Exploiting Location-Based Data Sources for Link Prediction in an Online Social Network

Thesis for the Award of the Academic
Degree of a Doctor of Technology

submitted by Michael Steurer
Graz, April 2014

Graz University of Technology
Institute for Information Systems and Computer Media

Supervisor: Prof. Dr. Frank Kappe
Second Reviewer: Prof. Dr. Vanessa Chang

"All of old. Nothing else ever. Ever tried. Ever failed. No matter. Try again. Fail again. Fail better."

—Samuel Beckett, Worstward Ho

Abstract

"Similarity breeds connection" – with these words Miller McPherson expressed the principle of homophily that connections between individuals are based on shared traits.

With the advent of online social networks users were able to interact with real-world friends on Web-based platforms without geographic constraints. Although this seemed to overcame the limitations of user's locations for cultivating relations, the spatial distance between users still plays an important role for the creation of new links. Location information of users represents the natural behaviour of their daily contacts and therefore allows a researcher to model their real-world contacts without joining explicit friends lists on the Web. Subsequently, data from online social networks and location-based sources can not only be used to model existing links but also to predict upcoming links.

The goal of this research is to investigate the extent to which different sources of location-based data can support data from an online social network for the prediction of upcoming interactions and the tie-strength of these interactions. We overcome the problem of missing large scale real-world data by collecting data from various sources in the virtual world of Second Life. In the first part of the dissertation we focus on the problem of link prediction in the online social network supported by one single source of location-based data. This approach sheds light onto valuable features to model social proximity between users and suitable machine learning techniques. Based on those results we next investigate the differences between three different but related location-based data sources for the prediction of interactions in the online social network. Finally, we combine all available data sources and tighten the link prediction problem to predict the tie-strength between already connected users.

Our analysis reveals that user-pairs with interactions are more similar than user-pairs without interactions and this homophily even grows with increasing tie-strength. Connected users share more groups, have a smaller spatial distance between them and visit more common places. These signs of intimacy are also supported by topological features that model the closeness of users from the network's perspective. For the actual prediction tasks we identify features based on the homophily as being most valuable in the online social network and the location-based data sources. Although the performance of the machine learning algorithms depends on the character of the actual data source and the availability of data, a combination of the online social network and the location-based sources could be identified as most valuable for the prediction tasks.

Location information turned out as a promising source of data for the prediction of interactions and the tie-strength between users. We complemented online social networks with three different sources of location-based data and demonstrated the benefit of this approach for several prediction tasks. To the best of our knowledge this is the first work that combines online social networks and three different location-based data sources obtained from the same group of users for the prediction of interactions and tie-strength between users.

Keywords: Online Social Networks ◇ Location-Based Social Networks ◇ Link Prediction ◇ Tie-Strength Prediction ◇ Data Mining ◇ Machine Learning ◇ Virtual Worlds ◇ Second Life

Kurzfassung

"Similarity breeds connection". Die Beziehungen zwischen Personen beruhen auf ihrer Ähnlichkeit – mit diesen Worten wurden von Miller McPherson die Prinzipien der Homophilie beschrieben.

Durch die Popularität von Web-basierten sozialen Netzwerken können Personen ohne geographische Einschränkungen miteinander kommunizieren und interagieren. Obwohl es den Anschein erweckt, dass dadurch Beziehungen über große Entfernungen geführt werden können, spielt die geographische Distanz zwischen Personen bei der Entstehung neuer Beziehungen noch immer eine große Rolle. Genaue Informationen über die Aufenthaltsorte von Personen entsprechen dabei deren natürlichem und alltäglichem Verhalten. Somit ist es möglich, Listen mit persönlichen Kontakten ohne explizite Freundeslisten von Web-basierten Plattformen zu erstellen. Mit diesen Informationen können Kontaktnetzwerke erstellt werden und in weiterer Folge auch Kontakte, die in der Zukunft entstehen, vorhergesagt werden.

Das Ziel dieser Dissertation ist es, den Einfluss von drei unterschiedlichen Quellen mit personenbezogenen Positionsinformationen auf ein Web-basiertes soziales Netzwerk zu evaluieren und zukünftige Text-Interaktionen zwischen Personen und die Intensität dieser Verbindungen vorherzusagen. Da es in der realen Welt keine geeigneten Datenquellen für diese Untersuchungen gibt, wurden die Daten im Umfeld der virtuellen Welt von Second Life erhoben. Im ersten Teil der Dissertation wird der Einfluss von Positionsdaten einer einzelnen Datenquelle auf die Vorhersagbarkeit von Text-Interaktionen in einem sozialen Netzwerk untersucht. Aus den gewonnen Resul-

taten werden Rückschlüsse auf geeignete Metriken zur Modellierung von Beziehungen zwischen Personen und die dafür geeigneten maschinellen Lernalgorithmen gezogen. Darauf aufbauend werden drei unterschiedliche Quellen mit personenbezogenen Positionsinformationen hinsichtlich der Vorhersehbarkeit von zukünftigen Text-Interaktionen zwischen diesen Personen verglichen. Im letzten Teil der Dissertation werden alle verfügbaren Datenquellen verknüpft um die Intensität der Beziehungen von Personen vorherzusagen die bereits im sozialen Netz verbunden sind.

Die durchgeführten Analysen zeigen eine sehr hohe Ähnlichkeit von Personen mit bestehenden Text-Interaktionen und mit wachsender Intensität dieser Verbindungen steigt auch deren Ähnlichkeit. Verbundende Personen gehören mehr gleichen Gruppen an, haben eine geringere geographische Distanz zueinander und besuchen öfter dieselben Orte. Unterstützt werden diese Beobachtungen durch Eigenschaften, die aus der topologischen Struktur des sozialen Netzwerks abgeleitet werden. Die Analyse von Metriken für die Vorhersage zeigt, dass die Ähnlichkeit zweier Personen, die aus dem sozialen Netzwerk und den Positionsdaten abgeleitet wurden, am einflussreichsten ist. Die bei der Untersuchung verwendeten Algorithmen zum maschinellen Lernen variierten mit der Verfügbarkeit und dem Charakter der Daten, wobei durch die Kombination des sozialen Netzwerks mit den Positionsdaten die besten Resultate für die Prognosen gefunden wurden.

In den durchgeführten Untersuchungen haben sich Positionsdaten von Personen als aussagekräftige Datenquelle zum Vorhersagen von Text-Interaktionen zwischen Personen und deren Intensität erwiesen. Dafür wurden Daten aus unterschiedlichen Quellen kombiniert, um den positiven Einfluss von Positionsdaten zu testen. Das ist die erste Arbeit, in der ein Text-basiertes soziales Netzwerk mit drei unterschiedlichen Positionsdatenquellen verknüpft wird, um eine Vorhersage für zukünftige Text-Interaktionen und deren Intensität zu untersuchen.

Schlüsselwörter: Online Social Networks ◇ Location-Based Social Networks ◇ Link Prediction ◇ Tie-Strength Prediction ◇ Data Mining ◇ Machine Learning ◇ Virtual Worlds ◇ Second Life

Acknowledgments

This thesis would not have been possible without the contribution and support of working colleagues, friends and my family. I owe my gratitude especially to the following: First and foremost I would like to thank my supervisor Professor Frank Kappe for making this research possible. He raised my interest in virtual worlds and provided me with support during the entire time of my PhD. He gave me the freedom to explore my own ideas and at the same time guided my thoughts into the right direction. I also want to express my deepest appreciation to Professor Vanessa Chang for reviewing my thesis and welcoming me in such a friendly way during my research stay in Perth, Australia.

I would like to thank Christoph Trattner for the collaboration and useful discussions on research matters. I am grateful to Marie-Louise Lampl and Gabriele Leitner who made all the administrative issues during my PhD very easy. I also wish to express my gratitude to Christian Safran for the great collaboration and Lukas Eberhard for being an awesome office mate. I am also very grateful to my friends and colleagues at Graz University of Technology for their funny and useful discussions during coffee and lunch breaks. Thank you for making my time at the institute so pleasant and for your help during various stages of my PhD.

Words can not express how thankful I am to all of my friends (the list would be too long but you know who you are) who helped me stay sane during the last years. Thanks for all the moments when you were there to talk, go out for a beer and help me recharge myself – you never let me doubt that there is a world outside of my PhD. Last but not least, I would like to thank my family, especially my mother and father, for always believing in me – I would not have made it this far without them. Thanks for your unconditional support throughout this endeavour. I owe you everything.

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, _____ _____

Place, Date Signature

Eidesstattliche Erklärung

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommene Stellen als solche kenntlich gemacht habe.

Graz, am _____ _____

Ort, Datum Unterschrift

Contents

CHAPTER 1

Introduction

It is in the nature of humans to build groups and socialize with others but with the evolution of humanity the social cohesion between individuals changed and became more complex. Reasons for that were technological advances in travel-abilities, global communications and personal interactions [Easley and Kleinberg, 2010]. Although it seems that the spatial influence of relations has weakened, geographic information still plays an important role in the creation of new relations. The rising popularity of online social networks like *Facebook* and location-based social networks like *FourSquare* implicated the availability of information about relations between individuals as well as position data of these individuals at a large scale. This data allows to represent existing relations between users as well as the prediction of upcoming relations with high accuracy.

1.1. Motivation

Boyd and Ellison [2007] defined the main characteristics of social networks as follows: 1) users maintain a profile that describes themselves and make the profile visible to the public or only to a subset of friends. 2) users have a list of friends or acquaintances to define their social relations to others. These friends-lists evolve and change over time and form the actual "social network". 3) users can view the profiles of their friends, interact with them and respond to updates or changes of their friends' profiles. Although this global definition of a social network is valid for most available networks, the aim and nature of these platforms are manifold: business networks like *LinkedIn*[*], hybrid information-entertainment networks like *Twitter*[†] or networks for special target groups like *Catsters*[‡]. The main motivations for the use of these social networks range from mapping existing friends to social networks, maintaining these friends and even making new friends [Lampe et al., 2006; Ellison et al., 2007]. These new links in the network form such as they do in the real world – between users that share similar interests or other habits [McPherson et al., 2001; Mislove et al., 2010].

To analyse and describe the pair-wise relations in a social network Bondy and Murty [1976] suggested the mathematical structure of *graphs*. A graph $G\langle V,E\rangle$ is defined as a set of vertices V, representing the actual users, and a set of edges E that connect users in this network. An example for a social network that models the relations among users is the network of interactions between the characters of Victor Hugo's *Les Misérables* as depicted in Figure 1.1. The characters in this network are represented as nodes and the relations between these characters are represented as edges.

Since the social proximity between users in a social network is not equal among all user-pairs, we can use two different approaches to measure the "distance" between them: topological features and homophilic features. Topological features affect the structure of the entire network from a global or local perspective. An example for a local network feature is the number of common neighbours a user-pair has, i.e. the more common neighbours a pair of users has, the higher is the probability that these users will form a connection in the future [Papadimitriou et al., 2011]. An example for a global network feature is the average shortest path between two users, i.e. the shorter the average path between two users, the closer these users are in the network [Romero and Kleinberg, 2010]. In contrast, homophilic features do not consider the structure of the surrounding network but only measure the "alikeness" or "similarity" of a user-pair. Examples are shared interests of two users or the number of locations two users visited concurrently [McPherson et al., 2001]. It is in the nature of humans that relations between users are not of equal strength and as

[*] www.linkedin.com
[†] www.twitter.com
[‡] www.catsters.com

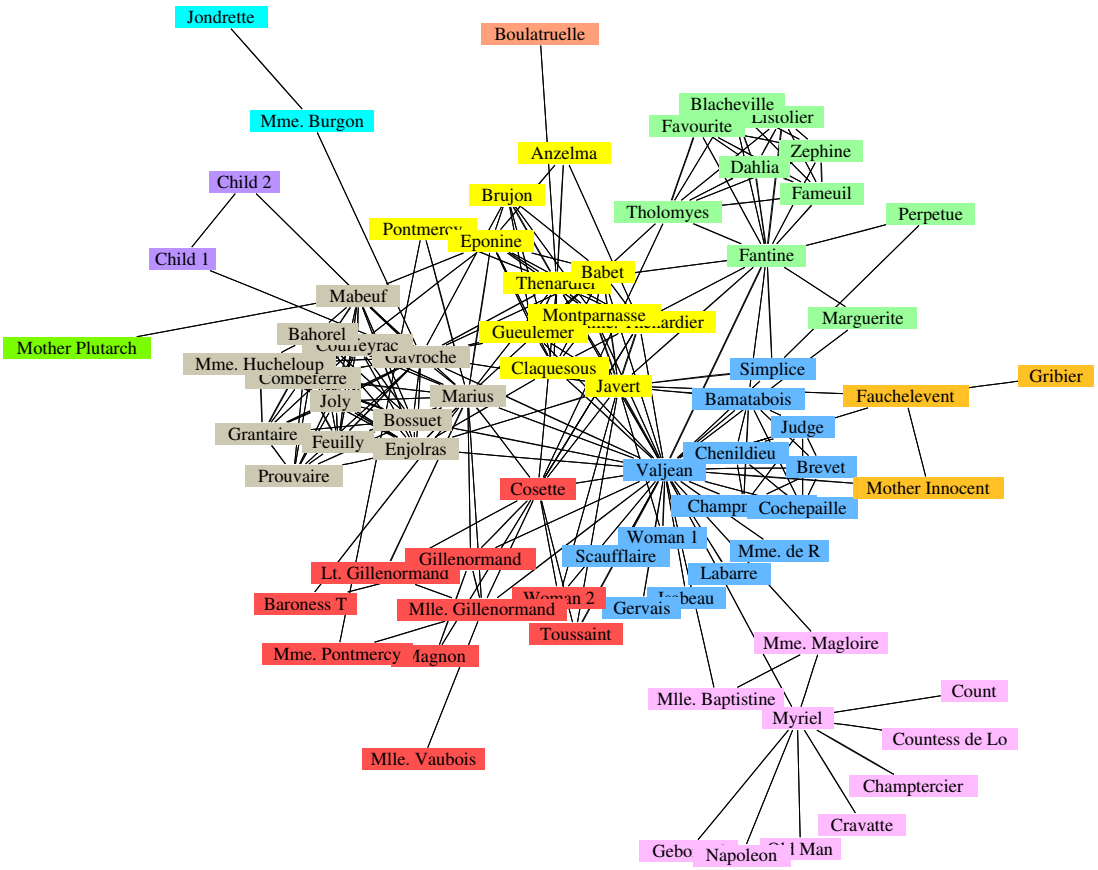


Figure 1.1.: Network of interactions between characters of Victor Hugo's *Les Misérables* [Newman and Girvan, 2004].

a consequence Granovetter [1973] introduced the *tie-strength* of a relationship by proposing the terms "strong tie" and "weak tie". An example for strong and weak ties in a network are *close friends* and *acquaintances* in the online social network of Facebook [Kahanda and Neville, 2009]. Although weak ties only model relations between slightly connected nodes in the network they are essential for the structure of the network and the coherence of its nodes [Granovetter, 1973; Gilbert and Karahalios, 2009; Jones et al., 2013; Onnela et al., 2007; Kahanda and Neville, 2009].

Although the possibilities of online social networks seem to overcome the limitations of a user's location for maintaining interpersonal relations, the spacial distance between user-pairs still plays an important role for the creation of new links [Lewis et al., 2008]. Up until a few years ago it was infeasible to collect location information of users on a large scale but with the advent of mobile GPS devices, e.g. mobile phones, and platforms to share position information, e.g. FourSquare, this information became available. Although people are concerned about the implications of the collection and analysis of this sensitive data [Iqbal and Lim, 2010; Consolvo et al., 2005], they

find benefits in some cases [Michael et al., 2006; Abbas et al., 2011]. From a user perspective the main motives to use location-based social networks are that 1) users want to show their current location to a befriended user, 2) users utilize the current location as a method of self representation and 3) users exploit location information to coordinate with friends [Lindqvist et al., 2011]. The final decision to share data depends on the actual user or service that requests data, the reason why the information is desired and finally the accuracy and resolution of the requested data [Michael et al., 2006].

Similar to online social networks where social or textual interactions represent links between users [Boyd and Ellison, 2007], location-based information provided by users complies with their relations as well [Li et al., 2008; Scellato et al., 2011; Liu et al., 2012; Mok and Wellman, 2007]. A detailed analysis of users' movement shows that 10-30% of these movements can be explained by location based data whereas 50-70% of these movements were influenced by periodic habits [Cho et al., 2011]. This is a clear indication that the locations users visit are more than they seem at a first glance. They contain information about work place and home location [Cho et al., 2011], the actual age of users [Popescu and Grefenstette, 2010] or health and relational status [Dong et al., 2011]. Based on this data it is easy to create user profiles and assign the users to groups that reflect their interests [Joseph et al., 2012]. Fusco et al. [2011] defined social network that are based on location data of users in a formal way: "A location-based social network is the convergence between location based services and online social networking."

The alikeness of user-pairs and the structural closeness of users can not only be used to model existing relations but also to predict relations that occur in the future. This "link prediction problem" was defined and described by Liben-Nowell and Kleinberg [2002] to predict co-authorship of scientific publications in the future. Ongoing work has shown that structural features [Romero et al., 2011; Fire et al., 2011] as well as homophilic features [Golder and Yardi, 2010; Rowe et al., 2012] can be employed as indicators for the probability that two users form a connection in the future. In general we can state that the closer two users are in the network (e.g. more common neighbours), the higher is the chance that they form a link in the future. Experiments applied to different online social networks and location-based social networks proved the efficiency of various features [Li and Chen, 2010; Golder and Yardi, 2010]. A common observation in these experiments was that more available data yields in a higher predictability of future links: A higher amount of information can be used to describe the relations between users more detailed or compute additional features.

Existing research either focuses on online social networks *or* location-based social networks for prediction of links and their tie-strength and there is no research that investigates the consequences and benefits of different sources of location-based information for the prediction of new links in an online social network. A possible explanation is that it is nearly infeasible to collect real-world

data on a large scale. Popular Web platforms like *Google+* or *Facebook* do not provide interfaces to automatically download this data and due to privacy restrictions user profiles are in general not publicly accessible.

1.2. Problem Statement and Research Questions

The aim of this thesis is to explore the benefits of online social networks enhanced by different sources of location information. In particular we aim to analyse the extent to which data from different domains, i.e. online social networks and location-based social networks, enhance the predictability of future links and the tie-strength of these links.

The main purpose of this thesis can be outlined with the overall problem statement:

> *How can different sources of location-based data be exploited to predict links and their strength in a related online social network and to what extent does the combination of location-based data and online social network data enhance this prediction?*

In order to start with the prediction of relations between users, i.e. link prediction or tie-strength prediction, we are interested in features to measure the social proximity between users that can be derived from an online social network and location-based data sources. Although literature suggests several measures, they strongly depend on the quality and the type of the used datasets. Based on the results of the evaluated features, we are further interested in the prediction of interactions in an online social network supported by location-based data. For the link prediction tasks literature in general recommends unsupervised machine learning approaches for feature engineering and supervised machine learning approaches for prediction but again we only found work that evaluates either one or the other domain, i.e. the online social network domain or the location-based domain and not a large-scale combination of both. To pursue with the prediction of links using an online social network supported by location-based data we are interested in the predictability of interactions using different sources of location-based data. Literature only evaluates different sources that are not related to each other and this circumstance makes it nearly impossible to compare these sources. Finally we are interested in the prediction of the tie-strength of social relations with social proximity features applied to the online social network combined with three location-based data sources.

According to this, we can state the four research questions as follows:

Research Question 1

> *Which social proximity features can be derived from an online social network and location-based data sources and how do they differ for different types of relations between users?*

Research Question 2

How can a combination of social proximity measures derived from an online social network and a location-based social network predict interactions between users?

Research Question 3

Can different location-based data sources be used to predict interactions in a related online social network and which source is the most valuable?

Research Question 4

To what extent can a combination of an online social network and three different location-based data sources support the prediction of tie-strength of links between users?

1.3. Scientific Contributions

In this thesis we overcome the problem of the missing large scale real-world data and replace it with virtual data: The datasets used in this thesis origin from the virtual world of Second Life but as they are created by humans they can be mapped to the real world. For the experiments we exploit data from two different domains: the Facebook-like online social network "My Second Life" where users of Second Life can interact with each other using text postings, comments, and loves, and three different location-based data sources of these residents: 1) "Shared Locations" – users can virtually check-in at specific locations and share this information on their Web-profile, 2) "Favoured Locations" – users can specify their top 10 locations within the virtual world on their Web-profile and 3) "Monitored Locations" – the in-world movement trajectories of Second Life Residents. The benefits of this approach are the public availability of the data, the absence of technical restrictions to collect the data and the anonymity of the users. Nevertheless, the data was created by real persons and can therefore be compared with real-world data of social and location-based networks.

We harvested the necessary information with Web crawlers respectively in-world robots over a period of 12 months and created a data model from the raw data. We analysed the data of the online social network as well as the data from the three location-based data sources. To determine the relations among all users we combined the collected information and applied metrics based on the homophily of user-pairs and the structure of the created network. We explored the differences and commonalities among different relation types of user-pairs and based on the obtained results we examined the predictability of future links and the tie-strength of these links using supervised and unsupervised machine learning algorithms. To the best of our knowledge there is no research that combines online social networks and different location-based social networks obtained from the same user group for the prediction of links and the tie-strength of these links at large scale.

1.4. Structure of the Thesis

1.4.1. Paper Contributions

Most parts of this thesis have been published in the proceedings of conferences, as journals or books and all chapters are annotated with the publications they are based on. The publications and the contributions can be summarised as follows:

- Steurer, Michael; Trattner, Christoph; *Predicting Partnership with Location-based and Online Social Network Data* submitted to *Elsevier Journal of Neurocomputing*.

  The idea for this paper was initiated by the author and Christoph Trattner. Christoph Trattner supported the author with discussion about the experiment's setup and the results. The author conducted the experiments and evaluated the results. The paper was mainly written by the author and the co-author contributed to the conclusion.

- Steurer, Michael; Trattner, Christoph; Helic Denis *Predicting Social Interactions from Different Sources of Location-based Knowledge*, The Third International Conference on Social Eco-Informatics, Lisbon, Portugal, 2013.

  The idea to write this paper was initiated by the author. The writing of the paper as well as the experiments and the evaluation of the results were done by the author. The second and third author contributed with discussing the results and contributed to the abstract.

- Steurer, Michael; Trattner, Christoph; *Acquaintance or Partner? Predicting Partnership in Online and Location-based Social Networks*; Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE/ACM, Niagara Falls, Canada, 2013.

  The idea for this paper was initiated by the author and Christoph Trattner. The experiments and the evaluation of the results were done by the author. The co-author contributed with useful discussion and support to interpret the results of the experiments. The paper was mostly written by the author and the co-author wrote the related work and contributed to the introduction and conclusions.

- Steurer, Michael; Trattner, Christoph; *Who will Interact with Whom? A Case-Study in Second Life using Online and Location-based Social Network Features to Predict Interactions between Users*; In Proceedings of the MUSE-MSM Post-Proceedings, 2013.

  The idea for writing this paper was initiated in equal parts by the author and Christoph Trattner. The experiments and the evaluation of the results were done by the author. The paper was mainly written by the author and the co-author wrote the related work and contributed to the introduction and conclusion.

- Steurer, Michael; Trattner, Christoph; *Predicting Interactions In Online Social Networks: An Experiment in Second Life*; Proceedings of the 4th International Workshop on Modeling Social Media, Paris, France, 2013.

  The idea for writing this paper was initiated by Christoph Trattner and the author. The experiments and the evaluation of the results were done by the author. Christoph Trattner supported the author with discussions about the experiment's setup and the results. The paper was written by the author and the co-author.

- Steurer, Michael; Trattner, Christoph; Kappe, Frank; *Success Factors of Events in Virtual Worlds A Case Study in Second Life*; Workshop on Network and Systems Support for Games, Venice, Italy, 2012.

  The idea for writing this paper was originated in equal parts by the author and Christoph Trattner. The experiments and the evaluation of the results were done by the author. The paper was written by the author and Christoph Trattner.

- Kappe, Frank; Zaka, Bilal; Steurer, Michael; *Automatically Detecting Points of Interest and Social Networks from Tracking Positions of Avatars in a Virtual World*; ASONAM, Athens, Greece, 2009.

  The idea for writing this paper was initiated by Frank Kappe. The experiments were conducted by Bilal Zaka and Michael Steurer. The paper was was mainly written by Frank Kappe and Bilal Zaka. The author contributed in writing about data retrieval social methods and the proximity analysis.

1.4.2. Organization

The thesis is organised as follows:

Chapter 1: Introduction

In this chapter we give an overview of the topic and motivate the aim for this research. We present the scientific contributions and formulate the problem statement and the research questions.

Chapter 2: Related Work

In this chapter we present an extended overview of existing literature in this field. We cover link prediction and tie-strength prediction in online social networks as well as in location-based social networks.

Chapter 3: Success of Events in Second Life

In this chapter we present the approaches to harvest data from a Second Life Web resource

and in-world position data at a large scale. Based on the collected information we evaluate the success of events in Second Life based on features derived from their meta description.

Chapter 4: Prediction of Interactions

In this chapter we describe the prediction of interactions between users and the reciprocity of these interactions. We employ data from a location-based social network and an online-social network of the same users to do the experiments.

Chapter 5: Compare Region Sources

In this chapter we compare three related but independent sources of location-based data regarding the predictability of social interactions. We apply different homophilic features to the location-data and use supervised and unsupervised machine learning approaches to evaluate them.

Chapter 6: Prediction of Partnership

In this chapter we analyse aspects of tie-strength of relations between users – defined as partners and acquaintances – in an online social network supported by location-based data obtained from three different sources.

Chapter 7: Research Results and Conclusions

In this chapter we summarize the found results and draw the conclusion. We present answers to the research questions defined in the introductory chapter of the thesis.

References

R. Abbas, K. Michael, M. G. Michael, and A. Aloudat. Emerging forms of covert surveillance using GPS-enabled devices. *Journal of Cases on Information Technology (JCIT)*, 13(2):19–33, 2011.

J. A. Bondy and U. S. R. Murty. *Graph theory with applications*, volume 290. Macmillan London, 1976.

D. Boyd and N. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge. Location disclosure to social relations: why, when, & what people want to share. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 81–90. ACM, 2005.

W. Dong, B. Lepri, and A. S. Pentland. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, pages 134–143. ACM, 2011.

D. Easley and J. Kleinberg. *Networks, crowds, and markets*, volume 8. Cambridge Univ Press, 2010.

N. B. Ellison, C. Steinfield, and C. Lampe. The Benefits of Facebook " Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.

M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici. Link prediction in social networks using computationally efficient topological features. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 73–80. IEEE, 2011.

S. J. Fusco, K. Michael, A. Aloudat, and R. Abbas. Monitoring people using location-based social networking and its negative impact on trust: an exploratory contextual analysis of five types of "friend" relationships. 2011.

E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 211–220. ACM, 2009.

S. Golder and S. Yardi. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 88–95. IEEE, 2010.

M. S. Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

M. U. Iqbal and S. Lim. Privacy implications of automated GPS tracking and profiling. *Technology and Society Magazine, IEEE*, 29(2):39–46, 2010.

J. J. Jones, J. E. Settle, R. M. Bond, C. J. Fariss, C. Marlow, and J. H. Fowler. Inferring tie strength from online directed behavior. *PloS one*, 8(1):e52168, 2013.

K. Joseph, C. H. Tan, and K. M. Carley. Beyond local, categories and friends: clustering foursquare users with latent topics. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 919–926. ACM, 2012.

I. Kahanda and J. Neville. Using Transactional Information to Predict Link Strength in Online Social Networks. In *ICWSM*, 2009.

C. Lampe, N. Ellison, and C. Steinfield. A face(book) in the crowd: social Searching vs. social browsing. In *Conference on Computer Supported Cooperative Work*, pages 167–170, 2006.

K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties, and time: A new social network dataset using Facebook.com. *Social networks*, 30(4):330–342, 2008.

N. Li and G. Chen. Sharing location in online social networks. *Network, IEEE*, 24(5):20–25, 2010.

Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 34. ACM, 2008.

D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2002.

J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman. I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2409–2418. ACM, 2011.

X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han. Event-based social networks: linking the online and offline social worlds. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1032–1040. ACM, 2012.

M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

K. Michael, A. McNamee, M. G. Michael, and H. Tootell. Location-based intelligence-Modeling behavior in humans using GPS. In *Technology and Society, 2006. ISTAS 2006. IEEE International Symposium on*, pages 1–8. IEEE, 2006.

A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM, 2010.

D. Mok and B. Wellman. Did distance matter before the Internet?: Interpersonal contact and support in the 1970s. *Social networks*, 29(3):430–461, 2007.

M. E. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos. Friendlink: Link prediction in social networks via bounded local path traversal. In *Computational Aspects of Social Networks (CASoN), 2011 International Conference on*, pages 66–71. IEEE, 2011.

A. Popescu and G. Grefenstette. Mining user home location and gender from Flickr tags. *ICWSM'10*, 2010.

D. M. Romero and J. M. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 138–145, 2010.

D. M. Romero, C. Tan, and J. Ugander. Social-Topical Affiliations: The Interplay between Structure and Popularity. *arXiv.org*, Dec. 2011.

M. Rowe, M. Stankovic, and H. Alani. Who will follow whom? Exploiting semantics for link prediction in attention-information networks. 2012.

S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM, 2011.

CHAPTER 2

Related Work

In this chapter we review existing literature in the field of online social networks and location-based social networks. We summarize previous work and concentrate the results and findings of previous experiments. A more detailed elaboration of related work can be found in the individual chapters.

This chapter can be divided into two sections: In Section 2.1 we present an overview of existing work in the general field of online social networks. In Section 2.1.1 we focus on link prediction in these online social networks and tighten this problem in Section 2.1.2 where we also consider the strength of links between users for the prediction task. In Section 2.2 we broadly cover literature on location-based social networks and work that is related to position data in general. Section 2.2.1 covers future place prediction whereas Section 2.2.2 focuses on the prediction of future links in location-based social networks.

2.1. Online Social Networks

It's in the nature of humans to form relations and graph theory defines a model to describe these relations in a mathematical structure: A graph is a mathematical structure that models the pair-wise relations between objects.

> A graph G is an unordered triple $\langle V(G), E(G), \psi_G \rangle$ consisting of a non-empty set $V(G)$ of *vertices*, a set $E(G)$, disjoint from $V(G)$, of *edges* and an *incidence function* ψ_G that associates with each edge of G an unordered pair of (not necessarily distinct) vertices of G. [Bondy and Murty, 1976]

The analysis of social networks started back in the 1960s (see [Barnes, 1969] or [Mitchell, 1969]), where users of the social network are represented as vertices and the relation between these users is indicated by edges between these vertices. Due to the limited availability of real word data, the analysis of online social networks started with small datasets and one of the first sets was introduced by Zachary [1977] that focused on an anthropological approach of detecting and analysing conflicts in a small karate club. The authors analysed the structure of the social network and explored the implications of the conflict that finally yielded in the separation into two clubs.

With the advent of the World Wide Web, the availability of larger dataset became feasible and starting with the *GeoCities* network, the *SixDegrees* network [Boyd and Ellison, 2007] and *MySpace*, *Facebook* finally became the most successful online social network with over 845 Million users in February 2012 [Wilson et al., 2012]. Due to it's popularity the Facebook social network is in the interest of researcher and so there are several datasets available under the umbrella of the Stanford Network Analysis Project [Leskovec, 2012]. Since it was introduced in 2004, the usage of Facebook changed in various ways: Lampe et al. [2006] state that the online social network of Facebook became more and more part of user's daily life. One reason for this is the increasing popularity of mobile phones and the according ease-of-use to access the social networks ubiquitously [Kisekka et al., 2013]. Users adapt the social activities and social environment on the network upon their personal changes using these new technological features, e.g. making new friends or moving to a different location [Lampe et al., 2008].

In most cases the social network of Facebook aims at users with an existing offline bonding who want to stay in touch with each other and with all their already known friends [Lampe et al., 2006]. Once a link in the network is created, the main motivations are the preservation of the real-world friendship and the access to information about users they met socially in class or dormitory [Lampe et al., 2006]. These users try to model their personal online profile according to their real personal profile and get the attention of their existing contacts from the real life [Lampe et al.,

2006]. This interplay between the real world and the social network is not one-way only but the effects are reciprocal: As reported by Ellison et al. [2007], relationships in Facebook have an effect onto the real world ties of users as well. In their paper they even found positive influence on the self-esteem and the life satisfaction of Facebook users. Further, the usage of an online social network even helps users to maintain their old relations and create new relations when they change their real world community due to the move to a different location [Ellison et al., 2007].

Per se, social networks are not statical but instead evolve over time. Users declare friendship and communicate to each other for a certain timespan [Chun et al., 2008]. In their paper Chun et al. [2008] analysed the differences between a social network i.e. friendship network, and the actual activity network i.e. communication network, of comments in an online guest books. Although they found similarities in the overall structure in both network types, the activity network contained more information due to its weighted and directed character and the additional factor of time. This implies that the actual interactions between users allow to draw a more accurate and detailed picture of the overall relations between users. Similar observations were reported by Wilson et al. [2009] who analysed the interactions and social relations of approximately 10 Million Facebook users. They reported differences between the activity network and the social network and suggested to "design social applications with the interaction graph in mind" as they "reflect the real user activity rather than the social linkage alone".

"Similarity breeds connection" - in networks of any type the homophily, i.e. alikeness or similarity, between users forms connections [McPherson et al., 2001]. Mislove et al. [2010] reported a high homophily of befriended users in the social network of Facebook which implies that user groups of befriended users share similar attributes. Based on this observation Mislove et al. [2010] inferred the attributes of all users in a group from a subset of only 20% of the users with over 80% accuracy. This indicates that users with similar attributes and characteristics form communities in social networks. In a subsequent work Ugander et al. [2011] investigated in the anatomy of the entire Facebook graph and even revealed strong age homophily on a local level and a strong nationality homophily on a global level. Surprisingly, they did not find any signs of a gender homophily between users. This observation goes inline with Lewis et al. [2008] who experimented with users from different social environments and areas. Their analysis unveiled differences in the characteristics of users with different status and location. Golbeck et al. [2011] investigated in the personality of Twitter users and used the publicly available information on their Web-profiles to classify users into the "Big Five" personality traits: Openness to Experience, Conscientiousness, Extroversion, Agreeableness and Neuroticism. Although there were differences between the classes, they could overall predict these classes with high accuracy using machine learning algorithms. An extension to this work was done by Adali and Golbeck [2012] who used the semantics of exchanged messages of Twitter users and compared it to the actual relations between the users.

Based on the similarity and the influence of their directed followers they tried to find the characteristics of personality. Their results go inline with earlier experiments by Golbeck et al. [2011] who unveiled the social environment of users as a good predictor for their personality.

Online social networks can not only be used to represent the profiles of single users and their personal information but also to infer the social relationship between them. In the next section we first focus on the analysis of these interactions and further on the prediction of social relations and interactions.

2.1.1. Link Prediction

One of the most prominent research topics in online social networks is the prediction of new links between users that evolve in the future. The idea is to create a graph model to represent the social network and use it to apply measures that map the relations between single users. These relations can depend on information that is intrinsic to the network, i.e. topological features like the common neighbours of two single users, on extrinsic information, i.e. homophilic features like the mutual interests of two users, or on a combined set of topological and homophilic features.

One of the well-known experiments to predict new links in an online social network using topological features was conducted by Liben-Nowell and Kleinberg [2002]. They defined the "link prediction problem" and applied their approaches to five co-authorship networks of scientific research papers. The examined networks changed their structure dynamically over time as more and more links formed new and so they applied a supervised learning approach to predict these evolving links. For their experiments they used intrinsic network information, i.e. information that can be derived from the topological structure of the network and their analysis revealed that this is a very valuable source for the prediction of future links between users. They identified the number of common neighbours two unconnected users have as one of strongest indicators for a possible future link. This concept was pursued by Romero and Kleinberg [2010] who investigated in these *triadic closures*, i.e. users that are linked through common neighbours tend to form a link as well. For their experiments they exploited information from the directed network structure of the Twitter network. On the one-hand side they could verify that the existence of a common neighbour in the network is a strong indicator of future connections in directed social networks but on the other-hand side they also found that this features is not homogeneously distributed over the network. The actual in-degree (number of neighbours with a link to the user in the directed network) is not the main predictor for a triadic closure but instead the in-degrees of all neighbours of a user is the best metric to predict future links. This goes in-line with Ugander et al. [2011] who found a clear degree associativity which means that users with higher activity connect to users with a higher activity as well.

Topological features can be broadly divided into local features that exploit the actual neighbourhood of users within the network and global features that take the entire structure of the network into account. Local features only consider the distance between two users using their direct neighbours, e.g. common neighbours, whereas global features consider the structure of the entire network, e.g. Katz Status Index [Papadimitriou et al., 2011]. Global features require in general more computational resources but they have a higher predictive power if compared to local features. As a consequence, there were attempts to combine local and global features.

Zhou et al. [2009] exploited nine of the most popular metrics that rely on the structure of the network and compared six different networks with respect to these features. Due to the fact that global based algorithms are very slow but effective and local based algorithms are fast but with a worse performance they tried to find a trade-off between simplicity and performance. As a consequence they introduced a new metric that has the performance of a global feature but the structure of the Adamic-Adar measure. Overall the introduced measure had a better performance than existing local features with a slightly higher computational costs. Another attempt to reduce the complexity of global features was done by Fire et al. [2011]. They introduced a replacement for the Katz Status Index referred to as Friends-Measure and evaluated this measure on five large online social networks using supervised machine learning algorithms. Similar attempts to combine different feature sets where made by Papadimitriou et al. [2011] who also tried to boost the performance of features based on the local network structure using global features. They introduced a new global measure called Friend Link that improves other local features and also other existing global features.

A first attempt to isolate homophilic features from profiles in the Twitter online social network was done by Golder and Yardi [2010]. They blanked out the topological and structural properties between users and let users choose to follow others only upon their social profiles. Although they compared it to existing topological techniques, the profile information alone was a significant predictor for new links in the network and even outperformed existing network-structure based algorithms. The network of Twitter gained attraction of other researchers as well due to it's hybrid information-entertainment architecture and the corresponding information richness. Yin et al. [2011a] states that existing approaches for the prediction of new links based on the network topology alone are not sufficient as the structures and relations between users are too dynamic and complex. In a subsequent paper Yin et al. [2011b] unveiled that 90% of upcoming links are triadic closures, i.e. users are not connected but have common neighbours and hence are not more than two hops away. Among the homophilic features they used for the analysis, the age of the user account was a crucial factor for the prediction of future links. This analysis revealed the particular importance of the profile information and Rowe et al. [2012] went a step further and exploited the semantic content of tweets to predict future followers. Their experiments showed that the seman-

tic information of the text messages even outperform features that are based on local topological information in the network. They further stated that social aspects play an important role in the creation of new links and users with a high degree, i.e. users that are well connected to other users in the network, are even more driven by these social factors.

Lee and Brusilovsky [2010] used the citation network of *CiteULike* to show the influence of homophily between users with varying social proximity in the network. They used different features based on profile metadata, tags and information items in general. Their experiments showed that connected users share more items with each other and are therefore more similar to each other than users that are not connected to each other. This observation can even be extended because the similarity between users even decreases with increasing distance in the network. For instance, users with a distance of two in the social network are less similar than users that are connected. Identical observation where made by Schifanella et al. [2010] who investigated the social networks of *Last.FM* and *Flickr*. In particular they looked for the similarities of users considering the groups they joined and interests they specified. They analysed the groups users participated in and the shared tags based on their music taste. Again, they found a strong correlation between friendship between users and the homophilic of their profiles. For the *Last.FM* dataset they even found that these similarities are more valuable for the prediction of friendship than the listening patterns of users.

Topological and homophilic features are both valuable predictors for the prediction of future links and Shibata et al. [2012] evaluated the differences between them. They investigated in a citation network to find new co-authors and used homophilic features (attribute features, semantic features) and topological features to predict this new co-authorships. With their experiments they could identify Jaccard's Coefficient and differences in the betweenness centrality from the topological features and the similarity between documents of two users from the homophilic features as most valuable. Their work showed that both types of features have high potential to predict new links.

However, there is still the unresolved questions on the actual method to combine homophilic and topological features to get the best prediction results. Backstrom and Leskovec [2011] used Supervised Random Walks to combine the structural information hidden in the network and the information of the actual user. Their results showed that this approach identifies positive links better than existing supervised and unsupervised learning approaches on either feature set. Al Hasan et al. [2006] also used a combination of topological and homophilic features to predict links in a co-authorship network of scientific publication data. As their focus was on computational efficiency, they omitted global topological network features and only used homophilic and local topological features. With the Support Vector Machine learning algorithm they identified a subset of features like the sum of common neighbours or similar keywords in the profile as most suitable

for the prediction of new links. In contrast, Chelmis and Prasanna [2012] evaluated the directed communication of users in the social networks of Facebook and Twitter using an unsupervised learning algorithm. They used topological features and combined them with context information of user-pairs using textual and temporal features, i.e. content and date similarity of interactions. Their analysis showed that this approach outperforms existing solutions and they finally state that the prediction gets better the more homophilic and topological features are available. Zheleva et al. [2010] extended this idea and used an existing online social network and added the information of the family structure of users. They derived homophilic features from the base network and extended it with the structural information of the family network. In particular they used the strength of ties between family members to achieve better results for the prediction. This additional source of knowledge yielded in a significantly higher results for the prediction of future links if compared to traditional approaches.

2.1.2. Influence of Tie-Strength

The strength of a relation between users can not only be used as an additional source of knowledge [Zheleva et al., 2010] but it can also be seen as an extension to the link prediction problem. Granovetter [1973] explored the different strength of links between users in a network and his studies revealed that the structure of networks highly depend on weak connections between users and so does the information transport within networks: he identified groups of users in the network and identified the weak links between these groups as essential for information transportation and the coherence of the network.

With the advent of larger social network datasets, the proof of Granovetter's theory became more and more feasible and for instance Gilbert and Karahalios [2009] used a Facebook dataset to investigate in the tie strength problem. They asked users to manually assess their relationships to other users and differentiate between user-pairs with a strong tie and user-pairs with a weak tie. For their network model, they used seven classes e.g. intimacy, intensity or duration, to define the features between users. Their experiments showed that the intimacy (number of friends or intimate words used) and intensity (number of wall words or number of outbound posts) were most suited to predict the tie strength. Jones et al. [2013] also used the social network of Facebook to predict the tie strength between users and conducted interviews with the users to assess the findings. For the actual prediction they exploited features obtained from the private and public messages sent between users. Their analysis revealed that the frequency of interactions is the most successful predictor for strong ties - the higher the frequency of communication, the higher is the probability that the user relation is a strong tie. Surprisingly, public communications on the wall (postings, comments, loves) is more valuable for the prediction of tie strength than the private communication

is. Another interesting proof of Granovetter's theory was conducted by Onnela et al. [2007] who used the data of a mobile phone network to investigate in the tie strength between users. They unveiled that the actual structure of the network is very robust if strong ties are removed but the network completely collapses as soon as weak ties are deleted. In contrast to this global effects upon the removal of weak ties, they observed that strong ties are vital to local structures in the network as weak links build around them.

As mentioned in Chun et al. [2008] there is a difference between the social network and the activity network of users. The activity network contains more information with all the interactions whereas the online social network has a broader meaning especially for sociological studies [Kahanda and Neville, 2009]. Kahanda and Neville [2009] focused on the tie-strength in an activity network with postings, comments and loves. For their supervised learning approach they used transactional information to model the information transfer between two users, topological features to model the local and global structures in the network and network-transactional features to model the information transfer between users with respect to other interactions in the entire network. Their experiments revealed that these network-transactional features are very valuable for the prediction of strong ties as they model the actual number and types of all interactions between users instead of simple counting the exchanged interactions between two users. Viswanath et al. [2009] used a Facebook dataset collected over a year to investigate in the changes of the activity network over time. They found significant differences as the activity network is highly time dependent whereas the social network maintains structural properties over time. Due to the fact that the tie strength is directly related to the activity network, there is a strong deviation of weak and strong ties.

The link prediction problem can be tightened to predict the tie-strength of an existing link between two users. Leskovec et al. [2010] applied topological features to several datasets and focused on the prediction of positive respectively negative links, e.g. *Slashdot* with links that are either "friends" or "foes". They state that the information about the negative links in a social network are an additional source of knowledge and therefore have significant influence on the characteristics of the network and thus to the prediction of positive links. Another example is Cheng et al. [2011] who determined whether users of the Twitter network have a reciprocal communication or not. Among all the features they examined, they identified the similarity of users as the most powerful feature. They state that users with a similar status have a high probability that they have a reciprocal communication as well. Among the number of interactions, the actual sign of the communication (uni-directional or bi-directional) is an additional source of knowledge and therefore a valuable source for link or tie-strength prediction tasks.

Granovetter [1973] stated that the information diffusion in a network highly depends on the type of the connection between users. This was examined by Romero et al. [2011] who tried to

predict the popularity of hashtags in the Twitter network. In their first analysis they found that the local social structure of people in a network that use a certain hashtag has a crucial influence on the future popularity of these hashtags. As a consequence they inferred that similarities in hashtags are a valuable source to predict future links within a network. Indeed, according to Granovetter's theory of information transmission among weak links, they found remarkable results for the prediction of weak links using the similarity of hashtags. This results were even topped when they combined the homophilic information of the hashtags with the structural information of the local social network. Buccafurri et al. [2013] defined the term tie-strength in a slightly different way and defined the connections between users in different social networks as weak ties. They introduced the term "bridge" for users that participate in more than one social network and hence connect users from different networks. They compared the properties of these bridge-users with power users of single networks and found that both cause a higher network degree in the combined network although there was no correlation between the two groups. Overall they confirmed that the combination of several different networks is a valuable source of knowledge which yields in better insights into the structure and characteristics of the network. A similar approach was chosen by Gilbert [2012] who tried to map the tie strength of users from one social medium to another. They used the Twitter network to learn about the relations between users and validated the found prediction model using a Facebook dataset of the same users. One of their main results was that the Twitter-built model generalizes to Facebook and can therefore be used to predict the tie-strength between users in other networks.

The more information available, the higher is the predictability of links or tie strength between users. So far we have seen that these sources of information or knowledge can be derived from the structure of networks, the homophilic similarity of user-pairs or other social network users. Another type of social networks that became available on large scale during the last years are *location-based social networks*. In the next section we will present an overview of literature in the fields of location-based services and location-based social networks.

2.2. Location-Based Social Networks

A native approach to cover the relations between users is the collection of the places they visit and their moving trajectories. In the past it was nearly impossible to collect this data but with the advent of mobile GPS devices (e.g. mobile phones) and Web platforms with focus on location data (e.g. FourSquare), it became feasible to harvest this information.

From a privacy perspective the collection of tracking data is very ambivalent. Michael et al. [2006] examined the benefits and drawbacks of location based tracking data and equipped users

with GPS enabled devices. After the data collection the users where interviewed about their personal concerns from the perspective of law enforcement and health aspects. This interviews revealed that users think this information can be useful for monitoring people with health problems but can also be very dangerous if the collected data is in the wrong hands. These observations go inline with Iqbal and Lim [2010] who also investigated in privacy issues with location-based data. They equipped volunteer users with GPS enabled devices and compared these data with self reports. They tried to rise the awareness of issues with the privacy of location based data and emphasize the need for ethical and legislative regulations to protect from abuse. Similar approaches and findings where made by Consolvo et al. [2005] who investigated the demanding problem of trust and privacy of location based data. In contrast to Michael et al. [2006] they did not ask for concerns in the collection process itself but for the sharing behaviour of users. They found out that users make the decision to share data depending on three factors: 1) it depends on the actual user or service that requests the data, 2) it depends on the actual reason why the information is desired and finally, 3) it depends on the the accuracy and resolution of the requested data. Further they found that the actual social relation between requester and user is not an indicator for the willingness to share location information. Li and Chen [2010] collected data over a period of 21 months from the location-based social network of *Brightkite* and found privacy concerns correlating with the age, gender, mobility and the geographic regions of users. They inferred that the privacy settings are highly influenced by other users that are are connected with a strong-tie, i.e. friends. According to Sadeh et al. [2009] users have problems in articulating their privacy preferences for sharing their location data through different applications. Privacy preferences change over time and the usage of applications. Hence, these settings can not be assigned at one time but need to be adopted over time. As a solution they presented machine learning techniques to automatically pre-select these settings for users.

One of the main concerns users had was the in-correctness of the collected data and the consequential implications. Abbas [2010] equipped users with GPS enabled devices and interviewed the participants upon their attitudes and raised socio-ethical questions about this collection process. The concerns were basically related to the inaccuracy of the datasets and it's implications, respectively the privacy concerns when sharing sensitive information with others. This goes along with a subsequent study by Abbas et al. [2011] to unveil the concerns of users when they are tracked by mobile devices. Although users can benefit from the tracking of position data, all of the participants had doubts about the vulnerability of the tracking system and the incorrectness of the data. The missing integrity of the data could have negative impacts and could result in incorrect evidence for incrimination [Abbas et al., 2011]. This even confirmed a previous study by Tsai et al. [2009] who reported user concerns in sharing location data with others. Although users identified some useful applications (tracking in case of emergency) users believed that the

drawbacks outweigh the benefits.

Another paper that investigates in trust and privacy concerns of users utilizing location based tools was conducted by Fusco et al. [2011]. They examined the different relational types between related users, e.g. friends or co-workers, upon their willingness to share location information. Although most of the people did not exactly know what the consequences of revealing their location information were, most of them believed that this kind of information would have a big impact on the future relations. Besides the concerns about privacy and security of these settings, some users also believed that sharing location information could strengthen their social links in two ways. Disclosing location information to friends was interpreted as a sign of trust, respectively backing off this information was also interpreted as sign of trust: "if you trust me then why the need to do lookups on my real-time or historical physical whereabouts? You should just believe me when I tell you where I am, where I have been and where I am about to go." [Fusco et al., 2011]

Similar to online social networks the raw data can be used to establish a network that models the relations between users. Fusco et al. [2010] defined location-based social networks as:

> "A Location Based Social Network is the convergence between location based services and online social networking."

One of this location-based social networks that enables users to share their locations is FourSquare and Lindqvist et al. [2011] conducted interview with users of this network. These users stated the main motives to use the service as: 1) show the current location to a user's friends, 2) use the check-ins at certain locations as a method of self presentation and finally 3) exploit the location information to coordinate with friends. Surprisingly, they also identified a group of users that exploited the Web service for safety reasons to inform others about their current location. Lindqvist et al. [2011] also reported that users are eager to check in at special places and locations that are unique and interesting instead of usual places. This behaviour sets them apart within their user group. In contrast to previous publications by Consolvo et al. [2005] or Michael et al. [2006] only a minor group of their interview partners had concerns on privacy. The authors believed that this could be explained by the selection of interview partners as they already used the FourSquare Web service and a therefore biased.

Examples for commercial applications of location-based data where made by Traynor and Curran [2012] who identified their business values for advertisements and user profiling. They addressed the security and privacy challenges of users which (in most cases) do not go along with the business cases. As a consequence they suggested the difficult challenge of a negotiation between all involved parties which is mandatory for the acceptance of these services. On the other hand, not all applications that are based on location data have these tight privacy demands. Wang et al. [2010] for instance, used internal social ties in a company based on location data to organise

meeting rooms schedule and connect the users through this interactions.

A requirement for all applications based on position information is the analysis of the structure of the location-based social networks and the relations between users. Li and Chen [2009] did a large scale analysis of the *Brightkite* network which is a location-based social network that allows users to "check-in" at certain places. In their analysis they classified users according to the locations they have visited and found high degree users with a higher mobile activity. It is obvious that the higher location diversity and frequency of a location change make future locations of power-users harder to predict regardless of the additional data. Scellato et al. [2011a] tried to infer the actual friendship between users from the geographic properties of their social relations. In their experiments they found that users with only a few friends have on average a shorter spatial distance to their friends than people with a lot of friends. They assumed that this positive correlation between the number of friends and distance is caused by social triads that are geographically wider. A deeper analysis of the tie-strength also revealed that the distance plays a crucial role for friendship as well-connected users have a shorter geographical distance. These results were also supported by Atzmueller et al. [2012] who investigated in the spatial information of users at scientific conferences. They tried to infer the relations among participants based on the sessions they attended and the face-to-face contact between users. They could find a correlation between the duration of the face-to-face contacts and the sessions users attended in common. Using the relationship information and the spatial information they could even identify the different roles of users within the conference. The relation between interactions and co-locations was also investigated by Liu et al. [2012] who worked on the correlation between attended events of users and the social interactions in online social networks. They found a high correlation between the social interactions and the visited events, and with a detailed location analysis they identified 70% of all online friends and nearly 85% of all location friends living within a distance of 10 miles. One of the first attempts to unveil the social information of users hidden behind location-based data sources was done by Popescu et al. [2010]. They examined the social information within the social network of *Flickr* and analysed the tags users attached to their pictures. Interestingly, the experiments did not only unveil the preferred locations of users but also their age.

The spatial distance between users is one of the most limiting factors of human relations. Experiments conducted by Mok and Wellman [2007] revealed a maximum distance for face-to-face contacts in our everyday life of around 5 miles. They further found indicators that the frequency of contacts drops off after 50 miles (equivalent to an hour by car) for personal contacts and 100 miles for telephone contacts. The datasets for these experiments were collected in the Nineteen Seventies but the authors of the paper inferred that the advent of social networking sites and the according ease of communication simply moved these limits further away. Mok and Wellman [2007] concluded that humans need the physical contact to others for continuous relationships but

the possibilities of a long distance communication with social media opens new directions for long distance relationships. In a similar paper by Cho et al. [2011], the authors found out that short distance travels in general are not influenced by the social relations whereas long distance travels are highly influenced by the social relations a users has. Overall, 10-30% of all movements of users could be explained by periodic movement whereas 50-70% of all movements were influenced by periodic movement.

Joseph et al. [2012] classified groups of users not upon their role or position the network but just by the places they visit. They found that geo-spatial properties and homophilic information of users are strong indicators and powerful predictors for the groups they are associated to. They state that the knowledge of classifying users can be useful to sharpen the check-in behaviour of users – an example application would be to identify tourists upon their movement and make appropriate suggestions for future locations to visit. A similar approach was done by Hegde et al. [2013] who used the profile information of users to classify places and assign tags to these places. These tags are derived from the interests and habits of users that visited that place. They could finally derive meaningful tags for places and these tags stabilize as more users visit a certain place. Li et al. [2008] investigated in the similarity between users and the probability of a contact if users share not only the same regions but also the same mobility patterns. They found out that similar users also share the same trajectories of movement and they could even observe an increasing user-alikeness the longer a common trajectory was. Similar to activity networks, the additional consideration of the time sequences resulted in an enormous boost for the prediction of user contacts. This results where confirmed by Gonzalez et al. [2008] who used the trajectories of mobile phone users and created patterns of their daily movement over a period of 6 months. Their analysis uncovered that users, despite their home and work location, follow reproducible patters. They identified the epidemic prevention of diseases and urban planning as a potential application for their findings. Another paper that investigates in urban planning was done by Giannotti et al. [2011] who equipped cars with GPS sensors and collected trajectory data. They also reported that these mobility patterns are a very powerful source of information but it requires a complex model to analyse the raw information to bring it into a useful and valuable resource.

2.2.1. Predict Future Places

The actual prediction of places users visit in the near future and new occurring links highly depends on the used source of data. Cho et al. [2011] compared a location-based social network where users specify their location using *check-ins* with location data provided by a telephone company. They analysed both data sources and found that check-in data is more detailed and valuable because users provide more information about their actual place if compared to the mobile phone data, e.g.

the fast food restaurant in the first floor and the company in the third floor of the same building can not be distinguished using mobile phone data but it can be distinguished using the meta-information of a check-in.

As stated in Cho et al. [2011] the mobility patterns of users are highly influenced by periodic movement of users. With this fact in mind, Song et al. [2010] investigated in the predictability of humans and found surprising results. Although only around 50% of all movements follow daily routines [Cho et al., 2011], they conducted experiments to predict the future locations of users with 93%. The conclusion of their observations was that these results combined with the regularities of users could have a positive influence on traffic engineering and urban planning. This regularities were further investigated by Gao et al. [2012] who evaluated the socio-historical ties in location-based social networks. They focused on the social contacts between users in the FourSquare network and compared it to the places users actually visited. They found that users in general "visit few places many times and many places a few time" which follows a power-law distribution. Further they showed that the location history of users influences their further places only for a short period of time, e.g. it is very likely that people go to a coffee shop right after they had a lunch break. Their experiments revealed that user-pairs with a social relation tend to visit similar places. This observation was confirmed by Lerman et al. [2012] who investigated in events users attended with respect to their role in a social network. Their experiments revealed a positive correlation between the number of common friends two users have and the probability that they will attend the same events. Based on these correlations they tried to predicted the future events users will visit. They proved that a higher similarity of two users in the online social network, results in a higher similarity for attended events and visited locations. Another approach to identify users that visit similar events was done by Zhang et al. [2012] who tried to find the "geo-social influence" of users in a given location-based social network. They attempted to identify the influence of social connections when visiting events and the identification of these events. Their experiments resulted in the spatial distance as most prominent feature for the relation between users, i.e. physically close users are more similar.

The moving behaviour of users and their visits to places was further examined by Noulas et al. [2012b] who investigated in two different location based social services. They found that 60-80% of all users checked-in at places they did not visit during the last 30 days and up to 80% of all locations were visited for the first time. In their approach they tried to recommend places to users they did not visit before and learned from "social ties and venue-visit data simultaneously". Their results revealed that a combination of social network data, the history of users and place features like frequency were most valuable for the recommendation task. In a subsequent work Noulas et al. [2012a] compared the prediction of a user's next place using unsupervised and supervised machine learning approaches. In the unsupervised prediction approach, implemented as a ranked list, they

used single features based on the previous location of users and their friends to predict user's next places. For the supervised approach they combined these features and reduced the problem to a binary classification problem. They state that the supervised learning approach outperformed the unsupervised approach due to the combination of features and the according information gain. Backstrom et al. [2010] found evidence (in particular the spatial distance between users) that the locations users visit have a high influence on the social communications. They showed that the probability of friendship falls monotonically with the distance and more specifically they state that "the probability of friendship is roughly proportional to the inverse of distance" for medium to long range distances. Based on this observations they also tried to predict users future locations by adding the available social relations to their models and outperformed other models significantly.

Zhou et al. [2012] used the history of user check-ins in the *Gowalla* location-based social network and explored different collaborative filtering approaches for the prediction of new locations. In contrast to previous studies they did not use any domain knowledge but only the actual user's location history. The recommenders they used were based on the exploitation of three different check-in types: 1) recommendations based on the location history of other users, 2) recommendations based on previous locations of the actual user and finally, 3) recommendations based on a semantic analysis of locations. They found the approach with semantic information about previous locations as the most valuable as it showed the best results for the prediction. Besides Gowalla and FourSquare, *Facebook* as one of the most successful online social networks introduced "Places" in August 2010 which allows users to share location information with their friends. Chang and Sun [2011] analysed the location history of Facebook users and tried to predict future check-in locations of users and the response of their friends upon this check-in. For the actual prediction of future places they identified the location history of users and their friends as most suitable. Then they predicted the responses of other users for these future check-ins and among others they identified the physical distance between the actual user and the responder as most valuable metric. Li et al. [2013] combined the prediction of future locations and the prediction of new links that occur in the future using the spatial-temporal components of the Gowalla location-based social network. Their analysis showed that visited locations have a high influence on future locations and the prediction of future social interactions as well. They used the inferred social-temporal features for supervised learning approaches and found remarkable results as they could predict future relations with over 90% just by using the social-tie features and similarities in the behaviour of users.

In the next section we will focus on this link prediction problem in location-based social networks and give a rough overview of existing literature.

2.2.2. Predict Future Links

In the previous section we have already seen that similar users tend to visit the same places and this knowledge can be used to predict future places of these users. Eagle et al. [2009] used the alikeness of users to extend the prediction of places to the prediction of new links between users. They collected data of users with mobile devices and compared it to self reported data. Experiments based on this dataset showed that location data generally correlates with the self reported data and even unveiled that about 95% of all friendship data could be predicted with the observed data. Their experiments were a first step into location based network analysis and brought first insights into the predictive power of the data. The two applications of future place and link prediction were also investigated by Ye et al. [2010] who applied unsupervised machine learning techniques using a simple collaborative filtering. Their approach was two-fold: 1) predict social links upon common visited places and 2) predict new places upon the location information of social ties. They validated their approaches using a dataset from the location-based social network of FourSquare and showed that prediction in both directions is feasible. Another paper that investigates in the correlation between the social proximity and the mobility traces of users was done by Wang et al. [2011]. They used mobility trajectories collected with mobile phones and tried to infer their proximity in the network. With their supervised learning algorithm they could outperform traditional prediction algorithms based on the social structure alone by adding location based information and data. Besides similarity features, i.e. places visited in common, the time sequence of these visit plays an important role in the prediction of future places and links. Based on this assumption, Lauw et al. [2005] found out that two individuals are likely to know each other if they have a co-occurrence in the same place at the *same time*. They used these "spatio-temporal co-occurrences" to model relationships between users and create a network from this data. They further used the intensity and the number of co-occurrences to make more precise measures about these relationships, i.e. they added a weight to differ between strong and weak ties. Finally they compared their approach to another similarity based approach to link users and found promising results that open further direction for research. One of these projects were simple predictions upon the actual places two users visited by Xiao et al. [2012]. They separated the prediction algorithms into three different modes: geographic overlap and semantic overlap and location sequence. They state that the probability that two users share the same interests does not only depend on the actual places they visit but also on the semantic information behind the location, e.g. a tag or category assigned to these places and the actual time sequence when users visits these places, e.g. first museum, then shopping mall and finally café. They applied their model to a location based dataset and showed that a model that considers these time sequences outperforms models without this information.

The prediction of new links can not only be applied to large online social networks like FourSquare but also to smaller datasets like halls of residence or scientific conferences. In their paper

Dong et al. [2011] tracked students at their dormitory rooms using their mobile phones and inferred the actual social relations between them from their location trajectories. Their results upon movement and social interactions with other students revealed that the collected information correlates with the health and relational status of these students. Further they found that the mobility traces of users change over the years due to changes in their lifestyle habits. Choi et al. [2013] conducted a similar experiment with trajectories of students and their communication habits. Although they only had few participants in their experiments they could identify two types of relationships: formal contacts and informal contacts. Chin et al. [2012b] examined the homophily of research-conference attendees by measuring their locations, i.e. attended sessions, for a recommendation system of new social links. They identified physical interactions as very useful as users attending the same sessions have similar research interests and therefore have a strong homophily between them. In a subsequent work Chin et al. [2012a] used position data and interaction data to focus on the activity network and investigated in the time before and after an actual tie was created. They found strong evidence that users tend to create ties ("friend", "follower" or "exchanged contacts") in online social networks after a co-occurrence at a certain location but the interaction activity (exchanging messages) on the online social network significantly drops after this link has been created. Scholz et al. [2012] followed this direction and employed the communication duration of conference participants to classify their relations into strong and weak ties. They identified topological network features a most useful to predict new links and to determine strong ties between users. Based on these results they predicted recurring links during the conference and found the contact duration combined with the strong link information as feature outperforming all other network topological features.

Besides homophilic features and topological features alone, a combination of both can support the prediction of links and tie-strength. Scellato and Mascolo [2011] showed that future links are highly influenced by location trajectories of users and in a subsequent study Scellato et al. [2011b] even found that the simple "common places visited" approach could identify 30% of all users as place friends. To improve these results they expanded the used geo-location features with social features (e.g. neighbours in the network) and global features (e.g. distance from home) for the prediction task. Using this combination of features they could identify 66% of all links that evolve in the future which is an tremendous improvement of predictability. Another interesting study that combines homophilic and topological metrics by Guy et al. [2010] considered nine different features which were classified into three different classes. 1) the common *friends* of a user-pair, 2) the *things* two users are interested in and 3) the *places* two users visited in common. Although the nine features alone produced varying results, the combination of features into classes was more successful: They found that the *things* two users are interested in is the most useful feature set for the prediction of similarity between users which is comparable to the prediction of a social

tie. Allamanis et al. [2012] used the spatial distance between users and their social information within the network to predict future social relations. Their analysis revealed that each source can predict future social links alone but a combination, referred to that as "gravitational attachment process", outperforms either sources. This combination takes the triadic closure of users in the social network and the spatial location of users for the prediction of new links into account. Further they used features and metrics of the places users visit to enhance the prediction of future links.

Bischoff [2012] showed that there is a correlation between homophilic similarities and structural relations between users in an online social network and the locations these users visit. They classified the links between users according to the events they visited and found a strong correlation to the tie strength in the online social network. Cranshaw et al. [2010] employed data collected from the location-based social network *Loccacino* which combines position information of users (check-ins) with the online social network of Facebook (text interactions). They found a strong correlation between the location trails of users and their social interactions. They employed features that measure the actual locations visits of users and the entropy at these places to support topological features of the online social network. This yields in "model of friendship" that highly correlates with the actual online social relations of these users. Another attempt to combine the position information of users with an online social network was done by Pan et al. [2011]. They combined the knowledge of an online social network and a location-based social network and analysed the interactions between users with respect to the tagging of places on Facebook. In their experiments they measured the influence of either networks in the combined network and found interesting results: The directed and weighted interaction network has a strong influence on the community structure of the entire network whereas the location based social network, referred to as *co-presence network*, has a strong influence on the degree and path metric. They concluded that a combined network of interaction and location information shows different aspects of users interaction and social relation but overall they complement each other.

Volkovich et al. [2012] collected data from an online social network and a location-based social network and separated the links according to Granovetter's theory into strong and weak ties. They observed that user pairs that are close to each other in the interaction network, i.e. strong ties, also have a shorter spatial distance between them if compared to weak ties. Further they observed that users with shorter spatial distances are more likely to form sub networks with a core structure. These results can be compared to Pappalardo et al. [2012] who tried to find indicators for the tie strength between users in multi-modal networks. The number of text interactions between users has already been identified as valuable measure but in their work they applied the tie strength to location based data. Besides the actual visits of locations they found that the tie strength of a link is also highly influenced by actual place of the interactions.

References

R. Abbas. Location-based services: An examination of user attitudes and socio-ethical scenarios. In *Technology and Society (ISTAS), 2010 IEEE International Symposium on*, pages 357–365. IEEE, 2010.

R. Abbas, K. Michael, M. G. Michael, and A. Aloudat. Emerging forms of covert surveillance using GPS-enabled devices. *Journal of Cases on Information Technology (JCIT)*, 13(2):19–33, 2011.

S. Adali and J. Golbeck. Predicting Personality with Social Behavior. In *Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on*, pages 302–309. IEEE, 2012.

M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM'06: Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

M. Allamanis, S. Scellato, and C. Mascolo. Evolution of a location-based online social network: analysis and models. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 145–158. ACM, 2012.

M. Atzmueller, S. Doerfel, A. Hotho, F. Mitzlaff, and G. Stumme. Face-to-face contacts at a conference: dynamics of communities and roles. pages 21–39, 2012.

L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 635–644. ACM, 2011.

L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, pages 61–70. ACM, 2010.

J. A. Barnes. Networks and political process. *Social networks in urban situations*, pages 51–76, 1969.

K. Bischoff. We love rock'n'roll: analyzing and predicting friendship links in Last. fm. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 47–56. ACM, 2012.

J. A. Bondy and U. S. R. Murty. *Graph theory with applications*, volume 290. Macmillan London, 1976.

D. M. Boyd and N. B. Ellison. Social network sites: Definition, history, and scholarship. *Journal of Computer-Mediated Communication*, 13(1):210–230, 2007.

F. Buccafurri, V. D. Foti, G. Lax, A. Nocera, and D. Ursino. Bridge analysis in a Social Internetworking Scenario. *Inf. Sci.*, 224:1–18, 2013.

J. Chang and E. Sun. Location3: How users share and respond to location-based data on social networking sites. In *Proceedings of the International Conference on the Weblogs and Social Media (ICWSM'11)*, 2011.

C. Chelmis and V. K. Prasanna. Predicting Communication Intention in Social Networks. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 184–194. IEEE, 2012.

J. Cheng, D. M. Romero, B. Meeder, and J. Kleinberg. Predicting reciprocity in social networks. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 49–56. IEEE, 2011.

A. Chin, B. Xu, H. Wang, and X. Wang. Linking people through physical proximity in a conference. In *Proceedings of the 3rd international workshop on Modeling social media*, pages 13–20. ACM, 2012a.

A. Chin, B. Xu, F. Yin, X. Wang, W. Wang, X. Fan, D. Hong, and Y. Wang. Using proximity and homophily to connect conference attendees in a mobile social network. In *Distributed Computing Systems Workshops (ICDCSW), 2012 32nd International Conference on*, pages 79–87. IEEE, 2012b.

E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM, 2011.

J. Choi, S. Heo, J. Han, G. Lee, and J. Song. Mining social relationship types in an organization using communication patterns. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 295–302. ACM, 2013.

H. Chun, H. Kwak, Y.-H. Eom, Y.-Y. Ahn, S. Moon, and H. Jeong. Comparison of online social relations in volume vs interaction: a case study of cyworld. In *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement*, pages 57–70, New York, NY, USA, 2008. ACM.

S. Consolvo, I. E. Smith, T. Matthews, A. LaMarca, J. Tabert, and P. Powledge. Location disclosure to social relations: why, when, & what people want to share. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 81–90. ACM, 2005.

J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 119–128. ACM, 2010.

W. Dong, B. Lepri, and A. S. Pentland. Modeling the co-evolution of behaviors and social relationships using mobile phone data. In *Proceedings of the 10th International Conference on Mobile and Ubiquitous Multimedia*, pages 134–143. ACM, 2011.

N. Eagle, A. S. Pentland, and D. Lazer. Inferring Social Network Structure using Mobile Phone Data. *Proceedings of the National Academy of Sciences*, 106(36):15274–15278, Sept. 2009.

N. B. Ellison, C. Steinfield, and C. Lampe. The Benefits of Facebook " Friends:" Social Capital and College Students' Use of Online Social Network Sites. *Journal of Computer-Mediated Communication*, 12(4):1143–1168, 2007.

M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici. Link prediction in social networks using computationally efficient topological features. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 73–80. IEEE, 2011.

S. J. Fusco, K. Michael, M. Michael, and R. Abbas. Exploring the Social Implications of Location Based Social Networking: An inquiry into the perceived positive and negative impacts of using LBSN between friends. In *Mobile Business and 2010 Ninth Global Mobility Roundtable (ICMB-GMR), 2010 Ninth International Conference on*, pages 230–237. IEEE, 2010.

S. J. Fusco, K. Michael, A. Aloudat, and R. Abbas. Monitoring people using location-based social networking and its negative impact on trust: an exploratory contextual analysis of five types of "friend" relationships. 2011.

H. Gao, J. Tang, and H. Liu. Exploring social-historical ties on location-based social networks. In *Proceedings of the 6th International AAAI Conference on Weblogs and Social Media*, 2012.

F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *The VLDB Journal*, 20(5):695–719, 2011.

E. Gilbert. Predicting tie strength in a new medium. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1047–1056. ACM, 2012.

E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 211–220. ACM, 2009.

J. Golbeck, C. Robles, M. Edmondson, and K. Turner. Predicting personality from twitter. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 149–156. IEEE, 2011.

S. A. Golder and S. Yardi. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 88–95. IEEE, 2010.

M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

M. S. Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

I. Guy, M. Jacovi, A. Perer, I. Ronen, and E. Uziel. Same places, same things, same people?: mining user similarity on social media. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 41–50, New York, NY, USA, 2010. ACM.

S. Han, D. He, P. Brusilovsky, and Z. Yue. Coauthor Prediction for Junior Researchers. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 274–283. Springer, 2013.

V. Hegde, J. X. Parreira, and M. Hauswirth. Semantic tagging of places based on user interest profiles from online social networks. pages 218–229, 2013.

M. U. Iqbal and S. Lim. Privacy implications of automated GPS tracking and profiling. *Technology and Society Magazine, IEEE*, 29(2):39–46, 2010.

J. J. Jones, J. E. Settle, R. M. Bond, C. J. Fariss, C. Marlow, and J. H. Fowler. Inferring tie strength from online directed behavior. *PloS one*, 8(1):e52168, 2013.

K. Joseph, C. H. Tan, and K. M. Carley. Beyond local, categories and friends: clustering foursquare users with latent topics. In *Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, pages 919–926. ACM, 2012.

I. Kahanda and J. Neville. Using Transactional Information to Predict Link Strength in Online Social Networks. In *ICWSM*, 2009.

V. Kisekka, S. Bagchi-Sen, and H. Raghav Rao. Extent of private information disclosure on online social networks: An exploration of Facebook mobile phone users. *"Computers in Human Behavior"*, 29(6):2722–2729, 2013.

C. Lampe, N. Ellison, and C. Steinfield. A face(book) in the crowd: social Searching vs. social browsing. In *Conference on Computer Supported Cooperative Work*, pages 167–170, 2006.

C. Lampe, N. B. Ellison, and C. Steinfield. Changes in use and perception of facebook. In *Conference on Computer Supported Cooperative Work*, pages 721–730, 2008.

H. W. Lauw, E.-P. Lim, H. Pang, and T.-T. Tan. Social network discovery by mining spatio-temporal events. *Computational & Mathematical Organization Theory*, 11(2):97–118, 2005.

D. H. Lee and P. Brusilovsky. Social networks and interest similarity: the case of CiteULike. In *Proceedings of the 21st ACM conference on Hypertext and hypermedia*, pages 151–156. ACM, 2010.

K. Lerman, S. Intagorn, J. Kang, and R. Ghosh. Using Proximity to Predict Activity in Social Networks. CoRR, Vol. abs/1112.2755. Oct. 2012.

J. Leskovec. Stanford network analysis project (snap). 2012.

J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650. ACM, 2010.

K. Lewis, J. Kaufman, M. Gonzalez, A. Wimmer, and N. Christakis. Tastes, ties, and time: A new social network dataset using Facebook. com. *Social networks*, 30(4):330–342, 2008.

N. Li and G. Chen. Analysis of a Location-Based Social Network. In *IEEE International Conference on Computational Science and Engineering*, pages 263–270, 2009.

N. Li and G. Chen. Sharing location in online social networks. *Network, IEEE*, 24(5):20–25, 2010.

Q. Li, Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. Mining user similarity based on location history. In *Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems*, page 34. ACM, 2008.

R.-H. Li, J. Liu, J. X. Yu, H. Chen, and H. Kitagawa. Co-occurrence prediction in a large location-based social network. *Frontiers of Computer Science*, 7(2):185–194, 2013.

D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2002.

J. Lindqvist, J. Cranshaw, J. Wiese, J. Hong, and J. Zimmerman. I'm the mayor of my house: examining why people use foursquare-a social-driven location sharing application. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2409–2418. ACM, 2011.

X. Liu, Q. He, Y. Tian, W.-C. Lee, J. McPherson, and J. Han. Event-based social networks: linking the online and offline social worlds. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1032–1040. ACM, 2012.

M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

K. Michael, A. McNamee, M. G. Michael, and H. Tootell. Location-based intelligence-Modeling behavior in humans using GPS. In *Technology and Society, 2006. ISTAS 2006. IEEE International Symposium on*, pages 1–8. IEEE, 2006.

A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM, 2010.

J. C. Mitchell. *Social networks in urban situations: Analysis of personal relationships in central African towns*. Buy this book, 1969.

D. Mok and B. Wellman. Did distance matter before the Internet?: Interpersonal contact and support in the 1970s. *Social networks*, 29(3):430–461, 2007.

A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining User Mobility Features for Next Place Prediction in Location-based Services. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1038–1043. IEEE, 2012a.

A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. A random walk around the city: New venue recommendation in location-based social networks. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 144–153. IEEE, 2012b.

J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

S. J. Pan, D. J. Boston, and C. Borcea. Analysis of fusing online and co-presence social networks. In *IEEE International Conference on Pervasive Computing and Communications*, pages 496–501, 2011.

A. Papadimitriou, P. Symeonidis, and Y. Manolopoulos. Friendlink: Link prediction in social networks via bounded local path traversal. In *Computational Aspects of Social Networks (CASoN), 2011 International Conference on*, pages 66–71. IEEE, 2011.

L. Pappalardo, G. Rossetti, and D. Pedreschi. "How Well Do We Know Each Other?" Detecting Tie Strength in Multidimensional Social Networks. In *ASONAM*, pages 1040–1045, 2012.

A. Popescu, G. Grefenstette, et al. Mining user home location and gender from Flickr tags. *ICWSM'10*, 2010.

D. M. Romero and J. M. Kleinberg. The directed closure process in hybrid social-information networks, with an analysis of link formation on twitter. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pages 138–145, 2010.

D. M. Romero, C. Tan, and J. Ugander. Social-Topical Affiliations: The Interplay between Structure and Popularity. *arXiv.org*, Dec. 2011.

M. Rowe, M. Stankovic, and H. Alani. Who will follow whom? Exploiting semantics for link prediction in attention-information networks. 2012.

N. Sadeh, J. Hong, L. Cranor, I. Fette, P. Kelley, M. Prabaker, and J. Rao. Understanding and capturing people's privacy policies in a mobile social networking application. *Personal and Ubiquitous Computing*, 13(6):401–412, 2009.

S. Scellato and C. Mascolo. Measuring user activity on an online location-based social network. In *IEEE Conference on Computer Communications Workshops, INFOCOM Wksps*, 2011.

S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *Proceedings of ICWSM*, 11:329–336, 2011a.

S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM, 2011b.

R. Schifanella, A. Barrat, C. Cattuto, B. Markines, and F. Menczer. Folks in folksonomies: social link prediction from shared metadata. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 271–280. ACM, 2010.

C. Scholz, M. Atzmueller, and G. Stumme. On the Predictability of Human Contacts: Influence Factors and the Strength of Stronger Ties. In *Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, pages 312–321. IEEE, 2012.

N. Shibata, Y. Kajikawa, and I. Sakata. Link prediction in citation networks. *Journal of the American society for information science and technology*, 63(1):78–85, 2012.

C. Song, Z. Qu, N. Blumm, and A.-L. Barabasi. Limits of predictability in human mobility. *Science*, 327(5968):1018–1021, Feb. 2010.

D. Traynor and K. Curran. Location-Based Social Networks. *From Government to E-Governance: Public Administration in the Digital Age*, page 243, 2012.

J. Tsai, P. Kelley, L. Cranor, and N. Sadeh. Location-sharing technologies: Privacy risks and controls. TPRC, 2009.

J. Ugander, B. Karrer, L. Backstrom, and C. Marlow. The anatomy of the facebook social graph. *arXiv preprint arXiv:1111.4503*, 2011.

B. Viswanath, A. Mislove, M. Cha, and K. P. Gummadi. On the Evolution of User Interaction in Facebook. In *Proceedings of the 2nd ACM SIGCOMM Workshop on Social Networks (WOSN'09)*, Aug. 2009.

Y. Volkovich, S. Scellato, D. Laniado, C. Mascolo, and A. Kaltenbrunner. The length of bridge ties: structural and geographic properties of online social interactions. *Proceedings of ICWSM*, 12, 2012.

D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108, New York, NY, USA, 2011. ACM.

H. Wang, L. Zhu, and A. Chin. An Indoor Location-Based Social Network for Managing Office Resource and Connecting People. In *Autonomic and Trusted Computing*, 2010.

C. Wilson, B. Boe, A. Sala, K. P. Puttaswamy, and B. Y. Zhao. User interactions in social networks and their implications. In *Proceedings of the 4th ACM European conference on Computer systems*, pages 205–218. Acm, 2009.

R. E. Wilson, S. D. Gosling, and L. T. Graham. A review of Facebook research in the social sciences. *Perspectives on Psychological Science*, 7(3):203–220, 2012.

X. Xiao, Y. Zheng, X. Xie, and Q. Luo. Inferring Social Ties between Users with Human Location History. 2012.

M. Ye, P. Yin, and W.-C. Lee. Location recommendation for location-based social networks. In *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 458–461. ACM, 2010.

D. Yin, L. Hong, and B. D. Davison. Structural link analysis and prediction in microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1163–1168. ACM, 2011a.

D. Yin, L. Hong, X. Xiong, and B. D. Davison. Link formation analysis in microblogs. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*, pages 1235–1236. ACM, 2011b.

W. W. Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research*, pages 452–473, 1977.

C. Zhang, L. Shou, K. Chen, G. Chen, and Y. Bei. Evaluating geo-social influence in location-based social networks. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1442–1451. ACM, 2012.

E. Zheleva, L. Getoor, J. Golbeck, and U. Kuter. Using friendship ties and family circles for link prediction. In *Advances in Social Network Mining and Analysis*, pages 97–113. Springer, 2010.

D. Zhou, B. Wang, S. M. Rahimi, and X. Wang. A study of recommending locations on location-based social network by collaborative filtering. pages 255–266, 2012.

T. Zhou, L. Lü, and Y.-C. Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630, 2009.

CHAPTER 3

Success of Events in Second Life

Parts of the contents of this chapter have been published in the paper *Success Factors of Events in Virtual Worlds A Case Study in Second Life* presented at the *Workshop on Network and Systems Support for Games* [Steurer et al., 2012].

In this chapter we describe the terminology of the virtual world of Second Life and the process to collect in-world data with robots. We exploit publicly available events in the virtual world and send autonomous robots to the event locations that collect position data of other residents. Based on the collected meta information of the events, e.g. categories, maturity rating or duration, we predict whether a future event will be successful or not.

The remainder of this chapter is organized as follows. Section 3.1 gives a general introduction to the topic and in Section 3.2 we discuss the related work in this area. In Section 3.3 we present details about the retrieval methods to collect data from a public event calendar with a web crawler and an in-world robot that moves through the virtual environment to harvest the location data of avatars. In Section 3.4 we describe the collected dataset and in Section 3.5 we present the set up and the results of the prediction experiments. Finally, Section 3.6 concludes the chapter and gives a perspective for future work.

Abstract

In this paper we present the results of a study that aims to analyse publicly announced event data in the virtual world of Second Life with the goal to predict whether or not an event will be successful by terms of increasing the average traffic of a region. To that end, we collected data from a publicly accessible event calender in Second Life and the in-world position data of avatars visiting these events. Based on the statistical analysis of features such as event category, duration or maturity rating, provided by the Second Life event calendar, we built a simple predictive model that can decide upon the success of an event with an accuracy of over 92%.

3.1. Introduction

Over the past 10 years social networks evolved and users spend more and more time with their online friends. The most popular social network *Facebook* has approximately 800 Million users by now and would be, if compared to residents of the world's countries, the third largest country in the whole world. Users have different ideas why they are using this new media, starting from being in contact with their friends to playing online games. Virtual worlds provide the same features as these networks do but enrich the user experience with a three dimensional representation of their environment. In contrast to online role-playing games, virtual worlds are not games or quest-oriented per se but focus on user interaction and creativity.

The virtual land of Second Life is divided into fixed-size quadratic regions with 65,536 square meters each that can be bought by users for real money. In the second quarter of 2011, the overall land size was over 2000 km^2 which equals about 30,500 regions [Linden Lab, 2011]. With an average of 64,000 online avatars in an area of 2000 km^2 (31.25 avatars per km^2) the main problem for new users is to find other avatars for interaction [Voyager, 2011]. One approach to meet new people are events, activities or parties carried out in a particular location in the virtual world. Just as in real life, residents of Second Life can host events and announce location and time of the event on a public event calender. Further they can add a description, set a category, e.g. *Education* or *Music*, and rate the event according to their maturity level. According to Figure 3.1 the number of events is approximately 900 per day and hence it is hard for new users to find events that fit their interests where they can meet new people.

In this paper we focus on these publicly accessible events hosted by residents of Second Life and investigate in their influence on the avatar traffic. We collected event information and combined it with the position information of avatars prior, during and after the event. Overall we have harvested approximately 80,000 events over a period of three months and about 110 Million data samples of avatars position information.

With a statistical analysis of the combined data we can answer questions about the usefulness of events to increase the average avatar traffic. Based on these results we could also find features and success factors for region owners to make their regions more popular. Further, we try to answer the question on the predictability of the success of an event to give both region owners and residents an estimator for future events.

3.2. Related Work

With the advent of mobile phones and location based services spatial data brings innovative perspectives to analyse users movements. Chen and Roy [2009] presented an approach to detect photos of events by using spatial and temporal information. One step further in this direction is to link position data to predict links between users. [Cranshaw et al., 2010] employed mobility patterns of users and the structure of their social network to predict friendship by comparing their location trajectories. Location based analysis can be even extended to an accurate prediction of users behaviour to predict future movements. Wang et al. [2011] showed that mobile phone data can be employed to predict the future movements of users with a probability of 92%. Although these papers all show the value of position data it is nearly impossible to collect real world position data of users on a large scale.

Due to their closeness to the real-world virtual environments have become a valuable field of research for different areas [Bainbridge, 2007; Hendaoui et al., 2008], starting from e-learning [Hefley et al., 2012], to a playground for real-word innovations [Kohler et al., 2009] and economy [Eisenbeiss et al., 2012]. In virtual worlds researchers are able to conduct experiments that are typically hard to perform in the real-world because of a lack of data. To overcome the problem of data collection, literature suggests several approaches to detect users in-world behaviour. The most simple solution is to place scripted sensors into the virtual world, monitor the surrounding users and send this information to a database [Kappe et al., 2009; Metaverse Business, 2011; Yee and Bailenson, 2008]. This approach turned out as not very sufficient because it relies on the owner of the land to place sensors and detect users. Further, in-world sensors can only detect at most 16 avatars concurrently within a range of 96 meters [La and Michiardi, 2008]. Other approaches intercept the communication protocol between the client software and the Second Life servers to do network traffic analysis and monitor avatars and objects [Varvello et al., 2008; Zhang et al., 2010]. For example Cranefield and Li [2009] integrated a tool into the Second Life client to monitor social expectations of other avatars.

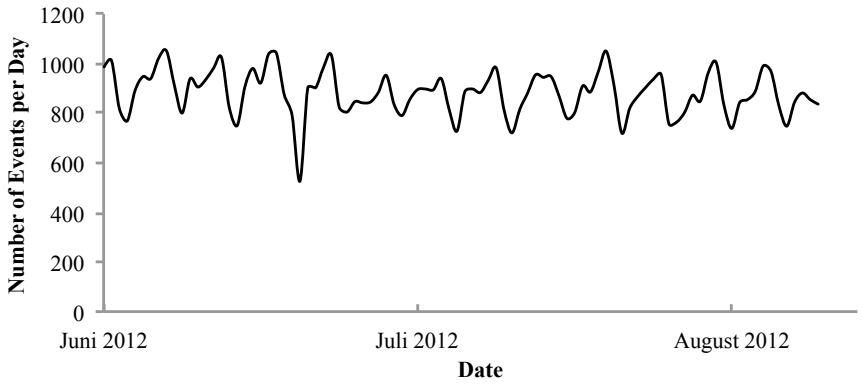


Figure 3.1.: Number daily events over a period of three months in Second Life.

3.3. Data Collection

3.3.1. Web Crawler

Linden Lab allows users to create events and announce them to the public* (see Figure 3.2). To do so, a user must log into its Second Life account on the Web page, specify the name of the event, a description and the actual location with date and time. Events can be assigned to predefined categories and rated according to their maturity level which is "general", "mature" or "adult". Every registered Second Life user can host events and hence the number of available events is very large no matter how many people are interested in the actual activity. Due to maturity restrictions users that are not logged in or did not yet confirm their age can only search for events with a G-rated maturity level. To access all events Linden Lab requires users to confirm that they are at least 18 years old by logging into the Web service and explicitly confirm their age. To actually find events, users then enter a query that matches their interests into a Web form and refine this query by selecting categories and the actual date of the activity (or search for currently on-going events). Depending on this settings, the user is provided with a list of all events that match the specified query. A query without any keywords or specified categories returns all available events ordered by start time.

To automatically harvest events from the Web page we have implemented a Web-crawler that runs daily to fetch all events from the public calendar. The procedure is straightforward because we imitated the browser's behaviour and parsed the required data from the HTML response. First, we registered a new avatar with Second Life and manually confirmed that the user is at least 18 years of age in the avatars profile to access all available events. To reverse engineer the communication protocol between web browser and server we employed the standard developer tools provided by

*https://secondlife.com/my/community/events/index.php

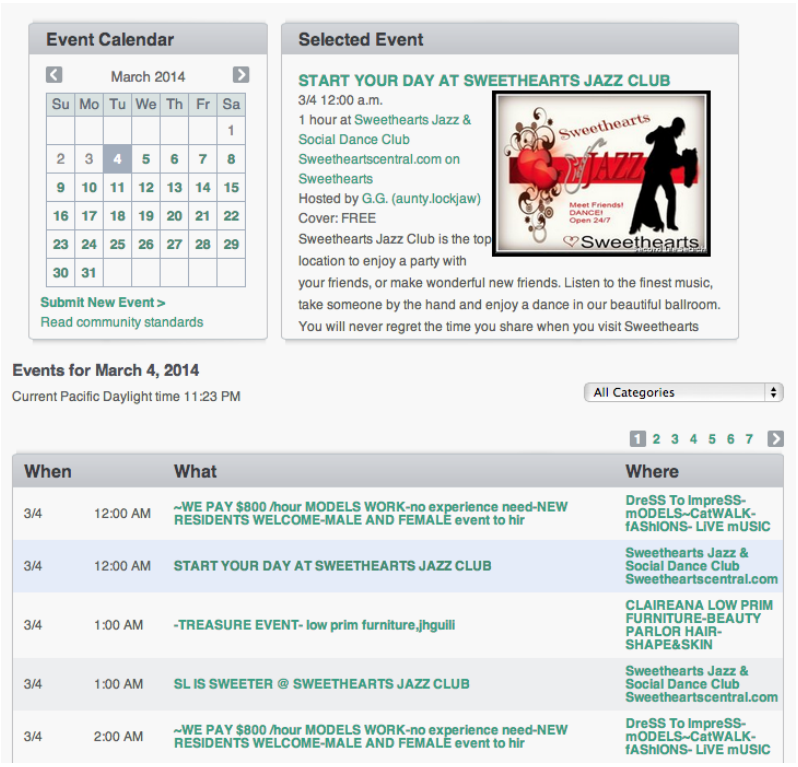


Figure 3.2.: Residents of Second Life can host events and announce them publicly accessible on a Web page.

Google's Chrome Web browser and recorded the HTTP requests and responses while manually logging into Second Life's Web page. After transmitting the user credentials the Web page replied with a session cookie which is necessary for all further interactions to prove the identity of the logged in user. After reverse engineering the Web requests to fetch all the event data with an empty query string (to get all available events), we were provided with an HTML page that contained the desired information. We parsed the Document Object Model tree of the response and extracted the required event information from the raw HTML data. This data was then stored in a database for further processing and evaluation. Unfortunately, the Web page responded only with 20 events and so all further events could only be accessed by sending AJAX requests. Google's developer tools again reveal the structure of these requests. To disguise the automated requests and to prevent from being banned by Linden Lab we added some fuzziness to the requests. We tried to imitate human behaviour by varying the interval between the requests and so it took about one hour to get the entire event data for a single day.

During a three-month's period we were able to extract about 84,000 events from the Web page which yields in about 935 events per day. Figure 3.1 depicts the distribution of events per day over the period of three months. The dataset contains information about the event with title and

description, the actual location of the event with region name and accurate coordinates, the name of the event's host, the time when the event starts and the duration, the category of the event, the required maturity and a potential admission fee.

Besides Second Life's public event calender, the described data collection approach can also be used to harvest data from other Web-resources of Second Life. Among others this includes the social network "My Second Life", a Facebook-like social network platform that aims at Second Life users. This platform allows users to interact with each other with postings, comments and loves, share their favourite locations, and specify their interests and potential partners. The log in process and the collection process with HTML requests are similar but only the raw HTML response is different. As a consequence, we only have to adopt the data parsers and the data model in the database to fetch other resources.

3.3.2. In-World Robots

To harvest the position data of residents of Second Life we have implemented two different approaches: First, a general approach that allows us to collect position data from a vast amount of regions. The drawback of this solution is that we were only able to collect information about the number of residents in regions and the rough coordinates of the present avatars but we could not identify the residents. In contrast, in the second approach we were able to determine the unique identity of the residents and their accurate location. Unfortunately, the computational resources for this collection process were higher and hence the collected amount of data was smaller. Figure 3.5 shows a typical region with avatars indicated as green dots: With the first approach we could identify the number of green dots and their approximate location with low computational resources, whereas we could identify the identity of the green dots and their accurate locations with high computational resources.

For both approaches we employed *LibOpenMetaverse*[†] which is a command line client to connect to Second Life. Basically the client has the same capabilities as the official Second Life viewer has but does not provide a graphical user interface. All interactions that are typically done by using mouse or keyboard are text-commands which makes it easier for developers to control the client and automate its behaviour [Zhang et al., 2010]. The needed computational resources to run a single instance of the client were very low and so several clients could be run concurrently on a standard Linux server.

The used clients should act completely autonomous and so it needed a few additional capabilities if compared to the original Second Life viewer to manage different tasks. The basic abilities

[†] `http://lib.openmetaverse.org`

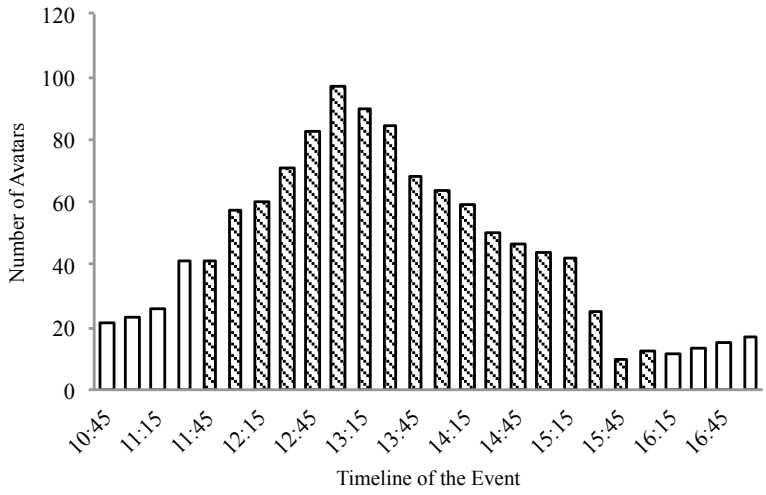


Figure 3.3.: Average number of avatars prior, during (indicated by the hatched bars) and after an event in Second Life.

for the bots were autonomous log in and the imitation of human behaviour to disguise the bot character:

- Autonomous login. Once the client was started, it requested a database for user credentials to log into Second Life. The database stored all the manually registered accounts which allowed a flexible management of user data as new users could be added easily. While logged in, the server disrupted the connection to the clients from time to time, e.g. restart of the region server, and automatically logged out the connected clients. If so, the client autonomously reconnected to the Second Life after a certain amount of time. In case of such a connection reset the client randomly selected an new region to log in again. This functionality was implemented with a Linux Cron job (a time-based job scheduler) that periodically checked for all running clients. If a client was not connected any more it tried to restart until it was connected again.
- Imitate human behaviour. After the log in process, the avatar did not move and stayed at the same place at the same location. In general region owners do not want any autonomous avatars in their regions and ban them as soon as they were detected. To avoid this banning, the bot imitated human behaviour, walked around in the virtual world and even teleported to arbitrary locations within the region.

In the next two sections we describe the collection process of the two approaches to collect data. With the first approach we could collect the location data of anonymous users with low computational costs whereas in the second approach we could collect the identity and the accurate location data of users with higher computational costs.

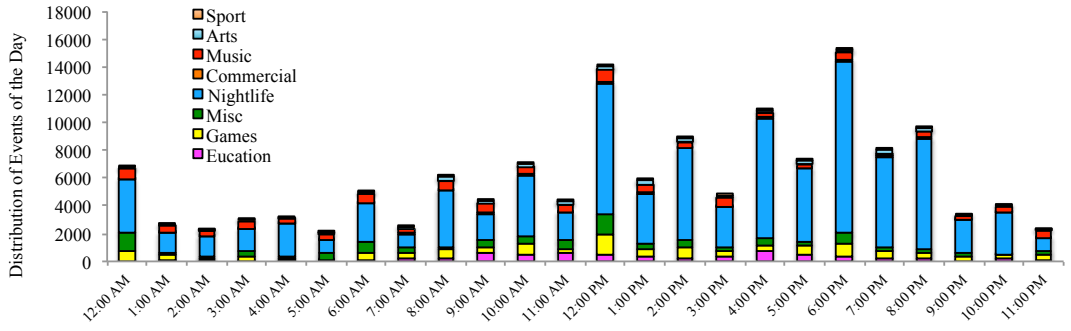


Figure 3.4.: Distribution of event durations for different event categories.

Harvest rough position data. The developed bot monitored regions without actually visiting these regions but only sent a request to each region to get the coordinates of all present avatars instead. The received data was parsed by the bot and then sent to a database where it is stored persistently for further processing. Unfortunately, the response did not contain the actual names of the avatars but with this approach we were able to fetch data from a large amount of regions. To reduce the number of requests and to reduce load on Linden Lab's server we created a simple metric that uses the number of currently online avatars for the interval between two requests. Details of this metric can be found in Table 3.1.

Harvest detailed position data. The second approach to collect the position information of avatars did not only harvest the location of avatars and the time stamp but also the unique identities of the present avatars. In contrast to the previous approach, it is required to visit the actual region in order to get all the information. The bot requested the database with on-going events for the next location to be visited and then teleported to this region. As soon at it arrived there, it fetched the information about surrounding avatars and their accurate locations and sent it to a database for further processing. Then it requested the database for the next location to teleport to and harvested information about avatars again. On average this took around one minute and so the harvesting capabilities of a single bot were very limited. As a consequence we employed a pool of bots (the user credentials were stored in a database) that simultaneously and independently visit regions, fetch the required information and move on to the next region. The regions to be visited were stored in a queue-like data structure obtained from a database. The actual frequency of region visits depended on the actual number of regions to be visited and the number of bots that harvest this data.

# Avatars in Region	Request Interval
0	45 min
$1-5$	15 min
$6-10$	10 min
>10	5 min

Table 3.1.: The update frequency for a rough estimation of the number of avatars in region depends on the actual number of present avatars in the actual region.

3.4. Dataset Description

In this section we evaluate the data collected event data and the avatars position data collect with the bots.

3.4.1. Event Data

We have collected data from Second Life's event calendar over a period of three months and the dataset contained 84,234 events hosted in 2,756 different regions. We could identify the most common words in the events titles as "DJ", "club", "rock", "party" and "music". Events can be filtered from 10 different categories with five categories containing 94% of all events: Nightlife/Entertainment 47%, Live Music 30%, and Commercial, Games and Contests 7% each. The duration of events varied from 10 minutes to 720 minutes with 32% lasting less or equal 60 minutes, 45% lasting between 61 minutes and 120 minutes, and 10% of all events lasting 720 minutes. The maturity of events depended on the location where they were hosted. Regions respectively events, can be rated as *General* with no age verification, *Mature* with the restriction to be at least 16 years old and finally *Adult* with no admittance below 18 years of age. The eventset contained 12.1% events rated as general accessible, 75.4% of the events rated as mature and 12.5% of all events are rated as adult. Figure 3.4 shows the distribution of events over a day with two peaks at noon and 6 pm. Hosts could demand for an entrance fee for avatars to attend an event but it was not widely used as only 385 events demanded for money ranging from 1 to 15,000 Linden Dollars. The collected events were created by 3,900 different avatars which is on average 22 events per avatar. A more detailed analysis showed that 3250 avatars created events only in one single region, 435 created events in two different regions, and 138 avatar created events in three different regions. Only 22 avatars created events in more than 5 different regions.

Due to the high number of events, we have limited the dataset of the events for all further computations according to the following considerations: we did not consider events with a duration of more than 4 hours because region owners create events up to the maximum of 8 hours just to be

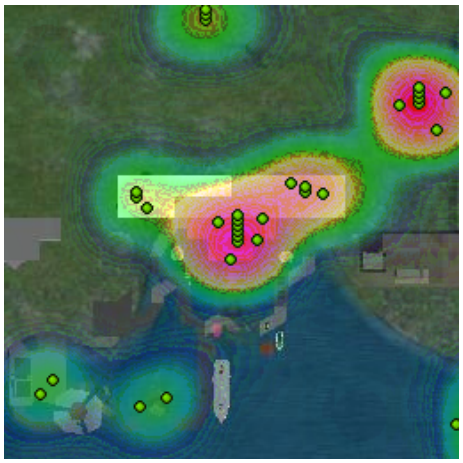


Figure 3.5.: A detail of Second Life's map with green dots representing avatars and a computed heatmap to indicate areas with high traffic.

visible in the list of on-going events on Linden's Web-page. Similar to the event duration we omitted recurring events (an event starts right after another event with the same name ends) because region owners create succeeding events only to be present in the list of on-going events.

3.4.2. Position Data

For the experiments in this paper we have collected 110 Million data samples in over 21,000 regions that contain the number of present avatars in a certain region and the time stamp. The number of avatars was not even distributed because most of the regions had less than 5 concurrently present avatars on average. In particular we identified around 3,000 regions with an average traffic of less than 1 avatar per day, 16,000 regions with an average traffic between 1 and 5 avatars per day and 2,000 regions with an average traffic between 6 and 10 avatars per day. The remaining regions had an average avatar traffic of more than 10 per day. Figure 3.5 shows a region where the green dots indicate avatars present in the regions. The heat map visualizes the most attractive locations within the region.

3.5. Experiments and Results

In order to predict the success of public events we combined the event data with the location data collected in the virtual world. We extracted basic information like location, start time and duration from all events and matched every event with the number of present avatars. To see the effect of a specific event on the number of concurrent online avatars we used this information of participants

from one hour and 15 minutes before the event started until one hour and 15 minutes after the event ended. Using this data we computed the average traffic as the average number of online avatars for every event prior, during and after the incident separately.

In this section we present the results of experiments we have conducted on the combined dataset of avatar traffic information prior, during and after the event. First, we provide an descriptive-statistic overview of events and their implications on the avatar traffic in a region. We divide the events into different time slots and show the effects of avatar traffic. Second, we present the results of a predictive model to forecast whether an event increases the traffic in a region or not.

3.5.1. Descriptive Statistics

To get a rough overview of the avatar traffic change during an event we compared the average avatar traffic during an event with the average avatar traffic one hour prior and past the event. On average we observed over 14 avatars per region prior an event (M = 14.07, SE = 13.85), over 19 avatars per region during an event (M = 19.67, SE = 14.16) and finally around 16 avatars per region after the event ended (M = 16.29, SE = 13.32). This is an increase of $+33.84\%$ avatars per region during an event if compared to the time interval prior the event. The significance of these differences was shown with a paired students t-test: $t(82,951) = -51.13,\ p \leq .01$. After the event ended, the average number of avatars dropped -16.9%. Again we computed the significance of this drop-off with a paired students t-test: $t(82,951) = 35.42,\ p \leq .01$.

For a more detailed analysis we split the events according to their categories, duration, and maturity to see the implications on the avatar traffic. The baseline for all measures is the average of the overall events with an increase of +33.84% of avatar traffic. Figure 3.6 shows the average increase of the avatar traffic for different categories: *Arts* and *Discussion* perform best whereas in contrast the most frequent categories *Music* and *Nightlife* were on average. Figure 3.7 depicts the average increase of avatar traffic for different durations of events if compared to the overall avatar traffic. It can be seen that there is no relation between the duration of events and the average gain. Figure 3.8 depicts the average increase of avatar traffic for the maturity rating of events. One can see that adult-rated events cause a higher average traffic boost than mature- or general rated events. We omitted the representation of traffic change for different weekdays as it did not influence the traffic during events at all.

3.5.2. Success of Events

During our analysis on the dataset we noticed that not all events had a positive effect on the avatar traffic for a region. In particular we saw that 40.6% of all events had a positive effect on the traffic

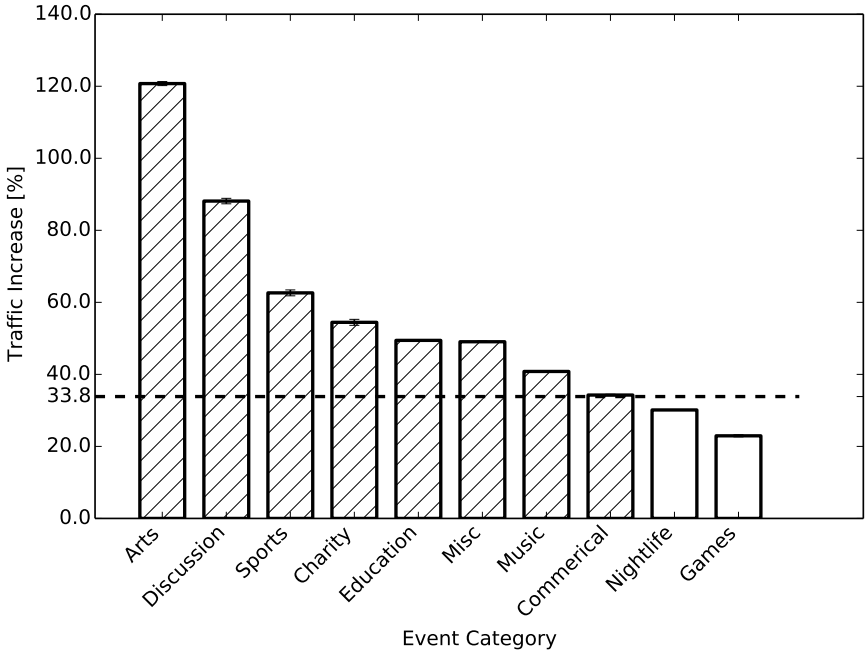


Figure 3.6.: The average traffic increase during an event segmented into the different event categories. Hatched bars indicate an increase above the overall average of 33.84%.

of a region and 59.4% of all events had a negative effect.

Based on this observation the question arises if it is possible to build a model (= train a classifier) to correctly identify good (= events that increase the traffic in a regions) and bad events (= events that decrease the traffic in a region). More formally, given a list of events as input samples for our model $I = \{e_1, e_2, \ldots, e_n\}$, we want to learn the function $f : I \rightarrow D$ which maps each event e_i correctly to a corresponding class $D = \{good, bad\}$. Due to the binary classification the baseline for these experiments is 50% (or 0.5 AUC).

In machine learning this is seen as a supervised learning setting with a range of possible learning algorithms. To find the best learning algorithm we conducted the experiments using for instance Support Vector Machines (SVM), Logistic Regression or Stochastic Gradient Descent and compared the average F1-scores and AUC values with each other. We could finally identify a *Naïve Bayes* classifier as most valuable and hence it was used for our final prediction model and the present results in this paper. To analyse the performance of the used classifiers and validate the found results we chose a 10-fold cross validation.

In Table 3.2, we show the performance of our model based on different features. The features *Maturity* (= maturity rating of an event), *Duration*, *Weekday* (= the day of the week an event takes

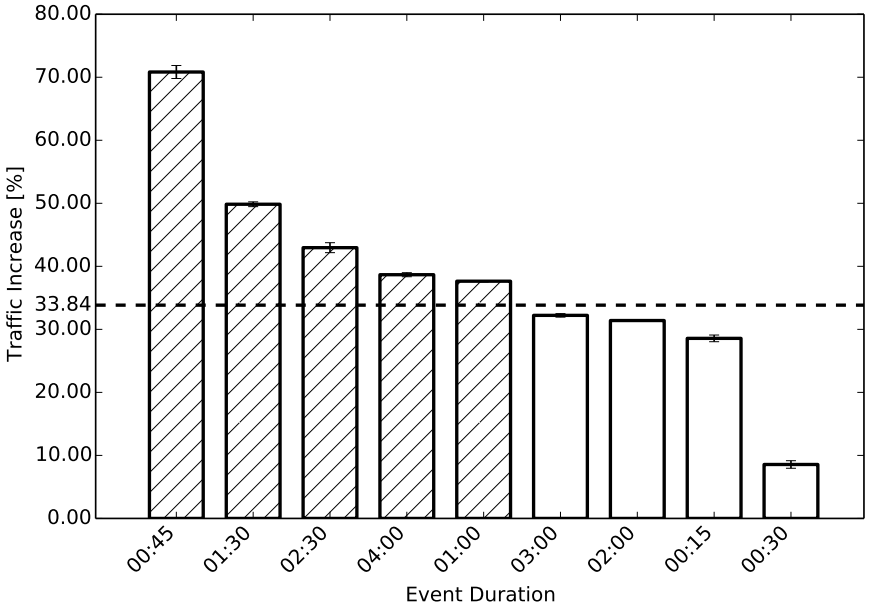


Figure 3.7.: The average traffic increase during an event segmented into the different event durations. Hatched bars indicate an increase above the overall average of 33.84%.

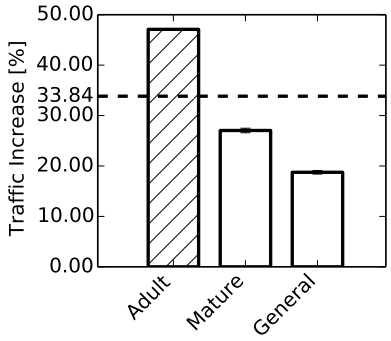


Figure 3.8.: The average traffic increase during an event segmented into different maturity ratings. Hatched bars indicate an increase above the overall average of 33.84%.

place) and *Start* (= the time of a day a event takes place) *alone* have very low classification power, i.e. the *AUC* values are close to the baseline. The feature *Category* showed the best results with 0.6 AUC. However, if we combine all these features (= Combined), we can see that we significantly outperform the baseline, which means that we classify an event as good or bad in 83.4% of the cases correctly if we look at the AUC value.

Feature	Precision	Recall	F1	AUC
Baseline	0.353	0.594	0.442	0.500
Category	0.618	0.627	0.619	0.601
Maturity	0.353	0.594	0.442	0.540
Duration	0.380	0.585	0.439	0.550
Weekday	0.353	0.594	0.442	0.521
Start	0.562	0.593	0.518	0.599
Combined	**0.766**	**0.765**	**0.766**	**0.834**
Prior Avatars	0.860	0.853	0.849	0.921
Region	0.758	0.760	0.758	0.835
Host	0.730	0.729	0.717	0.791
All	**0.846**	**0.845**	**0.845**	**0.929**

Table 3.2.: Results of the event prediction experiment using our best performing Naive Bayes classifier.

Apart from the standard features, we also checked the features *Host* (= name of the host), *Prior Avatars* (= maximum number of avatars one hour before an event takes place) and Region (= the region where event is hosted). As shown in Table 3.2, the highest classification power can be archived with the number of avatars prior an event. Interestingly, if we combine all features of the table (= All) we could predict 92.9% of all events correctly using this model.

3.6. Conclusion and Outlook

In this research paper we focused on public accessible events hosted by residents of a virtual world and their influence on the traffic of a region. For that purpose, we collected event information of the Second Life event calender and combined it with the position information of avatars prior, during and after an event. With a statistical analysis of the combined data we could answer questions about the usefulness of events to increase the traffic of a region and could extended this to find features and success factors for region owners to make their regions more popular. Further, we introduced a predictive model that classifies “good” and “bad” events prior to an event with a accuracy of over 92%. Among several features we have identified the number of avatars that are present prior an event as the most valuable parameter to predict the success of an event: present avatars attract more visitors.

References

W. S. Bainbridge. The Scientific Research Potential of Virtual Worlds. *Science*, 317(5837): 472–476, July 2007. doi: 10.1126/science.1146930. URL http://dx.doi.org/10.1126/science.1146930.

L. Chen and A. Roy. Event detection from flickr data through wavelet-based spatial analysis. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 523–532, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646021. URL http://doi.acm.org/10.1145/1645953.1646021.

S. Cranefield and G. Li. Monitoring social expectations in second life. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2*, AAMAS '09, pages 1303–1304, Richland, SC, 2009. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 978-0-9817381-7-8. URL http://dl.acm.org/citation.cfm?id=1558109.1558264.

J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Ubicomp '10, pages 119–128, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-843-8. doi: 10.1145/1864349.1864380. URL http://doi.acm.org/10.1145/1864349.1864380.

M. Eisenbeiss, B. Blechschmidt, K. Backhaus, and P. A. Freund. "the (real) world is not enough:" motivational drivers and user behavior in virtual worlds. *Journal of Interactive Marketing*, 26 (1):4 – 20, 2012. ISSN 1094-9968. doi: 10.1016/j.intmar.2011.06.002.

B. Hefley, W. Murphy, L. Macaulay, K. Keeling, D. Keeling, C. Mitchell, and Y. L. Tan. Using virtual world technology to deliver educational services. In L. A. Macaulay, I. Miles, L. Zhao, J. Wilby, Y. L. Tan, and B. Theodoulidis, editors, *Case Studies in Service Innovation*, Service Science: Research and Innovations in the Service Economy, pages 125–128. Springer New York, 2012. ISBN 978-1-4614-1972-3.

A. Hendaoui, M. Limayem, and C. Thompson. 3d social virtual worlds: Research issues and challenges. *Internet Computing, IEEE*, 12(1):88 –92, jan.-feb. 2008. ISSN 1089-7801. doi: 10.1109/MIC.2008.1.

F. Kappe, B. Zaka, and M. Steurer. Automatically detecting points of interest and social networks from tracking positions of avatars in a virtual world. In *Social Network Analysis and Mining, 2009. ASONAM '09. International Conference on Advances in*, pages 89 –94, july 2009. doi: 10.1109/ASONAM.2009.66.

T. Kohler, K. Matzler, and J. Füller. Avatar-based innovation: Using virtual worlds for real-world innovation. *Technovation*, 29(6-7):395–407, June 2009. ISSN 01664972. doi: 10.1016/j.technovation.2008.11.004. URL http://dx.doi.org/10.1016/j.technovation.2008.11.004.

C. A. La and P. Michiardi. Characterizing user mobility in second life. In *WOSP '08: Proceedings of the first workshop on Online social networks*, pages 79–84, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-182-8. doi: 10.1145/1397735.1397753. URL http://dx.doi.org/10.1145/1397735.1397753.

Linden Lab. The second life economy in q3 2011. http://community.secondlife.com/t5/Featured-News/The-Second-Life-Economy-in-Q3-2011/ba-p/1166705, 2011. [accessed on 14-September-2012].

Metaverse Business. Second life metrics. http://www.metaverse-business.com/secondlifemetrics.php, 2011. [accessed on 14-September-2012].

M. Steurer, C. Trattner, and F. Kappe. Success factors of events in virtual worlds a case study in Second Life. In *NetGames*, pages 1–2, 2012.

M. Varvello, F. Picconi, C. Diot, and E. Biersack. Is there life in second life? In *Proceedings of the 2008 ACM CoNEXT Conference*, CoNEXT '08, pages 1:1–1:12, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-210-8. doi: 10.1145/1544012.1544013. URL http://doi.acm.org/10.1145/1544012.1544013.

D. Voyager. Sl user concurrency stays below 70,000 during 2011. http://danielvoyager.wordpress.com/2011/11/30/sl-user-concurrency-stays-below-70-000-during-2011/, 2011. [accessed on 14-September-2012].

D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 1100–1108, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0813-7. doi: 10.1145/2020408.2020581. URL http://doi.acm.org/10.1145/2020408.2020581.

N. Yee and J. N. Bailenson. A method for longitudinal behavioral data collection in second life. *Presence: Teleoper. Virtual Environ.*, 17(6):594–596, Dec. 2008. ISSN 1054-7460. doi: 10.1162/pres.17.6.594. URL http://dx.doi.org/10.1162/pres.17.6.594.

Y. Zhang, X. Yu, Y. Dang, and H. Chen. An integrated framework for avatar data collection from the virtual world: A case study in second life. *Intelligent Systems, IEEE*, PP(99):1, 2010. ISSN 1541-1672. doi: 10.1109/MIS.2010.106.

CHAPTER 4

Prediction of Interactions

This chapter is based on the paper *Predicting Interactions In Online Social Networks: An Experiment in Second Life* published and presented at the *4th International Workshop on Modeling Social Media* and the book chapter *Who will Interact with Whom? A Case-Study in Second Life using Online and Location-based Social Network Features to Predict Interactions between Users* published in *MUSE-MSM Post-Proceedings* [Steurer et al., 2013; Steurer and Trattner, 2013].

In this chapter we employ data from a directed online social network and an undirected location-based social network, model the relations between users by applying features to both networks and finally combine the networks. These features can be broadly divided in topological features that reflect the relation between users with respect to the network structure and homophilic features that indicate the actual similarity or alikeness of a user-pair. We employ these features to first predict interactions between users and then predict whether this interaction is reciprocal or not. In both experiments we determine the most suitable features for the prediction tasks in either networks and a combination of both. Overall the chapter is structured as follows: In Section 4.2, we discuss related work. In Section 4.3 we shortly introduce the dataset used for our experiments. In Section 4.4 we outline the set of features used for our experiments in Section 4.5. Section 4.6 presents the results of our study. Finally, Section 4.7 discusses the findings and concludes the chapter

Abstract

Although considerable amount of work has been conducted recently of how to predict links between users in online social media, studies inducing features from different domain data are rare. In this paper we present the latest results of a project that studies the extent to which interactions – in our case directed and bi-directed message communication – between users in online social networks can be predicted by looking at features obtained from online and location-based social network data. To that end, we conducted a number of experiments on data obtained from the virtual world of Second Life. As our results reveal, location-based social network features outperform online social network features if we try to predict interactions between users. However, if we try to predict whether or not this communication was also reciprocal, we find that online social network features seem to be superior.

4.1. Introduction

As a part of the recent hype on social network research, a high amount of attention and research activity was devoted to the problem of predicting links between users [Liben-Nowell and Kleinberg, 2007], e.g. the issue of forecasting whether or not two users u and v of a given online social network $G\langle V,E\rangle$ will interact with each other in the future. While considerable amount of work has been recently conducted of how to predict links between users in online social media, studies comparing different sources of knowledge are rare.

To contribute to this research, we present in this paper the latest results of a research project that aims to study the extent to which interactions – in our case directed and bi-directed message communications –in online social networks can be predicted inducing features from online social network and location-based social network data. To tackle this issue we trained a binary classifier that learned the relations between users u and v based on a number of features induced from online social network and location-based social network data. For the purpose of our study we furthermore differentiated between two types of feature sets – network topological features and homophilic features [Wang et al., 2011]. Since it is nearly impossible to obtain rich large-scale real-world online social and location-based data, our investigation focused on the virtual world of Second Life, where we could easily find and mine both sources of data. We obtained data from a resource called *My Second Life* which is a large-scale online social network for residents of Second Life. This social network can be compared to Facebook but aims at a different target group: residents of Second Life who interact with each other by sharing text messages, comments, and loves. Additionally, we were able to collect location-based social network data of residents in the virtual world by implementing so-called in-world bots.

Overall, it is our interest to answer the following research questions:

- *RQ1:* To what extent do user pairs – interacting or not interacting with each other – differ based on social proximity features induced from the online social network and the location-based social network?

- *RQ2:* To what extent can we predict interactions between users and reciprocity of these interactions inducing features from both domains?

- *RQ3:* Which feature set (homophilic or topological) is most suitable to predict interactions between users and the reciprocity of these interactions.

To that end, we conducted a number of experiments using statistical methods and supervised learning approaches. As our statistical analysis reveals, there are many significant differences between user pairs with interactions and user pairs without interactions. For instance, users with an interactions on the online social network have a shorter average distance between them in the location-based social network. To predict these interactions with supervised learning, we find that location-based social network features outperform online social network features to a great extent. However, if we try to predict reciprocal message communication between users, online social network features seem to be superior. Finally, we find that there are no clear patterns whether or not homophilic or network topological features perform better to predict interactions or reciprocity between users.

4.2. Related Work

Although considerable amount of work has been recently conducted of how to predict links between users in online social media, studies exploiting different kinds of knowledge sources for the link prediction problem are rare. An example is a study conducted by Cranshaw et al. [2010] where the authors collected location data and Facebook friendship data through a mobile app. Based on a number of experiments they show that the so-called place-entropy features are best suited to predict friendship between users in Facebook. Interestingly and contrary to our study, Cranshaw et al. [2010] only looked at the mobile side, i.e. they did not investigate features induced directly from the social network. Furthermore, they only considered friendship links and did not look at communication links as we do in our study. Another related work in this context are the studies of Guy et al. [2008, 2009, 2010] where the authors investigate the similarity between users exploiting 9 different sources of data classified into three different classes: *people*, *things*, and

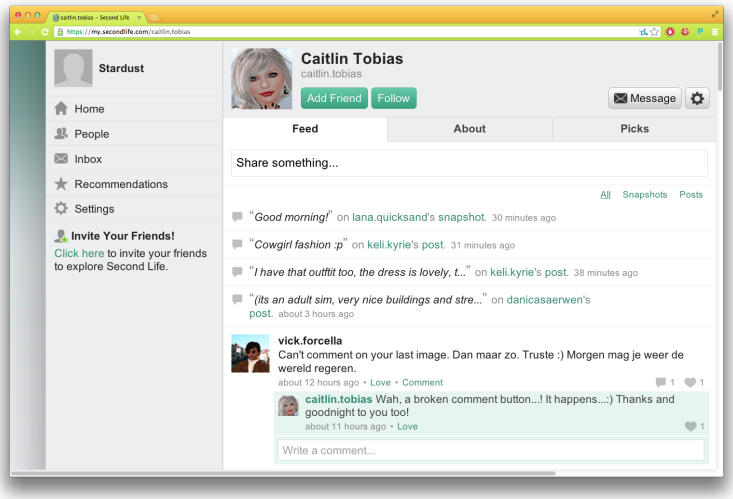


Figure 4.1.: Sample of a user profile in the online social network *My Second Life*. Users can *post* text message on their wall or can communicate with each other by *commenting* or *loving* onto each other's posts.

places. Looking at only semantic features such as tags, they find that the so-called "tagged-with" feature performs well in all three different data category sources.

Probably one of the first projects investigating the link prediction problem from the network topological perspective in the context of online social media is a work conducted by Golder and Yardi [2010]. In their paper they study the micro-blogging service Twitter and find "that two structural characteristics, transitivity and mutuality, are significant predictors of the desire to form new ties". The first paper investigating the extent to which reciprocity could be predicted in the online social media is a recent paper by Cheng et al. [2011]. By applying a rich set of network based features including link prediction features from Liben-Nowell and Kleinberg [2007], they show that the so-called out-degree measure of a user in Twitter is the best feature to predict reciprocity. Another interesting work in this context is a study conducted by Yin et al. [2011]. In their paper they investigate the link prediction problem within the micro-blogging system Twitter. The main contribution, apart from studying the performance of well established link prediction methods, is the introduction of a "novel personalized structure-based link prediction model" which "noticeably outperforms the state-of-the-art" methods. The first work studying the computational efficiency of network topological features in the online domain is a paper written by Fire et al. [2011]. In their work they apply a rich set of over 20 features on a set of 5 different online social network sites with respect to their computational efficiency. Their study reveals that the so-called friends measure shows a good trade-off between accuracy and computational efficiency.

Another study in this context is a recent study conducted by Rowe et al. [2012] In their work they

study the link prediction problem, or the question who will follow whom, in the micro-blogging system *Tencent Weibo*. Looking at both – semantic and network topological features – they show that the predictability of links can be significantly improved by training a classifier that uses both. Although the work of Rowe *et al.* has considerable amount of overlap with our own work, their study only looked at features which could be directly induced from the online media site Tencent Weibo. Hence, contrary to our own work they did not include external knowledge such as location-based social network data as we do in our study. Finally, the last study to be mentioned is a work conducted by Scellato et al. [2011]. Similar to our work they tried to exploit features from the location-based social network of *Gowalla* to predict links between users. However, in contrast to our work, they only focused on location-based social data and did not combine online social network and location-based social network data as we do in this paper. In their analysis over a period of three months they found that most of the links are formed between users that visit the same places or places that share similar properties.

4.3. Datasets

As stated in the introductory part of this paper we conducted our experiments on two types of datasets – online social network and location-based social data – both obtained from the virtual world of Second Life. The reasons for choosing Second Life over other real world sources are manifold: First, in contrast to networks such as Facebook, the online social network *My Second Life* does not restrict extensive crawling of user profiles. Second and contrary to real world online social networks, most profiles in *My Second Life* are public, i.e. we can mine a large fraction of the network. Third, in virtual worlds the location information of users can be harvested in an automated way whereas it is nearly impossible to obtain large-scale tracking data of users in the real world. In this section we describe the collection process for the data as used in our experiments.

4.3.1. Location-based Social Network Dataset

The collection of the location-based social network dataset in Second Life was a two stage process: First a list of popular locations from the Second Life Event calendar* was crawled. Second, overall 15 in-world agents so-called in-world-bots were implemented to teleport to these locations and gather location information of the users at place.

In detail the procedure was the following: In order to harvest all events in Second Life we implemented a Web-crawler that runs on a daily bases to obtain all publicly announced events on

*http://secondlife.com/community/events/

the Second Life Event calendar. Allover, we were able to obtain data of 218,245 unique events during a period of ten months starting in March 2012.

In order to collect location data of the users we implemented overall 15 in-world agents on the basis of the open source command-line client *libopenmetaverse*[†]. Due to the modularity of the tool, we were able to enhance the functionality of our agents to teleport around in the virtual world to collect location data of all surrounding users in a region. This location information comprised the current region, x and y coordinates of the location within this region, and a time stamp. The pool of agents was controlled by a centralized instance sending our in-world bots to ongoing events. Due to the large amount of concurrent events in several regions of Second Life and the constraint that a bot was only able to obtain data of one single region at the same time, our sampling rate was set to a limit of 15 minutes. All in all, we were able to obtain over 13 Million data samples of 190,160 unique users visiting events with this kind of approach [Steurer et al., 2012].

4.3.2. Online Social Network Dataset

In July 2011 Linden Labs introduced an online social network called *My Second Life*[‡] similar to other social networks such as Google+ or Facebook. Residents of the virtual world can log-in with their in-world credentials, access their friend lists and have a so-called *Feed* that can be compared to the Google+ Stream or the Facebook Wall. The social interaction with other users is done by sharing text messages, screenshots, comments and so-called loves which can be seen equally to a Like on Facebook or a Plus in Google+ (see Figure 4.1). Furthermore, users can enhance their profiles by adding personal information such as interests, groups, etc.

We attempted to download the profile data of all 190,160 users found by the avatar-bots. In the next step we parsed the interaction-partners of the these users and downloaded the profile information of the missing ones. This procedure was repeated until no new users could be found by our crawler anymore. Finally, this yielded in a dataset of 311,959 users with 300,657 of them opened to the public, and 135,181 with interactions on their feed.

4.4. Feature Sets

As already outlined, it is our interest to predict interactions between users in online social networks based on features induced from online social network and location-based social network data. To

[†] http://lib.openmetaverse.org/
[‡] https://my.secondlife.com/

Table 4.1.: Basic metrics of the two networks and their combination used for the experiments.

Name	Location-based G_M	Online G_F	$G_{FM} = G_F + G_M$
Type	undirected	directed	directed
Nodes	131,349	135,181	37,118
Edges	2,343,683	209,653	1,043,172
Degree	35.7	3.1	56.2

that end, we induced two different types of feature sets from our data sources: network topological and homophilic features [Wang et al., 2011]. In order to start with the description of the different features calculated for our experiments we first describe the networks derived from the collected data.

The first network, referred to as *online social network*, was based on data obtained from the users profile where every edge in this directed network indicates communication between two users. This yielded in a network with 135,181 users and 209,653 edges. The second network, referred to as *location-based social network*, was based on the users location data where every edge in this undirected network indicated that two users were seen concurrently in the same region on two different days. This yielded in a network with 142,507 nodes and 3,773,316 edges. A summery of both networks can be found in Table 6.1 and Figure 4.2 shows the degree distribution of the social network and location-based social network. Both networks show power-law qualities with an alpha of 1.55 and a corresponding fitting error of 0.13 for the online social network and and alpha value of 2.67 and a fitting error of 0.16 for the location-based social network [Clauset et al.].

4.4.1. Online Social Network: Topological Features

In social networks such as Facebook or Google+ the friendship of users is based on a mutual agreement where both confirm each other. In contrast to this, users of the online social network *My Second Life* can post onto each others' walls without this mutual agreement. Hence, as a consequence, we considered the social network as a directed graph $G_F\langle V_F, E_F\rangle$ with V_F representing the users and $e = (u,v) \in E_F$ if user u posted, commented, or liked something on the feed of user v.

We defined the set of the neighbors of a node $v \in G_F$ as $\Gamma(v) = \{u \mid (u,v) \in E_F \text{ or } (v,u) \in E_F\}$ and based on this definition of neighborhood we used the following topological features:

- *Common Neighbors* $F_{CN}(u,v)$. This represented number of interaction-partners two users had in common.

$$F_{CN}(u,v) = |\Gamma(u) \cap \Gamma(v)|$$

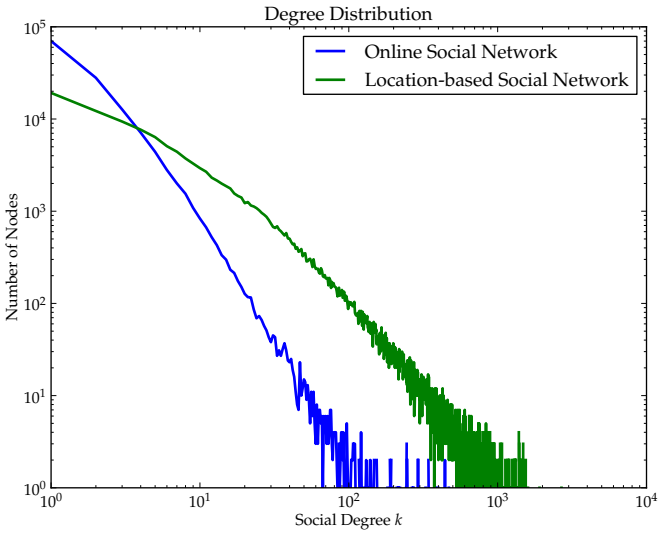


Figure 4.2.: Degree distributions for the online and the location-based social network.

For a directed network we split this into the number of common users $F_{CN}^{+}(u,v) = |\Gamma^{+}(u) \cap \Gamma^{+}(v)|$ to which both users sent messages to and the number of users $F_{CN}^{-}(u,v) = |\Gamma^{-}(u) \cap \Gamma^{-}(v)|$ from which both users received messages.

- *Jaccard's Coefficient* $F_{JC}(u,v)$. The ratio of the total number of neighbors and the number of common neighbors of two users was taken from [Jain and Dubes, 1988] and is defined as follows.

$$F_{JC}(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

For directed networks this could be split into two coefficients for received messages $F_{JC}^{-}(u,v) = \frac{|\Gamma^{-}(u) \cap \Gamma^{-}(v)|}{|\Gamma^{-}(u) \cup \Gamma^{-}(v)|}$ and sent messages $F_{JC}^{+}(u,v) = \frac{|\Gamma^{+}(u) \cap \Gamma^{+}(v)|}{|\Gamma^{+}(u) \cup \Gamma^{+}(v)|}$.

- *Adamic Adar* $F_{AA}(u,v)$. Instead of just counting the number of common neighbors with Jaccard's Coefficient in a network, this feature adds weights to all neighbors of a pair of users [Adamic and Adar, 2003].

$$F_{AA}(u,v) = \sum_{z \in \Gamma(u) \cap \Gamma(v)} \frac{1}{log(|\Gamma(z)|)}$$

According to Cheng et al. [2011] *et al.* this can be transformed into $F_{AA}^{-}(u,v) =$

$\sum_{z \in \Gamma^-(u) \cap \Gamma^-(v)} \frac{1}{log(|\Gamma^-(z)|)}$ for directed networks.

- *Preferential Attachment Score* $F_{PS}(u,v)$. This feature took into account that active users, i.e. users with many interaction partners, are more likely to form new relationships than users with not so many interactions [Barabási and Albert, 1999].

$$F_{PS}(u,v) = |\Gamma(u)| \cdot |\Gamma(v)|$$

  The score was applied to a directed network with two different features: $F^+_{PS}(u,v) = |\Gamma^+(u)| \cdot |\Gamma^-(v)|$, respectively $F^-_{PS}(u,v) = |\Gamma^-(u)| \cdot |\Gamma^+(v)|$ [Cheng et al., 2011].

4.4.2. Online Social Network: Homophilic Features

As stated before, users in Second Life can enhance their online social network profile by adding additional meta-data information such as interests or groups. As observed by a number of previous studies in this area [Rowe et al., 2012; Wang et al., 2011], homophily is an important variable in the context of the link prediction problem. To account for factor, we defined a set of homophilic features which we calculated as group and interest similarity between users u,v. Formally, we defined the groups of a user u as $\Delta(u)$, respectively her interests as $\Psi(u)$.

- *Common Groups* $G_C(u,v)$. The first feature we induce is the so-called common groups measure. It is calculated as follows.

$$G_C(u,v) = |\Delta(u) \cap \Delta(v)|$$

- *Jaccard's Coefficient for Groups* $G_{JC}(u,v)$. The second feature, is the so-called Jaccard's coefficient for groups. It was calculated in the following form.

$$G_{JC}(u,v) = \frac{|\Delta(u) \cap \Delta(v)|}{|\Delta(u) \cup \Delta(v)|}$$

- *Common Interests* $I_C(u,v)$. The third homophilic feature, was the number of interests two users had in common.

$$I_C(u,v) = |\Psi(u) \cap \Psi(v)|$$

- *Jaccard's Coefficient for Interests* $I_{JC}(u,v)$. And finally the last feature, which is a combination of total interests and common interests of the users.

$$I_{JC}(u,v) = \frac{|\Psi(u) \cap \Psi(v)|}{|\Psi(u) \cup \Psi(v)|}$$

4.4.3. Location-based Social Network: Topological Features

We applied the same network topological feature calculations to the location-based social network as we did for the online social network. The network had edges between users that met on at least two days. Using this relation between in-world users defined the topological features similar to Section 4.4.1. Here, the neighbors of a node in the undirected location-based social network $G_M\langle V_M, E_M\rangle$ were defined as $\Theta(u) = \{v \mid (u,v) \in G_M\}$ and starting with this we defined the topological features as follows.

- *Common Neighbors* $M_{CN}(u,v)$.

$$M_{CN}(u,v) = |\Theta(u) \cap \Theta(v)|$$

- *Jaccard's Coefficient* $M_{JC}(u,v)$.

$$M_{JC}(u,v) = \frac{|\Theta(u) \cap \Theta(v)|}{|\Theta(u) \cup \Theta(v)|}$$

- *Adamic Adar* $M_{AA}(u,v)$.

$$M_{AA}(u,v) = \sum_{z \in \Theta(u) \cap \Theta(v)} \frac{1}{log(|\Theta(z)|)}$$

- *Preferential Attachment Score* $M_{PS}(u,v)$.

$$M_{PS}(u,v) = |\Theta(u)| \cdot |\Theta(v)|$$

4.4.4. Location-based Social Network: Homophilic Features

These features were based on the actual distance between users, the regions they visit, and the number of days where they co-occurred concurrently. Let $O(u,v)$ be the co-locations of user u and user v, when both users resided in the same region concurrently. An observation $o \in O(u,v)$ was 4-tuple of region r, time stamp t, location coordinates of user u: $l_u = (x_u, y_u)$ and user v: $l_v = (x_v, y_v)$.

- *Physical Distance* $A_D(u,v)$. Whenever two users were observed concurrently, we measured the distance between them based on their x and y coordinates. As a indicator for their overall physical closeness, we therefore computed the average physical Euclidean distance between

two users for all observations where both were present in the same region concurrently.

$$A_D(u,v) = \frac{1}{|O(u,v)|} \sum_{o \in O(u,v)} \|o(l_u) - o(l_v)\|$$

- *Days Seen* $A_S(u,v)$. This feature indicated the number of days when two users have been observed in the same region concurrently.

The regions of a user were defined as $P(u) = \{\rho \in P \mid \text{user } u \text{ was observed in region } P\}$ and so we computed the region properties of users as follows:

- *Common Regions* $R_C(u,v)$. The number of regions two users visited, not necessarily at the same time.
$$R_C(u,v) = |P(u) \cap P(v)|$$

- *Regions Seen Concurrently* $R_S(u,v)$. In contrast to the Common Regions feature, this feature took only the regions into account where both users were observed in the same region concurrently.

- *Observations Together* $R_O(u,v)$. This feature was taken from Cranshaw et al. [2010] and represented the number of total regions of two users divided by the sum of each user's number of regions.
$$R_O(u,v) = \frac{|P_u \cup P_v|}{|P_u| + |P_v|}$$

4.5. Experimental Setup

All in all, we conducted two different experiments to study the extent to which interactions between users in online social networks can be predicted. Both experiments were based on the combination of the *online social network* G_F and the *location-based social network* G_M described in Section 4.4. To that end, we followed the approach of Guha et al. [2004] in both experiments who suggest to create two datasets with an equal number of "positive edges" and "negative edges" for the binary classification problem. This results in balanced datasets for the test- and the training data and therefore in a baseline of 50% for the prediction when guessing randomly. For the evaluation of the binary classification problem we employed different supervised learning algorithms and used the area under the ROC curve (AUC) as our main evaluation metric to determine the performance of our features [Huang and Ling, 2005; Ling et al., 2003]. We justified our findings with a 10-fold cross validation approach using the WEKA machine-learning suite [Hall et al., 2009].

In this section we describe in detail how the trainings and test data set for both experiments was generated.

4.5.1. Predicting Interactions

The task here is to predict whether or not two users interacted with each other on the feed by using topological and homophilic information of the online social network and the location-based social network. In the first step we computed the edge-features for the user-pairs as described in Section 4.4 for both networks independently. Then, in the second step we created the intersection of both networks as directed graph $G_{FM}\langle V_{FM}, E_{FM}\rangle$ where $V_{FM} = \{v | v \in V_F, v \in V_M\}$, and $E_{FM} = \{(u,v) | (u,v) \in E_M, (u,v) \in E_F, v \text{ and } u \in V_{FM}\}$. This newly created network consisted of 37,118 nodes and 1,014,352 pairs with location co-occurrences $((u,v) \in E_M)$, 36,213 pairs with social interaction $((u,v) \in E_F)$, and 7,393 edges with both $((u,v) \in E_M, E_F)$.

For the binary classification problem we uniformly selected 2,500 user-pairs with a social interaction and a location co-occurrence ("positive edges") $\{e^+ = (u,v) | e^+ \in E_{FM},\ e^+ \in E_F,\ e^+ \in E_M\}$ and 2,500 user-pairs with a location co-occurrence but without a social interaction ("negative edges") $\{e^- = (u,v) |\ e^- \notin E_F,\ e^- \in E_M\}$. These edges, i.e. pairs of users, and the according edge features from both domains were used as datasets for all further evaluations and experiments.

4.5.2. Predicting Reciprocity

The task here is to predict whether two users had mutual activities on each other's wall, i.e. reciprocal interactions, by exploiting topological and homophilic information of the online social network and the location-based social network. We defined a reciprocal edge as $e'' = (u,v) | (u,v) \in G_F, (v,u) \in G_F$, a non-reciprocal edge as $e' = (u,v) | (u,v) \in G_F,\ (v,u) \notin G_F$, and used this difference for the binary classification problem. In contrast to the previous experiment we considered the online social network as undirected network for the computation of the edge-features but retained information about the reciprocity of the interactions. The edge features for the location-based social network were again considered as undirected. For the actual experiment we combined the online social network and the location-based social network to a new undirected network referred to as $G'_{FM}\langle V'_{FM}, E'_{FM}\rangle$ where $V'_{FM} = \{v | v \in V_F, v \in V_M\}$, and $E'_{FM} = \{(u,v) | (u,v) \in E_M, (u,v) \in E_F$ or $(v,u) \in E_F, v \text{ and } u \in V'_{FM}\}$. Out of the 7,393 user-pairs with a social interaction and a location co-occurrence we identified 1,431 reciprocal edges and 4,531 non-reciprocal edges in the online social network. For the binary classification task we uniformly selected pairs of users from the undirected network G'_{FM} with 1,000 reciprocal edges ("positive edges") and non-reciprocal edges ("negative edges") each. These edges, i.e. user-pairs with the according features, were used for all further evaluations and experiments.

Table 4.2.: Means and standard errors of the features in the online social network and the location-based social network for the group of users having interactions with each other vs. the groups of users having no interactions (***=significant at level 0.001) .

	Features	Have Interactions	Have No Interactions
Online Social Network	Common Neighbors (in) $F_{CN}^{-}(u,v)^{***}$	2.81 ± 0.32	0.02 ± 0.00
	Common Neighbors (out) $F_{CN}^{+}(u,v)^{***}$	2.39 ± 0.27	0.01 ± 0.00
	Adamic Adar $F_{AA}(u,v)^{***}$	14.65 ± 1.28	1.71 ± 0.18
	Jaccard's Coefficient (in) $F_{JC}^{-}(u,v)^{***}$	0.05 ± 0.00	0.00 ± 0.00
	Jaccard's Coefficient (out) $F_{JC}^{+}(u,v)^{***}$	0.04 ± 0.00	0.00 ± 0.00
	Preferential Attachment (in) $F_{PS}^{-}(u,v)^{***}$	1566.55 ± 239.31	3.88 ± 0.64
	Preferential Attachment (out) $F_{PS}^{+}(u,v)^{***}$	2088.94 ± 441.14	4.92 ± 1.53
	Common Groups $G_C(u,v)^{***}$	1.92 ± 0.07	0.40 ± 0.02
	Jaccard's Coefficient $G_{JC}(u,v)^{***}$	0.05 ± 0.00	0.01 ± 0.00
	Common Interests $I_C(u,v)$	0.07 ± 0.01	0.02 ± 0.00
	Jaccard's Coefficient $I_{JC}(u,v)$	0.00 ± 0.00	0.00 ± 0.00
Location-based Social Network	Common Neighbors $M_{CN}(u,v)^{***}$	52.48 ± 4.98	83.61 ± 2.31
	Jaccard's Coefficient $M_{JC}(u,v)^{***}$	0.20 ± 0.00	0.10 ± 0.00
	Preferential Attachment $M_{PS}(u,v)^{***}$	218341.22 ± 164510.35	530640.88 ± 50352.29
	Adamic Adar $M_{AA}(u,v)^{***}$	26.89 ± 3.19	36.43 ± 0.98
	Regions Seen $R_S(u,v)^{***}$	2.81 ± 0.09	1.41 ± 0.02
	Common Regions $R_C(u,v)^{***}$	3.59 ± 0.34	3.03 ± 0.08
	Observations Together $R_O(u,v)^{***}$	0.22 ± 0.00	0.10 ± 0.00
	Distance $A_D(u,v)^{***}$	10.32 ± 0.36	38.13 ± 0.95
	Days Seen $A_S(u,v)^{***}$	7.34 ± 0.21	3.98 ± 0.09

4.6. Results

Before we start with the analysis of how to predict interactions between users, we show the differences between user pairs with and without interactions in the social network, respectively user pairs with reciprocal and non-reciprocal interactions for both domains. Both the Anderson-Darling test and the one-sampled Kolmogorov-Smirnov test showed that none of the distributions of the features described in Section 4.4 were normally distributed. Hence, and similar to Bischoff [2012], we compared the variances of all features using a Levene test ($p < 0.01$). To test for significant differences of the means, we employed Mann-Whitney-Wilcoxon test in case of equal variances and a two-sided Kolmogorov-Smirnov test in case of unequal variances. The differences of the means between the groups of users regarding their interaction type can be found in Table 4.3 and 4.2. Overall, we found the following:

- *Interactions:* Mean values of topological features in the online social network were significantly higher for user pairs with interactions compared to users without interactions. For homophilic features, a significant difference between user pairs was observed for features

Table 4.3.: Means and standard errors of the features in the online social network and the location-based social network for the group of users having reciprocal interactions vs. the groups of users having no reciprocal interactions with each other (*=significant at level 0.1, **=significant at level 0.01, and ***=significant at level 0.001).

	Features	Reciprocal	Non Reciprocal
Online Social Network	Common Neighbors $F_{CN}(u,v)^{***}$	10.20 ± 1.10	0.80 ± 0.10
	Adamic Adar $F_{AA}(u,v)^{***}$	6.46 ± 0.61	0.71 ± 0.06
	Jaccard's Coefficient $F_{JC}(u,v)^{***}$	0.08 ± 0.00	0.04 ± 0.00
	Preferential Attachment $F_{PS}(u,v)^{***}$	12544.28 ± 2066.82	403.15 ± 93.73
	Common Groups $G_C(u,v)$	2.04 ± 0.11	1.81 ± 0.10
	Jaccard's Coefficient $G_{JC}(u,v)$	0.06 ± 0.00	0.05 ± 0.00
	Common Interests $I_C(u,v)$	0.12 ± 0.02	0.05 ± 0.01
	Jaccard's Coefficient $I_{JC}(u,v)$	0.01 ± 0.00	0.00 ± 0.00
Location-based Social Network	Common Neighbors $M_{CN}(u,v)^{***}$	42.59 ± 2.67	61.29 ± 11.96
	Jaccard's Coefficient $M_{JC}(u,v)^{**}$	0.2 ± 0.01	0.19 ± 0.01
	Preferential Attachment $M_{PS}(u,v)^{*}$	41663.58 ± 4547.60	473151.99 ± 411215.48
	Adamic Adar $M_{AA}(u,v)$	21.01 ± 1.30	32.25 ± 7.79
	Regions Seen $R_S(u,v)$	2.82 ± 0.10	2.71 ± 0.18
	Common Regions $R_C(u,v)$	3.25 ± 0.12	4.00 ± 0.83
	Observations Together $R_O(u,v)$	0.23 ± 0.00	0.21 ± 0.00
	Distance $A_D(u,v)^{**}$	9.35 ± 0.48	11.19 ± 0.57
	Days Seen $A_S(u,v)^{**}$	7.22 ± 0.31	6.96 ± 0.33

based on group affiliation whereas features based on specified interests did not show significant differences. Topological features in the location-based social network also showed significant differences between users but contrary, users with no interactions had a higher number of common neighbors, preferential attachment score, and Adamic Adar score. Users with interactions had more common regions and observations, and they saw each other on more days. Furthermore, user pairs with interactions in the online social network had a significantly shorter average distance between them.

- *Reciprocity:* The differences between user pairs with reciprocal interactions and non-reciprocal interactions can be found in Table 4.3. The results revealed significant differences between users in the online social network for all topological features but no significant differences for homophilic features. Comparing differences between user pairs also showed significant differences in the topological features of the location-based social network (Common Neighbors, Jaccard's Coefficient and Preferential Attachment Score) but only the average distance between users and the number of days they saw each other was significantly different for the homophilic features

Table 4.4.: Overall results AUC and (ACC) of the Logistic Regression learning approach for predicting interactions between users and their reciprocity in the online social network of Second Life using online social network and location-based social network features.

	Feature Sets		Interaction	Reciprocity
Logistic Regression	Online Social Network	Topological	0.878 (71.8%)	0.676 (64.9%)
		Homophilic	0.640 (63.4%)	0.507 (52.5%)
		Combined	**0.863 (76.8%)**	**0.679 (64.8%)**
	Location-based Social Network	Topological	0.858 (76.7%)	0.530 (51.2%)
		Homophilic	0.885 (80.6%)	0.556 (54.4%)
		Combined	**0.919 (84.8%)**	**0.551 (53.5%)**
	All Features		**0.953 (89.6%)**	**0.709 (65.2%)**

In the remainder of this section we present the results obtained from the two supervised learning experiments described in Section 4.5. As learning strategy we used the *Logistic Regression* learning algorithm since it can be easily implemented and interpreted [Jones et al., 2013].

4.6.1. Predicting Interactions: Online Social Network vs. Location-based Social Network Features

The results of the first experiment can be found in Table 4.4 where we present the outcome of the prediction model for two different sources of knowledge and the according feature sets.

The values in the table represent the area under the ROC curve (AUC) and the accuracy of the prediction (ACC) as metrics for the predictability with a baseline for the binary classification problem of 0.5 AUC. As we can see, using topological features from the online social network improved the predictability of interactions between users by +37.8% whereas homophilic features (groups and interests) enhanced the baseline by +14.0%. In contrast to this, topological features from the location-based social network improved the baseline by +35.8% whereas homophilic features improved it by +38.5%. The combined topological and homophilic features from either networks resulted in a predictability of 0.953 AUC outperforming the baseline by +45.3%.

Overall, and interestingly, looking at the feature set in Table 4.4 we can see that location-based features were a great source to predict interactions between users in online social networks and they even outperformed online social network features. To evaluate the predictability of interactions of features separately, we present the coefficients of the Logistic Regression algorithm and their levels of significance when all features were used simultaneously. Table 4.5 shows that Preferential Attachment Score for incoming messages $F_{PS}^{-}(u,v)$ in the online social network and the average distance between users $A_D(u,v)$ in the location-based social network were most impact-

Table 4.5.: Coefficients of the Logistic Regression when all topological and homophilic features from both domains are used simultaneously in the dataset (***=significant at level 0.001).

	Features	Interactions	Reciprocity
Online Social Network	Common Neighbors (in) $F^{-}_{CN}(u,v)$	-1.782615***	–
	Common Neighbors (out) $F^{+}_{CN}(u,v)$	0.138448***	–
	Common Neighbors $F_{CN}(u,v)$	–	**-0.658291*****
	Adamic Adar $F_{AA}(u,v)$	0.196078	-0.108824***
	Jaccard's Coefficient (in) $F^{-}_{JC}(u,v)$	0.025060***	–
	Jaccard's Coefficient (out) $F^{+}_{JC}(u,v)$	2.416276 ***	–
	Jaccard's Coefficient $F_{JC}(u,v)$	–	0.495911***
	Preferential Attachment (in) $F^{-}_{PS}(u,v)$	**7.405495*****	–
	Preferential Attachment (out) $F^{+}_{PS}(u,v)$	-0.000097	–
	Preferential Attachment $F_{PS}(u,v)$	–	-1.107698
	Common Groups $G_C(u,v)$	-0.000066***	-0.000040***
	Jaccard's Coefficient $G_{JC}(u,v)$	0.216582***	-0.046399
	Common Interests $I_C(u,v)$	-1.230746	1.732937
	Jaccard's Coefficient $I_{JC}(u,v)$	0.932973	7.158616
Location-based Social Network	Common Neighbors $M_{CN}(u,v)$	-0.019859***	-0.004276
	Jaccard's Coefficient $M_{JC}(u,v)$	-0.001736***	-0.000470***
	Preferential Attachment $M_{PS}(u,v)$	0.000551***	0.000574
	Adamic Adar $M_{AA}(u,v)$	0.000001***	0.000000
	Regions Seen $R_S(u,v)$	0.294520	-0.101258
	Common Regions $R_C(u,v)$	0.717518***	0.093925
	Observations Together $R_O(u,v)$	0.022711***	-0.064381
	Distance $A_D(u,v)$	**10.570453*****	**1.158166*****
	Days Seen $A_S(u,v)$	-0.010596***	-0.002153

ing features. Struck-out values in the table indicate a significance value of $p > 0.05$. To give an overview of the correlation of the features, we calculated the pair-wise Spearman-rank correlation of the used features from both domains as shown inTable 4.8.

4.6.2. Predicting Reciprocity: Online Social Network vs. Location-based Social Network Features

The results of the second experiment can be found in Table 4.4 where we present the area under the ROC curve (AUC) and the accuracy of the prediction (ACC). As in the previous experiment the baseline for randomly guessing is 0.5 AUC due to the balanced dataset.

Using topological features from the online social network increased the predictability of reciprocity by +17.6% whereas homophilic features alone (groups and interests) performed as bad as the baseline. Due to the little predictive power of the homophilic features the combination of all features in the online social network results in a prediction gain of +17.6% which is equal to topological features alone. In contrast to this, topological features from the location-based social network improved the baseline approach by +3.0% for the topological features and by +5.6% for the homophilic features. The combination of feature sets in the location-based social network boosted the predictability by +5.1%. The combination of features from either domains elevated the predictability of the reciprocity between two users up to 0.709 AUC, which is a boost of +20.9% if compared to the baseline of 0.5 AUC. Similar to the previous experiment, we computed the coefficients of the Logistic Regression algorithm in Table 4.5. In the online social network domain the Common Neighbors feature $F_{CN}(u,v)$ and in the location-based social network domain the distance between users $A_D(u,v)$ had the highest and most significant values. Again, struck-out values indicate a significance $p > 0.05$.

4.6.3. Verification of Stability: Predicting Interactions and Reciprocity with SVM and Random Forrest

The results of the conducted experiments based on LogisticRegression clearly showed that features from the location-based social network are better suited to predict interactions between users, whereas features from the online social network are better suited to predict reciprocity of interactions. However, to verify the stability of these findings we employed two additional learning algorithms: *Random Forest* and *Support Vector Machine* which are well suited for high dimensional, numeric and inter-dependent attributes (see Table 4.8) [Bischoff, 2012; Jones et al., 2013]. The results of these learning algorithms are presented in Tables 4.6 and 4.7. Overall, the results can be interpreted as follows:

- *Predicting Interactions:* Using Logistic Regression, features from the location-based social network outperformed features from the online social network and similar results were observed for *Support Vector Machine* and *Random Forest*. In both cases features of the location-based social network resulted in a better prediction of interactions than features from the online social network. Overall, the performance of the combined feature set using Support Vector Machine was 0.882 AUC and using Random Forest was 0.979 AUC.
- *Predicting Reciprocity:* For the prediction of reciprocity of interactions between users using Logistic Regression, online social network features outperformed location-based social network features. For other learning algorithms we found similar results as features from the online social network also outperformed features from the location-based social network.

Table 4.6.: Overall results AUC and (ACC) of the SVM learning approach for predicting interactions between users and their reciprocity in the online social network of Second Life using online social network and location-based social network features.

	Feature Sets		Interactions	Reciprocity
SVM	Online Social Network	Topological	0.669 (66.9%)	0.646 (64.6%)
		Homophilic	0.638 (63.8%)	0.522 (52.2%)
		Combined	**0.737 (73.7%)**	**0.639 (63.9%)**
	Location-based Social Network	Topological	0.793 (79.3%)	0.529 (52.9%)
		Homophilic	0.761 (76.1%)	0.515 (51.5%)
		Combined	**0.849 (84.9%)**	**0.539 (53.9%)**
	All Features		**0.882 (88.2%)**	**0.638 (63.8%)**

Table 4.7.: Overall results AUC and (ACC) of the Random Forrest learning approach for predicting interactions between users and their reciprocity in the online social network of Second Life using online social network and location-based social network features.

	Feature Sets		Interactions	Reciprocity
Random Forest	Online Social Network	Topological	0.893 (79.7%)	0.628 (62.2%)
		Homophilic	0.624 (62.8%)	0.488 (50.4%)
		Combined	**0.910 (82.5%)**	**0.635 (60.5%)**
	Location-based Social Network	Topological	0.852 (77.9%)	0.530 (52.2%)
		Homophilic	0.872 (80.3%)	0.479 (49.2%)
		Combined	**0.916 (85.7%)**	**0.550 (53.2%)**
	All Features		**0.979 (93.0%)**	**0.684 (62.8%)**

The combination of all features from both domains predicted reciprocity of interactions with 0.652 AUC using Support Vector Machine respectively 0.684 using Random Forest.

4.7. Discussion and Conclusions

In this work we harvested data from two Second Life related data sources: an online social network with text-based interactions and a location-based social network with position data. We modeled the social proximity between users with topological and homophilic network features and conducted two experiments.

- *RQ1:* To answer the first research question, we compared different features of user pairs regarding their interactions and the reciprocity of these interactions. This analysis revealed that pairs with interactions were tighter connected in the online social network but the opposite was observed for the location-based social network. A possible explanation is that users

in Second Life are allowed to directly "jump" to different regions in the whole virtual world but see the present users only upon arrival. We believe that users are more likely to stay in a region if they know present users, i.e. they have interactions on the online social network. This mobility activity could explain the tight connections in the location-based social network. This assumption is supported by homophilic features from both networks: users with interactions had more common groups, regions, and they saw each other on more days. Furthermore, the average distance was significantly shorter than users without interactions. All observed features were significantly different except interest based features but we assume this is due to the sparse data. The found results for predicting reciprocity of interactions was similar to the prediction of interactions themselves. User pairs with reciprocal interactions had tight connections in the online social network but the opposite was observed for the location-based social network. Again, homophilic features of user pairs with reciprocal interactions indicated a higher alikeness in both networks.

- *RQ2:* For the second research question we predicted interactions and the reciprocity of these interactions. To do so, we chose Logistic Regression because it is easy to implement and interpret. We observed that interactions can be better predicted with features from the location-based social network than with features from the social network. Surprisingly, the opposite was observed for the reciprocity of interactions. In both experiments we found the combination of features from both networks outperforming either networks: Interactions could be predicted with 0.953 AUC and the reciprocity of these interactions with 0.709 AUC. The Logistic Regression coefficients of the features unveiled that a short average distance between users is a good indicator for interactions and their reciprocity. To verify our results that online social network features outperform features from the location-based social network for the prediction of interactions and vice versa for the prediction of reciprocity, we used two additional learning algorithms: Support Vector Machines and the Random Forest learning approach. Both algorithms approved the observations made in the experiment with Logistic Regression.

- *RQ3:* To answer the third research question, we compared homophilic features and topological features regarding the predictability of interactions and their reciprocity. Interestingly, we could not find a stable pattern over all experiments, as it was for instance proposed by Rowe et al. [2012]. Although topological features of the online social network outperformed homophilic features in all three learning algorithms we found variation of the results for the location-based social network. Using Logistic Regression homophilic features performed better than topological features but in contrast, the opposite was observed for Support Vector Machines. With Random Forest homophilic features were better suited for the prediction of interactions but homophilic features were better suited for the reciprocity of interactions.

For future work, it is planned to dig deeper into the data and to address issues such as the variety of time (which we did not address in this study) or the issue why reciprocal links seem to be better predicted with social network features than with position data. Furthermore, we plan to extend our approach to predict other relations between users besides communicational interactions such as for instance partnership which can be also mined from the social network of Second Life. Finally, it is our interest to switch from supervised to unsupervised learning.

References

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

A. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

K. Bischoff. We love rock'n'roll: analyzing and predicting friendship links in Last. fm. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 47–56. ACM, 2012.

J. Cheng, D. Romero, B. Meeder, and J. Kleinberg. Predicting reciprocity in social networks. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 49–56. IEEE, 2011.

A. Clauset, C. R. Shalizi, and M. Newman. Power-law distributions in empirical data, 2007. *arXiv preprint arXiv:0706.1062*, 64.

J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 119–128. ACM, 2010.

M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici. Link prediction in social networks using computationally efficient topological features. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 73–80. IEEE, 2011.

S. Golder and S. Yardi. Structural predictors of tie formation in twitter: Transitivity and mutuality. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 88–95. IEEE, 2010.

R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, pages 403–412. ACM, 2004.

I. Guy, M. Jacovi, E. Shahar, N. Meshulam, V. Soroka, and S. Farrell. Harvesting with sonar: the value of aggregating social network information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1017–1026, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-011-1. doi: 10.1145/1357054.1357212. URL http://doi.acm.org/10.1145/1357054.1357212.

I. Guy, N. Zwerdling, D. Carmel, I. Ronen, E. Uziel, S. Yogev, and S. Ofek-Koifman. Personalized recommendation of social software items based on social relations. In *Proceedings of the third ACM conference on Recommender systems*, RecSys '09, pages 53–60, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-435-5. doi: 10.1145/1639714.1639725. URL http://doi.acm.org/10.1145/1639714.1639725.

I. Guy, M. Jacovi, A. Perer, I. Ronen, and E. Uziel. Same places, same things, same people?: mining user similarity on social media. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, CSCW '10, pages 41–50, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-795-0. doi: 10.1145/1718918.1718928. URL http://doi.acm.org/10.1145/1718918.1718928.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

J. Huang and C. X. Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Trans. on Knowl. and Data Eng.*, 17(3):299–310, Mar. 2005. ISSN 1041-4347. doi: 10.1109/TKDE.2005.50. URL http://dx.doi.org/10.1109/TKDE.2005.50.

A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

J. J. Jones, J. E. Settle, R. M. Bond, C. J. Fariss, C. Marlow, and J. H. Fowler. Inferring tie strength from online directed behavior. *PloS one*, 8(1):e52168, 2013.

D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

C. X. Ling, J. Huang, and H. Zhang. Auc: a statistically consistent and more discriminating measure than accuracy. In *International Joint Conference on Artificial Intelligence*, volume 18, pages 519–526. LAWRENCE ERLBAUM ASSOCIATES LTD, 2003.

M. Rowe, M. Stankovic, and H. Alani. Who will follow whom? exploiting semantics for link prediction in attention-information networks. 2012.

S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM, 2011.

M. Steurer and C. Trattner. Who will interact with whom? a case-study in second life using online and location-based social network features to predict interactions between users. 2013.

M. Steurer, C. Trattner, and F. Kappe. Success factors of events in virtual worlds a case study in second life. In *NetGames*, pages 1–2, 2012.

M. Steurer, C. Trattner, and D. Helic. Predicting social interactions from different sources of location-based knowledge. In *SOTICS 2013, The Third International Conference on Social Eco-Informatics*, pages 8–13, 2013.

D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. ACM, 2011.

D. Yin, L. Hong, and B. Davison. Structural link analysis and prediction in microblogs. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1163–1168. ACM, 2011.

Table 4.8.: Spearman's Correlation Matrix (*=significant at level 0.1, **=significant at level 0.01, and ***=significant at level 0.001).

		F^-_{CN}	F^+_{CN}	F_{AA}	F^-_{JC}	F^+_{JC}	F^-_{PS}	F^+_{PS}	G_C	G_{JC}	I_C	I_{JC}	M_{CN}	M_{JC}	M_{PS}	M_{AA}	R_S	R_C	R_O	A_D	A_S
Online Social Network	F^-_{CN}	1.00																			
	F^+_{CN}	0.55***	1.00																		
	F_{AA}	0.49***	0.41***	1.00																	
	F^-_{JC}	0.99***	0.51***	0.47***	1.00																
	F^+_{JC}	0.53***	1.00***	0.40***	0.50***	1.00															
	F^-_{PS}	0.48***	0.47***	0.47***	0.46***	0.46***	1.00														
	F^+_{PS}	0.44***	0.49***	0.56***	0.41***	0.47***	**0.30****	1.00													
	G_C	0.16***	0.08***	0.09***	0.17***	0.09***	0.19***	0.06***	1.00												
	G_{JC}	0.16***	0.08***	0.08***	0.17***	0.09***	0.18***	0.06***	0.99***	1.00											
	I_C	0.11***	0.13***	0.12***	0.10***	0.12***	0.13***	0.12***	0.04**	0.03*	1.00										
	I_{JC}	0.11***	0.13***	0.12***	0.10***	0.12***	0.13***	0.12***	0.04*	0.03*	1.00***	1.00									
Location-based Social Network	M_{CN}	0.13***	0.09***	0.08***	0.14***	0.10***	**0.22****	0.06***	**0.20****	**0.21****	0.02	0.02	1.00								
	M_{JC}	0.03*	0.01	0.01	0.04*	0.01	0.05***	0.02	0.11***	0.10***	0.01	0.01	0.74***	1.00							
	M_{PS}	0.18***	0.14***	0.07***	0.19***	0.15***	**0.27****	0.07***	**0.24****	**0.26****	0.03*	0.03*	0.45***	0.32***	1.00						
	M_{AA}	-0.16***	-0.11***	-0.10***	-0.17***	-0.12***	**-0.30****	-0.07***	**-0.19****	**-0.20****	-0.04**	-0.04**	-0.38***	-0.19***	-0.45***	1.00					
	R_S	**-0.20****	-0.16***	-0.14***	**-0.21****	-0.16***	**-0.35****	-0.10***	**-0.22****	**-0.23****	-0.03*	-0.03*	-0.22***	0.10***	-0.53***	0.52***	1.00				
	R_C	-0.16***	-0.13***	-0.08***	-0.17***	-0.14***	**-0.24****	-0.09***	-0.19***	**-0.20****	-0.01	-0.01	-0.22***	-0.05***	-0.51***	0.32***	0.52***	1.00			
	R_O	-0.18***	-0.14***	-0.12***	-0.19***	-0.15***	**-0.32****	-0.08***	-0.18***	-0.19***	-0.03*	-0.03*	-0.18***	0.16***	-0.47***	0.50***	0.94***	0.34***	1.00		
	A_D	-0.07***	-0.05***	-0.06***	-0.07***	-0.06***	-0.17***	-0.02*	-0.02*	-0.03*	-0.01	-0.01	-0.08***	0.14***	-0.16***	0.33***	0.61***	-0.14***	0.78***	1.00	
	A_S	0.14***	0.11***	0.08***	0.14***	0.12***	0.16***	0.08***	**0.20****	**0.20****	0.02	0.02	0.38***	0.31***	0.21***	-0.06***	0.10***	-0.30***	0.23***	0.51***	1.00

CHAPTER 5

Compare Region Sources

This chapter has been published in the paper *Predicting Social Interactions from Different Sources of Location-based Knowledge*, presented at the *Third International Conference on Social Eco-Informatics* [Steurer et al., 2013].

In this chapter we employ the location data of users from three different sources to predict their interactions in a directed online social network. We define a set of features to model the user relations that can be applied to all location sources. In order to start with the actual prediction we compare user-pairs with interactions and user pairs without interaction among all three location sources. We employ a simple Collaborative Filtering algorithm to determine the most suitable features for the prediction task and then apply different supervised learning algorithms to predict interactions between users. The derived results are then compared to each other with respect to the feature selection and the results of the prediction.

In detail the chapter is structured as follows: In Section 5.2 we shortly discuss related work in the area. In Section 5.3 we introduce the collected datasets and introduce the features computed to predict social interactions between users in Section 5.4. The setup of the experiments is depicted in Section 5.5 followed by the results in Section 5.6. Finally, Section 5.7 discusses the findings and concludes the chapter.

Abstract

Recent research has shown that digital online geo-location traces are new and valuable sources to predict social interactions between users, e.g. check-ins via FourSquare or geo-location information in Flickr images. Interestingly, if we look at related work in this area, research studying the extent to which social interactions can be predicted between users by taking more than one location-based knowledge source into account does not exist. To contribute to this field of research, we have collected social interaction data of users in an online social network called My Second Life and three related location-based knowledge sources of these users (monitored locations, shared locations and favoured locations), to show the extent to which social interactions between users can be predicted. Using supervised and unsupervised machine learning techniques, we find that on the one hand the same location-based features (e.g. the common regions and common observations) perform well across the three different sources. On the other hand, we find that the shared location information is better suited to predict social interactions between users than monitored or favoured location information of the user.

5.1. Introduction

There is no doubt that tomorrow's world will be mobile and social. It is therefore not surprisingly that recent research has rigorously followed this trend to study new methods to for instance predict social ties or links between people in such an environment. Interestingly, if we look at related work in this area (e.g. [Scellato et al., 2011b; Cranshaw et al., 2010; Bischoff, 2012]), research studying the extent to which social links can be predicted between users typically takes just one knowledge source into account, e.g. online social network data from Facebook, or location-based social network data from FourSquare. To contribute to this emergent and still sparse field of research, we have recently started a project (see [Steurer et al., 2012; Steurer and Trattner, 2013a,b]) with the overall goal to predict links and tie strength between users from various sources of social and mobile data. Since it is nearly impossible to obtain a complete dataset containing both kinds of knowledge sources in the real world, we focused with our experiments on a virtual environment called My Second Life. This allowed us to easily mine any kind of information needed for such a type of a project on a large scale. So far, we have studied the extent to which partnership [Steurer and Trattner, 2013b] and in general interactions can be predicted [Steurer and Trattner, 2013a] by looking at homophilic features such as for instance common interests, common groups, or common-places visited and network topological features where we investigated common friends features such as Adamic Adar, Jaccard's coefficient etc. Interestingly, we find, that the location information of the user is to a great extent useful to predict tie strength for the interactions be-

tween users in the virtual world of Second Life, most of the time outperforming online social network features. While we only used one particular type of location-based knowledge source about users, namely monitored locations, in our previous research, in this paper we are interested to overall study three different types of knowledge sources: monitored locations, shared locations and favoured locations. We employed 10 different features to predict social interactions between users and unveil what type of location-based knowledge source and what types of features are the most valuable. Overall, we would like to answer the following research questions in this paper:

- *RQ1:* Are there any statistically significant differences between the users having and not having social interaction with each other based on the features induced from our three different kinds of location-based knowledge sources?
- *RQ2:* Which features perform best across those three types of location-based knowledge sources?
- *RQ3:* What kind of knowledge sources is in the end the most valuable to predict social interactions between users?

To answer the first question we analysed the datasets with statistical methods according to our features. This evaluation showed that there were significant differences between user-pairs with a social interaction and users without an social interaction across all computed features and all three sources of location information. For instance, user-pairs with a social interaction share more common regions compared to user-pairs without social interaction. To answer the second research question, we employed Collaborative Filtering for each feature independently to predict the social interactions between the users to find the most valuable features. Among others we found that common regions and common observations of two users were a good indicator for an social interaction between them. For the last question we combined the best features for each region source and showed that the user's Shared Locations are more valuable to predict social interactions than Monitored or Favoured locations.

5.2. Related Work

Approaches by Ling et al. [2003] or Al Hasan et al. [2006] for link prediction using features obtained from online social networks where greatly enhanced with the advent of user's location data. On of the first studies in this field was conducted Cranshaw et al. [2010] who combined the interaction of the online social network *Facebook* with the location-based social network of *Loccaccino*. They introduced various metrics to compute users homophily and found a significant correlation between social interactions and location-based features. Similar observations

were made by Thelwall [2009] who revealed significant homophily between interacting users in *MySpace* and even inferred an real-life friendship from the online social network. This goes inline with Bischoff [2012] who found relations between connections in *Last.FM* and visited music concerts based on demographic, structural and taste-related attributes. Scellato et al. [2011b] investigated in the location-based social network of *Gowalla* and found 30% of newly created links as "place friends". Research by Wang et al. [2011] follows this direction. They investigated in the prediction of social relations using mobility data obtained from mobile phones and found mobile information significantly outperforming simple network measures. Another paper by Scellato et al. [2011a] focuses on the structural differences between the three location-based social networks of *Brighknight*, *Foursquare*, and *Gowalla*. In contrast to our work, they did not have different location sources for one single online social network and their focus was on the actual spatial distance between user.

5.3. Data Sets

Our experiments were based on a social interaction dataset of users in an online social network and three independent location-based knowledge sources: *Monitored Locations*, *Shared Locations*, and *Favoured Locations* from a virtual world. In particular, we focused in our experiments on a virtual environment called Second Life, where we could easily mine the necessary information needed for the experiments on a large scale (see [Steurer et al., 2012; Steurer and Trattner, 2013a,b] for more details).

5.3.1. Social Interaction Dataset

The online social network *My Second Life* was introduced by Linden Labs, the company behind Second Life, in July 2011. It is a social network that can be compared to Facebook regarding postings and check-ins but aims only at residents of the virtual world: just as in Facebook, residents can interact with each other sharing text messages, and comment or love (similar to a "like" in Facebook) these messages. Figure 5.1 depicts a typical profile of a user with postings, comments, and loves from others. The profile of users can be accessed with a unique URI derived from the user name and we attempted to download the profile data of over 400,000 users with a web-crawler. We extracted their interaction partners and downloaded the missing profiles iteratively. With this approach, we found 152,509 profiles with interactions on their wall and identified 1,084,002 postings, 459,734 comments and 1,631,568 loves.

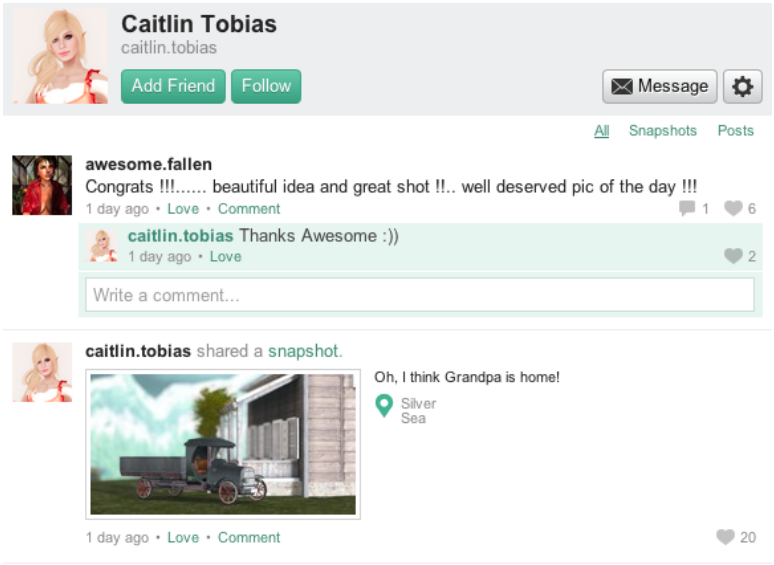


Figure 5.1.: User profile of an Second Life resident in the online social network My Second Life showing a posting, a shared snapshot with location information, and a comment.

5.3.2. Location-based Dataset

To predict the social interactions between users we employed location information obtained from three different sources of data.

Monitored Locations As in the real life, residents of Second Life can host events in the virtual world for other residents and publicly announce this information on an event calendar. We implemented a web-crawler that harvested this calendar periodically to extract all events with accurate event-location and start time. Based on this information we have implemented 15 avatar-bots that visited these events with an interval of 15 minutes and collected the accurate location of the participating users. Starting in March 2012 we were able to collect 262,234 events over a period of 12 months yielding in a dataset of nearly 19 million data samples, i.e. user-location tuples, of over 410,616 different users in 4,132 unique locations.

Shared Locations Users of *My Second Life* can not only interact with each other using postings, comments, or loves, they can also share location information about their current in-world location as well through in-world pictures. The idea of sharing these locations can be compared to pictures uploaded to *Flickr* or *Facebook* enriched with GPS information (see Figure 5.1). Overall, we identified 496,912 snapshots in 13,583 unique locations on 45,835 profiles.

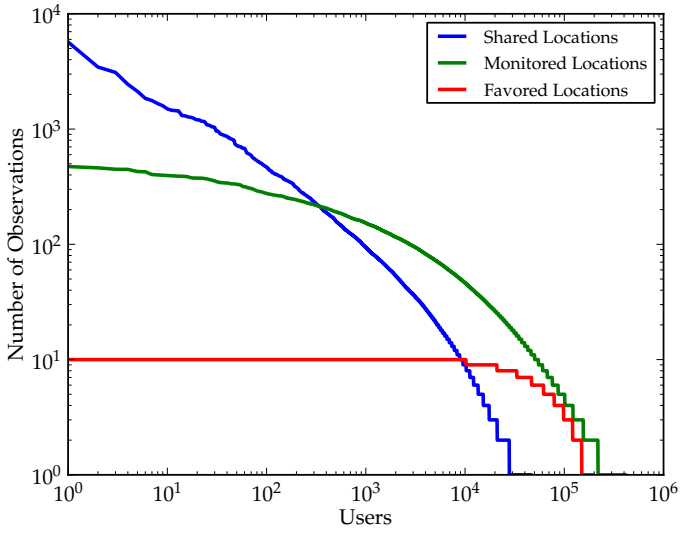


Figure 5.2.: The number of user observations in the three different location-based knowledge sources.

Favoured Locations Every resident of Second Life can specify up to 10 so-called "Picks" on it's profile representing the favourite locations of users. User can enhance these picks with a picture and personal text note. These favoured locations are visible to other users and hence it can be easily accessed with a Web browser using a URI derived from the user's name. We found 191,610 profiles with favoured location information sharing 811,386 locations in 25,311 unique regions.

Figure 6.2 depicts the number of observations of the collected users for the three location sources. Both, Shared and Monitored Locations show power law qualities which is in contrast to the Favoured Locations due to Linden's limitations of 10 picks per user.

5.4. Features

Based on the collected location-based user data we induced overall 10 different features in order to measure the homophily between the users and to predict social interactions between them [Cranshaw et al., 2010; Steurer and Trattner, 2013a,b]. For the reminder of this paper the sequence of observations $O(u)$ of a user u are denoted as 1) $O_m(u)$ for Monitored Locations, 2) $O_s(u)$ for Shared Location, and 3) $O_f(u)$ for Favoured Locations. In contrast, the set of locations where a user was observed is defined as $P(u) = \{\rho \in O(u)\}$. The actual features we used in our experiments are as follows:

Common Locations $R_C(u,v)$ The simplest metric to determine the homophiliy between two users u and v is the number of regions they have visited in common. In particular this can be computed as $R_C(u,v) = |P(u) \cap P(v)|$.

Total Locations $R_T(u,v)$ Analogous to the common regions, one can also define the regions two users have in total and use it as a homophilic feature $R_T(u,v) = |P(u) \cup P(v)|$.

Jaccard's Coefficient $R_{JC}(u,v)$ A combination of the common regions of two users and their total regions is the overlap of locations which is defined as the fraction of common locations and locations visited by both users Cranshaw et al. [2010]. This feature is also known as Jaccard's Coefficient $R_{JC}(u,v) = \frac{|P(u) \cap P(v)|}{|P(u) \cup P(v)|}$.

Location Observations $R_O(u,v)$ Another feature taken from Cranshaw et al. [2010] is the location observations that is similar to the Jaccard's Coefficient between two users. It is computed as the number of locations two users have in common divided by the sum of locations either user have $R_O(u,v) = \frac{|P(u) \cap P(v)|}{|P(u)| + |P(v)|}$.

Location User-Count $R_U(u,v)$ The following three features were first introduced by Cranshaw et al. [2010] and try model the location diversity of regions two users visited in common. The first and most simple feature to include the popularity of a region is the overall number of observations of unique users at a certain region. According to this we calculated the mean user-count $R_{U,\mu}(u,v)$ and the standard deviation of the mean $R_{U,\sigma}(u,v)$ of all regions two users visited in common $P(u) \cap P(v)$.

Location Frequency $R_F(u,v)$ Similar to the previous feature of counting users at a certain location, we can also compute the frequency at the actual location. Again we calculated the mean frequency $R_{F,\mu}(u,v)$ and the according standard deviation $R_{F,\sigma}(u,v)$ of the frequency of regions two users u and v have in common.

Location Entropy $R_E(u,v)$ A refinement of the two previous features, is the entropy that also takes the probabilities of observations at a location L into account. The probability that a user has visited a certain region is defined as the number of observations of the actual users divided by the overall number of observations at this regions. Let $O_{u,L}$ be the observations of a user u at a location L and O_L be all observations at the location L. The probability can then be computed as $prob_L(u) = \frac{|O_{u,L}|}{|O_L|}$. Based on this we can compute the entropy of a certain location L as $E_L =$

$-\sum_{u \in U_L} P_L(u) \cdot log(P_L(u))$ with U_L representing all users observed a the location L. With this definition we computed the mean entropy $R_{E,\mu}(u,v)$ of the locations two users visited in common and the according standard deviation of the mean $R_{E,\sigma}(u,v)$.

5.5. Experimental Setup

Overall, we conducted different kinds of experiments to study the social interactions between users based on the three different sources of location information.

In order to conduct these experiments we created a network from the social interactions obtained from the online social network of Second Life. In this network, nodes represented the users and edges indicate the social interactions between them. These edges were considered as unweighted and so we add an edge between two users no matter how often they communicated with each other. Further we did not distinguish the actual type of interaction and considered text messages, comments and loves equally. This finally yielded in a network of 152,509 users connected by 270,567 edges. Formally this can be written as $G'_O\langle V'_O, E'_O\rangle$ with V'_O representing the users with an interaction on their feed and $e = (u,v) \in E'_O$ if user u interacted with user v (comment, posting, love). Then we enriched the nodes with the observations $O(u)$ from all three location data sources and removed nodes from the network if this data was not available in all three sources. Formally this new network can be defined as $G_O\langle V_O, E_O\rangle$ where $V_O = \{u \mid u \in V'_O,\ u \in O_M,\ u \in O_S,\ u \in O_F\}$ and $e = (u,v) \in E_O$ if user u interacted with user v (comment, posting, love). This reduced the network size to 14,508 nodes and 23,446 edges. For the actual experiments we followed Guha et al. [2004] who suggest to create a balanced set of user-pairs with an interaction and without interaction for the prediction task. In particular we randomly selected 15,000 user-pairs with interaction $\{e^+ = (u,v) | (u,v) \in E_O, u \text{ and } v \in V_O\}$ connecting users V_O^+. The remaining 15,000 edges without interaction in between were created by selecting random user-pairs from the network without interaction $\{e^- = (u,v) | (u,v) \notin E_O, u \text{ and } v \in V_O\}$. Using this network we computed the features described in Section 5.4 for all 30.000 user-pairs and each location source separately. This network-setup implies a baseline of 0.5 for the prediction task in case of random guessing user-pairs with interactions or without interactions.

5.5.1. Analysis of Homophily

In the first experiment we analysed similarities and dissimilarities of user-pairs with interaction e^+ and user-pairs without interaction e^- for each location source separately. We computed the mean values of the features and the according standard error in either sources separately. Using a one-sampled Kolmogorov-Smirnov and a Anderson-Darling test showed that none of the distributions

of the features was from the family of normal distribution. As a consequence and similarly to Bischoff [2012], we compared the variances of all features between interacting and non-interacting user-pairs with a Levene test $(p < 0.01)$. To test for significant differences of the means, we employed Mann-Whitney-Wilcoxon test in case of equal variances and a two-sided Kolmogorov-Smirnov test in case of unequal variances.

5.5.2. Feature Engineering

In order to utilize the supervised machine learning algorithms to predict whether or not a user-pair interacted with each other we had to determine the features that are most suited for this task. To assess the impact of each feature separately we used a simple Collaborative Filtering algorithm for a first rough overview and implemented a method proposed by Liben-Nowell and Kleinberg [2002]: For every user in the network we created ranked lists of the remaining users in the network based on the homophily obtained from the single features. To evaluate the performance of this approach we compared lists with different length to the actual interaction partners of each user. This was computed as the fraction of correctly identified interaction partners divided by the length of the overall retrieved users also referred to as the positive predictive value or precision. To validate the results of this approach we additionally employed the built-in Information Gain and the Correlation-Based Feature Subset Selection of the WEKA learning suite [Hall et al., 2009] to find the most valuable features for supervised learning.

5.5.3. Predicting Social Interactions with Supervised Learning

Based on the most valuable features determined for every region source separately, we tried to predict whether two users have a social interaction in the online social network. We combined features selected by the Correlation-Based Feature Subset Selection for each location source separately and obtained the three feature sets used for supervised learning algorithms. Due to the split into 15,000 user-pairs with interactions and 15,000 user-pairs without interactions we reduced the experiment to a binary classification problem. To compare the different location-based knowledge sources against each other, we applied the WEKA machine learning suite onto the combined set of features obtained with feature engineering for each region source separately. To do so, we applied three learning algorithms: “Logistic Regression” as it can be easily interpreted, and “Random Forest” and “Support Vector Machine” as both of them are suited for high-dimensional data. For the verification of the results provided by the machine learning tool, we used a ten-fold approach for the split of training set and test set.

Table 5.1.: Means and standard errors of features applied to the three sources of location data comparing user-pairs with and without interactions ($^{*}p < 0.1$, $^{**}p < 0.01$, and $^{***}p < 0.001$).

	Features	Have Interactions	Have No Interactions
Monitored Locations	$R_C(u,v)^{***}$	0.49 ± 0.01	0.12 ± 0.00
	$R_{U,\mu}(u,v)^{***}$	179.02 ± 1.49	211.51 ± 2.55
	$R_{U,\sigma}(u,v)^{***}$	188.64 ± 1.17	215.04 ± 1.48
	$R_{E,\mu}(u,v)^{***}$	1.52 ± 0.00	1.60 ± 0.01
	$R_{E,\sigma}(u,v)^{***}$	0.53 ± 0.00	0.56 ± 0.00
	$R_{F,\mu}(u,v)^{***}$	637.50 ± 6.22	755.68 ± 10.96
	$R_{F,\sigma}(u,v)^{***}$	787.40 ± 5.66	894.85 ± 7.71
	$R_{JC}(u,v)^{***}$	0.05 ± 0.00	0.01 ± 0.00
	$R_O(u,v)^{***}$	0.04 ± 0.00	0.01 ± 0.00
	$R_T(u,v)^{***}$	12.24 ± 0.10	10.03 ± 0.07
Shared Locations	$R_C(u,v)^{***}$	1.01 ± 0.02	0.02 ± 0.00
	$R_{U,\mu}(u,v)^{***}$	22.78 ± 0.23	38.10 ± 3.23
	$R_{U,\sigma}(u,v)^{***}$	28.91 ± 0.25	37.09 ± 1.52
	$R_{E,\mu}(u,v)^{**}$	0.80 ± 0.00	0.86 ± 0.02
	$R_{E,\sigma}(u,v)^{***}$	0.44 ± 0.00	0.46 ± 0.01
	$R_{F,\mu}(u,v)^{***}$	92.99 ± 0.82	144.85 ± 11.18
	$R_{F,\sigma}(u,v)^{*}$	160.61 ± 1.14	180.70 ± 7.31
	$R_{JC}(u,v)^{***}$	0.03 ± 0.00	0.00 ± 0.00
	$R_O(u,v)^{***}$	0.02 ± 0.00	0.00 ± 0.00
	$R_T(u,v)^{***}$	63.70 ± 0.59	15.11 ± 0.19
Favoured Locations	$R_C(u,v)^{***}$	0.11 ± 0.00	0.00 ± 0.00
	$R_{U,\mu}(u,v)^{***}$	12.90 ± 0.33	18.57 ± 1.76
	$R_{U,\sigma}(u,v)^{***}$	13.23 ± 0.37	22.26 ± 2.17
	$R_{E,\mu}(u,v)^{**}$	0.71 ± 0.01	0.81 ± 0.03
	$R_{E,\sigma}(u,v)^{***}$	0.40 ± 0.00	0.51 ± 0.02
	$R_{F,\mu}(u,v)^{**}$	16.17 ± 0.37	21.55 ± 1.92
	$R_{F,\sigma}(u,v)^{**}$	15.79 ± 0.40	25.01 ± 2.28
	$R_{JC}(u,v)^{***}$	0.02 ± 0.00	0.00 ± 0.00
	$R_O(u,v)^{***}$	0.02 ± 0.00	0.00 ± 0.00
	$R_T(u,v)^{***}$	8.04 ± 0.03	6.95 ± 0.03

5.6. Results

In this Section we present the results of the conducted experiments.

Table 5.2.: Feature Engineering with Collaborative filtering and the according Information Gain. Highlighted features were derived from Correlation-Based Feature Subset Selection.

	Features	Info Gain	Collaborative Filtering		
			Pre@5	Pre@10	Pre@20
Monitored Locations	$\mathbf{R_C}(\mathbf{u},\mathbf{v})$	0.048	0.081	0.062	0.048
	$R_{U,\mu}(u,v)$	< 0.01	0.047	0.041	0.039
	$R_{U,\sigma}(u,v)$	< 0.01	0.046	0.040	0.037
	$R_{E,\mu}(u,v)$	< 0.01	0.025	0.029	0.029
	$R_{E,\sigma}(u,v)$	< 0.01	0.046	0.037	0.033
	$R_{F,\mu}(u,v)$	< 0.01	0.047	0.043	0.037
	$R_{F,\sigma}(u,v)$	< 0.01	0.046	0.040	0.035
	$\mathbf{R_{JC}}(\mathbf{u},\mathbf{v})$	0.051	0.071	0.063	0.052
	$\mathbf{R_O}(\mathbf{u},\mathbf{v})$	0.051	0.071	0.063	0.052
	$\mathbf{R_T}(\mathbf{u},\mathbf{v})$	0.012	0.077	0.043	0.023
Shared Locations	$R_C(u,v)$	0.211	0.280	0.252	0.208
	$R_{U,\mu}(u,v)$	< 0.01	0.133	0.119	0.104
	$R_{U,\sigma}(u,v)$	< 0.01	0.185	0.161	0.137
	$R_{E,\mu}(u,v)$	< 0.01	0.122	0.089	0.074
	$R_{E,\sigma}(u,v)$	< 0.01	0.192	0.164	0.129
	$R_{F,\mu}(u,v)$	< 0.01	0.115	0.099	0.091
	$R_{F,\sigma}(u,v)$	< 0.01	0.109	0.108	0.101
	$\mathbf{R_{JC}}(\mathbf{u},\mathbf{v})$	0.208	0.221	0.187	0.157
	$\mathbf{R_O}(\mathbf{u},\mathbf{v})$	0.208	0.221	0.187	0.157
	$\mathbf{R_T}(\mathbf{u},\mathbf{v})$	0.234	0.159	0.121	0.107
Favoured Locations	$\mathbf{R_C}(\mathbf{u},\mathbf{v})$	0.040	0.104	0.085	0.060
	$R_{U,\mu}(u,v)$	< 0.01	0.079	0.075	0.055
	$R_{U,\sigma}(u,v)$	< 0.01	0.082	0.074	0.058
	$R_{E,\mu}(u,v)$	< 0.01	0.082	0.076	0.056
	$R_{E,\sigma}(u,v)$	< 0.01	0.086	0.077	0.057
	$R_{F,\mu}(u,v)$	< 0.01	0.081	0.075	0.055
	$R_{F,\sigma}(u,v)$	< 0.01	0.074	0.071	0.056
	$\mathbf{R_{JC}}(\mathbf{u},\mathbf{v})$	0.040	0.108	0.086	0.059
	$R_O(u,v)$	0.040	0.108	0.086	0.059
	$\mathbf{R_T}(\mathbf{u},\mathbf{v})$	0.020	0.002	0.002	0.003

5.6.1. Analysis of Homophily

We computed the mean values and standard errors for all features of 15,000 user-pairs with interactions and 15,000 user-pairs without interactions in the online social network. Table 5.1 summarizes the differences for features applied to all three sources of location-based information.

5.6.1.1. Monitored Locations

On average user-pairs with interaction could be found in 0.5 common regions $R_C(u,v)$, had over 12 total regions $R_T(u,v)$, and Jaccard's Coefficient $R_{JC}(u,v)$ and observations $R_O(u,v)$ of around 0.05. For user-pairs with interaction we furthermore found an average user count $R_{U,\mu}(u,v)$ of over 179, an entropy $R_{E,\mu}(u,v)$ of 1.52 and a user frequency $R_{F,\mu}(u,v)$ of 637 for commonly visited regions. For user-pairs without interaction we observed less commonly visited regions and total regions as well as Jaccard's Coefficient and observations. In contrast, for features based on the location diversity, i.e. entropy, frequency, and user-count, we observed higher values. With the tests described in Section 5.5 we found significant differences for all applied features.

5.6.1.2. Shared Locations

The characteristics for the features applied to the Shared Locations were similar to the features applied to the Monitored Locations. For user-pairs with interaction we observed around 1 common region $R_C(u,v)$, 63 total regions $R_T(u,v)$, and a Jaccard's Coefficient $R_{JC}(u,v)$ and observations $R_O(u,v)$ in the same regions of around 0.03. For common regions we observed a average user-count $R_{U,\mu}(u,v)$ of 22, region entropy $R_{E,\mu}(u,v)$ of 0.8, and region frequency $R_{F,\mu}(u,v)$ of 92. Similar to the Monitored Locations dataset we observed higher values for common regions, Jaccard's Coefficient, observations, and total regions for user-pairs with interaction, whereas frequency, user-count and entropy were lower.

5.6.1.3. Favoured Locations

Again we observed similar results as already described for the previous locations dataset but due to the reduced number of picks per user the absolute values were lower. We observed 0.11 common regions $R_C(u,v)$ for users interacting with each other, respectively 0.02 for observations $R_O(u,v)$ and Jaccard's Coefficient $R_{JC}(u,v)$. In contrast these values were nearly 0 for user-pairs without interaction. Interacting users had around 8 total regions $R_T(u,v)$ whereas user-pairs without interaction had only around 7 total regions. For features that model the location diversity ($R_E(u,v)$, $R_F(u,v)$, $R_U(u,v)$) we again observed lower values for users interacting with each other if compared to users without interaction.

5.6.2. Feature Engineering

For a rough estimation of the predictability of interactions we employed a Collaborative Filtering algorithm using features applied to the three location-based knowledge sources. Previous results of

the analysis of homophily showed that user-pairs with interactions had higher values for common regions, total regions, Jaccard's Coefficient and observations. Hence, we rank this features in this experiment in descending order. Contrary, features based on the location diversity ($R_E(u,v)$, $R_F(u,v)$, $R_U(u,v)$) showed significantly lower values for interacting user-pairs and so we ranked them in ascending order. In addition to Collaborative Filtering, we used WEKA's Information Gain algorithm for verification of these results and finally a Correlation-Based Feature Subset Selection to find valuable features for further prediction. In Table 5.2 we present the results of Collaborative Filtering and the according values of the Information Gain algorithm for the features applied to the three location sources.

5.6.2.1. Monitored Locations

The Collaborative Filtering approach unveiled the common regions $R_C(u,v)$, total regions $R_T(u,v)$, respectively Jaccard' Coefficient $R_{JC}(u,v)$ and common observations $R_O(u,v)$ for different list lengths as most valuable. However, features modeling location diversity like user-count, entropy, frequency of user's common regions performed inferior. This results were inline with the Information Gain algorithm that showed similar results for the computed features. Additionally, Correlation-Based Feature Subset Selection identified these features as most valuable.

5.6.2.2. Shared Locations

Collaborative Filtering exposed common region $R_C(u,v)$, Jaccard's Coefficient $R_{JC}(u,v)$, and observations $R_O(u,v)$ as most valuable. These three features plus the total number of regions $R_T(u,v)$ were also identified as best features using the Information Gain algorithm. Similarily, the Correlation-Based Feature Subset Selection algorithm unveiled Jaccard's Coefficient $R_{JC}(u,v)$, common observations $R_O(u,v)$, and the total number of regions $R_T(u,v)$ as the most valuable features in the set.

5.6.2.3. Favoured Locations

Similar to the previous result the Collaborative Filtering approach identified the common regions $R_C(u,v)$, Jaccard's Coefficient $R_{JC}(u,v)$ and common observations $R_O(u,v)$ as most valuable. Information Gain additionally puts the total number of regions $R_T(u,v)$ on the list which is also inline with the previous result. Finally, Correlation-Based Feature Subset Selection found common regions $R_C(u,v)$, Jaccard's Coefficient $R_{JC}(u,v)$ and the total number of regions $R_T(u,v)$ to be best suited for further prediction tasks.

Table 5.3.: Predicting Interactions between user-pairs with supervised learning based on combined features of different location sources.

Feature Set	Logistic	SVM	Random Forest
Monitored Locations	**0.632**	0.605	0.618
Shared Locations	**0.849**	0.791	0.846
Favoured Locations	**0.630**	0.593	0.628

5.6.3. Predicting Social Interactions

Based on the results of the previous experiment we used features identified by Correlation-Based Feature Subset Selection for predicting whether two users have an interaction with each other or not. One can find these features highlighted in bold letters in Table 5.2 for different region sources. We combined these individual features to feature-sets for every location source separately and predicted the interaction between user-pairs with three different learning algorithms. We utilized *Logistic Regression*, *Support Vector Machine* (SVM), and *Random Forest* and used the Area under the ROC curve (AUC) as main evaluation metric. In Table 5.3 the results of these evaluations are shown and one can see that *Logistic Regression* outperforms the two remaining algorithms on each of the three location-datasets. In particular, we found that the feature-set applied to the Shared Location dataset predicted interactions between users with 0.849 AUC which is a boost of +34.9% if compared to baseline for random guessing. For the remaining two region sources we observed a predictability of around 0.63 which is +13% over baseline. Random Forest and SVM showed similar results but performed inferior.

5.7. Discussion and Conclusion

In this paper we have harvested data from different sources of the virtual world of Second Life: First we collected social interaction data between users from the online social network *My Second Life* and second, we collected data from three different and independent location sources, i.e. locations monitored while users were attending events, locations they share, and their favourite locations. For every single location source we computed 10 features representing the homophily between user-pairs and employed them to predict whether two users had social interaction with each other or not. This section concludes the paper and tries to give answers to the research questions from Section 5.1 and provides possible explanations for the results derived from the conducted experiments.

- *RQ1:* To answer the first research question, we evaluated the differences between user-pairs that had an interaction in the online social network and user-pairs without this interaction.

This analysis revealed statistically significant differences for nearly all features: User-pairs with interactions on average visited more common regions and had more common observations together. In contrast to this, they visited regions with a lower user-count, frequency, and entropy which can be interpreted as sign of intimacy: Users with interactions already know each other and therefore they meet in places that are less frequented by other users. We could observe this for all three data sources but due to the diverse datasets the characteristics were different: the Shared Locations dataset showed more distinct tendencies than, for instance the picks dataset with the limit of 10 picks per user.

- *RQ2:* To answer the second research question we employed Collaborative Filtering to predict the social interactions between the users based on 10 different features independently across all location sources. We found that the most valuable features over all the location-based knowledge sources were the number of common regions $R_C(u,v)$, the Jaccard's Coefficient $R_{JC}(u,v)$, and the total number of regions of two users $R_T(u,v)$. Although these characteristics were similar over all sources, we observed differences in the Information Gain. Features applied to the Shared Locations seemed best suited for predicting interactions as the Information Gain was higher if compared to Favoured or Monitored Locations.

- *RQ3:* Considering the Information Gain of features applied to the three location sources, we already had the premonition that data obtained from a user's Shared Locations has the highest potential to predict interactions. Indeed, a detailed look at the combined feature sets to predict interactions unveiled that the this dataset worked best among all sources. We believe that this is for the following two reasons: First, users can share message from everywhere within the virtual world over their social network and the data collection approach does not miss any data. Second, users explicitly share locations and places they like and spend time in. Other users that visit their profiles because they already know each other, see these locations, and also visit them. This can be seen as an explicit promotion of Shared Locations of a user. We believe that Monitored Location data performed inferior as we only have a clipping of the actual user's visited regions due to limited resources. A similar explanation can be made for the picks data source but here the limiting factor was not the crawling resources but the restriction to 10 picks per user. Overall, the three different learning algorithms applied to the datasets were stable and show similar results over all three sources – Logistic Regression showed the best results whereas Support Vector Machine and Random Forrest were inferior.

For future work we plan to also incorporate the number of social interactions in our predictive model to better account for the strengths of social ties between the users. Furthermore, we plan to account for the variation of time which we did not consider in this paper.

References

M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *SDM'06: Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

K. Bischoff. We love rock'n'roll: analyzing and predicting friendship links in Last. fm. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 47–56. ACM, 2012.

J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 119–128. ACM, 2010.

R. Guha, R. Kumar, P. Raghavan, and A. Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, pages 403–412. ACM, 2004.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2002.

C. X. Ling, J. Huang, and H. Zhang. AUC: a statistically consistent and more discriminating measure than accuracy. In *International Joint Conference on Artificial Intelligence*, pages 519–526. LAWRENCE ERLBAUM ASSOCIATES LTD, 2003.

S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *Proceedings of ICWSM*, 11:329–336, 2011a.

S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM, 2011b.

M. Steurer and C. Trattner. Predicting interactions in online social networks: an experiment in second life. In *Proceedings of the 4th International Workshop on Modeling Social Media*, MSM '13, pages 5:1–5:8, New York, NY, USA, 2013a. ACM. ISBN 978-1-4503-2007-8. doi: 10.1145/2463656.2463661. URL `http://doi.acm.org/10.1145/2463656.2463661`.

M. Steurer and C. Trattner. Acquaintance or partner?: Predicting partnership in online and location-based social networks. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 372–379, New York, NY, USA, 2013b. ACM. ISBN 978-1-4503-2240-9. doi: 10.1145/2492517.2492562. URL `http://doi.acm.org/10.1145/2492517.2492562`.

M. Steurer, C. Trattner, and F. Kappe. Success factors of events in virtual worlds a case study in second life. In *Network and Systems Support for Games (NetGames), 2012 11th Annual Workshop on*, pages 1–2. IEEE, 2012.

M. Steurer, C. Trattner, and D. Helic. Predicting social interactions from different sources of location-based knowledge. In *SOTICS 2013, The Third International Conference on Social Eco-Informatics*, pages 8–13, 2013.

M. Thelwall. Homophily in myspace. *Journal of the American Society for Information Science and Technology*, 60(2):219–231, 2009.

D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108, New York, NY, USA, 2011. ACM.

CHAPTER 6

Prediction of Partnership

This chapter is based on the paper *Acquaintance or Partner? Predicting Partnership in Online and Location-based Social Networks* published and presented at the *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* and *Predicting Partnership with Location-based and Online Social Network Data* submitted to the *Elsevier Journal of Neurocomputing* [Steurer and Trattner, 2013c].

In this chapter we evaluate the tie-strength between user-pairs – defined as partners or acquaintances – in an online social network and different location-based social networks. We differentiate between time-dependent and time-independent data sources and compute features that measure the relation between users. We apply these features to the data sources and evaluate them regarding the differences in tie-strength. Based on this analysis we use two approaches to predict the tie-strength between users. First, we employ various supervised learning algorithms to predict the tie-strength and in the second step we substantiate these findings using an unsupervised Collaborative Filtering approach.

In detail the chapter is structured as follows: In Section 6.2 we discuss related work. In Section 6.3 we shortly introduce the dataset used for our experiments. In Section 6.4 we outline the set of features used for our experiments described in Section 6.5. Section 6.6 presents the results of our study. Finally, Section 6.7 discusses our findings and concludes the chapter.

Abstract

Existing approaches to predict tie strength between users cover either online social networks or location-based social networks. However, there are no studies that combine these networks to unveil information about the intensity of the social relations between users. In this research paper we analyze aspects of tie strength -– defined as partners and acquaintances — in an online social network for residents of Second Life supported by location-based data obtained from three different sources. We compare user pairs according to their partnership and reveal significant differences between partners and acquaintances. Following these observations, we evaluate the social proximity of users with supervised and unsupervised learning algorithms and identified homophilic features as most valuable for the prediction of partnership.

6.1. Introduction

Social networks contain useful information about the relation between their participants and the understanding of their characteristics is a prerequisite to interpret social dynamics [Coleman, 1988]. The advent of online social networks afforded large-scale data and topological network features were complemented by homophilic features that model the alikeness of users in a network. Nevertheless, the social proximity between users in the real world is not only driven by online social networks but also by their mobility patterns. The availability of such data changed with the arrival of GPS aware mobile phones and location-based social network platforms like FourSquare, and opened new information sources.

As not all links in a network are equal, it is not sufficient to consider them merely as loosely coupled. Granovetter [1973] introduced the term “tie strength” to model the intensity of a user relation and proposed the overlap of neighbors derived from the network topology as an indicator. A considerable body of research investigating tie strength focused on either online social networks [Gilbert, 2012; Gilbert and Karahalios, 2009; Zheleva et al., 2010] or location-based social networks [Wang et al., 2011; Choi et al., 2013; Bischoff, 2012]. However, there are only few studies that combine both domains [Pan et al., 2011; Volkovich et al., 2012]. To further investigate into this combination, in this paper we analyze the tie strength between users – defined as *partners* and *acquaintances* - with social proximity features derived from an online social network and location-based data obtained from three distinct sources: monitored locations, shared locations and favoured locations. To the best of our knowledge there are neither studies that aim at tie strength prediction using the combined data from an online domain and a location domain nor studies that compare different sources of location-based data in terms of the tie strength.

Since it is nearly impossible to collect large-scale social network and position data of the same users in a real-world scenario, we obtained datasets for the experiments from the virtual world of Second Life. Text-based interactions between residents (posts, comments, loves) and profile data (affiliated groups, specified interests, partnership information) were harvested from a Facebook-like online social network called "My Second Life". In addition to text-interaction data, these profiles also allow users to post pictures with attached location information that are visible to others, i.e. Shared Locations (see Figure 6.1(a)), and specify the top-10 locations they like to visit, i.e. Favoured Locations (see Figure 6.1(b)). Additionally, we monitored position and mobility patterns of users that are attending events in the virtual environment over a period of 12 months. We computed social proximity features based on user interactions and location information to model the user relations and answer the following research questions:

- *RQ1:* To what extent do partners and acquaintances differ from each other with respect to social proximity features induced from an online social network and three different sources of location-based data?
- *RQ2:* How well can we predict the partnership between users with social proximity features derived from the online social network and the different location-based data sources?
- *RQ3:* Which social proximity features across all domains offer the highest information gain and the highest accuracy for the prediction of partnership between users?
- *RQ4:* To what extent does the available time information in the location-based data support the computation of social proximity features and affect the prediction of partnership?

Based on these questions, we conducted a number of experiments using statistical methods and supervised respectively unsupervised learning approaches with the following results: A statistical analysis of the studied features showed that significant differences existed between partners and acquaintances. For instance, we discovered that partners were less interested in exploring new locations and they meet at less frequented locations compared to acquaintances. This goes in line with the observation that the number of text interactions and the average spatial distance between users shows signs of intimate contact. The learning algorithms identified time-dependent features, such as attended events and the spatial distance between users to be most valuable with regard to partnership prediction. Our experiments further showed that the combination of features from both domains (=the online social network merged with location-based data) outperformed the features of either domains.

The major contributions of this paper are as follows: 1) The introduction of a novel large-scale dataset that incorporates online social network data and three location-based datasets of the same users. 2) The analysis of a large set of social proximity features from an online social network supported by location-based datasets to predict the partnership between users.

6.2. Related Work

Relevant related work in this area can broadly be divided into the following two areas: Predicting links and predicting tie strength in online and location-based social networks.

6.2.1. Predicting links in Online and Location-based Social Networks

Liben-Nowell and Kleinberg [2002] formalized the problem of predicting new links in a network and developed an approach based on the topology of the network. They used information about direct neighbors and employed the ensemble of all paths from one user to another. This approach yielded in significantly better predictability of new links compared to a random approach. Computationally efficient methods of this structure-centric approach were evaluated by Fire et al. [2011]. Surprisingly, using only topological features they could successfully find new links that evolve within two hops in the network. However, topological features can only be applied if the structure of the actual network is known. If this is not the case, homophilic features such as a metric for the likeness of two users can be used instead. In their work Thelwall [2009] investigated the social network of *MySpace* using homophilic features. Their studies revealed a significant homophily of origin, marital status and the sexual orientation in existing links. Further they uncovered that a friendship in an online social network could even reflect an offline friendship. While these papers were of analytical nature, Mislove et al. [2010] extended the known attributes of a few users in a network to learn about other users. They used Facebook datasets with educational data and region information as attributes and found that one could infer the attributes of 80% of users from the remaining 20% with an accuracy of 80%. Rowe et al. [2012] combined topological features and homophilic features in the Chinese microblogging service *Tencent Weibo*. In their work they predict network links and show that homophilic features do not only significantly outperform a random baseline but also topological features.

Scientific work for the link prediction problem was mainly done for online social networks but as more and more position data became available, the combination of both worlds was investigated too. One popular work in this respect is for instance a study of Cranshaw et al. [2010] who examined data of the Facebook application *Locaccino* and analyzed the offline mobility data of 489 users. They used the position data, separated it into two categories with topological and homophilic information and tried to predict the online links with the position information. Homophilic data obtained from the location was identified as valuable information but in combination with topological features it performed even better. Scellato et al. [2011b,a] also revealed the importance of place related features and identified 30% of new links in the *Gowalla* network as place friends and 40% of all links within a range of 100 kilometers. Further they uncovered a

weak correlation between the number of friends and their spatial distance. Noulas et al. [2012] used this fact to predict venues of users in *Gowalla* network. They also achieved best results when combining information from social ties and visited venues.

6.2.2. Predicting Tie Strength in Online and Location-based Social Networks

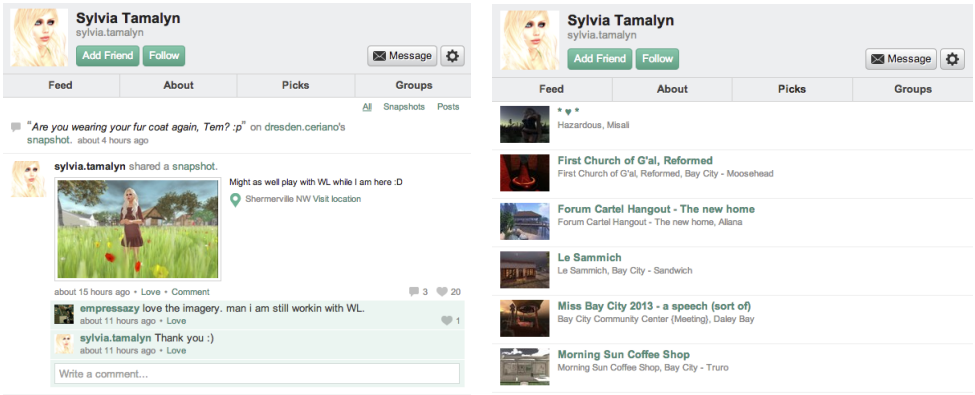
Not all links between nodes in a network are equal and Granovetter [1973] tightened the definition by introducing the *tie strength* of connections. Studies by Gilbert and Karahalios [2009]; Gilbert [2012] used Facebook to investigate tie strength between online friends. They combined communication features, topological information and the social distance, and correlated it with 2,000 users who specified their real friends in interviews and questionnaires. They could predict different tie strength with an accuracy of 85% and notably, they were even able to transfer information about the tie strength of two users from one social network to the other, i.e. from Facebook to Twitter.

While the tie strength between users in online social networks was extensively investigated, few studies combined offline networks and tie strength. The first to mention study is by Wang et al. [2011] who collected the mobile phone data of 6 Million users and measured the tie strength according to the number of calls between user pairs. They found that although new links could be predicted via mobility measures alone, combining them with topological information in the network yielded even better results. With regard to tie strength, they found that its correlation with the user mobility traces and social proximity was weak. This is in agreement with Choi et al. [2013] who analyzed communication patterns and indoor mobility tracking of 22 office mates. To define tie strength, the users formal and informal contacts were differentiated in their study. Via a supervised learning approach, they could identify links with an accuracy of 85%. Finally, the last to mention paper in this respect is the work of Bischoff [2012]. In her work the author studied the social network *Last.fm* in respect to the geo-location of user and attended events. Tie strength was defined as the number of commonly attended events, and online communications and music tastes were used to predict it. The results were in agreement with previous works, confirming a correlation between tie strength and data from the online social network.

6.3. Datasets

We conducted our experiments using online social network data obtained from *My Second Life*[*] and location-based data from three different sources: 1) Shared Locations, 2) Favoured Locations,

[*] http://my.secondlife.com/



(a) Text-Interactions and Shared Locations. (b) Favoured Locations.

Figure 6.1.: Users of Second Life can share text messages (posts, comments, and loves), location information using snapshots, and up to 10 so-called "Picks" that represent their favourite locations in Second Life.

and 3) Monitored Locations. There were several reasons for choosing Second Life as a fundament for our experiments: First, unlike networks such as Facebook, My Second Life does not restrict extensive crawling of the users profiles. Second, in contrast to real-world online social networks, most of the profiles in My Second Life are public, i.e. a large fraction of the network can be mined. Third, it is possible to automatically harvest location-based datasets of the same users from different sources at a large scale. In this section we describe the data collection process in order to conduct the experiments (see [Steurer et al., 2012; Steurer and Trattner, 2013b,c]).

6.3.1. Online Social Network Dataset

In 2007 Linden Labs introduced the online social network platform *My Second Life* which is similar to Facebook or Google+. The target group are residents of Second Life to share text messages, comments or loves (similar to Facebook's "Likes"). Users of Second Life automatically have a profile without additional registration and by default these profiles are opened for public access. In contrast to Facebook, there is no mutual friendship confirmation between users and every user can post onto the Feed of each other (similarly to Facebook's "Wall") without their explicit permission. Besides the interactions with others, users can enhance their profiles and describe themselves with a biography, interests, and their partnership status. Users can even marry in Second Life but a wedding is not free of charge and costs 10 Linden Dollars (Linden Dollar is the virtual currency used in Second Life – 1 US Dollar equals approximately 258 Linden Dollars). Though, nothing is forever and cancelling this partnership costs 25 Linden Dollars.

To harvest this data, we attempted to download the interaction data and profile information of

residents of Second Life with groups, interests, and partner. We extracted the user names of the interaction partners. Overall, we downloaded the profile data of 152,509 users with interactions on their walls and identified 1,084,002 postings, 459,734 comments, 1,631,568 loves and 285,528 unique groups. On average users joined 15.61 groups specified 6.5 interests. 39,936 users were in a partnership which resulted in 18,468 couples in the whole dataset. Formally, this network is defined as $G_O\langle V_O, E_O\rangle$, with V_O representing the users with interactions on their Feed, and $e = (u,v) \in E_O$ if user u interacted with user v (posting, comment, love). In Table 6.1 we present the basic properties of the network.

6.3.2. Location-Based Datasets

For the prediction task of a user-pair's relationship we employed location information of users obtained from three different location sources. *Shared Locations* and *Favoured Locations* could be extracted from the user profiles on the web-platform My Second Life, whereas *Monitored Locations* were collected directly in Second Life.

Shared Locations: Besides text-interactions, users can also share pictures, i.e. snapshots, of the virtual world on the Web-platform of My Second Life. Similar to applications like "FourSquare" or "Flickr" these pictures can be enriched with position information of virtual places. The position information consists of a *region* name (Second Life is parcelled into squared regions with 256x256 meters) and the accurate coordinates within this region. An example of a snapshot with attached position information can be found in Figure 6.1(a) with a check-in at region "Shermerville NW" – the actual coordinates within this regions are hidden in the meta-data. One can see that the time stamp of the actual check-in in this picture is not accurate as it is stated as "1 day ago". The older a posting is, the imprecisely this time stamp is, e.g. "a few days ago", "about 2 months ago", or "over a year ago". A deeper inspection of these pictures reveals that they are stored in Amazon's CloudFront network which is a web service to "distribute content to users with low latency, high data transfer speeds, and no commitments"[†]. The response of a simple HTTP *Head* request to the URI of the shared pictures unveils information about the storage location, the cache-control, and a time-stamp called "last-modified". Surprisingly, this time stamp correlates with the displayed inaccurate time stamp in the snapshot and turned out as the actual time stamp when the picture was shared. Using this method we could overall harvest 496,912 snapshots with accurate time stamps in 13,583 unique regions from 45,835 user profiles.

Similar to the online social network, we create a network using the time and place information of the shared snapshots. Whenever two users shared a picture with location information from a certain inworld location concurrently, we assume that they were somehow related to each other.

[†] aws.amazon.com/cloudfront/

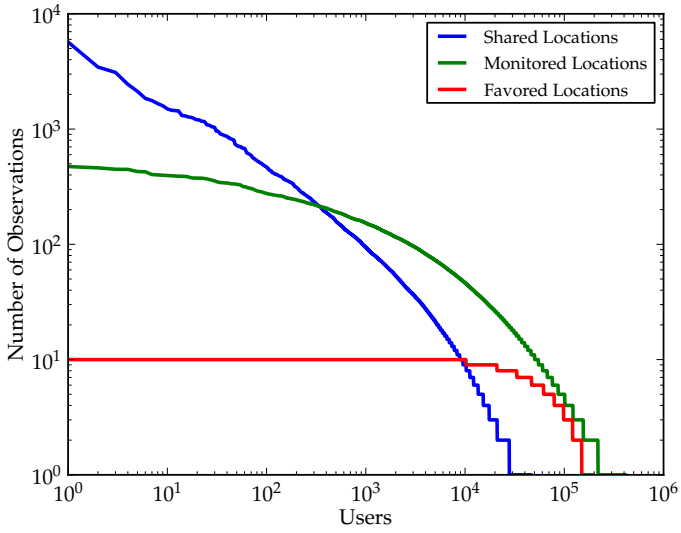


Figure 6.2.: The number of user observations in the three different location-based sources.

Hence, we create a network where users are represented as nodes and two users have an edge between them if they shared a picture from the same region within a time-interval of 2 hours. This approach yields in a shared locations network $G_S\langle V_S, E_S\rangle$ with 13,099 users connected by 23,251 edges (see Table 6.1 for further details).

Favoured Locations: Users of Second Life can specify up to 10 so-called "Picks" on their web profile that represent their favourite locations in Second Life. Besides the general description of the location and the accurate position (again with region name and coordinates within this region), users can enhance these picks with a picture and a personal text note. These favoured locations are visible to other users and hence they can be easily accessed with a Web browser and a Web crawler. Using this crawler, we automatically collected 191,610 profiles sharing 811,386 locations in 25,311 unique regions. Due to the nature of this kind of location data we are not able to fetch any time information from this source as it simply does not exist. As a consequence, it is further not possible to create a network structure of users as described in the previous section.

Monitored Locations: Similar to the real life, residents of Second Life can host events and announce them to the public[‡]. Users log into the Second Life Web page and create new events with *name*, *description*, *location* and *start time* and assign them to one out of ten predefined categories, e.g. *Nightlife* or *Live Music*. Further, events have three maturity ratings that depend on the rating of the location: *General* without any age restrictions, *Mature* with users at least 16 years old, and finally *Adult* accessible only for grown-up users.

[‡]`http://secondlife.com/community/events/`

One of the many advantages of using Second Life as testbed for our experiments is the fact that all events are announced to the general public on the Second Life website – called the Second Life event calendar. In order to harvest this kind of data we implemented a simple Web-based crawler which collected all relevant event information on a daily basis. Starting in March 2012, we collected 262,234 events over a period of 12 months [Steurer et al., 2012].

To participate in the virtual world, users register with Second Life, download the client software from Linden Labs and log in. Among other third party clients, *libopenmetaverse*[§] is an open-source client for the command-line to enter the environment. It can be run as a server process and the functionality can be easily enhanced due to the modular design. We added new capabilities to automatically move around in the virtual world and to collect information of surrounding users. These user-bots were controlled by a centralized server-instance that sent them to places with ongoing events. On average a bot needed 1 minute to move to a new location and collect the position data of surrounding users. To speed up the collection process and to visit more events concurrently, we employed a pool of 15 bots that alternately visited events. The collected information comprised user names, accurate position of the observed users, and a time stamp. Overall, we collected nearly 19 million data samples of 410,619 different users in 4,105 different locations from in-world Second Life.

The naive approach to create a network out of this huge amount of data would have been to inter-link two users with each other whenever they met. Since this would yield in a network with billions of edges, we applied a simple heuristic to prune our data. In particular, we only inter-linked two users with each other in the network if they were seen concurrently in the same region on two different days. With this simple approach at hand we were able to reduce the number of edges to 4,473,739. Formally we define this network as graph $G_M\langle V_M, E_M\rangle$ with V_M representing the users in the network and $e = (u,v) \in E_M$ if users u and v were concurrently observed in the same region on two different days. In Table 6.1 we present the basic properties of the network.

Figure 6.2 depicts the number of observations of the collected users for the three location sources. Both, Shared- and Monitored Locations show power law qualities which is in contrast to the Favored Locations due to Linden's limitations of 10 picks per user.

6.4. Feature Description

As already outlined in the introductory part of this paper, it is our aim to study the extent to which partnership between users can be predicted based on data from two different domains – online

[§] http://lib.openmetaverse.org/

Table 6.1.: Basic metrics of the used networks and their combination used for the experiments.

Name	Online G_O	Monitored G_M	Shared G_S	Combined G
Type	directed	undirected	undirected	directed
#Nodes	152,509	156,844	13,099	44,603
#Edges	270,567	4,473,739	23,251	1,419,543
Degree	3.54	57.05	3.55	63.65

social network data and location-based data. In this section we induce the features to model the relationship between users for these sets of data [Rowe et al., 2012; Steurer and Trattner, 2013b,a].

6.4.1. Online Social Network Features

Based on the interactions between users we created a social network, where users participating in this network were represented as nodes and the interactions between these users were modelled by edges. Using the structural information of this network and the information about the group affiliation the the specified interests we can compute topological and homophilic features as follows:

Topological Features We defined the neighbors of a user u in the network with respect to the communication direction: Neighbors that received a message from user u were denoted as $\Gamma(u)^+ = \{v \mid (u,v) \in E_O\}$ and neighbors that send messages to user u as $\Gamma(u)^- = \{v \mid (v,u) \in E_O\}$. The first and most simple measure was the number of common friends a pair of users had. Due to the different definitions of neighbors, we defined the common outgoing neighbors as $O^+_{CN}(u,v) = |\Gamma^+(u) \cap \Gamma^+(v)|$ and the common incoming neighbors as $O^-_{CN}(u,v) = |\Gamma^-(u) \cap \Gamma^-(v)|$. Similar to the common neighbors a pair of users has, we can also define the total number of neighbors a user-pair has. Analogous, we define the total number of outgoing neighbors as $O^+_T(u,v) = |\Gamma^+(u) \cup \Gamma^+(v)|$ and the total number of incoming neighbors as $O^-_{CN}(u,v) = |\Gamma^-(u) \cap \Gamma^-(v)|$. A simple combination of these two features is Jaccard's Coefficient and can be seen as a measure for exclusiveness of this relation. Again, we split it into two features $O^+_{JC}(u,v) = \frac{|\Gamma^+(u) \cap \Gamma^+(v)|}{|\Gamma^+(u) \cup \Gamma^+(v)|}$ and $O^-_{JC}(u,v) = \frac{|\Gamma^-(u) \cap \Gamma^-(v)|}{|\Gamma^-(u) \cup \Gamma^-(v)|}$.

In their paper Cheng et al. [2011] investigate in the reciprocity of user communication in a directed network and to take this bidirectional communication into account, we computed $O_R(u,v) = 1$ if $(u,v) \in E_O, (v,u) \in E_O$ and $O_R(u,v) = 0$ if $(u,v) \in E_O, (v,u) \notin E_O$. Furthermore they proposed a modification to the Adamic-Adar measure for directed networks which can be written as $O^-_{AA}(u,v) = \sum_{z \in \Gamma^-(u) \cap \Gamma^-(v)} \frac{1}{log(|\Gamma^-(z)|)}$.

"Preferential Attachment Score" takes the level of activity into account and due to the directed structure used $O^{+}_{PS}(u,v) = |\Gamma^{+}(u)| \cdot |\Gamma^{+}(v)|$, and one for received-message neighbors $O^{-}_{PS}(u,v) = |\Gamma^{-}(u)| \cdot |\Gamma^{-}(v)|$.

Homophilic Features In contrast to the topological features in the previous section, homophilic features directly represent the alikeness of user-pairs. These features do not dependent on their direct neighbors in the network or the structure of the network per se because they are only based on properties of either nodes. These features have been identified as valuable resource for link prediction in several studies [Thelwall, 2009; Mislove et al., 2010; Scellato et al., 2011a].

Users of Second Life can join groups and specify interests on their profiles to state their opinions. The structure of the data is quite similar for interests and groups, so we could apply the same mechanisms to indicate the similarity between a pair of users. Formally, we defined the groups of a user u as $\Delta(u)$ and the specified interests as $\Psi(u)$. For each pair of users in the network we defined the common interests and the common groups they share: $G_C(u,v) = |\Delta(u) \cap \Delta(v)|$, respectively $I_C(u,v) = |\Psi(u) \cap \Psi(v)|$. Further, we computed Jaccard's Coefficient to take the total number of groups and interests into account: $G_{JC}(u,v) = \frac{|\Delta(u) \cap \Delta(v)|}{|\Delta(u) \cup \Delta(v)|}$, respectively $I_{JC}(u,v) = \frac{|\Psi(u) \cap \Psi(v)|}{|\Psi(u) \cup \Psi(v)|}$. Users can share text messages, comments, or loves with others and the intensity of this communication could be an indicator of their partnership. As a consequence we measured the number of occurrences for each type of interaction and summed it up for the overall number of interactions between users. We defined $P_P(u,v)$ as the number of text messages, $P_C(u,v)$ as the number of comments, $P_L(u,v)$ as the number of loves, and $P_I(u,v) = P_P(u,v) + P_C(u,v) + P_L(u,v)$ as the number of interactions between user u and v.

Another measure for the proximity of users is the average message length of all interactions between them. Hence, we computed the average message length $P_A(u,v)$ as concatenation of all postings and comments between user u and user v and divided it by their quantity.

6.4.2. Location-Based Features

In Section 6.3.2 we have introduced the three different sets of location information. Based on the obtained location data of users we can compute features that are either *time-independent*, i.e. which places users visited, or *time-dependent*, i.e. when users visited which places.

6.4.2.1. Time-Independent Features

In this feature set we consider the places users visited without respect to time and induced overall 10 different features to measure time-independent relations between them [Cranshaw et al., 2010;

Steurer et al., 2013; Steurer and Trattner, 2013c].

For the reminder of this paper the sequence of observations $O(u)$ of a user u are denoted as 1) $O_s(u)$ for Shared Location, 2) $O_f(u)$ for Favoured Locations, and 3) $O_m(u)$ for Monitored Locations. In contrast, the set of locations where a user was observed is defined as $P(u) = \{\rho \in O(u)\}$. The actual features we used in our experiments are as follows: The simplest metric to determine the homophily between two users u and v is the number of regions they have visited in common. In particular this can be computed as $R_C(u,v) = |P(u) \cap P(v)|$. Analogous to the common regions, one can also define the regions two users have in total and use it as a homophilic feature $R_T(u,v) = |P(u) \cup P(v)|$. A combination of the common regions $R_C(u,v)$ of two users and their total regions $R_T(u,v)$ is the overlap of locations which is defined as the fraction of common locations and locations visited by both users [Cranshaw et al., 2010]. This feature is also known as Jaccard's Coefficient $R_{JC}(u,v) = \frac{|P(u) \cap P(v)|}{|P(u) \cup P(v)|}$.

A feature taken from Cranshaw et al. [2010] is the location observations that is similar to the Jaccard's Coefficient between two users. It is computed as the number of locations two users have in common divided by the sum of locations either user have $R_O(u,v) = \frac{|P(u) \cap P(v)|}{|P(u)| + |P(v)|}$.

The following three features were also first introduced by Cranshaw et al. [2010] and try model the location diversity of regions two users visited in common. The first and most simple feature to include the popularity of a region is the overall number of observations of unique users at a certain region. According to this we calculated the mean user-count $R_U(u,v)$ of all regions two users visited in common $P(u) \cap P(v)$. The second feature taken from Cranshaw et al. [2010] is similar to the previous feature of counting users at a certain location. We computed the frequency defined as the overall observations of users at a certain location. Again we calculated the mean frequency $R_F(u,v)$ of the frequency of regions two users u and v have in common. A refinement of the two previous features, is the entropy that also takes the probabilities of observations at a location L into account. The probability that a user has visited a certain region is defined as the number of observations of the actual users divided by the overall number of observations at this regions. Let $O_{u,L}$ be the observations of a user u at a location L and O_L be all observations at the location L. The probability can then be computed as $prob_L(u) = \frac{|O_{u,L}|}{|O_L|}$. Based on this we can compute the entropy of a certain location L as $E_L = -\sum_{u \in U_L} P_L(u) \cdot log(P_L(u))$ with U_L representing all users observed a the location L. With this definition we computed the mean entropy $R_E(u,v)$ of the locations two users visited in common.

6.4.2.2. Time-Dependent Features

For the two location sources of Shared- and Monitored Locations we collected the actual locations users visited but also the accurate time and date information of these visits. Using this additional

information we could create two location-based social networks G_S and G_M formed upon the co-occurrences of users. Hence, we can employ the structure of the networks to compute topological features and the actual position and time information of users to compute homophilic features.

Topological Features Based on the structure of location-based social networks we can compute undirected features that are similar to the topological features in the online social network. Users in online networks with small-world characteristics are clustered locally and the more neighbors two users have in common, the closer they are connected. With the formal definition of the neighbors of a node $u \in V_L$ as $\Theta(u) = \{v \mid (u,v) \in E_L\}$ this feature could be computed as $L_{CN}(u,v) = |\Theta(u) \cap \Theta(v)|$. This measure indicated the overlap of neighbors regardless of the total number of neighbors the users have. This total number of neighbors two users have can be computed as $L_T(u,v) = |\Theta(u) \cup \Theta(v)|$ As a combination of both metrics, we computed *Jaccard's Coefficient* as the number of common neighbors divided by the total number of neighbors of two users: $L_{JC}(u,v) = \frac{|\Theta(u) \cap \Theta(v)|}{|\Theta(u) \cup \Theta(v)|}$.

A refinement of this metric was proposed by Adamic and Adar [2003]. As not all neighbors in a network have the same tie strength, they added weights to the links and computed the relation between two users as $L_{AA}(u,v) = \sum_{z \in \Theta(u) \cap \Theta(v)} \frac{1}{log(|\Theta(z)|)}$.

Another feature to measure the structural overlap of two users was introduced by Cranshaw et al. [2010]. They introduced the "neigbourhood overlap" as the number of common neighbors divided by the sum of neighbors of either users. Formally, this can be written as $L_{NO}(u,v) = \frac{|\Theta(u) \cap \Theta(v)|}{|\Theta(u)| + |\Theta(v)|}$.

Active users within a network are more likely to form new interactions than users with less activity. "Preferential Attachment Score" was first mention by Barabási and Albert [1999] and is the product of the sum of neighbors of either users: $L_{PA}(u,v) = |\Theta(u)| \cdot |\Theta(v)|$.

Homophilic Features Previously mentioned topological features only depend on the actual network structure. In contrast, this is not possible for homophilic features as they depend on the available information which is different for Shared- and Monitored Locations.

Shared Locations: For this feature set we can only compute the number of days two users have seen each other, i.e. the metric $A_S(u,v)$ represents the number of days where a user-pair concurrently (within 2 hours) shared pictures from the same region.

Monitored Locations: As outlined before, we implemented user-bots that monitored the present users at event sites. Using the position data of users, respectively the location and time span of events, we identified all events a user u visited over a year: $\Pi(u) = \{e_1, \ldots, e_n\}$ where e_i represented the i'th event out of n visited. With this simple metric we computed the number

of events two users attended in common $E_C = |\Pi(u) \cap \Pi(v)|$, the total number of events $E_T = |\Pi(u) \cup \Pi(v)|$, and finally their fraction $E_{JC} = \frac{|\Pi(u) \cap \Pi(v)|}{|\Pi(u) \cup \Pi(v)|}$.

A refinement of this measure also takes the trajectory of the visited events into account. For each user pair (u,v) we created two vectors $\vec{\varepsilon}(u)$, $\vec{\varepsilon}(v)$ that represent their totally visited events. The j'th component of each vector $\vec{\varepsilon}$ was set to 1 if the user visited the actual event and was set to 0 if it did not. Then we computed the cosine similarity of these two vectors which is formally defined as $E_{CS} = \frac{\vec{\varepsilon}(u) \cdot \vec{\varepsilon}(v)}{||\vec{\varepsilon}(u)|| \cdot ||\vec{\varepsilon}(v)||}$ where $||\vec{\varepsilon}||$ represented the Euclidean length of the vector.

Similar measures can be based on the categories and the maturity rating of events. Events are assigned to different categories and for each user u we created a vector $\vec{\delta}$ of length, where every item represented the number of events attended in a category. We computed the cosine similarity of two users' vectors $\vec{\delta}(u)$ and $\vec{\delta}(v)$ as $E_{CCos} = \frac{\vec{\delta}(u) \cdot \vec{\delta}(v)}{||\vec{\delta}(u)|| \cdot ||\vec{\delta}(v)||}$. The same measure was applied to the maturity rating of events: $E_{MCos} = \frac{\vec{\gamma}(u) \cdot \vec{\gamma}(v)}{||\vec{\gamma}(u)|| \cdot ||\vec{\gamma}(v)||}$ with $\vec{\gamma}$ representing number of events a users visited with the according maturity level.

Finally, we present two features that reflected the user's activity. First, the already known number of days two users were concurrently seen in the same region $A_S(u,v)$ and second, we defined the average distance between them: with the accurate position of every user, we computed the Euclidean distance between them and averaged over all observations to get the spatial proximity $A_D(u,v)$ of the users u and v.

6.5. Experimental Setup

In the previous section we described different features that depend on the domain, i.e. online social network or location-based datasets, and the availability of time information. In this section we present the design of the experiments to answer the research questions stated in Section 6.1.

In Section 6.3.1, we created a network $G_O\langle V_O, E_O\rangle$ from the social text-interactions obtained from the online social network of Second Life . We enrich the nodes V_O of this network with groups-, interests- and interaction information of users and compute the topological and homophilic features as described in Section 6.4.1. Then we add time-independent location information from the *Shared Locations*, *Picked Locations*, and *Monitored Locations* (see Section 6.3.2) to this network and compute the according time-independent features as presented in Section 6.4.2. Further, we compute topological and homophilic time-dependent features (Section 6.4.2) for the networks $G_S\langle V_S, E_S\rangle$ based on the Shared Locations and $G_M\langle V_M, E_M\rangle$ based on the Monitored Locations and again add the derived data to the online social network.

As already outlined, we obtained data from two different domains: text-interaction data for the

online social network, the *Shared Locations* and the *Favoured Locations* of users from their profile pages on the Web-platform My Second Life and in contrast, the *Monitored Locations* dataset directly from inworld Second Life. Hence, the data origins from different domains and as a consequence, we create the overlap of users in both sources. In other words: for the experiments we only considered users with data available in both domains: the profile pages on the Web-platform and inworld Second Life. The overall number of users in this new network $G\langle V,E\rangle$ was 44,603 and the number of edges was 1,419,543 with 1,584 user pairs in a partnership (see Table 6.1 for basic characteristics of the network).

To actually answer the research questions, we describe the analysis to compare partners and acquaintances upon their features to determine significant differences. Then we show supervised and unsupervised learning approaches to evaluate these features regarding their predictability of partnership.

6.5.1. Comparing Partners and Acquaintances

To answer the first research question, we analyzed the similarities and dissimilarities between partners and acquaintances with respect to the features described in Section 6.4. We split the user pairs into balanced sets of partners and acquaintances, and computed mean values and standard errors of all features in either sets separately. The one-sampled Kolmogorov-Smirnov and the Anderson-Darling test showed that none of the distributions of the features were from the family of normal distribution. As a consequence and similarly to Bischoff [2012], we compared the variances of all features between partners and acquaintances using a Levene test ($p < 0.01$). To test significant differences of mean values, we employed Mann-Whitney-Wilcoxon test in case of equal variances and a two-sided Kolmogorov-Smirnov test in case of unequal variances.

6.5.2. Predicting Partnership

Residents of Second Life can marry their friends and the partnership information with the partner's name appears on their profiles. To answer the remaining research questions, we employed the social proximity features to predict whether a user pair is in a partnership or not.

Basically, we used two different techniques:

Predicting Partnership with Supervised Learning In this approach we applied different learning algorithms onto a training set to identify characteristics of partnership and then verified this in a test set. To do so, we reduced the prediction problem to a binary classification problem by selecting 1,500 partners and acquaintances from the network whose proximity features were fed

into the WEKA machine learning suite [Hall et al., 2009]. To validate the obtained results we used a ten-fold cross validation approach.

Predicting Partnership with Unsupervised Learning Due to the balanced data set of partners and acquaintances, the binary classification problem has a baseline of 0.5 when randomly guessing. However, to better estimation the performance and importance of the supervised learning algorithm features, it is recommended to compare the results with an unsupervised learning approach [Bischoff, 2012]. For that purpose, we used a simple Collaborative Filtering technique that was first proposed by Liben-Nowell and Kleinberg [2002]: for every user in a partnership, we rank all acquaintances according to the features described in Section 6.4. Next, we ranked potential partners for every feature separately and computed the success rate of finding the partner within a results list of length k.

6.6. Results

In this section we present the results of the conducted experiments.

6.6.1. Comparing Partners and Acquaintances

We computed the mean values and standard errors for all features of partners and acquaintances and used the Mann-Whitney-Wilcoxon test, respectively Kolmogorov-Smirnov test to determine whether they differ significantly. In Table 6.2, 6.3 and 6.4 we present the differences between partners and acquaintances for features from the online social network and the location-based data sources.

6.6.1.1. Online Social Network Features

At first glance, Table 6.2 reveals that partners were less connected in the network than acquaintances. In particular, we can see that acquaintances had approximately 11 common interaction partners $O^{+}_{CN}(u,v)$, $O^{-}_{CN}(u,v)$ whereas partners had about 1 partner in common. Similar observations were made for Jaccard's Coefficient $O^{+}_{JC}(u,v)$, $O^{-}_{JC}(u,v)$, Adamic-Adar $O_{AA}(u,v)$, and Preferential Attachment Score $O^{+}_{PS}(u,v)$, $O^{-}_{PS}(u,v)$. For the communication direction $O_{RE}(u,v)$ we examined bidirectional communication in nearly 50% of all partnerships but in only 30% of all acquaintances. All topological features were significantly different.

Looking at the homophilic features, partners had on average 2.30 common groups $G_C(u,v)$ versus 0.50 for acquaintances. This goes in line with the Jaccard's Coefficient $G_{JC}(u,v)$ that also

Table 6.2.: The mean values and standard errors for topological (white background) and homophilic (grey background) features in the online social network. (***=significant at level 0.001).

	Features	Partners	Acquaintances
Online Social Network Features	$O_{AA}(u,v)^{***}$	0.81 ± 0.10	6.30 ± 0.07
	$O^{+}_{CN}(u,v)^{***}$	1.12 ± 0.18	11.70 ± 0.13
	$O^{-}_{CN}(u,v)^{***}$	1.08 ± 0.23	11.93 ± 0.13
	$O^{+}_{JC}(u,v)^{***}$	0.06 ± 0.00	0.05 ± 0.00
	$O^{-}_{JC}(u,v)^{***}$	0.03 ± 0.00	0.05 ± 0.00
	$O^{+}_{PS}(u,v)^{***}$	367.21 ± 107.91	9854.15 ± 132.40
	$O^{-}_{PS}(u,v)^{***}$	361.17 ± 107.91	6921.79 ± 115.11
	$O_{RE}(u,v)^{***}$	0.49 ± 0.01	0.29 ± 0.00
	$O^{+}_{T}(u,v)^{***}$	10.62 ± 0.98	119.49 ± 0.79
	$O^{-}_{T}(u,v)^{***}$	9.66 ± 1.04	141.18 ± 1.08
	$G_{C}(u,v)^{***}$	2.30 ± 0.09	0.50 ± 0.01
	$G_{JC}(u,v)^{***}$	0.06 ± 0.00	0.01 ± 0.00
	$G_{T}(u,v)^{***}$	20.46 ± 0.43	28.44 ± 0.09
	$I_{C}(u,v)$	0.05 ± 0.01	0.06 ± 0.00
	$I_{JC}(u,v)$	0.00 ± 0.00	0.00 ± 0.00
	$I_{T}(u,v)^{***}$	3.94 ± 0.15	9.02 ± 0.05
	$P_{A}(u,v)^{***}$	27.25 ± 0.67	18.53 ± 0.13
	$P_{C}(u,v)^{***}$	5.00 ± 0.50	2.02 ± 0.07
	$P_{I}(u,v)^{***}$	19.40 ± 1.81	13.11 ± 0.29
	$P_{L}(u,v)^{***}$	12.33 ± 1.35	10.47 ± 0.24
	$P_{P}(u,v)^{***}$	2.07 ± 0.13	0.62 ± 0.06

differs significantly. The total number of groups two users joined was significantly different as well but this time partners joined on average less total groups than acquaintances (20.46 vs. 28.44). In contrast to the significant differences for group-based features, only the total number of interests $I_T(u,v)$ showed significant differences for interest-based features. Common interests $I_C(u,v)$ and Jaccard's Coefficient were not meaningful due to small values and insignificant differences.

Although the topological features would let us assume that partners did not actively participate in the online social network, the interaction data drew a different picture: on average partners had 19.40 interactions $P_I(u,v)$ which was significantly more than acquaintances with 13.11. This significant difference was observed for postings, comments, and loves as well. Accordingly, we noticed an average message length $P_A(u,v)$ of 27.25 characters per message for partners but only 18.53 characters for acquaintances.

Table 6.3.: The mean values and standard errors for time-independent features based on the three different location sources. (*=significant at level 0.1 and ***=significant at level 0.001).

		Features	Partners	Acquaintances
Time-Independent Location-based Features	Shared	$R_C(u,v)^{***}$	0.50 ± 0.05	0.46 ± 0.01
		$R_U(u,v)^{*}$	33.92 ± 1.62	29.12 ± 0.18
		$R_E(u,v)$	0.90 ± 0.02	0.88 ± 0.00
		$R_F(u,v)^{*}$	118.14 ± 5.93	115.55 ± 0.67
		$R_{LO}(u,v)^{*}$	0.03 ± 0.00	0.01 ± 0.00
		$R_O(u,v)^{*}$	0.02 ± 0.00	0.01 ± 0.00
		$R_T(u,v)^{***}$	6.08 ± 0.33	35.59 ± 0.22
	Favoured	$R_C(u,v)^{***}$	0.57 ± 0.02	0.08 ± 0.00
		$R_U(u,v)^{***}$	44.26 ± 1.76	48.17 ± 0.76
		$R_E(u,v)^{***}$	1.07 ± 0.01	1.14 ± 0.01
		$R_F(u,v)^{***}$	55.69 ± 1.99	61.09 ± 0.88
		$R_{LO}(u,v)^{***}$	0.12 ± 0.00	0.01 ± 0.00
		$R_O(u,v)^{***}$	0.08 ± 0.00	0.01 ± 0.00
		$R_T(u,v)^{***}$	4.85 ± 0.08	5.31 ± 0.02
	Monitored	$R_C(u,v)^{***}$	3.99 ± 0.10	0.98 ± 0.01
		$R_U(u,v)^{***}$	1319.20 ± 18.33	1382.99 ± 4.28
		$R_E(u,v)^{***}$	2.29 ± 0.01	2.35 ± 0.00
		$R_F(u,v)^{***}$	5217.94 ± 81.61	5545.56 ± 19.42
		$R_{LO}(u,v)^{***}$	0.44 ± 0.01	0.08 ± 0.00
		$R_O(u,v)^{***}$	0.28 ± 0.00	0.06 ± 0.00
		$R_T(u,v)^{***}$	10.53 ± 0.21	15.47 ± 0.06

6.6.1.2. Location-Based Social Network Features

Time-Independent Features Table 6.3 shows a detailed overview of the differences between partners and acquaintances for the time-independent features applied to the three different location sources.

Shared Locations: Among all features only the number of common- and total locations showed differences at a significance level of $p < 0.001$. Partners visited more common locations $R_C(u,v)$ (0.50 vs. 0.46) but had on average less total locations $R_T(u,v)$ (6.08 vs. 35.59). For the remaining features we could only identify negligible differences between partners and acquaintances.

Favoured Locations: All features based on this dataset showed significant differences. User-pairs in a partnership had on average more common locations $R_C(u,v)$ with 0.57 vs. 0.08 but had less total locations $R_T(u,v)$ with 4.85 vs. 5.31. The three features that consider the characteristics of places two users visited in common (Entropy $R_E(u,v)$, User Count $R_U(u,v)$, and Frequency

Table 6.4.: The mean values and standard errors for time-dependent topological (white background) and homophilic (grey background) features of the Shared- and Monitored dataset. (**=significant at level 0.01 and ***=significant at level 0.001).

		Features	Partners	Acquaintances
Time-Dependent Location-based Features	Shared	$L_{AA}(u,v)$	1.11 ± 0.13	1.73 ± 0.05
		$L_{CN}(u,v)$	0.11 ± 0.02	0.22 ± 0.01
		$L_{JC}(u,v)$	0.01 ± 0.00	0.00 ± 0.00
		$L_{NO}(u,v)^{***}$	0.07 ± 0.01	0.03 ± 0.00
		$L_{PS}(u,v)^{**}$	6.52 ± 1.66	33.94 ± 1.36
		$L_{T}(u,v)^{**}$	0.96 ± 0.12	1.67 ± 0.05
		$A_{S}(u,v^{***}$	0.37 ± 0.05	0.09 ± 0.00
	Monitored	$L_{AA}(u,v)^{***}$	77.87 ± 6.03	181.24 ± 3.15
		$L_{CN}(u,v)^{***}$	52.30 ± 5.48	53.33 ± 1.19
		$L_{JC}(u,v)^{***}$	0.29 ± 0.01	0.17 ± 0.00
		$L_{NO}(u,v)^{***}$	0.81 ± 0.00	0.87 ± 0.00
		$L_{PS}(u,v)^{***}$	92324.84 ± 42759.98	82591.90 ± 3771.87
		$L_{T}(u,v)^{***}$	153.51 ± 9.33	355.12 ± 5.62
		$E_{CCos}(u,v)^{***}$	0.82 ± 0.01	0.66 ± 0.00
		$E_{C}(u,v)^{***}$	9.51 ± 0.85	1.00 ± 0.02
		$E_{Cos}(u,v)^{***}$	0.43 ± 0.01	0.04 ± 0.00
		$E_{JC}(u,v)^{***}$	0.31 ± 0.01	0.02 ± 0.00
		$E_{MCos}(u,v)^{***}$	0.76 ± 0.01	0.19 ± 0.00
		$E_{T}(u,v)^{***}$	32.22 ± 1.60	41.45 ± 0.27
		$A_{S}(u,v)^{***}$	11.54 ± 0.44	6.74 ± 0.11
		$A_{D}(u,v)^{***}$	5.02 ± 0.27	11.70 ± 0.22

$R_F(u,v)$) showed significant differences as well. On average all three values were smaller for partners if compared to acquaintances.

Monitored Locations: The overlap of visited regions $R_C(u,v)$ with 3.99 was significantly higher than the according feature of acquaintances with 0.98 but again, the opposite was observed for the total number of regions $R_T(u,v)$ (10.53 vs. 15.47). Similar to Favoured Locations, the places partners visited in common had significantly less entropy $R_E(u,v)$, a lesser user frequency $R_F(u,v)$ and user count $R_U(u,v)$.

Time-Dependent Features As depicted in Table 6.4, we could only compute these features for Shared Locations and Monitored Locations as there is not time-information available in the Favoured Locations dataset.

Shared Locations: For the topological features in this dataset we only observed a significant difference of $p < 0.001$ for the feature considering the neighbourhood overlap $L_{NO}(u,v)$ were

partners had a higher overlap than acquaintances. Although the values were less significant, we observed a lower preferential attachment score $L_{PS}(u,v)$ and a lower total number of neighbours $L_T(u,v)$ for partners. The remaining features did not show any significant differences. The only homophilic feature in this dataset indicates that partners met on 0.37 days whereas acquaintances only met on 0.09 day which is a significant difference.

Monitored Locations: We observed over 52 common neighbors $L_{CN}(u,v)$ for partners and over 53 common neighbors for acquaintances with a significant difference. The results of the Adamic-Adar measure $L_{AA}(u,v)$ and Preferential Attachment Score $L_{PS}(u,v)$ go in line but Jaccard's Coefficient $L_{JC}(u,v)$ was slightly higher for partners. For the homophilic features we discovered a significantly higher number of total events $E_T(u,v)$ for acquaintances if compared to partners. In contrast, partners did not only attend more common events $E_C(u,v)$, these events also showed a higher similarity in sense of categories $E_{CCos}(u,v)$ and maturity $E_{MCos}(u,v)$. This results can be compared to the actual days user-pairs met: partners have seen each other on over 11 days compared to over 6 days of acquaintances and during their co-occurrence they had a significantly less spatial distance (5.02 vs. 11.70 meter) between them.

6.6.2. Predicting Partnership with Supervised Learning

We utilized popular supervised learning approaches such as *J.48*, *Logistic Regression* and *Support Vector Machine* (SVM) to predict partnership and used the area under the ROC curve (AUC) as our main evaluation metric [Ling et al., 2003; Huang and Ling, 2005]. The detailed results are depicted in Table 6.5 and can be described as follow:

Online Social Network In this dataset, we combined features that are based on the structure of the network, i.e. topological features, and on the other hand side we combined features that measure the alikeness of two users, i.e. homophilic features. Finally, we combined the two sets and determined the overall predictive power of online social network features. With Logistic Regression, topological features alone could predict the partnership between two users with 0.836 AUC and the homophilic features could predict it with 0.787 AUC. The combination of these two feature sets surpassed the prediction of either feature sets and predicted partnership with 0.864 AUC. Overall, Logistic Regression outperformed J.48 and the SVM in all datasets but the characteristics of the prediction were stable among all algorithms.

Location-based Datasets In these datasets, Logistic Regression also performed best for time-independent features if compared to the remaining algorithms. We could predict the partnership

Table 6.5.: Area under the ROC curve (AUC) to predict partnership with different supervised learning algorithms using feature sets from the online social network and three different sources of location.

Feature Sets		Logistic	J.48	SVM
Online Social Network				
Topological		0.836	0.823	0.647
Homophilic		0.787	0.759	0.689
Combined		0.864	0.814	0.730
Location-based Datasets				
Time Independent	Shared	0.771	0.737	0.623
	Favoured	0.705	0.648	0.659
	Monitored	0.901	0.846	0.831
Time Dependent	Shared	0.524	0.523	0.505
	Favoured	–	–	–
	Monitored	0.894	0.860	0.801

between two users correctly with an AUC of 0.771 for Shared Locations, 0.901 for Monitored Locations, and 0.705 for Favoured Locations using Logistic Regression. The application of learning algorithms to the time-dependent Shared Locations dataset showed poor prediction results with an AUC around 0.52 which is closed to flipping a coin. In contrast, the combination of topological features (0.728 AUC) and homophilic features (0.873 AUC) in the Monitored Locations dataset resulted in an AUC of 0.894 for Logistic Regression. Again, Logistic Regression outperformed J.48 and SVM although the results were stable among all three algorithms.

Combined Datasets Besides the in-depth analysis of the two domains, we also combined available features from different source to predict partnership. As depicted in Table 6.6 the results of all three algorithms were stable and again, Logistic Regression performed best among all of them. In all cases the combination of features from the online social network and a location-based dataset outperformed either sources. The combination of features from the online social network and features from the Monitored Location dataset performed best and could predict a partnership between users with 0.939 AUC.

As Logistic Regression performed best among all datasets obtained from the online social network and the location-based datasets we use this algorithm to determined the usefulness of each single feature alone. Additionally, we determined the information gain of the single features using WEKA's attribute evaluation algorithm. The results of these computations are presented in Ta-

Table 6.6.: Predicting partnership with supervised learning algorithms based on combined feature sets from the online social network and the three different sources of location.

Feature Sets	Logistic	J.48	SVM
Online Social Network	0.864	0.814	0.730
Shared Locations	0.771	0.737	0.599
Favoured Locations	0.705	0.648	0.659
Monitored Locations	0.917	0.883	0.842
Online + Shared	0.883	0.827	0.744
Online + Favoured	0.898	0.821	0.763
Online + Monitored	0.939	0.867	0.852

ble 6.7 for the online social network respectively in Table 6.8 and Table 6.9 for the location-based datasets.

6.6.2.1. Online Social Network Features

As depicted in Table 6.7, the Preferential Attachment Scores for messages from neighbors $O^{+}_{PS}(u,v)$ in the online social network had the highest information gain with 0.304 a and corresponding prediction factor of 0.842 AUC. For homophilic features, Jaccard's Coefficient for groups $G_{JC}(u,v)$ was around 0.6 AUC and features based on the interests of users $I_C(u,v)$, $I_{JC}(u,v)$ did not work at all. Communication based features $P(u,v)$ with number of postings and loves, respectively average message length could predict partnership with around 0.60 AUC.

6.6.2.2. Location-Based Social Network Features

Time-Independent Features The detailed results for the time-dependent features can be found in Table 6.8.

Shared Locations: The only noteworthy feature in this dataset is the total number of locations $R_T(u,v)$ two users have visited. We computed an information gain of 0.175 and predictive power of 0.740 for the area under the ROC curve, AUC. The remaining features can be considered as less important due to small values for AUC and information gain.

Favoured Locations: Considering the homophilic features of the favoured locations we found an information gain of 0.132 for the location overlap $R_{LO}(u,v)$ and the common observations $R_O(u,v)$, respectively 0.126 for the regions two users visited in common $R_C(u,v)$. This goes in line with the AUC values of these three features with around 0.66. The remaining features have negligible values for the information gain and the AUC.

Table 6.7.: Results of the supervised and unsupervised approach to predict partnership using topological (white background) and homophilic (grey background) features from the online social network

	Features	AUC	Gain	*SR*@1	*SR*@5	*SR*@10
Online Social Network Features	$O_{AA}(u,v)$	0.615	< 0.1	0.120	0.329	0.549
	$O^{+}_{CN}(u,v)$	0.609	< 0.1	0.120	0.318	0.593
	$O^{-}_{CN}(u,v)$	0.672	0.100	0.153	0.329	0.494
	$O^{+}_{JC}(u,v)$	0.435	< 0.1	0.186	0.461	0.637
	$O^{-}_{JC}(u,v)$	0.651	< 0.1	0.186	0.428	0.593
	$O^{+}_{PS}(u,v)$	0.842	0.304	0.044	0.186	0.450
	$O^{-}_{PS}(u,v)$	0.709	0.156	0.033	0.230	0.439
	$O_{RE}(u,v)$	0.527	< 0.1	0.131	0.406	0.637
	$O^{+}_{T}(u,v)$	0.805	0.248	0.054	0.153	0.428
	$O^{-}_{T}(u,v)$	0.829	0.281	0.033	0.197	0.417
	$G_{C}(u,v)$	0.595	< 0.1	0.252	0.483	0.604
	$G_{JC}(u,v)$	0.599	< 0.1	0.296	0.472	0.604
	$G_{T}(u,v)$	0.613	< 0.1	0.022	0.153	0.406
	$I_{C}(u,v)$	0.510	< 0.1	0.087	0.329	0.560
	$I_{JC}(u,v)$	0.510	< 0.1	0.087	0.329	0.560
	$I_{T}(u,v)$	0.669	< 0.1	0.044	0.241	0.505
	$P_{A}(u,v)$	0.632	< 0.1	0.076	0.516	0.725
	$P_{C}(u,v)$	0.534	< 0.1	0.395	0.692	0.813
	$P_{I}(u,v)$	0.557	< 0.1	0.318	0.648	0.846
	$P_{L}(u,v)$	0.578	< 0.1	0.263	0.604	0.747
	$P_{P}(u,v)$	0.688	< 0.1	0.461	0.747	0.868

Monitored Locations: Features based on the commonly visited events $R_C(u,v)$, $R_O(u,v)$, and $R_{LO}(u,v)$ showed results around 0.4 for information gain and around 0.88 for the AUC. The values computed using this dataset showed similar characteristics as it was in the Favoured Locations but performed in general superior.

Time-Dependent Features A detailed listing of all results for this feature sets can be found in Table 6.9. Due to the missing time information in the Favoured Locations dataset we can only compute these features for Shared- and Monitored Locations.

Shared Locations: Interestingly, none of the topological and the homophilic features in this dataset had an information gain over 0.1. Further, considering the AUC of the single features we observed negligible predictive power not far from flipping a coin, i.e. 0.5 AUC.

Monitored Locations: None of the topological features in the Monitored Locations dataset had an information gain over 0.1 but this time these features were outperformed by homophilic fea-

Table 6.8.: Results of the supervised and unsupervised approach to predict partnership using time-independent features obtained from three different sources of location data.

		Features	AUC	Gain	*SR*@1	*SR*@5	*SR*@10
Time-Independent Location-based Features	Shared	$R_C(u,v)$	0.530	< 0.1	0.142	0.417	0.626
		$R_U(u,v)$	0.516	< 0.1	0.022	0.153	0.164
		$R_E(u,v)$	0.502	< 0.1	0.044	0.109	0.164
		$R_F(u,v)$	0.509	< 0.1	0.044	0.153	0.175
		$R_{LO}(u,v)$	0.477	< 0.1	0.153	0.439	0.659
		$R_O(u,v)$	0.477	< 0.1	0.153	0.439	0.659
		$R_T(u,v)$	0.740	0.175	0.033	0.153	0.384
	Favoured	$R_C(u,v)$	0.663	0.126	0.175	0.428	0.615
		$R_U(u,v)$	0.563	< 0.1	0.131	0.230	0.263
		$R_E(u,v)$	0.522	< 0.1	0.120	0.230	0.263
		$R_F(u,v)$	0.557	< 0.1	0.131	0.241	0.263
		$R_{LO}(u,v)$	0.666	0.132	0.175	0.417	0.615
		$R_O(u,v)$	0.666	0.132	0.175	0.417	0.615
		$R_T(u,v)$	0.512	< 0.1	0.033	0.175	0.395
	Monitored	$R_C(u,v)$	0.829	0.274	0.406	0.725	0.835
		$R_U(u,v)$	0.546	< 0.1	0.087	0.395	0.637
		$R_E(u,v)$	0.521	< 0.1	0.076	0.428	0.648
		$R_F(u,v)$	0.553	< 0.1	0.098	0.351	0.593
		$R_{LO}(u,v)$	0.888	0.416	0.527	0.780	0.890
		$R_O(u,v)$	0.888	0.416	0.527	0.780	0.890
		$R_T(u,v)$	0.654	< 0.1	0.000	0.131	0.351

tures: Jaccard's Coefficient $E_{JC}(u,v)$ and the cosine similarity $E_{Cos}(u,v)$ of events had the highest information gain and correctly predicted partnership with around 0.84 AUC. The average distance between two avatars $A_D(u,v)$ had a predictive power of 0.744 AUC but we observed a remarkable low AUC of 0.350 for the days two avatars were concurrently seen in the same regions $A_S(u,v)$.

6.6.3. Predicting Partnership with Unsupervised Learning

Additionally, we compared the results of the supervised prediction algorithm with the outcome of an unsupervised learning algorithm. This is useful to better estimate the performance in real applications [Bischoff, 2012] and to support our previous findings. Hence, we implemented a simple Collaborative Filtering approach to rank potential partners of users according to their similarity. The success rates that the actual partner was found in lists of length 1 (SR@1), 5 (SR@5), and 10 (SR@10) are presented in Table 6.7 for the online social network and in Table 6.8 and Table 6.9 for location-based datasets. Obviously, we could observe an increasing hit rate with increasing number of suggested users, i.e. increasing list length.

Table 6.9.: Results of the supervised and unsupervised approach to predict partnership using topological (white background) and homophilic (grey background) features from the Shared- and Monitored Locations.

		Features	AUC	Gain	$SR@1$	$SR@5$	$SR@10$
Time-Dependent Location-based Features	Shared	$L_{AA}(u,v)$	0.525	< 0.1	0.087	0.120	0.142
		$L_{CN}(u,v)$	0.521	< 0.1	0.098	0.164	0.208
		$L_{JC}(u,v)$	0.508	< 0.1	0.098	0.186	0.208
		$L_{NO}(u,v)$	0.510	< 0.1	0.109	0.164	0.164
		$L_{PS}(u,v)$	0.524	< 0.1	0.087	0.120	0.175
		$L_{T}(u,v)$	0.525	< 0.1	0.087	0.120	0.142
		$A_{S}(u,v)$	0.501	< 0.1	0,131	0,197	0,208
	Monitored	$L_{AA}(u,v)$	0.741	< 0.1	0.230	0.626	0.681
		$L_{CN}(u,v)$	0.711	< 0.1	0.384	0.626	0.703
		$L_{JC}(u,v)$	0.486	< 0.1	0.417	0.615	0.703
		$L_{NO}(u,v)$	0.516	< 0.1	0.351	0.571	0.659
		$L_{PS}(u,v)$	0.177	< 0.1	0.241	0.626	0.681
		$L_{T}(u,v)$	0.727	< 0.1	0.230	0.626	0.681
		$E_{CCos}(u,v)$	0.666	< 0.1	0.175	0.417	0.604
		$E_{C}(u,v)$	0.813	0.290	0.483	0.747	0.791
		$E_{Cos}(u,v)$	0.841	0.368	0.538	0.758	0.802
		$E_{JC}(u,v)$	0.839	0.364	0.549	0.747	0.802
		$E_{MCos}(u,v)$	0.776	0.264	0.307	0.571	0.692
		$E_{T}(u,v)$	0.605	< 0.1	0.000	0.131	0.406
		$A_{S}(u,v)$	0.350	< 0.1	0.472	0.681	0.703
		$A_{D}(u,v)$	0.744	< 0.1	0.197	0.582	0.659

6.6.3.1. Online Social Network Features

Although 5 out of 10 topological features had an information gain > 0.1 the success-rate with a list length of 1 was considerable low. Among all features Jaccard's Coefficient $O^{+}_{JC}(u,v)$ performed best with $SR@1 = 0.186$, $SR@5 = 0.461$, and finally $SR@10 = 0.637$. The evaluation of homophilic and interaction based features showed that the number of postings performed better than topological and group- or interest-based features. We found out that the actual number of postings $P_P(u,v)$ and the number of comments $P_C(u,v)$ two users shared performed best. In group related features in the online social network, 29.6% of all partners were ranked on top of the list for $SR@1$.

6.6.3.2. Location-based Social Network Features

Time-Independent Features A detailed overview of the described features and results can be found in Table 6.8.

Shared Locations: Although all features performed inferior, we found that the common locations $R_C(u,v)$, location overlap $R_{LO}(u,v)$ and common observations $R_O(u,v)$ two users visited performed best.

Favoured Locations: Similar to the supervised learning approach, the commonly visited locations $R_C(u,v)$, location overlap $R_{LO}(u,v)$, and common observations $R_O(u,v)$ outperform other features.

Monitored Locations: We identified the location overlap $R_{LO}(u,v)$ and commonly visited locations $R_C(u,v)$ with an accuracy of around $SR@1 = 0.527$ as most valuable for the prediction which is in line with the according values of information gain. The remaining features performed inferior. Further, Monitored Locations resulted in the best results for the prediction of partnership among all three location sources.

Time-Dependent Features A detailed list of results for the unsupervised learning approach with time-dependent features can be found in Table 6.9.

Shared Locations: The results for this approach support the findings from the supervised learning approach: Neither topological nor homophilic features seem to be suitable for the prediction of partnership.

Monitored Locations: Similar to the supervised learning we could observe that homophilic features outperform topological features in the unsupervised learning approach. Among these features we identified cosine-similarity $E_{Cos}(u,v)$ and the Jaccard's Coefficient $E_{JC}(u,v)$ of commonly visited events as the most valuable for prediction. From the set of topological features only Jaccard's Coefficient $L_{JC}(u,v)$ showed promising results.

6.7. Discussion and Conclusion

In this work we harvested data from two Second Life related domains: an online social network with text-based interactions and three sources of location-based position data. We modelled the relations between users with social proximity features and conducted experiments to answer the research questions.

- *RQ1:* For the first research question, we evaluated the differences between partners and acquaintances in the online social network and the location-based datasets. Interestingly, this analysis revealed that partners had less common neighbors and communication partners than acquaintances in the location-based social networks and the online social network. Contrary to this observation, homophilic features revealed a strong affection between partners; we found evidence that partners shared more common groups, had more interactions between them and attended more events together. Besides this we observed similar characteristics for time-independent features in the three location sources *Monitored Locations*, *Shared Locations*, and *Favoured Locations*: partners tend to visit more regions in common and have more common observations together. In contrast to this, they visit regions with a lower user-count, frequency, and entropy which can be interpreted as sign of intimacy: users in a partnership are familiar with their environment and are not anxious to meet new users in unknown places. This is in line with the observation that partners were on average spatially closer than acquaintances during their co-occurrence.

- *RQ2:* For the second research question, we predicted the partnership between users and hence merged the data from the two domains and the according features into one network. We reduced the prediction problem to a binary classification problem and evaluated the features using three different learning algorithms. Although all of them showed similar characteristics, *Logistic Regression* performed best which is in line with related work in this area [Rowe et al., 2012; Leskovec et al., 2010]. Network topological features turned out as useful if sufficient data is available but nevertheless homophilic features like the number of common groups two users joined or the common events they attended outperformed these features. This result can be compared to the real world where the alikeness of two users, i.e. homophily, is a premise for a working partnership. We found that the most valuable features of time-independent features across all three location-based sources were the number of common regions $R_C(u,v)$, the Jaccard's Coefficient $R_{JC}(u,v)$, and the total number of regions of two users $R_T(u,v)$. Although these characteristics were similar over all sources, features applied to the Monitored Locations seemed best suited for predicting partnership if compared to Favoured or Shared Locations. We believe that the difference between the location sources is founded in the divergent location sources. Only 31% of all users have ever shared snapshots with location information and only 66% specified favoured locations on their profile. This sparseness of data seems to be the main reason why the Shared Locations and Favoured Locations were outperformed by the Monitored Locations. Overall, the combination of features from the online social network and the *Monitored Locations* could predict partnership with 0.939 AUC using Logistic Regression.

- *RQ3:* For the third research question, we compared the predictive power of single features with a simple Collaborative Filtering approach. To that end, we computed the predictability of partnership for every feature with Logistic Regression and ranked lists of users' similarity. In online social networks homophilic features outperform topological features when using unsupervised learning which is in contrast to a supervised learning approach. The results for time-independent features are similar over all three location sources and go in line with the supervised learning algorithms: *Monitored Locations* outperform *Shared Locations* outperform *Favoured Location*. Results of supervised learning algorithms using time-dependent features goes in line with supervised learning as well. Features like Jaccard's Distance $L_{JC}(u,v)$ or cosine similarity of attended events $E_{Cos}(u,v)$ have a high predictive power with both concepts. Overall this lets us assume that homophilic features have a better correlation for tie strength than topological features in general. In particular we identified features derived from the attitude of users, like events and groups, as features with the highest information gain. Further, interpersonal bonding with spatial distance and number of postings were detected as evidence for a partnership between two users.

- *RQ4:* For the last research question, we examined topological and homophilic time-dependent features based on the Shared- and Monitored Locations datasets. As already mentioned, time-dependent features derived from the Shared Locations dataset do not contribute to the prediction of partnership due to the sparseness of available data. However we found promising results for the Monitored Locations dataset: In the time-dependent dataset, homophilic features outperform topological features but their combination was even slightly outperformed by time-independent features. This results are surprising in several ways: Although time-dependent features perform worse than time-independent features, a combination of both outperforms either sources significantly. This lets us assume that the time-component is a very crucial factor for the prediction of tie-strength. Further, homophilic features performed better than topological features which is in contrast to the online social network where topological features outperformed homophilic features.

Conclusion Features in an online social network induced from text-interactions in Second Life perform well to foresee partnership between users and location-based data even supports this prediction. Overall, topological features were identified as useful if sufficient data is applicable whereas homophilic features were more robust against this sparseness as they do not depend on the missing data of other users in the network.

References

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

A. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286(5439):509–512, 1999.

K. Bischoff. We love rock'n'roll: analyzing and predicting friendship links in Last. fm. In *Proceedings of the 3rd Annual ACM Web Science Conference*, pages 47–56. ACM, 2012.

J. Cheng, D. Romero, B. Meeder, and J. Kleinberg. Predicting reciprocity in social networks. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 49–56. IEEE, 2011.

J. Choi, S. Heo, J. Han, G. Lee, and J. Song. Mining social relationship types in an organization using communication patterns. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 295–302. ACM, 2013.

J. S. Coleman. Social capital in the creation of human capital. *American journal of sociology*, pages S95–S120, 1988.

J. Cranshaw, E. Toch, J. Hong, A. Kittur, and N. Sadeh. Bridging the gap between physical location and online social networks. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, pages 119–128. ACM, 2010.

M. Fire, L. Tenenboim, O. Lesser, R. Puzis, L. Rokach, and Y. Elovici. Link prediction in social networks using computationally efficient topological features. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 73–80. IEEE, 2011.

E. Gilbert. Predicting tie strength in a new medium. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 1047–1056. ACM, 2012.

E. Gilbert and K. Karahalios. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 211–220. ACM, 2009.

M. S. Granovetter. The strength of weak ties. *American journal of sociology*, pages 1360–1380, 1973.

M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. Witten. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

J. Huang and C. X. Ling. Using auc and accuracy in evaluating learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, 17(3):299–310, 2005.

J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650. ACM, 2010.

D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2002.

C. X. Ling, J. Huang, and H. Zhang. AUC: a statistically consistent and more discriminating measure than accuracy. In *International Joint Conference on Artificial Intelligence*, pages 519–526. LAWRENCE ERLBAUM ASSOCIATES LTD, 2003.

A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM, 2010.

A. Noulas, S. Scellato, N. Lathia, and C. Mascolo. Mining User Mobility Features for Next Place Prediction in Location-based Services. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1038–1043. IEEE, 2012.

S. J. Pan, D. J. Boston, and C. Borcea. Analysis of fusing online and co-presence social networks. In *IEEE International Conference on Pervasive Computing and Communications*, pages 496–501, 2011.

M. Rowe, M. Stankovic, and H. Alani. Who will follow whom? exploiting semantics for link prediction in attention-information networks. In *Proceedings of the 11th international conference on The Semantic Web - Volume Part I*, ISWC'12, pages 476–491, Berlin, Heidelberg, 2012. Springer-Verlag.

S. Scellato, A. Noulas, R. Lambiotte, and C. Mascolo. Socio-spatial properties of online location-based social networks. *Proceedings of ICWSM*, 11:329–336, 2011a.

S. Scellato, A. Noulas, and C. Mascolo. Exploiting place features in link prediction on location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1046–1054. ACM, 2011b.

M. Steurer and C. Trattner. Predicting interactions in online social networks: an experiment in second life. In *Proceedings of the 4th International Workshop on Modeling Social Media*, MSM '13, pages 5:1–5:8, New York, NY, USA, 2013a. ACM. ISBN 978-1-4503-2007-8.

M. Steurer and C. Trattner. Predicting interactions in online social networks: an experiment in second life. In *MSM*, page 5, 2013b.

M. Steurer and C. Trattner. Acquaintance or partner?: Predicting partnership in online and location-based social networks. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 372–379, New York, NY, USA, 2013c. ACM. ISBN 978-1-4503-2240-9. doi: 10.1145/2492517.2492562. URL `http://doi.acm.org/10.1145/2492517.2492562`.

M. Steurer, C. Trattner, and F. Kappe. Success factors of events in virtual worlds a case study in second life. In *NetGames*, pages 1–2, 2012.

M. Steurer, C. Trattner, and D. Helic. Predicting social interactions from different sources of location-based knowledge. In *SOTICS 2013, The Third International Conference on Social Eco-Informatics*, pages 8–13, 2013.

M. Thelwall. Homophily in myspace. *Journal of the American Society for Information Science and Technology*, 60(2):219–231, 2009.

Y. Volkovich, S. Scellato, D. Laniado, C. Mascolo, and A. Kaltenbrunner. The length of bridge ties: structural and geographic properties of online social interactions. *Proceedings of ICWSM*, 12, 2012.

D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108, New York, NY, USA, 2011. ACM.

E. Zheleva, L. Getoor, J. Golbeck, and U. Kuter. Using friendship ties and family circles for link prediction. In *Advances in Social Network Mining and Analysis*, pages 97–113. Springer, 2010.

CHAPTER 7

Research Results and Conclusions

In this dissertation we collected data from an online social network that aims at residents of the virtual world of Second Life and three different location-based data sources of these residents. We computed metrics that model the social proximity between users and evaluated them for different types of user relations. Based on these results we applied supervised and unsupervised machine learning techniques to the location-based data sources and the online social network as well as a combination of both to predict links between users and the tie-strength of these links.

This chapter outlines the found results and can be divided into two sections: In Section 7.1 we summarize the main contributions presented in this thesis and answer the research results stated in the introductory part. Finally, we conclude the dissertation in Section 7.2.

7.1. Summary of Results

In this thesis we described approaches to collect user data in a virtual world from two different domains: online social network data which is a Facebook-like social network where users share text messages, specify interests and join groups, and location-based data that represents users' movement from three different sources: 1) "Shared Locations" – users check-in at specific locations, 2) "Favoured Locations" – users specify their top 10 locations and 3) "Monitored Locations" – users' in-world movement trajectories. We computed metrics that model the social proximity between users with topological features and homophilic features. Topological features (e.g. number of common neighbours) were obtained from the network structure of both the online social network data and the available location-based data sources. In contrast, homophilic features that model the alikeness or similarity of a user-pair were split into time-dependent features (e.g. number of days two avatars were seen concurrently) and time-independent features (e.g. number of commonly visited locations) depending on the information richness of the actual data source.

Using this test bed we conducted several experiments to answer the research questions stated in Chapter 1:

Research Question 1

Which social proximity features can be derived from an online social network and location-based data sources and how do they differ for different types of relations between users?

The analysis revealed that user-pairs with interactions are more tightly connected in the online social network than users without interactions. Interestingly, the opposite was observed for the location-based social network. A possible explanation is that users in Second Life are allowed to directly teleport to different locations in the whole virtual world but see the currently present users only upon arrival. We believe that users are more likely to stay in a location if they know any present users (i.e. they have interactions with these users in the online social network) or otherwise move on to the next location. This higher activity and the accordingly higher diversity due to the short-term visits explain the tighter integration of non-interacting users into the location-based social network. This assumption is supported by homophilic features from both networks: users with interactions had more common groups, visited more common locations and they saw each other on more days. Furthermore, the average spatial distance between interacting users was significantly shorter than the spatial distance between users without interaction. The evaluation of the reciprocity of interactions between users was similar as users with a reciprocal interaction also had tighter connections in the online-social network. Again, homophilic features of user-pairs with reciprocal interactions indicated a higher alikeness in both networks if compared to user-pairs with uni-directional interactions.

Interestingly, we found out that user-pairs tend to be less integrated into the network structure with increasing tie-strength which can be interpreted as a sign of intimacy. In other words, user-pairs that are in a relationship have less common friends if compared to user-pairs that are just acquaintances. This intimacy is even supported by the observation that homophilic features revealed a strong affection between users connected with a strong tie: they shared more common groups, had more interactions between them and attended more events together than user-pairs with a weak tie. Further, they tend to visit more locations in common and have more common observations in locations with a lower user-count, frequency and entropy. It seems that users connected by a strong tie are more familiar with their environment and are not anxious to meet new users in unknown places or interact with them in the online social network. This is in line with the observation that strong ties were on average spatially closer than weak ties during their co-occurrences.

Research Question 2

How can a combination of social proximity measures derived from an online social network and a location-based social network predict interactions between users?

We used unsupervised machine learning approaches to determine valuable and information rich features, and we applied supervised machine learning algorithms for the actual prediction task where we tried to predict whether or not a user-pair has an interaction in the online social network. We combined the topological and homophilic features from the online social network with the topological and homophilic features of the "Monitored Locations" dataset.

To determine the most promising features for the link prediction we implemented an unsupervised learning algorithm based on ranked lists of users' similarity, respectively used the Information Gain metrics of single features. We identified Jaccard's Coefficient and Preferential Attachment Score as most promising among the topological features in the online social network as well as in the location-based social network. Homophilic features in the online social network suffered from the sparseness of data and hence were not very valuable. In contrast, homophilic features obtained from the location-based social network were very promising: common visited locations, common observations and the average distance between users were identified as most valuable.

For the actual prediction of interactions between pairs of users, features obtained from the location-based social network performed better in general than features from the online social network. It is interesting to note that topological features outperform homophilic features in the online social network whereas the opposite could be observed for the location-based social network. Overall, with the combination of features from the online and the location-based social network we were able to predict whether two users have an interaction or not with more than

97% AUC. For the prediction of reciprocity of interactions we observed similar results with the difference that features from the online social network performed better than features from the location-based social network. Overall we could predict reciprocal interaction with features from both domains with an accuracy of nearly 68% AUC. Although the results among all supervised learning algorithms were stable in their prediction performance, Logistic Regression performed best.

Research Question 3

Can different location-based data sources be used to predict interactions in a related online social network and which source is the most valuable?

Based on the results using the online social network and one location-based social network, we were further interested in the differences between the three location-based data sources ("Shared Locations", "Favoured Locations" and "Monitored Locations") and hence compared their value for the prediction of interactions.

As the time information was not available for all the datasets we could not infer a network structure for all of them and as a consequence we only used homophilic time-independent features to model the social proximity between users for a fair comparison. In order to predict interactions, we evaluated the differences between user-pairs that had an interaction in the online social network and user-pairs without this interaction. The detailed analysis of the results revealed statistically significant differences for nearly all features: User-pairs with interactions on average visited more common locations and had more common observations. In contrast to this, they visited locations with a lower user-count, frequency and entropy which can be interpreted as a sign of intimacy: Users with interactions already know each other and therefore they meet in places that are less frequented by others. We could observe this for all three data sources but due to the diverse datasets the characteristics were different: the "Shared Locations" dataset showed more distinct tendencies than, for instance, the "Favoured Locations" dataset with its limitation of 10 picks per user.

For the prediction of interactions we found significant differences between the data sources and identified "Shared Locations" as most successful (84.9% AUC) outperforming "Monitored Locations" (63.2% AUC) and "Picked Locations" (63.0% AUC). This was again caused by the different characteristics as well as the sparseness of the data sets: First, users can share locations from everywhere within the virtual world on their profile and the data collection approach does not miss any data. Second, users explicitly share locations and places they like and spend time in. Other users who visit their profiles because they already know each other, see these locations and also visit them. This can be seen as an explicit promotion of a user's shared locations. An explanation why the others perform worse is that "Monitored Locations" only cover a clipping of

locations due to resource limitations and "Picked Locations" are limited per se to 10 per users. A more detailed analysis revealed that the most valuable features among all the location-based knowledge sources were the number of common locations, Jaccard's Coefficient and the total number of locations two users visited.

Research Question 4

To what extent can a combination of an online social network and three different location-based data sources support the prediction of tie-strength of links between users?

To tighten the problem of predicting interactions between user-pairs we aimed to even forecast the tie-strength between already connected user-pairs. Strong ties were defined as users that are in a partnership (residents of Second Life can get married) and have an interaction in the online social network whereas weak ties are user-pairs that just have an interaction in the online social network. To model the social proximity between users, we applied homophilic and topological features to the online social network as well as the three location-based data sources. We reduced the prediction problem to a binary classification problem and evaluated the features using different learning algorithms with Logistic Regression performing best. Although the characteristics were similar and stable over all three location-based data sources, features applied to the "Monitored Locations" data set were suited best for predicting partnership if compared to "Favoured Locations" or "Shared Locations". The reason is that the "Monitored Locations" data set is more detailed as accurate time information and context data are available if compared to the "Shared Locations" or "Favoured Locations" data sets. As a consequence we were able to compute topological features and even more homophilic features, e.g. average distance between users or commonly visited events, which supports the prediction of tie-strength. Overall we could predict partnership with over 93.9% AUC for the combination of features from the online social network and features from the "Monitored Locations" dataset.

Network topological features turned out as useful if sufficient data is available but nevertheless homophilic features like the number of common groups two users joined or the common events they attended outperformed these features. For the prediction of tie-strength with data from the online social network we identified the Preferential Attachment Score as the most valuable topological feature whereas features derived from the communication behaviour of users turned out as promising homophilic features. In the location-based data sources we identified the Adamic-Adar coefficient as most successful although these features were outperformed by homophilic features. We found the number of locations two users visited in common as most valuable among all location-sources. If available, time-dependent features like the days two users where seen concurrently in the same location or the average distance between two users were also identified as powerful metrics for the prediction task.

7.2. Conclusions

In this thesis we collected data from an online social network and three different sources of location-based data of Second Life residents. We computed features with the collected data that model the social proximity between users and evaluated differences between various types of user relations. Based on these results we predicted links between users and the tie-strength of these links using supervised and unsupervised machine learning techniques.

The conducted experiments for the prediction tasks clearly show the value of online social networks supported by location-based information. Although features derived from the online social network alone show promising and valuable results for the prediction of interactions and tie-strength, a combination of features from the online social network and location-based information outperforms these results. We compared three different location-based data sources to each other and all of them showed similar characteristics for the prediction tasks. Overall, we have clearly identified the value of homophilic features in the location-based domain and the consequences for the prediction tasks. These homophilic features performed well among all the three location-based sources but we have also seen that available time information allows the computation of additional features which are valuable for the prediction tasks.

With the conducted experiments and the results of these experiments in this thesis we shed light onto the benefit of location-based information combined with personal online social networks for the prediction of interactions between users and the prediction of tie-strength of these interactions. Although all these experiments are based on data collected from a virtual world, it is created by humans as they control their virtual character in the virtual world. As a consequence the data is a perfect test bed for the conducted experiments as all the data is publicly available and per se anonymised due to the anonymity of Second Life. Overall this thesis shows the relevance and benefits of location-based data but also potential risks as real-world position information of users becomes more and more available.

Own Publications

Peer-Reviewed Conferences

- Steurer, Michael; Trattner, Christoph; Helic Denis *Predicting Social Interactions from Different Sources of Location-based Knowledge*, The Third International Conference on Social Eco-Informatics, Lisbon, Portugal, 2013.
- Steurer, Michael; Trattner, Christoph; *Acquaintance or Partner? Predicting Partnership in Online and Location-based Social Networks*; Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, IEEE/ACM, Niagara Falls, Canada, 2013.
- Steurer, Michael; Trattner, Christoph; *Predicting Interactions In Online Social Networks: An Experiment in Second Life*; Proceedings of the 4th International Workshop on Modeling Social Media, Paris, France, 2013.
- Steurer, Michael; Trattner, Christoph; *Success Factors of Events in Virtual Worlds A Case Study in Second Life*; Workshop on Network and Systems Support for Games, Venice, Italy, 2012,
- Steurer, Michael; *A Webshop for Digital Assets in Virtual Worlds Supported by a 3D Object Representation*; The Sixth International Multi-Conference on Computing in the Global Information Technology, Luxemburg, Luxemburg, 2011.

- Kappe, Frank; Steurer, Michael; *A micropayment enabled webshop for digital assets in virtual worlds*; Proceedings of the 14th International Academic MindTrek Conference: Envisioning Future Media Environments, Tampere, Finland, 2010.
- Trattner, Christoph; Steurer, Michael; Kappe, Frank; *Socializing Virtual Worlds with Facebook - A prototypical implementation of an expansion pack to communicate between Facebook and OpenSimulator based Virtual Worlds*; Proceedings of the 14th International Academic MindTrek Conference: Envisioning Future Media Environments, Tampere, Finland, 2010.
- Kappe, Frank; Steurer, Michael; *The Open Metaverse Currency (OMC) A Micropayment Framework for Open 3D Virtual Worlds*; E-Commerce and Web Technologies, Bilbao, Spain, 2010
- Zaka, Bilal; Steurer, Michael; Kappe, Frank; *A Framework for Extending Plagiarism Detection in Virtual Worlds*; RCIS, Fes, Morocco, 2009.
- Kappe, Frank; Zaka, Bilal; Steurer, Michael; *Automatically Detecting Points of Interest and Social Networks from Tracking Positions of Avatars in a Virtual World*; ASONAM, Athens, Greece, 2009.

Journal Publications

- Steurer, Michael and Trattner, Christoph; *Predicting Partnership with Location-based and Online Social Network Data* submitted to *Elsevier Journal of Neurocomputing*.

Book Chapters

- Steurer, Michael and Trattner, Christoph; *Who will Interact with Whom? A Case-Study in Second Life using Online and Location-based Social Network Features to Predict Interactions between Users*; In Proceedings of the MUSE-MSM Post-Proceedings, Springer, 2013.

Others

- Tögl, Ronald; Steurer, Michael; *OpenTC WP3 Report: Java API and Library Implementation*; 2008
- Pirker, Martin; Steurer, Michael; Tögl, Ronald; *Trusted Computing Meets Java*: LinuxTag. Berlin, May 2008

List of Figures

List of Tables