threshold T is defined by the x-axis. We observe that the removal of weak links increases the prediction accuracy of most network proximity measures. Especially on the LWA 2010 and LWA 2012 datasets the AUC value for the unweighted rooted PageRank increases for more than 15% AUC, when we remove all links weaker than 200 seconds. Considering this threshold, we can also observe an increase of AUC for all weighted and unweighted neighbourhood-based network proximity measures. For the *weighted rooted PageRank* predictor, we observe the interesting trend that the removal of weak ties seems to have less influence concerning the prediction accuracy. This stability can be explained by the fact that the *weighted rooted PageRank* also uses the information of the first day's tie strength. Except for the LWA 2010 dataset, this result is also true for the weighted Katz predictor.

6. CONCLUSIONS

In this paper we analysed the predictability of recurring links in face-to-face contact networks. We compared path-based and neighbourhood-based measures and studied the *current tie strength* as predictor for recurring links. Considering stronger links, we observed that the weighted variants of the path-based network proximity measures perform much better in the prediction of recurring links than neighbourhood-based network proximity measures. The results also show, that the current tie strength performs better than the path-based measures on two of the three datasets. This is surprising, because path-based measures combine information from the current tie strength and the nodes' neighbourhood. Furthermore, we studied the predictability of recurring links, when weak links are removed from the network. We observed that removing links with weight (aggregated contact length) smaller than 200 seconds increases the AUC-values for most network proximity measures.

For future work, we aim to investigate social aspects further with respect to the analysis of recurring links. In addition *recurring links* may play a major role considering the prediction of new links. Hence, we plan to analyze the role of *recurring links* concerning the predictability of new links.

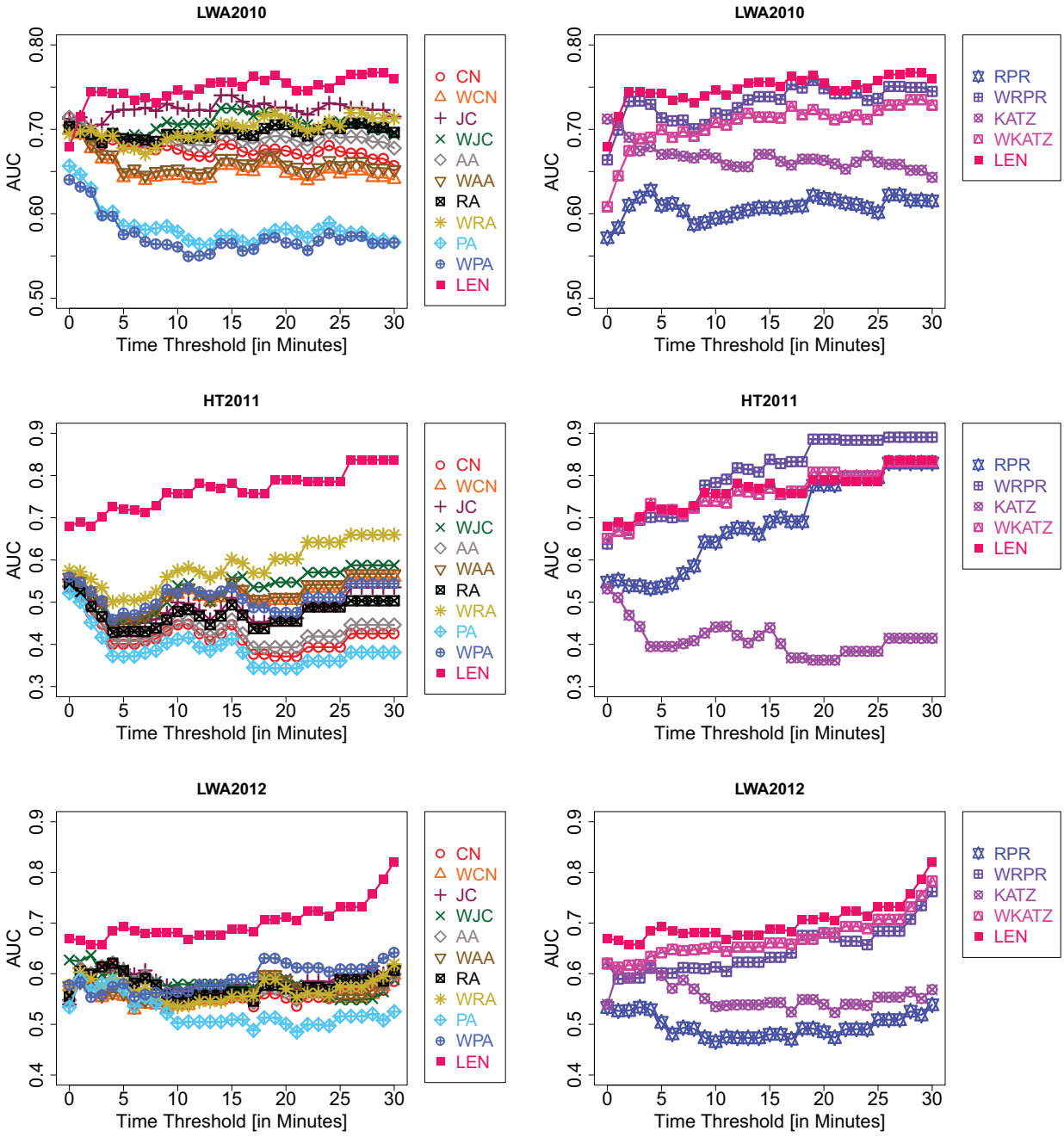


Figure 3: AUC values for recurring link prediction for different time thresholds. The y represents the AUC value for the given time threshold T (defined by the x-axis). We note here that we only consider future links with tie strength $>= T$ for the prediction task. LEN here indicates the *current tie strength* as predictor.

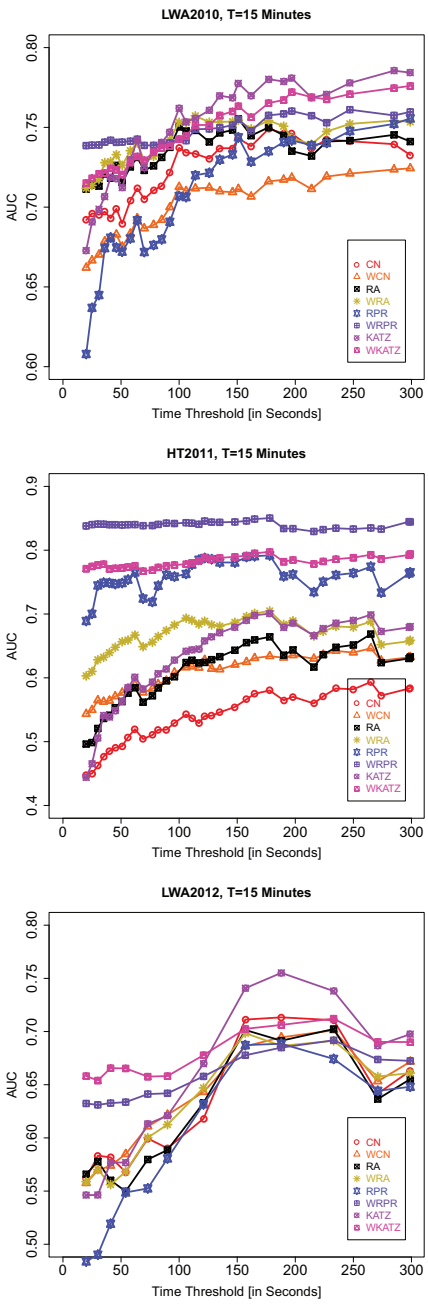


Figure 4: AUC value for several network proximity measures, when we delete all links that fall below time threshold T. The time threshold T is given on the x-axis.

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.

7. REFERENCES

[1] 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*. Springer Verlag, Heidelberg, Germany, 2012.

[2] L. Backstrom and J. Leskovec. Supervised Random Walks: Predicting and Recommending Links in Social Networks. In *Proc. WSDM*, pages 635–644, New York, NY, USA, 2011. ACM Press.

[3] A.-L. Barabasi. *Linked the New Science of Networks*. Perseus Pub., Cambridge, Mass., 2002.

[4] 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.

[5] S. Brin and L. Page. The Anatomy of a Large-Scale Hypertextual Web Search Engine. *Computer Networks*, 30(1-7):107–117, 1998.

[6] C. Cattuto, W. Van den Broeck, A. Barrat, V. Colizza, J.-F. Pinton, and A. Vespignani. Dynamics of Person-to-Person Interactions from Distributed RFID Sensor Networks. *PLoS ONE*, 5(7):e11596, 07 2010.

[7] Christoph Scholz and Martin Atzmueller and Alain Barrat and Ciro Cattuto and Gerd Stumme. New Insights and Methods For Predicting Face-To-Face Contacts. In *Proc. 7th Intl. Conf. on Weblogs and Social Media*, Palo Alto, CA, USA, 2013. AAAI Press.

[8] 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.

[9] L. Isella, J. Stehlé, A. Barrat, C. Cattuto, J.-F. Pinton, and W. V. D. Broeck. What's in a crowd? analysis of face-to-face behavioral networks. *Journal of Theoretical Biology*, 271:166–180, 2011.

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

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

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

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

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

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

[16] 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*, pages 196–210, 2010.

[17] S. Tsugawa and H. Ohsaki. Effectiveness of Link Prediction for Face-to-Face Behavioral Networks. *PLoS ONE*, 8(12):e81727, 12 2013.