On Structural Aspects of Unconnectedness in Knowledge and Social Networks

Julia Perl, geb. Preusse
jpreusse@uni-koblenz.de
Institute for Web Science and Technologies
University of Koblenz-Landau

Dezember 2014

Vom Promotionsausschuss des Fachbereichs 4: Informatik der Universität Koblenz-Landau zur Verleihung des akademischen Grades

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation.

Datum der wissenschaftlichen Aussprache:	5. Dezember 2014
Vorsitz des Promotionsausschusses:	Prof. Dr. Karin Harbusch
Gutachter:	Prof. Dr. Steffen Staab
Gutachter:	Prof. Dr. Markus Strohmaier

Abstract

Through the increasing availability of access to the web, more and more interactions between people take place in online social networks, such as Twitter or Facebook, or sites where opinions can be exchanged. At the same time, knowledge is made openly availably for many people, such as by the biggest collaborative encyclopedia Wikipedia and diverse information in Internet forums and on websites.

These two kinds of networks – social networks and knowledge networks – are highly dynamic in the sense that the links that contain the important information about the relationships between people or the relations between knowledge items are frequently updated or changed. These changes follow particular structural patterns and characteristics that are far less random than expected.

The goal of this thesis is to predict three characteristic link patterns for the two network types of interest: the *addition of new links*, the *removal of existing links* and the presence of *latent negative links*.

First, we show that the prediction of link removal is indeed a new and challenging problem. Even if the sociological literature suggests that reasons for the formation and resolution of ties are often complementary, we show that the two respective prediction problems are not. In particular, we show that the dynamics of new links and unlinks lead to the four link states of growth, decay, stability and instability. For knowledge networks we show that the prediction of link changes greatly benefits from the usage of temporal information; the timestamp of link creation and deletion events improves the prediction of future link changes. For that, we present and evaluate four temporal models that resemble different exploitation strategies.

Focusing on directed social networks, we conceptualize and evaluate sociological constructs that explain the formation and dissolution of relationships between users. Measures based on information about past relationships are extremely valuable for predicting the dissolution of social ties. Hence, consistent for knowledge networks and social networks, temporal information in a network greatly improves the prediction quality. Turning again to social networks, we show that negative relationship information such as distrust or enmity can be predicted from positive known relationships in the network. This is particularly interesting in networks where users cannot label their relationships to other users as negative. For this scenario we show how latent negative relationships can be predicted.

Zusammenfassung

Viele Menschen kommunizieren und interagieren zunehmend über soziale Online-Netzwerke, wie Twitter oder Facebook, oder tauschen Meinungen mit Freunden oder auch Fremden aus. Durch die zunehmende Verfügbarkeit des Internets wird auch Wissen für immer mehr Menschen offen verfügbar gemacht. Beispiele hierfür sind die Online-Enzyklopädie Wikipedia oder auch die vielfältigen Informationen in diversen Webforen und Webseiten.

Diese zwei Netzwerkkategorien – Soziale Netzwerke und Wissensnetzwerke – verändern sich sehr schnell. Fast sekündlich befreunden sich neue Nutzer in sozialen Netzwerken und Wikipedia-Artikel werden überarbeitet und neu mit anderen Artikeln verlinkt. Diese Änderungen an der Verlinkung von Menschen oder Wissensbausteinen folgen bestimmten strukturellen Regeln und Charakteristiken, die weit weniger zufällig sind als man zunächst annehmen würde.

Das Ziel dieser Doktorarbeit ist es, drei charakteristische Verlinkungsmuster in diesen zwei Netzwerkkategorien vorherzusagen: das *Hinzufügen von neuen Verlinkungen*, das *Entfernen bestehender Verbindungen* und das Vorhandensein von *latent negativen Verlinkungen*.

Zunächst widmen wir uns dem relativ neuen Problem der Vorhersage von Entlinkungen in einem Netzwerk. Hierzu gibt es zahlreiche soziologische Vorarbeiten, die nahelegen, dass die Ursachen zur Entstehung von Beziehungsabbrüchen komplementär zu den Gründen für neue Beziehungen sind. Obwohl diese Arbeiten eine strukturelle Ähnlichkeit der Probleme vermuten lassen, zeigen wir, dass beide Probleme nicht komplementär zueinander sind. Insbesondere zeigen wir, dass das dynamische Zusammenspiel von neuen Verlinkungen und Entlinkungen in Netzwerken durch die vier Zustände des Wachstums, des Zerfalls, der Stabilität und der Instabilität charakterisiert ist. Für Wissensnetzwerke zeigen wir, dass die Vorhersagbarkeit von Entlinkungen deutlich verbessert wird, wenn zeitliche Informationen wie der Zeitpunkt von einzelnen Netzwerkergeignissen mit genutzt werden. Wir präsentieren und evaluieren hierfür insgesamt vier verschiedene Strategien, die von zeitlichen Informationen Gebrauch machen.

Für soziale Netzwerke analalysieren wir, welche strukturellen Einflussfaktoren zur Entstehung und Löschung von Links zwischen Benutzern in Twitter indikativ sind. Auch hier zeigt sich, dass zeitliche Informationen darüber dass eine Kante schon einmal gelöscht wurde, die Vorhersagbarkeit von Verlinkungen und insbesondere Entlinkungen enorm verbessert. Im letzten Teil der Doktorarbeit zeigen wir, wie negative Beziehungen (beispielsweise Misstrauen oder Feindschaft) aus positiven Beziehungen zwischen Nutzern (etwa Vertrauen und Freundschaft) abgeleitet werden können. Dies ist besonders relevant für Netzwerke in denen nur positive Beziehungen kenntlich gemacht werden können. Für dieses Szenario zeigen wir, wie latent negative Beziehungen zwischen Nutzern dennoch erkannt werden können.

Acknowledgments

First I'd like to thank my family, my husband and my friends for supporting and encouraging me over the last years. This thesis is dedicated to you.

I'm very grateful for the opportunity to do my PhD at the WeST institute with Prof. Steffen Staab and my advisor Jérôme Kunegis. Without your academic guidance and valuable feedback, I could not have done this research! I have greatly benefited from talking to my colleagues at WeST and Gesis during my research. In particular I thank Klaas Dellschaft, Renata Dividino, Jérôme Kunegis, Cristina Sarasua, Felix Schwagereit, Claudia Wagner and Tina Walber for long and productive discussions.

Contents

List of Figures

List of Tables

1 Introduction

The study of social relationships between people has a long tradition in Sociology. The formation, maintenance, and dissolution of social relationships has been widely studied in social networks ranging from married couples to criminal networks and high-school students [Parks, 2007]. With the rise of social networking websites, new means to analyze interactions between people have emerged. Over 1.82 billion users[1] currently use *online social networks* such as Facebook[2], Google+[3] or Twitter[4] to keep in touch with friends and find new friends. Whereas sociological studies had to collect relationship data by questioning or observing individuals, there are now many datasets available which provide information on relationships between online users. If the patterns of online relationships are similar to offline relationships, then these online datasets provide new opportunities to analyze human behavior at a large scale. Conversely, if the patterns of online relationships are different from offline relationships, then it is worthwhile to observe the driving factors of relationships in an online context.

The study of relationships between knowledge items that are captured in so-called *knowledge networks* is also facilitated by the amount of available online data. A non-negligible part of social media is concerned, not with exchanging personal information, but with building knowledge bases. Such knowledge bases are for instance given by any part of the *Semantic Web*, in which knowledge is represented in a systematic manner. Most prominently, the online encyclopedia Wikipedia[5] represents one of the largest online communities dedicated to establishing a knowledge base. The knowledge contained in Wikipedia, rather than being arranged alphabetically or chronologically, as in paper encyclopedias, consists of articles connected by hyperlinks. These hyperlinks have the specific purpose to allow readers a simple navigation in the encyclopedia, and can thereby be considered to represent the linked structure of the knowledge itself.

Relationships between items in a dataset are traditionally modeled as a graph, i.e., a set of nodes that is connected by links. These relationships may represent friendships between users in a social network or semantic relatedness between knowledge items in a knowledge network.

Link Changes Many datasets of online *social networks* and *knowledge networks* are highly dynamic in the sense that existing content is frequently updated or new content is added. These content changes have implications on the structural level of a network; as not only the content is modified but also the relationships between content items are changed. The research field of network analysis seeks for characteristic change patterns that are consistent across various networks to understand the underlying mechanisms that drive the evolution of networks.

In *knowledge networks*, two types of link changes influence the structure of a network. On

[1] `http://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/`
[2] `http://www.facebook.com`
[3] `https://plus.google.com`
[4] `https://twitter.com`
[5] `http://www.wikipedia.org`

the one hand, new connections between knowledge items are established if they carry important information and thus improve the organization of knowledge. For instance, the topics of *Social Network Analysis*, *Graph Theory* and *Sociology* are highly related and should thus appear connected in a knowledge network that contains all academic disciplines. On the other hand, some connections may be wrong or not meaningful enough. Even for very different and unrelated concepts, a relationship can be formulated. Having many of such unimportant connections hinders the navigation of the content. Since knowledge connections in some knowledge networks such as Wikipedia and the web are created by humans, a connection may also be falsely established. An article on quantum physics may be falsely linked to the German soccer player Thomas Müller as opposed to the German physicist with the same name.

Many people are active in online social networks such as Facebook and Twitter or product review websites such as Epinions[6] and Slashdot[7]. Due to the high number of possible users to interact with, these platforms employ recommender systems that help users to find new relational partners or interesting content. The task of recommender systems can then be best explained by predicting links that are likely to appear in a network. On the other hand, users may not always be able to maintain all their online relationships, so they decide to dissolve relationships with people that are not important for them anymore. Sometimes, users may also have reasons to end a specific relationship because of a break-up or a special event such as a bad post or a controversy. In fact, these unlinkings are quite common across different networks; for instance around 25% of all Twitter relationships are terminated [Myers and Leskovec, 2014].

Many popular social networking services allow users only to form relationships with a positive connotation such as friends or followers. Even if users cannot explicitly mark relationships as negative, *latent negative* relationships exist in a social network. For instance, you would rather not add your biggest enemy as a friend on Facebook and would hence not form a friendship with him or her.

Prediction Problems With more and more data – in particular longitudinal datasets – becoming available, the evolution of knowledge and social networks can be observed on a larger scale. Many applications that make use of networks can be described as prediction of links. For instance, recommending friends, predicting friendship dissolution or finding latent negative links in a social network, predicting new connections and spurious or unimportant connections in a knowledge network can be modeled as link prediction problems. Given a current network, a link prediction problem predicts the location of new links or unlinks that will occur in the future network.

In this thesis, we study the problems of predicting new links (*the link prediction problem*), and the problem of predicting link removals (*the unlink prediction problem*) in social networks and knowledge networks. For social networks, we also study characteristic patterns that describe latent negative links – *the latent negative prediction problem*. These three prediction problems are visualized in a small toy network in Figure 1.1.

Whereas structural patterns for the appearance of links have been widely studied in many networks and contexts [Liben-Nowell and Kleinberg, 2003, Lü and Zhou, 2011], the removal of links has substantially received less attention. This is mainly due to the fact that there are only few datasets which contain unlink events and are thus suited for empirical evaluation.

[6] `http://www.epinions.com`
[7] `http://www.slashdot.org`

Figure 1.1: Visualization of the three studied prediction problems.

Nevertheless, unlinking is actually quite common in social networks and knowledge networks [Myers and Leskovec, 2014, Mislove et al., 2013]. Existing studies on link removal in social networks have used platform-specific information or the content that was produced in interactions [Kwak et al., 2012, Kivran-Swaine et al., 2012, Quercia et al., 2012], thus the results are not general and cannot be adapted for other platforms or networks.

The problem of predicting latent negative links is new and was defined by us [Kunegis et al., 2013] to overcome the absence of negative relationships in many social online platforms.

Structural Aspects of Unconnectedness All these relevant problems involve the states and state transitions of non-connected links. A given non-connected link in a network could be the result of an unlink, if the link was present in the network at some point in time. A currently non-connected link could also transform into a connected link, which will create a new link in the network. Some non-connected links in a trust or friend network may also be unconnected because they express a latent negative relationship that cannot be explicitly marked as negative because it is not possible in a given platform. Therefore, the corresponding prediction problems of link, unlink and latent negative prediction all target the finding of *structural characteristics of unconnectedness*. For this research we consider only the structure of the network. The reasons for this are threefold. First, many behavioral change theories relate the changes to structural characteristics. Second, structural models can be compared more easily across different networks, whereas the content diverges. Third, the structural models can easily be extended with additional information, such as content or even external knowledge.

1.1 Research Questions

In this thesis, we study the problem of predicting the appearance of new links, the removal of links and the existence of latent negative links. In the following, we present our four research questions: the first two research questions are related to link changes in knowledge network. The first one asks for characteristics of links and unlinks in knowledge networks, whereas the second research question addresses the influence of temporality for the predictability of links and unlinks. Research question number three is related to link and unlink prediction in social networks. The fourth main research question is related to the prediction of latent negative links in social networks. Our four research questions, which are individually tackled in one

chapter, are as follows.

Chapter 3: Predicting Link Additions and Removals in Knowledge Networks

In 2003, Liben-Nowell and Kleinberg were the first to define and tackle the link prediction problem [Liben-Nowell and Kleinberg, 2003]. Since then, many measures and models were developed to improve the prediction of new links. On the contrary, unlink prediction is a relatively new problem. Before we started our studies, this problem has not been tackled in a systematic and purely structural matter. Given only the structure of a network, the goal is to find structural measures to predict links that will be removed. Hence, the corresponding research question and subquestions are as follows.

RQ 1 *Which structural characteristics are indicative for the removal of links?*

In particular, we ask whether indicators for links can be used to characterize unlinks as well. In the past, link prediction has already been extensively studied, whereas the prediction of unlinks has only been researched in a handful of studies [Quercia et al., 2012, Kwak et al., 2012, Kivran-Swaine et al., 2012, Kwak et al., 2011]. Do we need to consider both problems or is one problem enough to draw conclusions about the other?

RQ 1.I *How are unlinks related to new links, i.e., can characteristics of new links be used to characterize unlinks?*

If one problem can be reduced to the other one, then classic link prediction measures can be used to predict unlinks as well. We hypothesize that the two problems are highly related: factors that drive the formation of new links should hinder the removal of links and vice versa.

RQ 1.II *What is the interplay of link and unlink dynamics?*

This question sets out to answer how numerical indicators of a link can be interpreted for link and unlink prediction. Both problems have so far only been considered separately, so this line of research will aim to provide a unified view of both problems.

Chapter 4: Temporal Models of Knowledge Networks

Many collaborative knowledge networks evolve rapidly – knowledge items are added, interlinked and revised constantly, thus reflecting the fast changes in many knowledge areas. Classic link prediction approaches represent the relationships in a network dataset by one single snapshot from which only static characteristics can be computed. Since change is an inherently temporal phenomenon, we may ask the question whether change in Wikipedia's hyperlink structure is mediated by temporal phenomena. Hence, we propose to employ a temporal approach to evaluate whether recent or long-lived connections have a bigger influence on the formation of links and unlinks. Hence, the corresponding research question and subquestions are as follows.

RQ 2 *Does the exploitation of temporal data improve the classification of new links and unlinks?*

If the timestamps of individual addition and removal events are given for a dataset, how can this temporal information be exploited for the prediction of links and unlinks?

RQ 2.I *What strategies would be adequate to exploit temporal informations as to classify new links and unlinks?*

Information of addition and removal events can be leveraged on different levels; one could use the specific timestamp of an event, use only the ordering of events or exploit the qualitative information how often a link has been added or deleted.

RQ 2.II *Does the exploitation of temporal data improve the classification of new links and unlinks?*

Will the prediction results be significantly better than without temporal information? The snapshot representation of a dataset does not provide any evidence to whether links that are not in the snapshot have been present before. We hypothesize that information on unlinks, that can be extracted from temporal data, should improve the predictability of new unlinks.

Chapter 5: Predicting Link Additions and Removals in Social Networks

There are various sociological constructs that explain the formation and dissolution of social relationships [Parks, 2007]. Building upon this existing body of work, we want to translate these constructs to predict the formation and dissolution of directed relationships between users in a social network, where latent or explicit user groups are given. Due to social mechanisms such as group conformity [Bernheim, 1994], users are influenced by their friends and groups (e.g. teams, organizations, parties) that they belong to.

RQ 3 *Which structural characteristics predict link formation and dissolution in directed social networks with latent or explicit groups?*

In the related work, some characteristics were shown to be indicative for the formation of a tie, while other characteristics were found to correlate with the dissolution of a tie. One can then asses which influence factors have the highest impact on the prediction of new links and unlinks.

RQ 3.I *Which influence do structural characteristics have on the prediction of new links and unlinks?*

While many datasets provide only a snapshot of the network that does not lent itself to derive unlinks [Kunegis, 2013], a dataset consisting of multiple snapshots can be used to derive links *and* unlinks.

RQ 3.II *What is the added value of unlink data for link and unlink prediction?*

This research question targets the question of how useful this additional unlink information is, i.e., how much the prediction of new links and unlinks is improved when unlink data is exploited.

Chapter 6: Latent Negative Links in Social Networks

Many social networking websites such as Facebook, Google+ or Twitter prohibit the user to label relationships as negative; users can only be added as friends or followers. Although it is not possible in these platforms to explicitly label other users as foes or distrusted, users implicitly

have negative relationships or opinions about other users. We refer to these relationships as *latent negative*. Some platforms allow users to sign their social network, i.e., to define friends and foes. We will use datasets of two such platforms to evaluate how negative links are embedded in the network of positive links.

RQ 4 Which structural characteristics are indicative for latent negative links in social networks?

For the first scenario, we assume that only the positive links, e.g. all friendships in a network, are given. The goal of this research is to find characteristic patterns for negative links in the network consisting of only positive relationships.

RQ 4.I *Which structural indicators infer negative links from only positive links?*

Some networks do not allow the user to label relationships as negative. Therefore we ask what the added value of the negative link feature for the prediction of negative links is. For that we compare two settings: How much easier is it to predict latent negative ties, when some negative information is used in contrast to the sole usage of only positive ties.

RQ 4.II *What is the added value of the negative link feature?*

The predictive performance of the two prediction settings will be compared to obtain the added value of the negative link feature.

1.2 Contributions

The contributions of the work in this thesis are twofold. First, we propose several new models and approaches for link prediction problems. Second, we perform several experiments with overall seven knowledge networks and three social networks and obtain new insights into factors that drive the formation and dissolution of links and the existence of latent negative links. The specific contributions for each chapter are as follows.

Chapter 3: Predicting Link Additions and Removals in Knowledge Networks

Complementarity We have proposed two different transformations of unlink prediction problems as link prediction problems, the *complement score* and the *complement network* model. In an empirical evaluation, we found that the *complement score* model is superior over the *complement network* model and that the *complement network* model performs worse than a random baseline for most datasets [Preusse et al., 2012]. With this research, we have shown that unlink prediction cannot be understood as a simple transformation of link prediction.

Interplay To study the interplay of unlink and link prediction, we have defined a unified view for both problems [Preusse et al., 2013]. In networks links are added and removed, links can be classified into the four states of *growth*, *decay*, *stability* and *instability*. Whereas growing links are likely to be added, decaying links are likely to be removed. The distinction between stable and instable links classifies, whether the state of a node pair will not change or is likely to change between growing and decaying which we defined as unstable. We have presented structural indicators for each category and refined the link and unlink indicators as indicators of growth and decay [Preusse et al., 2013].

Characteristics of Unlinks As the previous contributions reveal, link and unlink prediction are not symmetric problems. Having further specified the problem of unlink prediction as the problem of predicting links that are likely to be removed but not to be added again, we have evaluated the predictive performance of several structural characteristics at the task of unlink prediction. Features of the embeddedness of a relationship, e.g. common neighbors or common neighbors of neighbors, have shown to perform well for several knowledge network datasets [Preusse et al., 2013, Preusse et al., 2012].

Chapter 4: Temporal Models of Knowledge Networks

Temporal Models Using temporal information on link additions and removals, we presented and implemented four models of temporal change [Perl et al., 2014a]. In contrast to the time-agnostic setting, the *qualitative model* captures which links have been removed. The *decay model* exploits the ordering of changes and the *neighborhood evolution model* uses the evolution of an article's neighborhood to reason about an article's future.

Added-Value of Temporal Data We have shown that temporal information improves the classification of links and unlinks significantly [Perl et al., 2014a]. Data on unlinks should not be discarded, but serves as valuable indicator for new links and unlinks. Further, we have demonstrated the theoretical feasibility of unlink prediction by using the actual neighborhood size as opposed to an estimation for the neighborhood evolution model.

Chapter 5: Predicting Link Additions and Removals in Social Networks

Computational Social Science Approach for Link Changes Given an overview of social theories that aim to explain the formation and dissolution of social ties in a network, we have presented a computational approach for quantifying new links and unlinks [Perl et al., 2014b]. For that, we developed a model of influence factors that describes the network effects that lead to the formation and dissolution of social ties by means of the network's structure and information on latent or explicit group associations for users. Our model can be applied to any directed social network where explicit or latent group memberships are given.

Predictive Performance of Influence Factors We have demonstrated the utility of our approach in a case study about the evolution of the social network of German politicians on Twitter and present our empirical results on the impact of different theoretical influence factors on the formation or dissolution of ties in a social network of politicians [Perl et al., 2014b]. Our results show amongst others, that the tie formation behavior of a user is more in line with the tie formation behavior of his friends or group members than a user's tie dissolution behavior.

Added Value of Unlink Information We have shown that measures based on information about past links are extremely valuable for predicting the dissolution of social ties, while measures based on the link network are sufficient for the prediction of new social ties [Perl et al., 2014b]. For that, we have compared two classifiers that utilize only link information respectively link and unlink information.

Chapter 6: Latent Negative Links in Social Networks

Definition of the Latent Negative Prediction Problem We have defined a new and interesting prediction problem that has many applications for networks with positive and negative relationships as well as for networks where only positive relationships can be expressed [Kunegis et al., 2013]. Further, we have defined the prediction set up to evaluate the predictive performance of any measure for the task of predicting latent negative links.

Characteristics of Latent Negative Links When only positive links in a network are given, we have measured that a combination of page-rank and cosine-similarity performs best to predict all known negative links [Kunegis et al., 2013]. Further, we have demonstrated that the added value of the negative link feature that is employed in only few platforms, is only minor for the prediction of negative links. This implies that negative links can only be predicted slightly better in platforms with negative and positive links than in platforms with only positive links.

1.3 Publications

This thesis contains work that was reported in five papers.

[Preusse et al., 2012] In this submission to arXiv, two possible transformations from the unlink prediction into the link prediction problem were evaluated. The implementation and the main paper work was done by me.

[Preusse et al., 2013] The analysis of the interplay of link and unlink prediction is published in a paper at the International AAAI Conference on Weblogs and Social Media 2013. I implemented my own idea and performed the analysis on my own. Regarding the publication, I wrote the majority of the paper.

[Perl et al., 2014a] Models that exploit temporal information for new links and link removal were proposed and analyzed in this work that is yet only published in a technical report. I implemented my own idea and performed the analysis on my own. The majority of the paper was written by me.

[Perl et al., 2014b] In this submission to the International Conference on Social Informatics, I have analyzed link formation and link dissolution behavior of users in directed social networks using a case study of German politicians on Twitter. I implemented my own idea and performed the analysis on my own. For the submission, I wrote the majority of the paper.

[Kunegis et al., 2013] Analysis on Features that predict latent negative links in signed networks was published at the International World Wide Web Conference 2013. I had the idea for this work and developed the methodology and evaluation together with Jérôme Kunegis. I wrote large parts of the paper, but the majority of text was written by Jérôme Kunegis.

1.4 Outline of the Thesis

The thesis is structured as follows.

Chapter 2, *Foundations*, introduces the basic concepts and notations of networks. Link state prediction problems are introduced and formalized as a class of link mining problems and the prediction framework is described. The linking behavior of individuals is summarized, including studies on the formation, maintenance and dissolution of personal relationships. Prediction models to solve link state prediction problems are presented.

Chapter 3, *Predicting Link Additions and Removals in Knowledge Networks*, studies whether an unlink prediction problem can be transformed into a link prediction problem. The interplay of both prediction problems, resulting in four states of a link, is evaluated experimentally for the knowledge network Wikipedia.

Chapter 4, *Temporal Models of Knowledge Networks*, evaluates how much temporal information improves the prediction of links and unlinks in Wikipedia. For that, four temporal models are described and evaluated.

Chapter 5, *Predicting Link Additions and Removals in Social Networks*, studies the linking and unlinking behavior of users in directed social networks. Nine sociological influence factors are translated to structural measures and evaluated on a Twitter dataset of German politicians.

Chapter 6, *Latent Negative Links in Social Networks*, examines the problem of predicting latent negative links in two social datasets. For that, the prediction of negative links from positive links is evaluated.

Chapter 7, *Conclusions and Future Directions*, concludes the work of this thesis and shows limitations and future directions.

2 Foundations

In this chapter, we familiarize the reader with the mathematical concept of a network, network properties and evolution of networks. We give an overview of general link mining problems and describe the three link state prediction problem which are treated in this thesis. Related sociological work on changes in relationships is surveyed. Mathematical methods and models for the prediction of link state changes are presented.

2.1 Foundations of Networks

The relationships between objects are commonly represented using the formalism of a *network*. A network N is formally defined as a tuple $N = (V, E)$ of objects or vertices V and relationships or links between objects E. An example of a network is the structure of Wikipedia, where articles link to each other so that users can navigate the article pages in Wikipedia. In this example articles represent the objects and the relations between articles are formed by article links. Throughout this thesis, we use the terms link, edge and tie as well as the terms vertices and nodes interchangeably.

2.1.1 Networks Types

Directed versus Undirected Networks In a directed network, each link has a particular direction. For instance the fact that user i follows user j on Twitter is modeled by an edge (i, j). Therefore all edges $e \in E$ are node tuples, which is commonly indicated by the notation (i, j) if i relates to j. In an undirected network, the direction of edges can be dropped, because all relationships are bidirectional. An example is a friendship network; if person i is a friend of person j, then j is also a friend of person i. Edges in undirected networks are therefore commonly represented by sets of node pairs, where $\{i, j\} \in E$ if the two entities i and j are in a relationship with each other, which is indicated by the notation $i \sim j$. In this thesis, we study *state networks* only and thus introduce the distinction between networks *state* and *event* networks as follows.

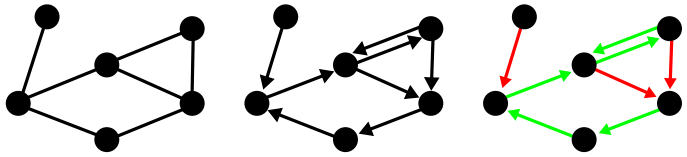
State versus Event Networks In general, *state networks* reflect the state of a relationships between two nodes, i.e., either there is a relationship between two nodes or there is none. Therefore state networks have no parallel edges; two nodes are either connected or not. Examples of a state network are friendship networks, the hyperlink network of Wikipedia articles, or an employment network that contains the information which worker is employed at which company.

Contrarily, *event networks* contain events or actions between two nodes, such as the postings of two users, paper-collaboration of two authors or online users that reply to each other in an online forum. These network types allow parallel edges, since more than one action can take place between two nodes. Event-networks are fundamentally different from state networks. For

a state network, the state of a given relationship is always defined: two nodes are either in a relationship or not. On the contrary, the relationship between two nodes in an event network is only defined for the specific time point of the event. If two users have exchanged mails, what is the state of this relationship two hours after the message exchange? Are the users still connected? What about two years after the message exchange?

Signed Networks Whereas the networks described before only capture whether two nodes are in a relationship or not, *signed networks* additionally designate a sign – either positive or negative – to the relationship. Signed networks can be undirected, but are mostly directed in real networks. For instance, in a network of persons, one can thereby express friends and foes and who trusts or distrusts whom. Signed networks are a specific kind of *weighted* network, where in general a weight is assigned to a relationship.

Figure 2.1 shows an example for each network type considered in this thesis.



(a) undirected (b) directed (c) signed

Figure 2.1: Example networks for (a) undirected networks, e.g. friendship networks, (b) directed networks, e.g. hyperlink networks, and (c) signed networks, e.g. trust networks.

2.1.2 Network Characteristics

The network structure of an *unsigned* network can be represented with the network's *adjacency matrix* A, which is a binary matrix $A \in \{0,1\}^{|V|\times|V|}$ where an entry $A(i,j)$ is defined as

$$A(i,j) = \begin{cases} 1, & \text{if } i \text{ links to } j, \\ 0, & \text{else.} \end{cases}$$

The adjacency matrix of an undirected network is thus symmetric, since

$$\{i,j\} \in E \Leftrightarrow A(i,j) = A(j,i) = 1.$$

For a signed network, the values in the adjacency matrix take either 0, 1 or -1 depending on the sign of the relationship.

There are several interesting network statistics and characteristics that describe an individual network. We describe them in the following.

Density The density of a network is a measure of the sparseness of the adjacency matrix, which is computed by the ratio of the actual number of edges in the network and the number of possible edges in the network. A network with n nodes is maximally connected when it is

complete, i.e., every node is adjacent to each other. Thus, the density of a network N is defined as

$$\text{density}(N) = \frac{2|E|}{n(n-1)}.$$

Generally speaking, most large networks are sparse, i.e., only a small fraction of the possible edges exist [Kunegis, 2011].

Degree Distribution The degree distribution of a network is a function $P(d(x) = k)$ which describes the fraction of the nodes x which have degree $d(x)$ of k. The degree distribution describes how the links in the graph are distributed among the nodes. For many networks, the degree distribution is heavy-tailed, i.e., only a small fraction of the nodes have a very high degree, while the majority of nodes have a small degree. The degree distribution of many networks more specifically follows a power-law [Barabási and Albert, 1999], i.e., the probability that a node x has degree k is given by

$$P(d(x) = k) \sim k^{-\alpha},$$

where α is the so-called power-law exponent. For many real-world networks, power-law coefficients between 2 and 3 have been observed, e.g. [Barabási and Albert, 1999]. Despite the popular usage of this descriptive characteristic, some fundamental criticism has been raised because only few degree distributions seem to significantly follow a power-law [Lima-Mendez and van Helden, 2009, Clauset et al., 2009]. Since many studies seem to compute the power-law exponent only to indicate the skewness or inequality of a distribution, other measures such as the gini coefficient[1] seem more appropriate and can be universally measured even for non-power-law distributions [Kunegis and Preusse, 2012].

Connectivity An undirected network is called *connected* if each node can be reached from each other node. Directed networks are called *strongly connected* if each node can be reached from each other node on a correctly directed path. If each node is only reachable when ignoring the direction of edges, then the network is called *weakly connected*. Since most networks contain several isolated nodes or groups of nodes, they are are *not* connected. Then, the *largest connected component* is an inclusion-maximal set of nodes for which a path between all pairs of nodes exists. For directed networks the weakly largest connected component is then defined for undirected paths between node pairs, whereas the strongly largest connected component is defined for directed paths between all node pairs.

Radius and Diameter The eccentricity of a node is the distance of the longest shortest path between it and any other node. The radius is the minimum eccentricity among all nodes, whereas the diameter is defined as the maximum eccentricity. Expressed otherwise, the diameter is the longest of the shortest paths in a network. Following this definition, radius and diameter are defined as ∞ in an unconnected network. Henceforth, these two network measures are usually computed only for connected node pairs in the largest connected component. Since

[1] The gini coefficient is defined as twice the area under the Lorenz curve, where the Lorenz curve is defined as the set of points (x,y), where the share x of nodes with the lowest degree covers a share y of all edges.

the diameter is very sensitive to outliers such as only several long link chains in the network, the effective diameter is then used. The effective diameter of a graph is the characteristic number for which 90% of the graphs maximum distances between two nodes are smaller than or equal that value [Leskovec et al., 2005].

Clustering Coefficient The clustering coefficient of a node is the ratio of existing links between the node's neighbors and possible ones. It is given by the ratio of triads and the number of possible triads and defined for an undirected network N as follows:

$$\text{Clusco}(N) = \frac{|\{\{j,k\} \in E \mid \{i,j\} \in E \wedge \{i,k\} \in E\}|}{|\{\{i,k\} \in E \mid \{i,j\} \in E\}|}.$$

The clustering coefficient of a network thus reflects the amount of transitivity in a network; if nodes i and j form a relationship and a relationship between i and k is also present, how likely is there a relation between j and k, as well?

2.1.3 Network Models

To describe and understand the effects that lead to the structure of real-world networks, different network evolution models have been proposed. These models are described to match common global characteristics of networks, such as the power-law of the degree distribution or a high-clustering coefficient. In the following, we give an overview over some well-known network models.

The simplest graph model is the *Random Graph Model* [Erdös and Rényi, 1959]. Every edge in this model exists with the same global probability γ. Random graph models fail to produce a power-law degree distribution and fail to capture the amount of clustering that is observed in many networks and particularly social networks [Watts and Strogatz, 1998].

The *Preferential Attachment Model* suggests that the likelihood of a node to form new links is proportional to its in-degree (the number of its neighbors) [Barabási and Albert, 1999, Barabási et al., 1999]. Thus, the more links a node has already received, the more it will receive in the future – the "rich get richer" phenomenon. This network model has been shown to produce networks with power-law degree distribution [Barabási et al., 1999].

The *Assortative Mixing Model* implements that nodes are more likely to form links with nodes of similar degree [Newman, 2002]. Whereas the preferential attachment model tends to produce networks where mostly low-degree nodes are connected to high-degree nodes, the assortative mixing model produces networks that follow real-world observations in which high-degree nodes in networks tend to connect to other high-degree nodes [Newman, 2002, Mislove, 2009].

The *Small-World Model* results from randomly replacing a fraction p of the links of an n-dimensional ring lattice with random links [Watts and Strogatz, 1998]. It has been shown that this model reproduces the clustering coefficient and the characteristic path length – the average of all shortest paths between all node pairs – better than the random graph model [Watts and Strogatz, 1998].

The *Copying Model* expresses the probability of a new edge in terms of the probability of copying one of the neighboring node's neighbors or the degree of a node in the network [Kleinberg et al., 1999, Kashima and Abe, 2006] . For instance, when writing a paper one finds a new related work and cites some of the same papers that are cited within.

The *Forest Fire Model* is an extension of the copying model which copies only out-going links of a node in that it also considers incoming links of other nodes [Leskovec et al., 2005]. In the forest fire model, a new node randomly connects to existing nodes and then *burns* links outwards or inwards from this node, meaning that each link found on the out-going or in-coming path is copied with a certain probability. The forest fire model particularly captures the evolution of two characteristics: the shrinking diameter and a faster growth of links than of nodes that can be observed in many networks. Using the same example as for the copying model, one does not only look for the related work that is cited within one paper, but also considers papers that cite the work.

In the *p1 Model*, the probability of each edge is defined by a log-linear combination of features of tie characteristics [Holland and Leinhardt, 1981]. Importantly, the p1 model assumes the independence between ties, i.e., the existence of one tie does not influence the existence of other ties.

Exponential Random Graph Models allow a generalization beyond the restrictive dyadic independence assumption of the p1 model class [Frank and Strauss, 1986, Anderson et al., 1999]. Accordingly, this class permit models to be built from a more realistic interpretation of the structural foundations of social behavior. Exponential random graphs focus on local statistics of the graph, such as the number of triads or reciprocated dyads. Each statistic is weighted by a parameter that can be interpreted as the log-odds of a tie conditional on the other statistics which are fixed. Thus, a negative value indicates that the statistic is observed less often than by chance, whereas a positive value signals that the feature occurs more often than expected by chance.

2.1.4 Evolving Networks

We consider a scenario of evolving networks N_t, where

$$N_t = (V_t, E_t)$$

for $t \in N$ is the network N_t at time t with V_t being the set of nodes of N_t and $E_t \subseteq V_t \times V_t$ the set of links of N_t. Without loss of generality, we assume that $V_t = V_{t'}$ for all $t, t' \in N$, otherwise we could define $V_1 \cup V_2 \ldots$ to be the set of nodes for each network.

Given an evolving network, structural changes in a network can be studied on two levels: changes on the micro-level and changes on the macro-level of the network.

Micro-Level Micro-level changes are defined as changes on the node-level or changes on the link level of the network structure. New nodes may either enter the network or be deleted. User-focused research has studied the evolution of user characteristics within the network structure, e.g., when users leave a network [Karnstedt et al., 2010] or how users change their activity over time [Rowe, 2013]. In our research, we disregard node changes and keep a constant set of nodes throughout our methods. Since we study state networks, changes on the link level are transformations of states - a link that was previously not present can either be added or a link that is present can be removed. There are two lines of work that study the evolution of links. The first line of research aims to determine the characteristics of new links, i.e., how new links are embedded within the current network. For instance large sequences of link additions were analyzed to understand the driving factors of this process [Leskovec et al., 2008]. This

study shows that most new edges span very short distances, typically closing triangles or new links are attached to higher-degree nodes. Changes in the link structure have been analyzed to understand how communities or networks evolve (e.g. community evolution studies how dense subgraphs in a network change over time) or to understand what drives the decay of a whole network [Garcia et al., 2013].

Contrarily, the research area of link prediction problems seeks to accurately predict the addition and removal of links. Since the research in this thesis studies link prediction problems, we will explain this topic in more detail in Section 2.2.

Macro-Level Despite frequent micro-level changes in the network, the global characteristic of networks remain relatively stable [Viswanath et al., 2009, Parks, 2007]. Even though many ties are added or removed throughout the evolution of a network, descriptive network statistics such as the clustering coefficient, the average node degree or the size of the connected component change only slightly [Kossinets and Watts, 2006]. Many real-world networks have been shown to become slightly but significantly denser over time and that the effective diameter is shrinking [Leskovec et al., 2005]. The densification of networks means that the number of edges grows faster than the number of nodes and the shrinking diameter indicates that real-world networks consist of bridge nodes that succeed to connect previously unconnected nodes.

2.2 Link State Prediction Problems

The previous section described how relationships between actors are commonly represented as a network. Whereas information can also be stored in the form of node attributes, such as the age or gender of an actor, the structural information contained in the links alone often suffices to make educated guesses about the network's future. Links carry important information, such as the importance of an actor or which communities exist in a network. This very structured knowledge is leveraged by different problems.

In the following, we focus on link state prediction problems, in particular the *link prediction problem*, the *unlink prediction problem* and the *latent link prediction problem*. These link state prediction problems belong to the category of *Link Mining Problems*, which are defined as problems that solve network-related tasks by using the links of the network [Getoor and Diehl, 2005]. To understand the differences and similarities between these three problems, we first describe the general setup for a link state prediction problem and then describe the three prediction problems.

Link prediction problems try to predict the state of a link, given the state of other links. In general, all link state prediction problems have the same set up:

Input:

- node pairs in the *training set*
- node pairs in the *test set*, i.e., node pairs in the *true test set* that should be predicted and node pairs in *false test set* which should not be predicted.
- a prediction function

Procedure:

1. Compute the prediction function for all node pairs of the test set on the training set.
2. Rank node pairs in descending order of their value of the prediction function.
3. Measure the quality of the ranking.

Output:

- The quality of the ranked node pairs in the test set.

Optimizing Function:

- Choose a prediction function to maximize the quality of ranked node pairs in the test set such that node pairs in the true test set are ranked higher than node pairs in the false test set.
- The maximum is reached when *all* node pairs in the true test set are ranked better than any node pair from the false test set.

Hence, each particular link state prediction problem can be characterized by the specific choice of the training set, the test set consisting of the true and false test set and the prediction function. Note that the test set is solely used for testing the predictive performance of a prediction function. Thus, no information in this set is allowed to be used for the actual prediction.

We introduce the notation of

$$\mathcal{P} : \text{Training set} \rightarrow \text{True test set} \mid \text{False test set},$$

to formalize a prediction problem P by its training, true and false test set.

We distinguish two kinds of link state prediction problems: those that predict *state changes* from current to future links and those that predict the *status* of current links. Accordingly, the set-up for both prediction classes is slightly different: whereas the input data for state change problems is temporally split, the input data for status predictions is split into known and left-out edges.

This thesis will study three problems: the link addition prediction problem – abbreviated as *link prediction problem*, the link removal prediction problem –abbreviated as *unlink prediction problem*, and the *latent negative prediction problem*. The link prediction problem seeks to accurately predict edges that will appear in the future, given the current set of links. The unlink prediction problem targets the prediction of links that will be removed in the future, given the current links. Therefore, these two problems are considered as *state change prediction problems*. The latent negative problem infers negatively signed links from positively signed links disregarding the temporal dimension. Thus, we consider the latent negative problem as a *status prediction problem*. The three prediction problems are displayed in Figure 2.2. We define the two classes of link state prediction problems in the following.

2.2.1 Link State Change Prediction – Link and Unlink Prediction

Link state change prediction problems aim to predict how the state of a node pair will change from time t_1 to time t_2. Hence, for this problem category one needs to consider the temporal evolution of links, i.e., it is important whether a link was added before or after time t_1.

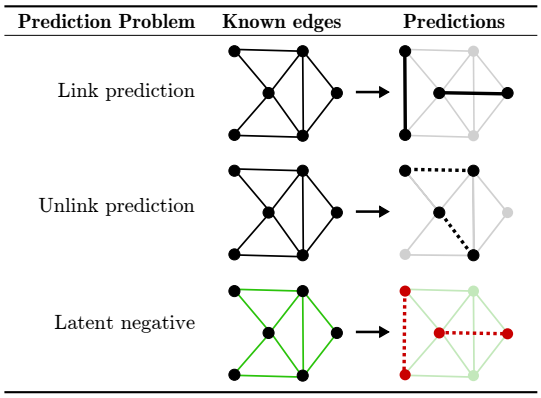


Figure 2.2: Overview of three link state prediction problems.

In order to predict the changes of links we consider a scenario of evolving networks. Let

$$N_t = (V_t, E_t)$$

for $t \in N$ be the network N_t at time t with V_t being the set of nodes of N_t and $E_t \subseteq V_t \times V_t$ the set of links of N_t. Without loss of generality, we assume that $V_t = V_{t'}$ for all $t, t' \in N$, otherwise we could define $V_1 \cup V_2 \ldots$ to be the set of nodes for each network. We also write $N_t = (V, E_t)$ for $t \in N$ and define $n = |V|$. Depending on whether one performs an unsupervised or supervised prediction, the network will be given for two or respectively three timepoints. Given the set of

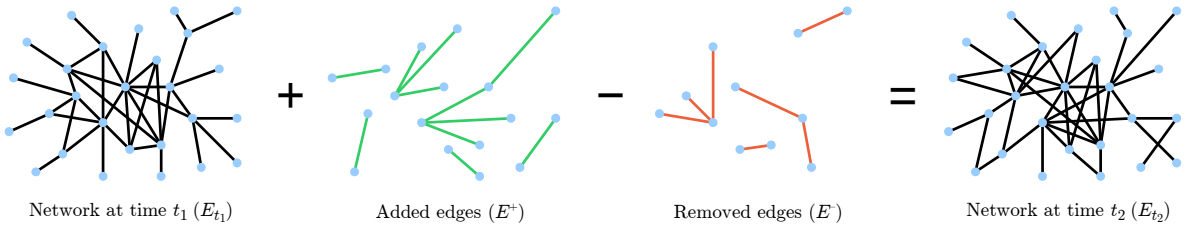


Figure 2.3: Schematic representation of the link addition and removal process. At time t_1, the network has the edge set E_{t_1}. After t_1, the set of edge E^+ is added and the set E^- is removed, giving the set of edges E_{t_2} at time t_2. Link directions are not indicated in the figure.

links E_{t_1} present at a particular time t_1, we want to solve the problem of how to predict state changes of node pairs between time t_1 and time t_2. Since there are two possible state changes – edges may either be added or removed – there are two different prediction problems. The first one is how to predict new edges E^+ and the second one targets the prediction of deleted edges E^-, where

$$E^+ = E_{t_2} \setminus E_{t_1},$$
$$E^- = E_{t_1} \setminus E_{t_2},$$

such that

$$E_{t_2} = (E_{t_1} \setminus E^-) \cup E^+.$$

The sets of added and removed links are illustrated in Figure 2.3. The problem of predicting new links E^+ is called the link addition prediction problem, or simply the *link prediction* problem [Liben-Nowell and Kleinberg, 2003]. Typically, the link prediction problem is solved by *link prediction functions*, i.e., functions that map node pairs to numerical scores, based on the known edges in the set E_{t_1}. The problem of predicting the removal of edges – called the *unlink prediction* problem – can then be solved analogously by *unlink prediction functions*.

Data Split For both prediction problems, the data is split into a training and a test set at time t_1. Commonly, t_1 is the temporal proportion of node pairs in the training set to node pairs in the test set is 3:1, which means that t_1 is chosen as $t_1 = \frac{3}{4} \cdot t_2$ which corresponds to a 75% : 25% data split. For link prediction, we consider only node pairs in the largest connected component of the network. The reason for this is that predictions for nodes that are not connected by any path hardly make sense, since the structural information that can be used for this prediction would be too weak. This step is not necessary for the unlink prediction problems, where all unlinked node pairs must have been connected before and thus enough structural information can be used. Figure 2.4 depicts the *temporal* data split for link state change prediction problems. For the link prediction problem, node pairs in the true test set E^+

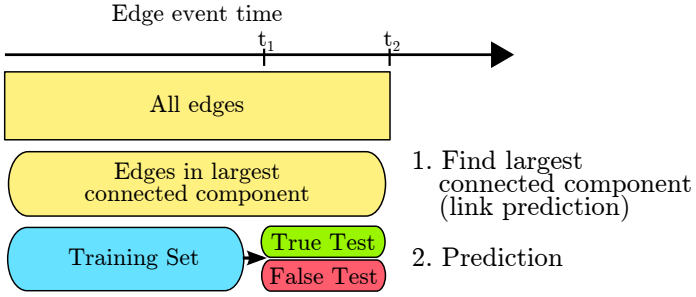


Figure 2.4: The data split for link state change prediction problem is depicted. First, the largest connected component of the network is computed. This set is then split into training, false and true test set to perform the prediction.

must be distinguished from those that where not added, i.e., those in the false test set E^+_{false}. Analogously, the prediction of link removal aims at distinguishing links that are removed, in the true test set E^-, from those that are not removed, in the false test set E^-_{false}. The set E^-_{false} is thus defined as

$$E^-_{\text{false}} = E_{t_1} \cap E_{t_2}.$$

The set E^+_{false} is defined as a random sample of node pairs from the set of node pairs which are neither connected at time t_1 nor at time t_2

$$\begin{aligned} E^+_{\text{FALSE}} &= V \times V \setminus (E_{t_1} \cup E_{t_2}), \\ E^+_{\text{false}} &\subset E^+_{\text{FALSE}} \text{ , with} \\ |E^+_{\text{false}}| &= |E^+|. \end{aligned}$$

Note that E^+_{false} is a sample of non-edges because most real-world datasets are very sparse which means that there are by far more non-edges than actual edges. Computing the predictive function for *all* non-edges would be too time-consuming. The link prediction problem P_L and the unlink prediction problem P_U are thus formalized as

$$\begin{aligned} \mathcal{P}_L &: E_{t_1} \to E^+ \mid E^+_{\text{false}} \\ \mathcal{P}_U &: E_{t_1} \to E^- \mid E^-_{\text{false}}. \end{aligned}$$

Parameter Training When the prediction function contains parameters, such as the weight of an individual feature in a feature regression model, these parameters need to be trained. For that, the data is split into a source and target set at time $t_0 = \frac{3}{4}t_1$, where the parameters are trained from the source to the target set. Having trained the parameter values, one can then apply the trained link prediction function as in the unsupervised scenario from the training to the true and false test set. Thus, the trained classifier is then evaluated on unseen data. Figure 2.5 depicts the data split for parameter training.

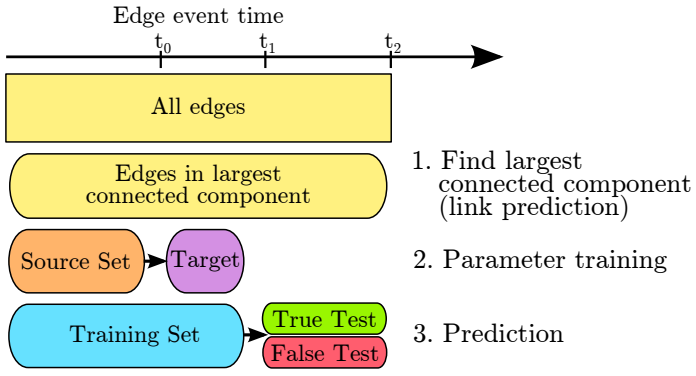


Figure 2.5: The data split for parameter training of state changes is illustrated. The dataset is additionally split into a source and target set to train the parameters of the prediction function.

New Link Prediction versus Repeated Link Prediction Since this work is only concerned with state networks which do not contain parallel edges, we study the prediction of *new* links. In *event networks* with parallel edges, an edge may be added that was present in the network before. We call these kinds of links *repeated* links and the corresponding problem *repeated link prediction*. If we consider a posting event network, the distinction between the two kind of link changes becomes clearer. Whereas the *new link prediction problem* seeks user pairs that have

not had a post-exchange so far, the *repeated link prediction problem* seeks user pairs that will interact again.

Edge-centric versus Node-centric Approaches Most link prediction frameworks measure the existence likelihood of a given set of links, thus these frameworks can be considered as edge-centric. The scenario for node-centric approaches takes a set of users as input and tries to predict which links are the likeliest to appear for these users. This task is relevant for user recommendations, where items or other users (e.g. friends) are recommended to users [Tylenda et al., 2009].

2.2.2 Link Status Prediction – Latent Negative Prediction

Whereas the time of an addition or removal event for link state change problems is considered, link status prediction problems disregard the temporal components in the data. Instead, link status prediction problems aim to correctly classify the status of a left-out set of node pairs.

The input network $N = (V, E_w)$ is defined as a time-independent set of nodes V and set of weighted links E_w. The set of weighted links is here defined as

$$E_w \subset V \times V \times \mathbb{R},$$

where w is an additional weighting function that assigns a weight to each node pair, $w : V \times V \to \mathbb{R}$.

In general, link status prediction problems then predict the weight of a left-out set of node pairs E_b given known node pairs E_a, where $E_a \cap E_b = \emptyset \wedge E_a \cup E_b = E$, i.e., both node pair sets are a disjunct decomposition of the set of all node pairs.

For the *latent negative* prediction problem, the set E_a consists of a subset of positive links and E_b consists of all negatively signed links and the remaining positively signed links. The problem of predicting the correct state of nodes in E_b is then solved by *link status prediction functions*. In the following, we define the prediction methodology for signed networks only. In *signed* networks, the weighting function more specifically assigns a value of -1 or 1 to all node pairs. Therefore, the set of edges E contains the disjunct edge sets of positively signed edges P and negative signed edges N. Note that we assume that both sets are disjunct, that means in particular that a link cannot change its sign.

Data Split Analogously to link state change prediction problems, the dataset is split into a training and a test set that consists of a true and a false test set. To ensure that sufficient structural information is available for the targeted prediction nodes, only nodes that are in the network's largest connected component are considered for the prediction problem. This is done because no structural information can be leveraged to predict the sign of node pairs that are not connected to others.

We define the data split for the latent negative problem as follows. First, the set of all positive links P is split into two sets P_a and P_b, such that $|P_a| = 3 \cdot |P_b|$, which corresponds to a 75% : 25% split. The set P_a corresponds to the training set, whereas P_b will be part of the false test set. The true test set N is then formed by all negatively signed links. The larger set of positively signed links P_a is then used to predict links that are signed negatively, N, against links that are signed positively P_b and non-links that we denote as O. Figure 2.6 depicts the

data split for the latent negative problem. Further, one also needs to ensure that negatively

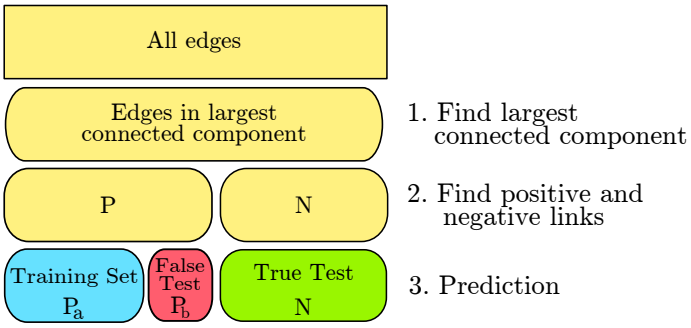


Figure 2.6: The data split for the latent negative problem is depicted. The set of positively signed links is split into a training and a false test set. The false test set then consists of all negatively signed links.

signed links are distinguished from positive links and negative links are distinguishable from non-links. Thus, the latent negative problem is defined as follows

$$\begin{aligned} \mathcal{P}_{LN} : P_a &\to N|P_bO, \\ P_a &\to N|P_b, \\ P_a &\to N|O. \end{aligned}$$

Link Sign Prediction The problem of predicting which positive and negative edges will appear is called the link sign prediction problem. In this prediction scenario, unlabeled links are given and are classified either as positive or negative. For example, positive links can be trust or friend links and negative links can express relationships of distrust or enmity. Hence, the problem setup is different from latent negative predictions in that positively signed links must only be distinguished from negatively signed links.

Link Completion Problem The task of the link completion problem is to identify which other links a node will attach to, given links that were formed by the node at the same time [Kubica et al., 2003a, Kubica et al., 2003b]. One example is that three people have a meeting, but only the name of two of them is known. Given all previous meeting events, the link completion problem targets to infer the most likeliest third participant. The link completion problem is different from the link prediction problem because the missing link is formed at the exact same time as other links. The link completion data for the problem input is not a state, but an event network; best-performing methods rely on re-occurrence measures [Kubica et al., 2003a].

2.2.3 Solving a Link State Prediction Problem

Having now defined the individual data splits for the two kinds of prediction problems, we now introduce the solution procedure. First, one needs to decide on a prediction function that is then evaluated on the dataset to asses its ability to distinguish node pairs in the true and false test set.

Prediction Functions Remember, that a network is typically represented by its adjacency matrix, which consists of 0s and 1s for a state network. To solve a prediction problem, one uses a function

$$f_m : \{0,1\}^{n\times n} \to \mathbb{R}^{n\times n}$$

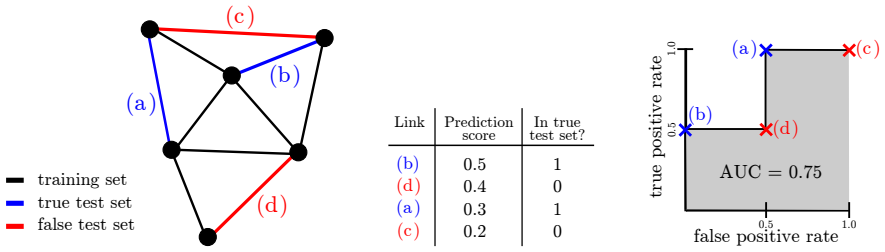
that takes the adjacency matrix of the *training set* and assigns for each node pair $i, j \in V$ a prediction score by computing measure m. For link prediction, this function gives a score for a node pair i, j to quantify its existence likelihood, for unlink prediction it quantifies the likelihood of an unlink and for latent negative it quantifies how likely a link is negatively signed.

In general, f_m is considered as a good prediction function when the scores of node pairs in the true test set are higher than the score of node pairs in the false test set.

For link prediction f_m is a good link prediction function when it gives node pairs in E^+ higher values than node pairs in E^+_{false}. Analogously, f is a good unlink prediction function when it gives edges in E^- higher values than edges that are not removed, in E^-_{false}. Conversely, f is a good latent negative function when it assigns higher function values to node pairs in N than for nodes in P_b or O. In Section 2.4.1, we present an overview over commonly used prediction measures m.

Evaluation of a Prediction Function To measure the accuracy of a prediction function, we use the *area under the curve* (AUC), defined as the area under the *receiver operating characteristic* (ROC) curve [Bradley, 1997]. The AUC-value is a robust measure in the presence of imbalance [Stager et al., 2006]. In the following we describe the ROC curve for link addition prediction; the definition is analogous for link removal prediction.

The evaluation procedure to obtain the AUC-value of a toy network is shown in Figure 2.7.



(i) Data split (ii) Compute ranking (iii) Compute AUC-value

Figure 2.7: The evaluation procedure of prediction function is depicted. (i) All edges in the true and false test set are (ii) ranked in decreasing order by the link prediction function. (iii) The ROC-curve is constructed from the ranking and the AUC-value is then computed as the area under the ROC-curve.

Let f be a prediction function. All node pairs in the combined true and false test set are sorted by descending values of f. Starting from the best-ranked position, for every position in the ranking, the false positive rate is plotted against the true positive rate. The true positive

rate equals the number of observed node pairs from the true test set divided by the overall number of node pairs in the true test set. Analogously, the false positive rate is computed as the number of observed node pairs of the false test set divided by the overall number of node pairs in the false test set. The ROC curve is always contained in the square $[0,1] \times [0,1]$. The AUC-value is then defined as the area under the ROC curve and is thereby a value in the interval $[0,1]$. For a random predictor, the ROC curve approximates the diagonal connecting the points $(0,0)$ and $(1,1)$, giving an AUC-value of 0.5, whereas a perfect predictor yields an AUC-value of 1. When a prediction function f is inverted to give $-f$, its AUC-value x is replaced by $1-x$. This observation allows to turn a below-random predictor into a predictor that performs above random by negating the chosen ranking measure. The AUC-value can then be interpreted as the probability that a randomly chosen node pair from the true test set is ranked higher than a randomly chosen node pair from the false test set.

Alternative measures of accuracy, which we do not use in this thesis, are the mean average precision [Najork et al., 2007] and the normalized discounted cumulative gain [Järvelin and Kekäläinen, 2002]. We choose the AUC-value since it is robust with respect to changes in the size of the split and an established prediction measure in the link prediction community.

Comparison of Classifiers The ROC curves of two prediction functions provide insights into a classifier as to how its prediction operates in which range. Thus, a prediction function A is superior over a prediction function B if all points of A's ROC-curve are above all points of B's ROC-curve [Lichtenwalter et al., 2010]. We will use the AUC-value to compare prediction functions in an aggregated fashion. In most link prediction scenarios it is more desirable to achieve high precision in the left half of the precision-recall curve, because in actual applications only the top K predictions are often desired.

2.3 Linking Behavior in Social Networks

The formation, maintenance and dissolution of personal relationships has been widely studied in many social networks ranging from married couples to criminal networks and high-school students [Parks, 2007]. Even if it is still not clear whether individuals behave similarly or differently in online and offline networks, the last century of sociological studies has developed several highly-interesting theories and discovered influence factors that are worthwhile to consider for online networks, too. According to the Dunbar number we are only capable of maintaining a certain number of relationships [Dunbar, 1992]. Therefore, we have to decide on a regular base which relationships to maintain and which to dissolve. Many of the studies that have targeted the understanding of personal relationships have focused on individual characteristics of actors rather than on describing actors as embedded in larger social networks. These two approaches correspond to *action theories* and *structuralist theories* [Parks, 2007].

Action Theories Action theories emphasize on the individual variability and choice of each actor to explain personal relationships. This theoretical branch assumes that individuals choose whom to interact with according to personal preferences to maximize their personal benefit. The *Social exchange theory* is a prominent representative of this category. Social Exchange theory, originally proposed by Homans in 1958, states that individuals choose to form the relationship they expect to profit from the most, or to have the lowest cost [Homans, 1958,

Garlaschelli and Loffredo, 2004, Emerson, 1976]. According to this theory, individuals will stick to these relationships if they are rewarded and no other relationships provide better opportunities at lower costs.

Structuralist Theories Structuralists explain individual behavior by the larger social structures that a person is embedded in. They see individual behavior not as the product of personal choice but rather as one's position held in a social network. People with the same position or function are assumed to behave similar regardless of personal traits.

Lazarsfeld and Merton introduced the concept of *homophily* which states that individuals are likely to bond with others that are similar to themselves [Lazarsfeld and Merton, 1954]. The similarity between two individuals may be present in the form of similar age, gender, class or organizational role. The effect of homophily has been found to hold in many diverse networks, such as friend networks, neighborhood networks or co-worker networks [McPherson et al., 2001]. At the same time, homophily also influences the dissolution of relationships, McPherson et al. also found that ties between non-similar individuals are also likelier to break. The theory of *assortative mixing* states that nodes of similar degrees, in particular higher-degree nodes, are likelier to get connected with each other than nodes with a highly dissimilar degree [Newman, 2002]. *Balance theory* [Heider, 1958] states that people tend to align their preferences with others. A structural consequence of this theory is that triangles are likely to be balanced, i.e., to contain an even number of negative edges [Harary, 1953]. On the other hand, if we dislike a partner's friends – and hence an unbalanced triad is formed – we may either disconnect with the partner or come up with some coping strategies to reduce the imbalance. Synthesizing this idea, Granovetter asserts the *strength of weak ties* which further develops the concept of *triadic closure* [Granovetter, 1973]. In his famous theory, he posits that if a person is connected by strong ties to two other people, these two people are likely to be connected themselves.

In this work we define changes of actors or entities in the network based on structuralist theories - i.e., we treat all users in the network the same and seek for general network mechanisms that explain the network evolution rather than focusing on individual differences.

2.3.1 Characteristics of Relationships

Network structuring is a combination of two processes: personal strategic decisions and unintended consequences of the behavior of others and oneself. Even if we decide to a certain extent how to form our personal network, some decisions and created structures are unintended. The network structure determines who we are likely to get in touch with because of a high number of shared friends. One might even make friends with someone with a high friend overlap so that activities with friends are not competing too much. Thus, even if individual choice plays an important role in the network structuring process, the network structure sets the frame of who we could meet or befriend. To characterize the structure that a personal relationship is embedded in, Parks gives the following seven influence factors on the relational life cycle of a tie [Parks, 2007].

Network Distance is defined as the distance between two people in a network and reflects the closeness of two people in a given network. A relationship is called *direct* if two people are directly connected with each other and *indirect* if they are only connected through a friend's friend or in general with a distance greater than one.

Network Overlap is defined as the degree of linkage between two peoples' networks. It is measured by ratio of the number of common neighbors and the number of possible common neighbors.

Cross-Network Contact quantifies the frequency of communication with the partner's other relational partners that are not among one's own contacts. It is measured by the number of people of the partner's network that one has had contact with divided by the size of the partner's network.

Cross-Network Density reflects to which extend the relational partners of two persons interact with or know each other. The proposed measure is computed by the proportion of actual to possible cross-links.

Attraction to Partner's Network is defined as the like or dislike for members of the partners network. A corresponding measure would try to capture one's attitude towards the members of the partner's network.

Support from Network Members is the extend to which other network members or one's relational partners support either the relationship with the partner or the partner. Note that these two dimensions are different, since network members could support the relationship with the partner but not the partner or vice versa.

In the following, we summarize the results of empirical studies that have observed characteristic effects for the formation of relationships, the maintenance of relationships and the dissolution of relationships in various domains.

2.3.2 Establishing Relationships

Analyzing one year of email exchanges between 43,553 students, faculty, and staff at a large university Kossinets and Watts found that the formation of new links is driven by the shared activities and affiliations of their members, by similarity of individuals' attributes, and by the closure of short network cycles [Kossinets and Watts, 2006].

In general, Parks found the following four conceptual reasons for establishing new relationships: physical proximity, group norms regarding partner choice, social proximity effects and third party effects [Parks, 2007].

Physical Proximity The physically closer two people are, the likelier they may run into each other. If people work at the same place together and also meet frequently they are likely to talk to each other and thus likelier to establish a personal relationship beyond the professional context. People have also found to befriend with others just because they live close-by and not even because they particularly like them [Parks, 2007].

Group Norms People tend to reach out to partners that are similar to them and the group that they are in. Even the general culture that a person lives in may dictate how to choose a partner. This dictations might be up to things as simple as *Men must be taller than women.*

Situational Generalization It has been recognized that the situation that two people get to know each other influences the chance of the two people to bond with each other. For example, it was shown that adventures bond people together, as well as situations with a successful outcome such as winning a competition [Parks, 2007].

Social Proximity Effects It has been generally recognized in many studies that one often gets to know a new partner via common friends [Parks, 2007, Kossinets and Watts, 2006, Martin and Yeung, 2006]. Social proximity effects relate changes in the network structure to changes in the network distance to other network members that may trigger the formation of new relationships. In an experimental study of four large online social networks it has been observed that between 30–60% of all new links are closing triads in the social network [Leskovec et al., 2008]. Thus social proximity plays a major role for explaining the formation of new links.

Third Party Effects These effects refer to 'little helpers' or 'matchmakers' and rather occur in a dating context. One often experienced example is that one's friends try to set you up with someone. A person may also be hired to find a suitable partner.

2.3.3 Maintaining Relationships

Studies on social networks show that the persistence of ties is influenced by several factors. We divide them into individual, relationship and network factors which are defined as follows. Individual factors are personal characteristics of an actor's personality traits and the current stage in life. Thus, they are independent of the relationship with other actors. Relationship factors characterize the personal relationship of two people and omit other relational partners and network members. Network factors then consider all effects that can be explained by the network structure that two actors are embedded in. Hence, relations to other network members are also used to characterize the relationship in question.

Individual Factors The likelihood that a tie persists increases with the age of the actor – an effect called *liability of newness* [Burt, 2000]. For example, this effect was observed when new employers enter a company or new members enter a sport team. Further, it has been observed that women are better than men in maintaining relationships [Kirke, 2009, Rubin, 1986], hence the gender has an influence on the relationship maintenance. Also, marriage and child-bearing have been shown to decrease the binding to existing friends and to favor locally close people [Martin and Yeung, 2006].

Relationship Factors The closer two individuals are to each other with respect to their individual traits, the likelier the tie is to persist [Martin and Yeung, 2006]. This homophily-effect does thereby not only influence the likelihood of the formation of a tie, it is also likelier that individuals with many shared interests will stay connected. Further the liability of newness has also been observed for relationships; the longer to people are relational partner the likelier the relationship persists [Martin and Yeung, 2006, Burt, 2000]. Martin and Yeung found that proximity is an important factor for the persistence of even strong social ties [Martin and Yeung, 2006]. Strong ties appear to be more persistent because partners in such relationships exchange intimate feelings, are mutually connected and have more frequent contact than weak ties [Wellman et al., 1997]. Once invested the time and effort, these partners are thus likely to

maintain their relationship. In particular kin-ship ties are likely to be long-lasting as phrased in the saying 'blood is thicker than water' [Wellman et al., 1997].

Network Factors Ties that are embedded in a long-lasting group are likelier to retain [Martin and Yeung, 2006]. For instance if people have worked together or played in a team together for a long time, they have also more time to form long-lasting relationships with others from the same community. The structural embeddedness of a tie plays a further very important role. Ties that are well-embedded, i.e. partners sharing many common friends, are likelier to persist over time. Densely knit networks are likelier to be durable, because they bind their members more strongly by social control and collective identity [Wellman et al., 1997]. A prevalence of imbalanced relationships was found in the following study, conducted with high school students [Parks, 2007]. 82% of the high school students reported that at least one close friend had a close friend who they disliked. Exemplary, a subject stated that the disliked person is known for three years and communication with him/her takes place around 3 times a week. Thus, attraction to a partner's network may play a big role in getting connected, but even if some network members are disliked, the tie may still persist. Instead individuals find appropriate coping strategies to handle the negative attraction. Additionally if the level of cross-communication is high, marital relationship have also been observed as more stable [Kearns and Leonard, 2004].

2.3.4 Dissolving Relationships

Most relationships end. We meet a lot of people and bond with a lot of people on an acquaintance-level, but we seem to stay in touch only with few [Parks, 2007]. Wellman et al. conducted two interviews with 33 people who had to name their strong and weak ties twice: in 1968 and ten years later [Wellman et al., 1997]. Only 27% of relationships considered as intimate remained so ten years later. This number is in accordance with the observations of Suitor and Keeton where only between one quarter and one third of all supporting relationships were maintained across a 10-year period [Suitor and Keeton, 1997].

Individual Factors Individuals with an introverted or neurotic personality where shown to be be more prune to a friendship resolution in Facebook [Quercia et al., 2012]. This corresponds with a longitudinal study of marriage stability that observed partnerships with partners that are high in neuroticism and low in extraversion to be likelier to dissolve [Karney and Bradbury, 1995].

Relationship Factors If a relationship is less highly-developed[2], the relational partners are likelier to disconnect [Parks, 2007]. But also highly-developed relationships exhibit high rates of instability; Burt found that 92% of close business partners were not connected any more three years later [Burt, 2000]. Romantic relationships have also been observed to end despite high levels of trust and a high amount of interaction. 65% of 38 studied couples that started dating broke up within the first 4 months [Berg and McQuinn, 1986] and 60% of marriages in the US end with divorce [Preston and McDonald, 1979]. A lack of reciprocity in the marital relationship, indicated by missing support from the partner or imbalance of contribution to the marriage, was also found to be an influential break-up factor [Karney and Bradbury, 1995].

[2]Highly-developed relationships are characterized by high levels of trust and a high amount of interaction.

Network Factors The process of *network structuring* offers an explanations why a relationship is likelier to break if the network between two relational partners is breaking apart. When the number of shared partner decreases, the barrier to the dissolution of the relationship decreases and partners are likelier to end their relationship [Milardo, 1987]. The same study also revealed that the dissolution of one relationship can also cause further dissolutions of ties in the network of common friends. This effect corresponds with the intuition that common friends have to choose sites. Changes in the cross-network density, i.e., the amount of a partner's communication and relationships with the partner's network, have also been observed as a reason for relationship dissolution [Parks, 2007]. That is because a high cross-network density works as a barrier to dissolve a relationship. A reduction of cross-network density then resolves the barrier of the two relational partners to break up and is thus correlated with a declining relationship. Little support from one's own as well as from the partner's network also carries some potential for breaking up. In particular if the disliking person is important to one of the partners and frequently seen, the conflict increases the tension within a relationship and may lead to the dissolution of the relationship [Cleek and Pearson, 1985].

Transformation of Relationships Even if romantic partners break up, only 27% of studied couples stated that they do not have any relationship with the former partner anymore [Parks, 2007]. The remaining 73% either returned to being friends or engaged in an unfriendly relationship or something in between. Therefore, it is worthwhile to consider that a relationship may not dissolve but rather transforms from one state (being married) to another (being friends).

2.4 Prediction Measures and Models

In Section 2.2, we have introduced the general framework for a link state prediction problem. The heart of the prediction problem is the prediction function which quantifies the existence likelihood for links in question. The previous chapter has shed some light as to which characteristics have found to be indicative for the formation of new relationships, the persistence of a relationship and its dissolution in social networks. The presented sociological studies are based on *observations* of personal relationships. On the contrary, the work of this thesis targets the *prediction* of the state of a tie. Some of the previously described characteristics have been described in terms of the underlying network of relationships. In the following, we review structural measures that characterize a tie (i, j) and more complex graph models that express the existence likelihood of a tie.

2.4.1 Prediction Measures

Various measures have been proposed and implemented to quantify the relationship (i, j), i.e., a link from node i to node j. We summarize characteristics that are the important for our work and divide them into the three categories of node-based, link-based and neighborhood-based measures. For a detailed survey on different link prediction measures, we refer to [Lü and Zhou, 2011]. Node-based measures describe the structural characteristics $c(i)$ of a single node i. To then quantify the tie (i, j), the product of the individual node characteristics is computed, i.e., $c((i, j)) = c(i) \cdot c(j)$. Link-based measures characterize a tie, whereas neighborhood-based characteristics describe the neighborhood of a tie, e.g., the shared neighborhood or paths of length three between two nodes. In the following table, we list some common measures from

the three feature classes along with the definition for a node i or a node pair (i, j) in an *undirected* network. For *directed* networks, the set of possible measures is bigger, because the direction of an edge can be considered. Note that the notation $\{j \mid i \sim j\}$ corresponds to all nodes j that are adjacent to i, i.e. all nodes that are connected with i in the network.

Feature class	**Feature**	**Definition**
Node-based	Degree	$d(i) = \lvert\{k \mid k \sim i\}\rvert$
Link-based	Reciprocity	$r(i,j) = 1 \leftrightarrow (j,i) \in E$
Neighborhood-based	Joint degree	$jd(i,j) = \lvert\{k \mid k \sim i \vee k \sim j\}\rvert$
	Common neighbors	$CN(i,j) = \lvert\{k \mid k \sim i \wedge k \sim j\}\rvert$
	Jaccard	$Jacc(i,j) = \frac{CN(i,j)}{jd(i,j)}$
	Cosine distance	$cos(i,j) = \frac{CN(i,j)}{\sqrt{d(i) \cdot d(j)}}$
	Adamic-Adar	$Adad(i,j) = \sum_{\{k \mid k \sim i \wedge k \sim j\}} \frac{1}{\log d(k)}$
	Paths of length 3	$P3(i,j) = \{k \mid i \sim k \wedge k \sim l \wedge l \sim j\}$

Table 2.1: The table gives an overview over some common prediction measures that are also used throughout the thesis. The node based-measures are defined for node i and the link- and neighborhood-based measures are defined for the node pair (i, j)

Ensemble Prediction Functions When quantifying the likelihood of the state different ties, one can use one of the *single* link measures listed before. Since the predictive expressiveness of a single feature is rather limited, combinations of multiple features come into play. We require an *ensemble* link prediction function that a) produces a numeric value that we can use to rank ties and compute the AUC-value, and that b) is easy to use with only few parameter to tune and c) does not make any assumptions on the distribution of the data, such as a normal-distribution of errors. Based on our requirements, we choose *logistic regression* as used for link prediction by others (e.g. [O'Madadhain et al., 2005, Potgieter et al., 2007, Raeder et al., 2011]) to obtain an ensemble prediction function and describe it in the following.

Logistic Regression Logistic Regression is a classification method that returns the probability that a binary dependent variable may be predicted from the independent input variables [Llullaku et al., 2009]. For the prediction set up considered in this thesis, the input variables are given by a set of *independent* prediction measures and the output variable is the likelihood that a particular state is true. In order to learn the regression weights, the training set is split into a source and target set as described in Section 2.2.

If $f_1, f_2, \ldots, f_k$ are the individual prediction functions,e.g., measures from Table 2.1, then the ensemble prediction function is given by

$$f_* = L(b + a_1 f_1 + a_2 f_2 + \ldots + a_k f_k),$$

where b and a_i are the parameters of the ensemble method, which are learned by logistic regression, and $L(x) = \frac{1}{(1+e^{-x})}$ is the logistic function.

The least squares optimization function is used as a statistical method for estimating the coefficients of the logistic regression model. Because logistic regression produces unstable results when the input variables are highly-correlated, one needs to remove the correlated variables before learning the parameters. Only if the input variables are completely independent, the learned parameter weights can be interpreted: If the weight of a parameter is positive then the effect on the outcome variable is positive, i.e., the input variable is positively correlated with the outcome variable. For even slightly correlated variables, one cannot interpret the weights.

2.4.2 Graph Models for Prediction

In the following, we summarize two approaches for graph models that can be applied to solve link state prediction problems.

Global Organizing Principle These methods assume some organizational principle, such as a generative model, for which the specific parameters are learned to maximize the likelihood of the current network. In other words, the parameters that are most likely for the current network are calculated. Having fitted the model on the data, one can then asses which of the potential new links will produce the most likeliest network in the next step. Maximum likelihood methods are computationally very expensive; they cannot handle large networks [Lü and Zhou, 2011]. A prominent example are stochastic block models [White et al., 1976, Faust and Wasserman, 1992] which partition a network into different blocks and characterize the likelihood that one actor interacts based on the likelihood that the respective two blocks that the users belong to interact with each other. These blocks can then be formed by formally defined roles, structural characteristics or detected communities in the network. These kinds of models give interesting insights in the underlying organizing principles but are not feasible for large networks [Lü and Zhou, 2011, O'Madadhain et al., 2005].

Probabilistic Relational Models Given a network, a probabilistic model with defined features will be trained to best explain the current network. Then, the probability of links in the test set can be estimated by the conditional probability of the links given the learned model. One well-studied example for probabilistic models are *Exponential Random Graphs* – also called p*models [Anderson et al., 1999, Snijders and Steglich, 2013]. A weight for each feature in the p* model is learned; when the weight of a feature is bigger than one, then the new graph with an increased value of this feature (e.g. more common neighbors) is more likely. If the weight is lower than one, the new graph is more likely if the value of the feature decreased. In general, probabilistic models are highly complex because the estimated probabilistic model greatly depends on the choice of the prior, the model of dependencies and the chosen inference model to learn the parameters. The parameter estimation is particularly difficult and inefficient for many dependent variables [Lü and Zhou, 2011].

2.5 Applications

Link state prediction problems have been widely used in many applications. This section gives an overview over applications for link state prediction tasks.

Recommender Systems In particular in online social networks such as Facebook, new friends can be hard to discover. Recommender systems help the user to overcome the problem of finding friends in the large set of users. Recommender systems are also developed to recommend items, e.g. movies or products, to persons. Amazons *Customers who bought X also bought Y* feature helps users discover new and personally relevant products. Netflix's recommender system suggests unseen movies to users based on which kind of movies the user has seen so far and which other movies similar users have watched [Koren, 2010, Koren, 2008]. Recommender systems are therefore applying link prediction methods. Methods that solve the latent negative prediction problem can be applied to detect hidden negative relationships that should then not be recommended to users.

User Navigation Users navigate in hyperlink networks such as Wikipedia. Given typical navigation paths, link prediction methods can then be used to propose new connections that facilitate user navigation [Perkowitz and Etzioni, 1997]. For instance, if users often navigate from A via B to C, a shortcut link from A to C could be the result of a link prediction method.

Storage of Big Graphs Popular social media platforms are too big to store their data on one partition only. Efficient algorithms to partition data efficiently have drawn a lot of attention in the research community. To perform well, algorithms need to consider the recency of interactions. For instance users that have interacted but that are not likely to interact again can be stored at different partitions [Carrasco et al., 2011]. Methods of the repeated link prediction can be used to detect these re-occurring user interactions.

Distributed Processing With the rapid growth of online social networks, a scalable architecture is required that can handle database queries and analysis of the data. Local queries, i.e., queries that operate within one partition are desired, hence one needs to store data items that are connected in the queries preferably together. For tasks such as news stream generation and friend recommendation, this implies that well-connected users should be placed on the same partition, but – since networks evolve over time – repartitioning might have to be performed and decreases the efficiency of the processing system. Thus, data partitioning heuristics should not only take existing edges into account, but also those that are likely to appear. These links can be detected with link prediction methods.

Biological Networks Biological networks may consist of interacting molecules or proteins, where the existence of an interaction – a link in the network – must be demonstrated by an experiment. Because biological experiments are very costly, researchers are interested in the most probable interactions that can than be proven in experiments [Lü and Zhou, 2011]. Link prediction methods can be used to predict the most likely chemical interaction.

Communication Surveillance The communication patterns of a group of target persons is observed over time to predict whether targets will communicate again and whether new communication relationships are formed [Huang and Lin, 2009]. Hence, methods from repeated link prediction and link prediction can be applied to this problem setting.

Information Cascades To predict how influence or goods will propagate in a network, current methods use the present network structure. Some of the links in the network are spurious or weak, thus information will not propagate as expected. Using unlink and repeated link prediction methods, one can identify which links are likely to be removed or decay to improve the performance of information flow prediction.

Interaction Suggestion If a friendship is detected as at risk, then social networks might suggest befriended users to interact again or even rank status posts of the user pair better, so that interaction is facilitated.

News Stream Ranking Many users on social networking sites such as Facebook or Twitter are overwhelmed by the sheer amount of content produced by friends or followees. Unlink prediction can help to identify potential non-friends and rank their content lower. In the same way, methods of repeated link prediction could be used to detect that users will not interact with each other's content again and then also rank the produced content lower.

Advertising Jobs Business services such as LinkedIn[3] or Xing[4] are booming. Employees are active in this websites to establish and maintain contact with other firms and business partners. Firms primarily want to recruit new employees on these platforms. Instead of sending job offers to anyone that matches the job description, firms can use unlink prediction methods to detect employers that may potentially leave their current firm.

[3] `https://www.linkedin.com`
[4] `https://www.xing.com/`

3 Predicting Link Additions and Removals in Knowledge Networks

3.1 Introduction

Since the appearance of the World Wide Web, creation of human knowledge has been increasingly collaborative and dynamic. On web sites such as Wikipedia, knowledge is aggregated and interlinked in a massively collaborative and parallel fashion: the number of participants in the creation of collaborative knowledge is virtually unlimited, and changes are made continuously and in parallel. As an example, the English Wikipedia[1] holds more than four million interlinked articles, and currently sustains more than 30,000 active users[2]. The knowledge collected in such knowledge bases is often represented as text, but also increasingly in the form of a knowledge network consisting of connections between concepts. In the case of Wikipedia, these connections are given in the form of links from one article to another, so-called wikilinks. In other cases, a knowledge network may be formed by other types of connections, for instance interactions between drugs and diseases in the Diseases Database[3]. In either case, a remarkable property of these networks is their connectivity: All concepts are related to all other concepts through one or more connections. Thus, the understanding of the underlying knowledge networks is of primary importance to understand the knowledge bases themselves.

While the addition of individual pieces of knowledge to knowledge networks has been studied, collaborative knowledge networks also allow the removal of edges. In fact, the collaborative nature of online knowledge bases results in differences of opinions, and therefore in a high number of removals and reverts of content. On Wikipedia for instance, between 20 and 30 percent of all edits remove one or more wikilinks[4]. Despite these numbers, the disappearance of relationships in knowledge networks is only rarely studied. To fill this gap, we proposes to investigate the structural signals leading to the appearance and disappearance of knowledge links between concepts. Our study is performed on the largest collaborative knowledge network in existence, the online encyclopedia Wikipedia, and consists in identifying structural features of a knowledge network that can be used to predict the appearance and disappearance of edges, and investigating in what way these features can be used as signals to understand the evolution of these networks.

Analysis of link structures is traditionally an important component of Web information systems, such as search engines, recommender systems, spam filters, content summarization tools, and many others. These applications are supported by a wide range of state of the art methods for link-based authority ranking, prediction of further network evolution, and detection of structural anomalies. Well-known properties of networks such as the Web are (1) highly

[1] `http://en.wikipedia.org/`
[2] `http://en.wikipedia.org/wiki/Wikipedia:Wikipedians`
[3] `http://www.diseasesdatabase.com/`
[4] See Table 3.6

imbalanced distributions of node degrees (in a broader sense of several existing models, node "authoritativeness"), and (2) high clustering coefficient, indicative for existence of multiple or tightly connected sub-components ("cliques") [Adamic, 1999]. Among many possible use cases, this knowledge can be used for suggesting new network edges that appear "reasonable" in an existing network structure, e. g., by connecting two nodes that have many neighbors in common. The prediction of such "missing links" (e. g., references between Web pages, friendships in social networks, followers and citations on Twitter, cross-references between articles in Wikipedia, etc.) can be seen as an established recommendation scenario that has been intensively discussed over the last decade.

Since the invention of written language, humans have aggregated knowledge in written form. In recent times, knowledge has been accumulated in encyclopedias, dictionaries, thesauri and other reference works. What these types of works have in common is their structure: They consist of individual items of knowledge such as concepts or words, connected by cross references. These links are not just additional information, but an integral part of the knowledge. Imagine an encyclopedic article about the city of Paris. This article will invariably mention that the city is located in France. Thus, a link is formed between the article *Paris* and the article *France*. In online encyclopedias such as Wikipedia, these links are represented explicitly: The article about Paris contains a hyperlink to the article about France. Thus, the hyperlinks in an online encyclopedia are a representation of the knowledge contained in that encyclopedia, and thus an analysis of the hyperlink structure can reveal much about the knowledge itself.

An online encyclopedia such as Wikipedia also differs in another important way from traditional encyclopedias: It is collaborative, i.e., written by many people simultaneously, and thus it changes much faster and much more often than a traditional encyclopedia. What is more, different authors often have different opinions about the topic at hand, and their edits will clash, resulting in one editor reverting the edits of another editor. This leads to a high amount of dynamism in the hyperlink structure, where links are added, but also removed, very frequently. In order to analyze the dynamics of these changes, we will thus resort to theories of network analysis.

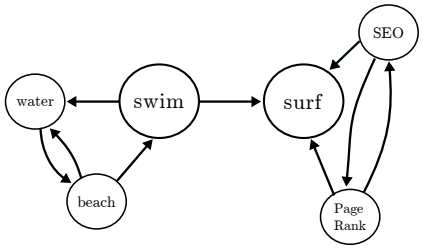


Figure 3.1: Sample network N of interlinked Wikipedia articles. The connection between articles 'swim' and 'surf' is intuitively wrong.

As a running example, we may consider the fictional network N of a sample of Wikipedia articles from Figure 3.1 which consists of the nodes V

$$V = \{water, swim, beach, surf, SEO, PageRank\}.$$

A link (i, j) indicates that article i links to article j.

The network N contains two tightly connected components

$$T_1 = \{swim, water, beach\},$$
$$T_2 = \{surf, SEO, PageRank\}.$$

The link (*swim*, *surf*) does not directly belong to structures of T_1 and T_2 and thus does not connect closely related resources, this can be recognized by the fact that (*swim*, *surf*) does not substantially contribute to the high clustering coefficient of G. Consequently, we may assume that the link (*swim*, *surf*) may demand critical reconsideration as a potential mistake and will be possibly removed in the future.

Conceptually, we study the hypothesis that knowledge of the structure of social networks and models allows for defining invariant indicators for "superfluous" links. More precisely, we consider different ways to solve the unlink prediction problem as a special case of link prediction, by introducing novel graph models and edge weighting metrics, customized for prediction of low-likelihood edges.

In our sample network introduced before, the wrong link has been set due to missing disambiguation of two meanings for 'surf'. In general, the decision to withdraw a link may have many different reasons and cannot be fully explained without domain-dependent knowledge about the particular network and without content resp. context analysis of affected nodes (users, web pages, postings). Our contribution aims to answer the fundamental question: to what extent can structural analysis contribute to the prediction of unlinks? The resulting domain-independent approach can be easily combined with case-specific content analysis and adopted for a variety of applications, such as advanced authority ranking, detection of link spam and manipulations, recommendations for re-organization of social graphs by users and content providers, and many others.

In the following, we investigate the problem of predicting the removal of links in networks in a general and formal manner.

Structural Link Predictions Depending on the type of a network, removal of links may be caused by different issues. In general, the reasons for a link being removed may be content-based reasons, e. g., a hyperlink from a Wikipedia page is removed as the articles' topics are not related, structural reasons, e. g., removing a network link in a telecommunications network, or a combination of both. For our treatment we consider only structural properties of the underlying network and we do this for two reasons. First, our objective is to find general domain-independent models, whereas content is clearly domain-dependent. Second, we hypothesize that several content-based reasons are also reflected in the network structure. Coming back to our introductory example from Figure 3.1, two different main topics can be found and are manifested in the two highly-connected components and only one link between. Although this hypothesis is, of course, not generally applicable we focus on structural properties in order to investigate how good we can predict decay of links without considering content.

Research Questions

This chapter will be concerned with the following research questions.

RQ 1 *Which structural characteristics are indicative for the removal of links?*

In 2003, Liben-Nowell and Kleinberg were the first to define and tackle the link prediction problem [Liben-Nowell and Kleinberg, 2003]. Since then, many measures and models were developed to improve the prediction of new links. On the contrary, unlink prediction is a relatively new problem. Before we started our studies, this problem has not been tackled in a systematic and purely structural matter. Given only the structure of a network, the goal is to find structural measures to predict links that will be removed.

In particular, we ask whether indicators for links can be used to characterize unlinks as well. In the past, link prediction has already been extensively studied, whereas the prediction of unlinks has only been researched in a handful of studies [Quercia et al., 2012, Kwak et al., 2012, Kivran-Swaine et al., 2012, Kwak et al., 2011]. Do we need to consider both problems or is one problem enough to draw conclusions about the other?

RQ 1.I *How are unlinks related to new links, i.e., can characteristics of new links be used to characterize unlinks?*

If one problem can be reduced to the other one, then classic link prediction measures can be used to predict unlinks as well. We hypothesize that the two problems are highly related: factors that drive the formation of new links should hinder the removal of links and vice versa.

RQ 1.II *What is the interplay of link and unlink dynamics?*

This question sets out to answer how numerical indicators of a link can be interpreted for link and unlink prediction. Both problems have so far only been considered separately, so this line of research will aim to provide a unified view of both problems.

The structure of this chapter is as follows. Section 3.2 will review existing research on predicting link removals. Section 3.3 discusses research question RQ 1.I and evaluates two possible transformations between link and unlink prediction. Since unlink prediction appears to be more than a simple transformation of link prediction, Section 3.4 provides a unified view on both prediction problems. For that, numerical indicators for both problems are combined in a second experiment to answer research question RQ 1.II. Section 3.5 then concludes on both empirical evaluations and summarizes the most indicative features for the removal of links.

3.2 Related Work

Predicting Link Additions The problem of predicting the appearance of links in networks has received substantially more attention than the problems of predicting their removal. Surveys on the link prediction problem are provided by [Liben-Nowell and Kleinberg, 2003] and [Lü and Zhou, 2011]. For many networks, the number of common neighbors, the degree of an actor and the ratio of the number of common neighbors and the actor-neighborhood sizes are good indicators for the formation of new links [Lü and Zhou, 2011]. Other algorithms for links prediction include the index of Katz [Katz, 1953], graph kernels [Ito et al., 2005] and diffusion models [Kondor and Lafferty, 2002].

Predicting Link Removals Work on the removal of links in networks has mainly focused on social networks and on explaining why users on particular social networking platforms such as Facebook and Twitter unfriend or unfollow each other [Kwak et al., 2012, Quercia et al.,

2012, Kivran-Swaine et al., 2012]. As these studies use very specific user information, such as personality traits, gender, or Twitter-specific interaction data, they cannot be used to classify the formation of new links and link removal in networks other than social networks. See Section 5.2 for a detailed survey on unlink prediction in online social networks.

One recent work on structural characteristics of unlinks performed a data analysis for friendship relationships between 32 freshmen over one year at seven different points in time [Snijders and Steglich, 2013]. This work trains a p* model to characterize unlinks and links in the dataset. Though that approach is very powerful, it does not scale well. The corresponding software package[5] is applicable to networks with 10 to 1,000 nodes and thus cannot even be applied to the dataset used in our study.

Strong versus Weak Ties Strong ties in communication networks are associated with a high amount of communication between two partners [Kossinets and Watts, 2006, Onnela et al., 2007]. Onnela et al. observed that the removal of weak ties in a mobile phone network, associated with a small amount of communication between two people, lets the network fall apart, i.e. the largest connect component is fragmented into smaller components [Onnela et al., 2007]. In contrast, the removal of strong ties has a minor influence on the connectedness of the network. The influence of the removal of ties is measured by the relative size of the largest connected component. This is in accordance with "the strength of the weak ties" which conjectures that structural information of strong ties is somehow redundant in a network, because strong ties appear in highly-clustered regions [Granovetter, 1973]. Hence, after the removal of strong ties, remaining ties still keep the network connected. The analysis of information spreading within the described mobile phone network yields that neither strong ties nor weak ties are important for the conduction of information. The authors explain this by weak ties offering too few opportunities for communication partners to exchange news and strong ties to be embedded in highly-clustered communities with little access to new information.

Related Problems

In the following, we discuss works on related problem types that are similar, but not identical to the prediction problem discussed in this chapter.

Link Decay In many networks, links cannot be removed but are rather considered to become *inactive* or to *decay*. Two studies that predict decay in mobile phone communication networks assume that links decay if no communication was exchanged between the actors for a particularly chosen time period [Raeder et al., 2011, Hidalgo and Rodriguez-Sickert, 2008]. Both works conclude that links are more likely to persist when the connection is reciprocated and when either both actors' degrees are low or both degrees are high. Raeder et al. observed that the "liability of newness" holds, i.e., the age of a tie is correlated with its persistence [Raeder et al., 2011]). Viswanath et al. found that the longer two users have engaged in wall-to-wall interactions on Facebook, the more likely they are to continue and thereby the less likely the interaction link in Facebook is to decay [Viswanath et al., 2009]. We cannot use features such as the interaction frequency, since links in a state network mostly change from present to removed and thus the history of a tie is not useful. Further, this line of research deals with derived link removals as the datasets themselves do not contain explicit unlinks.

[5] `http://www.stats.ox.ac.uk/~snijders/siena/`

Declining Participation Related works focus on two perspectives of declining participation: the user-perspective and the community perspective. The decay of groups in social networks has been studied by Kairam et al. through interaction patterns and the social structure of users [Kairam et al., 2012]. They observe that groups with a high rate of interaction of group members that are internally well-connected, are less likely to die. In [Garcia et al., 2013], cascading effects that lead to the decline of a community were studied. Given that users leave the community if they have less than k friends, they analyze how this contributes to the community decline. A user-related phenomenon is called *churn*, describing the situation in which a user quits a social community or quites using a service [Karnstedt et al., 2010].

Anomaly Detection A related problem is the identification of spurious links, i.e., links that have been erroneously observed [Guimerà and Sales-Pardo, 2009, Zeng and Cimini, 2012]. A related area of research is the detection of link spam on the Web, in which *bad* links are to be detected [Benczúr et al., 2005]. The problem of anomaly detection is structurally similar to the problem studied here, but do use content features as opposed to structural features of the dataset.

Infer Missing Links in Wikipedia Various works aim to complete Wikipedia's hyperlink structure. These works use textual analysis to predict which phrases in an article should be linked, and to which articles they should point. Some methods only use linguistic features to detect potential link targets and predict the links to target articles with the highest semantic relatedness to the source article [Milne and Witten, 2008] or by classifiers trained on textual features [Mihalcea and Csomai, 2007]. A more structural approach uses pre-processed n-grams and ranks them by structural characteristics of the network, inferred by similar links to other articles [Adafre and de Rijke, 2005] or a principal component analysis of the structure [West et al., 2009]. This line of work relies on textual features or textual preprocessing of an article. In particular, the presented approaches target the problem of predicting new links between Wikipedia articles and do not cover the prediction of unlinks.

Ontology Alignment An ontology is a knowledge representation of facts in a database. Ontology alignment is then the procedure to relate or map the concepts of two ontologies with each other. It becomes necessary, when information from different ontologies should be compared, merged or queried. For instance, two concepts from two different ontologies can be the same and just labeled differently or one instance can be a subclass of another instance from a different ontology. Ontology matching algorithms usually have three different approaches to compare entities: lexical analysis, structural analysis and semantic analysis. The lexical analysis compares the string values of the entity labels. Similarly to the structural approach of this thesis, the structural analysis focuses on the surrounding structure of two entities, such as their subclasses and super-classes, siblings and mapped entities. The information about structurally related entities can then be used to asses the similarity of two entities. Even if two entities might not be named similarly, the surrounding entities might be and therefore they can be mapped to each other. The semantic analysis of two entities requires the use of reasoning mechanisms and deduction of new assertions. For a state of the art survey on ontology matching, we refer the reader to [Shvaiko and Euzenat, 2013]. The structural analysis of ontologies uses ontology-specific class-relationships and is therefore different from the relationships between knowledge items that we consider.

Citation Analysis Another type of knowledge network is a citation network, which consists of scientific publications which are connected by citations. While this type of network fits our definition of a knowledge network, it grows in a very specific and simple way: The only possible change is the addition a new publication. This corresponds to a new node, added simultaneously with all its outgoing edges. The addition or removal of an edge between two existing nodes is not possible in such a network, and as such traditional link prediction methods are not applicable. Instead, research on these types of networks has focused on modeling measures of popularity and similarity.

3.3 Transformations from Link to Unlink Prediction

Intuitively, link prediction seeks to predict links that appear with a high likelihood, whereas unlink prediction targets the prediction of links with a low likelihood in the current network. Hence, both prediction problems seem highly related and we may therefore ask whether the unlink prediction problem can be understood as a simple transformation of the link prediction problem. If so, then all theories, models and methods developed to solve the link prediction problem can also be used to predict unlinks. In this section, we will evaluate two plausible transformation.

Problem Formalization

In order to predict the removal and addition of links, we consider a scenario of evolving networks. Let

$$N_t = (V, E_t)$$

for $t \in N$ be the network N_t at time t with V being the set of nodes of N_t with $n = |V|$ and $E_t \subseteq V \times V$ the set of links of N_t.

Typically, a network N_t is represented by its adjacency matrix $A(N_t)$, i. e., V is defined via $V = \{1, \ldots, n\}$ and $A(N_t) \in \{0,1\}^{n \times n}$ is defined as

$$A(N_t)_{ij} = \begin{cases} 1, & \text{if } (i,j) \in E_t, \\ 0, & \text{otherwise.} \end{cases}$$

If the actual network and evolution step is of no importance we usually write A instead of $A(N_t)$.

Prediction Functions A link prediction function f_m is a function

$$f_m : \{0,1\}^{n \times n} \to \mathbb{R}^{n \times n} \tag{3.1}$$

that takes the adjacency matrix of the network $A(N_{t_1})$ at time t_1 and assigns for each node pair $i, j \in V$ a link creation score by computing measure m. The bigger a link prediction score of an edge $(i,j) \notin E_{t_1}$ is, the more it is expected to actually be added to the network. Thus, good link prediction functions assign larger scores to links (i,j) that will appear until time t_2, i.e. $(i,j) \in E_{t_2} \setminus E_{t_1}$, than to others.

For the problem of predicting link removal, our aim is to define a unlink score function g_m of the form

$$g_m : \{0,1\}^{n \times n} \to \mathbb{R}^{n \times n} \tag{3.2}$$

that takes a matrix $A(N_{t_1})$ and computes for each node pair a decay score by measure m. More specifically, for edges $(i,j) \in E_{t_1} \setminus E_{t_2}$ we expect $g_m(N(A_{t_1}))_{ij}$ to be significantly larger than unlink scores of other edges.

3.3.1 Prediction Models

The problem of predicting whether a link will be removed can be viewed as the inverse problem of predicting the creation of links. The objective of our approach is to validate how far unlinks can be predicted with the same structural methods as new links. In the following, we propose two different approaches for answering this question. These approaches transform the link prediction problem into an unlink prediction problem by complementing the score (cf. Model 1: *Complement Score*) and the network (cf. Model 2: *Complement Network*), respectively.

Model 1: Complement Score

Using a link prediction function f_m from (3.1) that computes a score by measure m we define its *inverse link prediction function* g^1_m via

$$g^1_m(A) = -f_m(A) \qquad .$$

The rationale behind this complement model is that links that have a high link prediction score should not be removed, whereas links with a low score are expected to be deleted. In the literature a series of different approaches have been proposed for solving the link prediction problem [Lü and Zhou, 2011]. Here we consider the following measures as the basis for unlink prediction.

Preferential attachment Let $d(i)$ denote the degree of node i and let $d(j)$ denote the degree of a node j in A. Preferential attachment estimates that an edge (i,j) is added with a likelihood proportional to the product of the degree of i and the degree of j, i.e., we have $f_{PA}(A)_{ij} = d(i) \cdot d(j)$. Hence, the *complement score* score of (i,j) is

$$g^1_{PA}(A)_{ij} = -d(i) \cdot d(j). \tag{3.3}$$

Thus according to this method, links are likelier to be removed between two nodes of a low degree.

Common neighbors This link predication method implements the intuition that two nodes are to be linked if they share a lot of neighbors. The function f_{CN} is defined via $f_{CN}(A)_{ij} = (A^2)_{ij}$, where $(A^2)_{ij}$ is the number of paths of length 2 between i and j, i. e., the common neighbors. g^1_{CN} is therefore defined as

$$g^1_{CN}(A)_{ij} = -(A^2)_{ij} \tag{3.4}$$

Links in this model are expected to be removed if they have only few common neighbors.

Cosine similarity With the cosine similarity method, an edge (i, j) is estimated to be created with likelihood proportional to the angle between the degree vectors of node i and j. f_{cos} and g^1_{cos} are defined as

$$g^1_{cos}(A)_{ij} = -f_{cos}(A)_{ij} = -\frac{(A^2)_{ij}}{\sqrt{d(i)} \cdot \sqrt{d(j)}}. \tag{3.5}$$

If the two nodes are connected to the same nodes, the link between them is expected to stay.

Jaccard index Let $N(k)$ be the set of *neighbors* of node $k \in V$, i. e.,

$$N(k) = \{l \in V \mid A_{kl} = 1\}$$

With the Jaccard index, an edge is created with likelihood proportional to the number of common neighbors divided by the number of different neighbors of both nodes. The function f_{Jacc} and the corresponding function g^1_{Jacc} are defined via

$$g^1_{Jacc}(A)_{ij} = f_{Jacc}(A)_{ij} = -\frac{(A^2)_{ij}}{|N(i) \cup N(j)|}. \tag{3.6}$$

If two nodes are not connected to many nodes but share only few common nodes, the link between them is expected to be removed.

Adamic–Adar The measure used by the approach of Adamic and Adar [Adamic and Adar, 2001] counts the number of neighbors of nodes i and j, weighted by the inverse logarithm of each neighbor k's degree $d(k)$:

$$g^1_{Adad}(A)_{ij} = f_{Adad}(A)_{ij} = -\sum_{k \in N(i) \cap N(j)} \frac{1}{\log d(k)}. \tag{3.7}$$

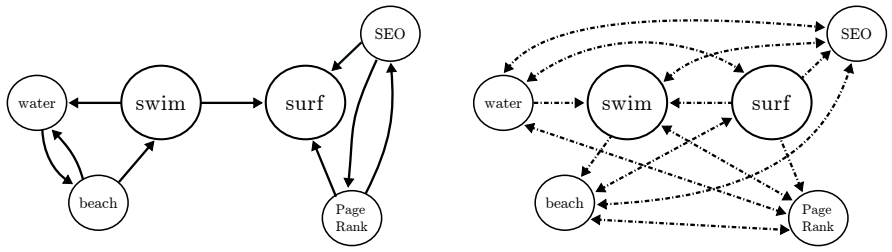
Thus, if two nodes share only few common neighbors with a high degree, the link between them is not expected to stay in the network.

Model 2: Complement Network

The second family of unlink functions we consider employs link prediction functions as well. But rather than inverting the prediction function we now invert the problem itself and consider predicting removal of links in a network by predicting creation of links in its *complement network*. Using a link prediction function f_m we define its complement link prediction function g^2_m via

$$g^2_m(A) = f_m(\bar{A}) \qquad .$$

Given a network $N = (V, E)$ its complement $\bar{N} = (V, \bar{E})$ is defined via $\bar{E} = \{(i, j) \mid i \neq j, (i, j) \notin E\}$, i.e., $\bar{N}$ contains only links between different nodes that are not connected in N. The complement network of the network in Figure 3.1 is shown in Figure 3.2. The rationale



(a) Original Network (b) Complement Network

Figure 3.2: The Complement network of network (a) is illustrated in Figure (b). It consists of all edges that are not present in the original network.

behind this complement model is that since it contains all non-edges, edges that are predicted in it, should not be present in the original network. Thus, we can conclude the likelihood with which they can be removed. The complement network is by far not sparse, thus we cannot represent the complement network as a matrix. Since link prediction methods compute a score of a network's adjacency matrix, we will use the following alternative that does not need the adjacency matrix of the complement network to be constructed. If $A = A(N)$ is the adjacency matrix of N then $\bar{A} = A(\bar{N})$ can be written as

$$\bar{A} = \mathbf{1} - I - A \tag{3.8}$$

where $\mathbf{1}$ is the 1-matrix (containing only 1s) and I is the identity matrix (containing 1s in the diagonal).

We expect that predicting creation of links in $\bar{A}$ also solves the problem of predicting removal of links in A. Considering Figure 3.2 again, we can see that predicting a link between nodes 'swim' and 'surf' is very likely, e. g., using the number of common neighbors. From the prediction of this edge in the complement network, its removal in the original network would be predicted.

In the following, we use Equation (3.8) to derive $g_m^2(A)_{ij}$ using the same link prediction measures m as in the previous section.

Preferential attachment An edge (i, j) is removed with a likelihood proportional to product of the degree of node i and degree of node j in the complement network $\bar{N}$.

$$\begin{aligned} g_{PA}^2(A)_{ij} &= f_{PA}(\bar{A})_{ij} \\ &= (n - 1 - d(i)) \cdot (n - 1 - d(j)) \end{aligned} \tag{3.9}$$

A link is therefore likely to be unlinked between low-degree nodes.

Common neighbors The unlink score of an edge (i, j) in the original network is then translated to the link prediction score in its complement network by

$$\begin{aligned} g^2_{CN}(A)_{ij} &= f_{CN}(\bar{A})_{ij} \\ &= n - d(i) - d(j) + (A^2)_{ij}. \end{aligned} \tag{3.10}$$

Thus, a link is likely to stay if the degrees of its incident nodes are big and share many neighbors.

Cosine similarity An edge is removed with a likelihood proportional to the angle between the complemented degree vectors

$$\begin{aligned} g^2_{cos}(A)_{ij} &= f_{cos}(\bar{A})_{ij} \\ &= \frac{n - d(i) - d(j) + (A^2)_{ij}}{\sqrt{(n-1-d(i))} \cdot \sqrt{(n-1-d(j))}}. \end{aligned} \tag{3.11}$$

Jaccard index Applied to the complement network, we obtain the following unlink score

$$\begin{aligned} g^2_{Jacc}(A)_{ij} &= f_{Jacc}(\bar{A})_{ij} \\ &= \frac{n - d(i) - d(j) + (A^2)_{ij}}{n - |N(i) \cap N(j)|}. \end{aligned} \tag{3.12}$$

According to this measure, an edge is expected to be removed if the degrees of its incident nodes are small and have more dissimilar neighbors.

Adamic–Adar The weighted variant of the Adamic–Adar score of the complement network is as follows

$$\begin{aligned} g^2_{Adad}(A)_{ij} &= f_{Adad}(\bar{A})_{ij} \\ &= \sum_{k \in V} \frac{1}{\log d(v)} - \sum_{k \in N(i)} \frac{1}{\log d(k)} - \sum_{k \in N(j)} \frac{1}{\log d(k)} \\ &+ \sum_{k \in N(i) \cap N(j)} \frac{1}{\log d(k)}. \end{aligned} \tag{3.13}$$

Under this model, if nodes i and j are adjacent to few and rather high-degree nodes the link (i, j) is likely to be removed.

A summary of the scoring methods is given in Table 3.1.

Predictions in Directed Networks

The link and unlink prediction methods in this section were aligned for undirected networks, so they used characteristics such as degree $d(i)$ and neighborhood $N(i)$ of a node i. We evaluate methods and models for link predictions on *directed* Wikipedia article-hyperlink networks. Instead of only one node degree for undirected networks, three different degrees of a node can be defined for directed networks: a node's out- and in-degree and its degree. Consider

Name	Link prediction function	Inverse	Complement
Preferential attachment	$f_{PA}(A)_{ij}=d(i)\cdot d(j)$	$-f_{PA}(A)_{ij}$ [Eq. (3.3)]	$f_{PA}(\bar{A})_{ij}$ [Eq. (3.9)]
Common neighbors	$f_{CN}(A)_{ij}=(A^2)_{ij}$	$-f_{CN}(A)_{ij}$ [Eq. (3.4)]	$f_{CN}(\bar{A})_{ij}$ [Eq. (3.10)]
Cosine similarity	$f_{cos}(A)_{ij}=\frac{(A^2)_{ij}}{\sqrt{d(i)}\cdot\sqrt{d(j)}}$	$-f_{cos}(A)_{ij}$ [Eq. (3.5)]	$f_{cos}(\bar{A})_{ij}$ [Eq. (3.11)]
Jaccard index	$f_{Jacc}(A)_{ij}=\frac{(A^2)_{ij}}{\lvert N(i)\cup N(j)\rvert}$	$-f_{Jacc}(A)_{ij}$ [Eq. (3.6)]	$f_{Jacc}(\bar{A})_{ij}$ [Eq. (3.12)]
Adamic–Adar	$f_{Adad}(A)_{ij}=\sum_{k\in N(i)\cap N(j)}\frac{1}{\log d(k)}$	$-f_{Adad}(A)_{ij}$ [Eq. (3.7)]	$f_{Adad}(\bar{A})_{ij}$ [Eq. (3.13)]

Table 3.1: Overview of all score methods for link and link decay prediction of an edge (i, j)

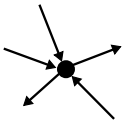


Figure 3.3: An arbitrary node i with incoming and outgoing edges.

the node shown in Figure 3.3. It's out-degree d_{out} is defined as the number of outgoing links from it and its in-degree d_{in} is defined as the number of incoming links. For the given node i, $d_{out}(i) = 2$ and $d_{in}(i) = 3$. The degree d is defined as $d_{out} + d_{in}$, so $d(i) = 5$. Further, the node neighborhood $N(i)$ of a node i can now be defined for outgoing and incoming links accordingly

$$N_{out}(i) = \{j \in V \mid (i, j) \in E\}$$
$$N_{in}(i) = \{j \in V \mid (i, j) \in E\}.$$

A common approach when predicting links in directed network is to use the same methods as for undirected networks but to test different degree combinations [Lichtenwalter et al., 2010]. Thus, all undirected degrees $d(i)$ and $d(j)$ are aligned with all given combinations from Table 3.2.

For better readability, the methods in this section were aligned with the 'sym' degree (column 1 in Table 3.2) version only. Other methods can be defined analogously and have been systematically tested in this work.

3.3.2 Methodology

By utilizing common link prediction methods we have defined two families of approaches to predict decay in networks. In this section we conduct an empirical evaluation on how good our approaches work on real datasets. In particular, we stipulate that, given the evolution of some network, links that are removed in a step of the evolution receive a high link decay score.

Name	sym	asym	in	out
$d_1(i)$	$d(i)$	$d_{out}(i)$	$d_{in}(i)$	$d_{out}(i)$
$d_2(i)$	$d(j)$	$d_{in}(j)$	$d_{in}(j)$	$d_{out}(j)$

Table 3.2: List of the combinations of degrees of node i and node j used.

Wikipedia	#Articles	Adds [$\times 10^6$]	Deletes [$\times 10^6$]
French	1,763,659	41.7	17.3
German	1,526,219	58.7	27.6
Italian	953,208	26.0	8.9
Polish	765,930	18.8	6.2
Dutch	751,888	15.3	4.7

Table 3.3: The datasets used in our evaluation. The number of articles also includes articles that were removed.

Furthermore, given that we approach the problem of predicting removal of links by using link prediction methods we ask the question of how related those two problems are in real datasets and if they can be solved using the same methods. We conduct our analysis using five directed large-scale networks from Wikipedia. As general practice, we evaluate link decay methods for directed networks with different combinations of in-degree and out-degree [Lichtenwalter et al., 2010]. Thus, we will explore which effects the different degree combinations have on the prediction quality and which prediction method provides the best precision.

Datasets

To evaluate our proposed decay models, we use the directed article-hyperlink networks of five of the six largest[6] Wikipedias. We choose Wikipedia because it's the biggest publicly available online encylopdia and its content is actively maintained which makes it a highly-dynamic large scale knowledge network. We skip the largest one, the English Wikipedia, due to its size and limited computational resources. In the directed article-hyperlink network of Wikipedia, a link between two articles i and j is present if article i links to article j. For our link decay prediction scenario we omit user pages and article discussion pages.

For each of the five Wikipedias we considered all creation and deletion events for links since their installment. An overview over the datasets is given in Table 3.3. The French Wikipedia is the biggest dataset used with around 1.8 million articles between which overall 41.7 million links where added and 17.3 million removed. Note that the number of articles includes also articles that where removed later. For these Wikipedias, link deletions make up about 24–31% of all link operations, thus accounting for a large part of structural changes.

Evaluation Methodology

In our evaluation we aim to compare how well we can distinguish edges that have been removed and edges that are not removed. An illustration of the temporal dynamics of an article network N is shown in Figure 3.4.

For that we split the datasets of a Wikipedia article network $N = (V, E)$ at time point $t_1 = 3/4t_2$ of the whole time interval t_2. We define the training set as all edges that are present

[6] `http://meta.wikimedia.org/wiki/List_of_Wikipedias`

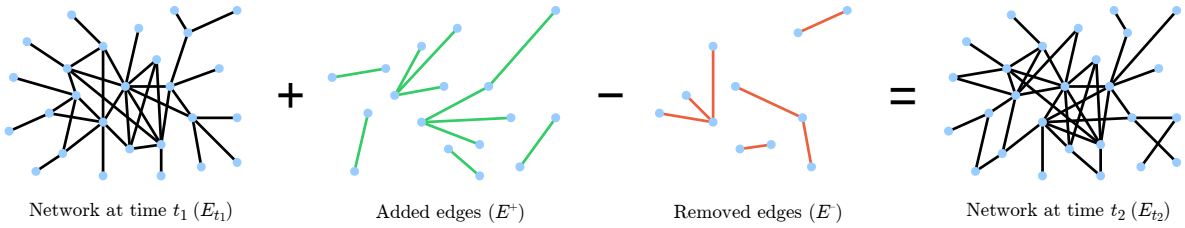


Figure 3.4: Schematic representation of the link addition and removal process. At time t_1, the network has the edge set E_{t_1}. After t_1, the set of edge E^+ is added and the set E^- is removed, giving the set of edges E_{t_2} at time t_2. Link directions are not indicated in the figure.

at time point t_1

$$E_{\text{Training}} = E_{t_1},$$

the true test set E^- as all edges from the training set that are not present anymore at time t_2

$$E^- = \{(i,j) \in E_{t_1} \setminus E_{t_2} \mid i,j \in V\},$$

and the false test set E^-_{false} as random sample of edges from the training set that are still present at time t_2 with size $|E^-_{\text{false}}| = |E^-|$

$$E^-_{\text{false}} = \{(i,j) \in E_{t_1} \setminus E_{t_2} \mid i,j \in V\}.$$

The three edge sets are illustrated in Figure 3.5.

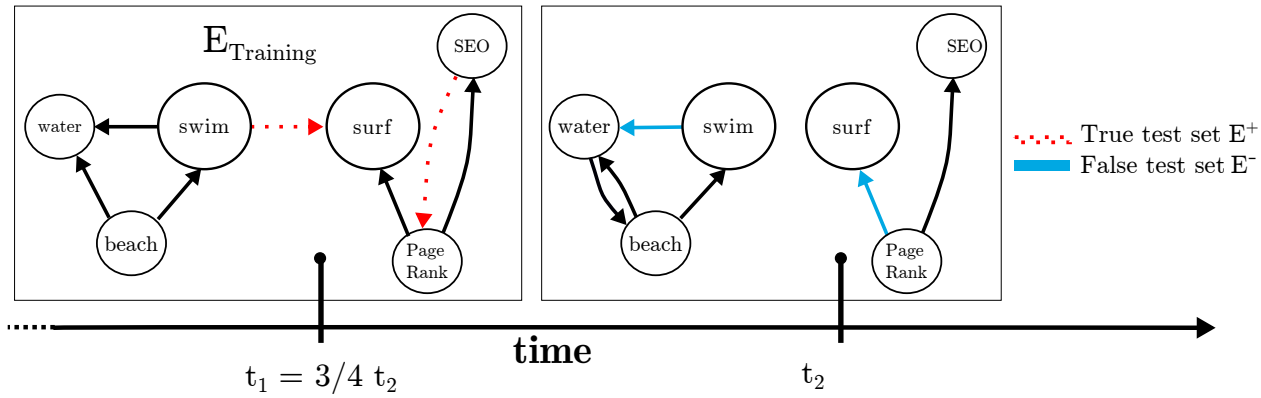


Figure 3.5: Split in training and test set.

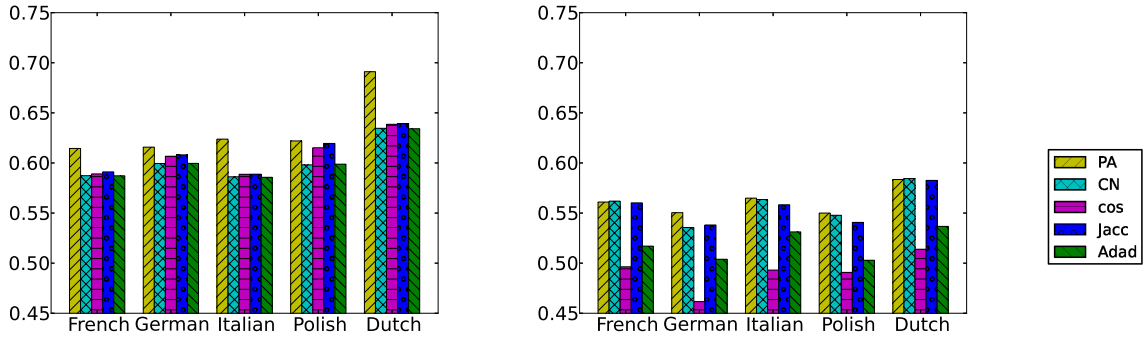
We compute the precision of our models with the AUC-value (cf. Section 2.2.3). By construction, the AUC-value is ranged between 0 and 1, where a perfect prediction would obtain an AUC-value of 1.0 and the random baseline is 0.5.

We compute the AUC-value for all combinations of unlink scores shown in Table 3.1 and the four combinations of degrees from Table 3.2 for the five largest Wikipedias.

3.3.3 Evaluation

In the following, we provide results of our empirical evaluations.

Precision of Unlink Models We have defined two link removal models that transform the link prediction problem to the problem of predicting unlinks. Each of the two unlink models computes scores of five classic link prediction methods: *preferential attachment* (PA), *common neighbors* (CN), *cosine* (cos), *Jaccard* (Jacc), and *Adamic–Adar* (Adad), which in turn are varied by four different out and in-degree combinations. Figure 3.6(a) and Figure 3.6(b) show



(a) Complement score

(b) Complement network

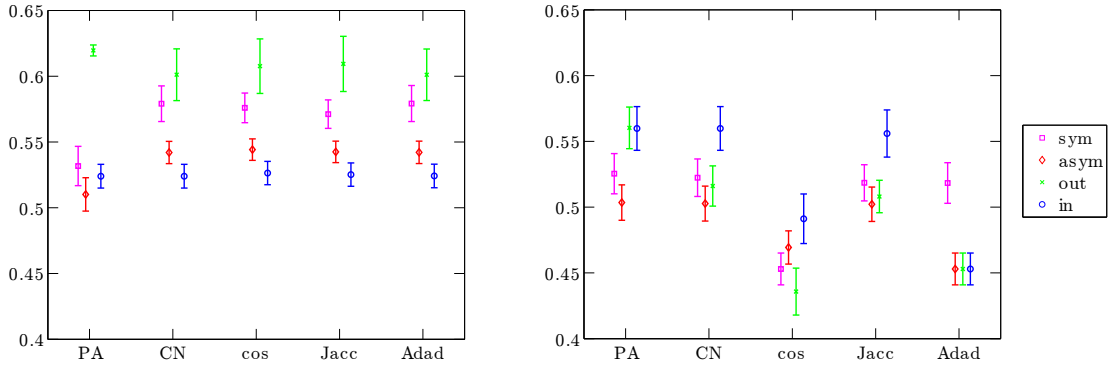
Figure 3.6: AUC-values of (a) *complement score* model and (b) *complement network* model. Only the AUC-values of the best-performing degree combination are depicted for each method.

the best AUC-values over all degree combinations of each method for the *complement score* model and the *complement network* model.

The *complement score* model performs significantly better than random, all methods have an AUC-value above 0.5. Preferential attachment is the top-performing method, superior over the four remaining methods on all five datasets. This means that the likelihood of an edge to be removed is bigger if the two adjacent nodes have a low degree. Up to 69.7% of all edges from the test set where correctly classified as to remove. Jaccard and cosine as well as common neighbor and Adamic–Adar perform very similar to each other with precisions above the random baseline, too.

The AUC-value of the *complement link prediction* model, shown in Figure 3.6(b), has lower precisions than the preceding approach. However, all methods, except cosine, out-perform the random baseline. PA, CN and Jaccard predict link removals with the highest precision, which leads to up to 58.4% of correct predictions for the test set. In comparison, the *complement score* approach does a better job in predicting link removals.

Effect of degree combinations Computing unlink scores for all edges (i, j) of the test set, we have tested four different degree combinations (cf. Table 3.2) of node i's and node j's in respectively out-degree.



(a) Complement score model

(b) Complement network model

Figure 3.7: Error plot of the four different degree combinations for (a) *complement score* model and (b) *complement network* model. The error is computed by the standard deviation of AUC-values of the five Wikipedia datasets.

In Figure 3.7 we compare the unlink prediction precisions of these four degree combinations across all methods for the *complement score* approach and the *complement network* approach. The error bars indicate the standard deviation across the five datasets.

Varying the types of degrees leads to a drastic deviation within each prediction method. For the *inverse link prediction* method, precision values go from slightly above the random baseline – when using in-degrees – up to 0.6 or more when out-degrees are considered. The precision values in Figure 3.7(a) are staggered: out-degrees perform best, followed by degrees, out-degree/in-degree and in-degrees. By construction, the *complement network* method thus performs best when considering node in-degrees. The deviation of precision is not as big as for the inverse method and the ranking of degree combinations is more mixed.

3.3.4 Conclusion

We investigated the problem of predicting the link removal in networks such as the knowledge network Wikipedia. We proposed two approaches that utilize link prediction methods and rely on inverted problem descriptions of the link prediction problem. While our first approach simply complements the prediction scores of a link prediction method our second approach applies link prediction to the complement network. Our evaluation showed that, in general, the first approach outperforms the second. However, despite the fact that our evaluation showed that our approaches both outperform the random baseline we discovered that the problem of predicting removal of links is generally harder than the problem of link prediction. This observation also justifies the need for further research on the problem of link removal.

Results Our evaluations show that structural analysis makes a meaningful contribution for the prediction of link removals. Using link prediction methods we have outperformed a random predictor. In our evaluations the *complement score* approach combined with preferential attachment performs best. Thus, an edge between nodes with a small degree is more likely to be removed. Reasons for this could be that these articles are still evolving, thus their network

structure changes because they are not 'settled' yet, or, that wrong connections were made caused by the lack of understanding of the article content. Using only out-going node characteristics, such as a node's out-degree and out-going neighborhood achieved the best precisions for the *complement score* approach. This could be interpreted as some kind of 'you are who you link to' rule. Two articles are more similar if they link to the same articles. For link removal this means, that two articles linking to very few common pages are likely dissimilar and thus they should not be connected by a link.

To ascertain whether unlink prediction is of the same difficulty as link prediction, we have also computed link prediction precisions for the five Wikipedia datasets. Actually, link predictions with the same methods are more accurate, precisions around the 0.85 mark were achieved. Thus, the problem of predicting link removals seems to be more difficult than link prediction.

Weak ties The best-performing decay prediction method does not use any community characteristics, such as the number of common neighbors or the union of neighbors. In the beginning, we have hypothesized that two nodes should not be connected anymore if they have a low degree or if they have a higher degree and have only very few neighbors in common. The first hypothesis is somehow verified by the good precision value of preferential attachment. On the other hand, few neighbors seem not to be a good indicator for link removal. Thus, the network data must contain a few nodes pairs with little common neighbors that stay connected. These links are weak links following Granovetter [Granovetter, 1973], that introduce shortcuts into the network which lead to the small-world phenomenon. Considering solely the structure, one cannot distinguish between links that should be removed and links that operate as weak ties.

Complement Score Outperforms Complement Network Our results show that the complement score approach outperforms the complement network approach for all methods and networks. The complement network model assumes that every non-edge is an intentionally unlinked node pair and thus translates it to a connected node pair in the complement network. This in praxis uncertain information – node pairs might just not be linked yet or they may be somehow connected even if the link is not established – is translated to certain information of connectedness in the complement network. We believe that this translation step makes some strong assumptions that are generally not fulfilled. Also some of the characteristics are not meaningful in the complement network. The diameter in the complement network is very small, for instance it is reduced from 4 in the original example network to 2 in the complement network in Figure 3.2.

3.4 Interplay of Link Addition and Link Removal

In the preceding section, we have seen that the problems of predicting link removals and predicting link additions are two distinct prediction problems. Some of the structural characteristics from the link prediction research are indeed useful to predict unlinks, but a deeper understanding of link removal processes is still needed. This section provides a unified view of both problems. Instead of considering each problem separately, we consider the interplay of both problems to reveal interesting dynamics in knowledge networks.

Remove \ Add	**Negative**	**Positive**
Positive	decay	instability
Negative	stability	growth

Table 3.4: Classification of indicators by their ability to predict link addition and link removal. "Add" and "Remove" refer to the type of event to be predicted. "Positive" and "Negative" refer to positive and negative predictive power for the type of event.

We compare the predictive ability of individual features at the task of predicting the addition and removal of individual edges, and are able to identify four classes of indicators: those that indicate growth of links, those that indicate decay of links, those that indicate the stability of links and those indicate the instability of links. We then use these insights to classify the individual addition and removal events, according to their role in the knowledge network's growth. In the following we state our model and describe the experiments that we performed.

3.4.1 Modeling Structural Changes in Knowledge Networks

In the business sciences, a knowledge network is defined as a correlational knowledge structure that is inherently symmetric because it connects entities that are related to another [Saviotti, 2009]. We align our definition of a knowledge network with the working definition of the semantic web community, which assigns directions to links between knowledge items. Thus, we define a knowledge network to be a directed graph $N = (V, E)$ consisting of a set of vertices V representing the knowledge items, and a set of edges E representing the links between them. Individual knowledge items will be denoted i, j, etc., and a link from i to j will be denoted (i, j). In general, links in knowledge networks are not symmetric, i.e., an edge (i, j) does not imply that the inverse edge (j, i) is present as well.

Problem Description

Our goal is to determine which indicators are useful to explain the formation of new edges and the removal of existing edges. Since we are not interested in modeling the appearance and disappearance of individual knowledge items, we consider the set of nodes to be invariant over time.

A way to model the growth and the decay of a network is to determine numerical indicators that correlate with observed growth and decay in actual networks. As an example, the number of common friends is used in social networks to predict the appearance of new ties. Thus, the number of common neighbors is a feature that is used for link addition prediction in social networks. Conversely, in the literature concerned with predicting the disappearance of links, other individual features are evaluated at that task. In order to take into account both the appearance and the disappearance of links, we will classify features by their performance on both tasks, resulting in four classes of features, as depicted in Table 3.4:

- **Stability** features are those indicating that neither link addition nor link removal will take place.

- **Instability** features are those indicating that both link addition and link removal are likely.
- **Growth** features are those indicating that link addition is likely whereas link removal is unlikely.
- **Decay** features are those indicating that link removal is likely whereas link addition is unlikely.

These four classes allow us to give a fine-grained characterization of individual features. For instance, a feature such as the number of common neighbors may be well-known to be an indicator for edge addition, but it is unknown whether it is also an indicator for the disappearance or for the non-disappearance of edges. The number of common neighbors may actually be a measure of growth, or of instability. Thus, the distinction of these four classes will also allow us to shed a new light on existing link addition prediction features, to tell whether they are indicators for the presence of edges or only for the change in the states of edges. In the following, we describe several potential signals for link addition and link removal from the literature.

Features

A large number of features for predicting link appearance and disappearance can be found in the literature [Liben-Nowell and Kleinberg, 2003, Raeder et al., 2011, Lü and Zhou, 2011]. These features can be grouped by the theory or model that explains how these features behave for the tasks at hand. Hypotheses (a)-(e) cover known models from the literature. The following list contains both node-level features and node pair-level features. To construct numerical indicators for node pairs from node-level features, we use the product of the feature values for both nodes, e. g., $d(i,j) = d(i) \cdot d(j)$.

(a) Preferential Attachment

The model of *preferential attachment* states that links are more likely to attach to nodes with a high degree [Barabási and Albert, 1999].
Hypothesis: *The number of adjacent nodes is a good indicator for link addition.*

Node degree: $d(i)$ is defined as the number of nodes adjacent to i, regardless of link direction.

Joint degree: $jd(i,j)$ is defined as number of nodes that are adjacent to node i or node j, regardless of link direction.

(b) Embedding

The embeddedness of a node pair measures to what extent two nodes are part of a larger cluster [Burt, 2000].
Hypothesis: *The embeddedness of a link is suitable to predict the appearance of links and the non-disappearance of existing links, i.e., it is an indicator for growth.*

Common neighbors: $CN(i,j)$ is defined as the number of common neighbors of node i and j.

Paths of length three: $P3(i,j)$ is defined as the number of paths of length three between node i and node j.

(c) Reciprocity

A link is reciprocated if the link in the opposite direction is present [Raeder et al., 2011].
Hypothesis: *The presence of a link makes the addition of a link in the opposite direction more likely and the removal of a reciprocal link less likely. Thus, it is an indicator for growth.*

Back-links: $back(i, j)$ is defined as a binary feature indicating whether a back-link exists, i.e., $back(i, j) = 1$ if $(j, i) \in E$ and $back(i, j) = 0$ otherwise.

(d) Liability of Newness

The principle termed *liability of newness* states that newly formed links are less likely to persist than older links and generally more volatile [Burt, 2000, Karney and Bradbury, 1995]. Also new nodes are likelier to form unstable ties.
Hypothesis: *The freshness of an edge or a node are good indicators for link change.*

Edge freshness: $eFresh(i, j)$ is defined as the time passed since the last add-event, i.e., the last time that node i has been linked to j. Since the freshness denotes the last add-event, links with a higher edge freshness value are considered as more fresh than others.

If an edge has never been present in the evolution of a network, the aforementioned feature is undefined. Thus, we elaborate on the idea of *liability of newness* and propose the following node feature.

Node freshness: We define $nFresh(i)$ as the freshness of node i, denoting the last time that any event related to node i occurred. Since the freshness of a node denotes its last add-event, nodes with a higher freshness value are also considered as more fresh than others.

(e) Stability of Oldness

We consider a node as stable if its content or its incident edges remain relatively unchanged. Generally, older nodes were found to be more stable with respect to their ties in the network [Burt, 2000, Karney and Bradbury, 1995].
Hypothesis: *The more stable nodes i and j are, the more stable the link (i, j) is, whether present or not.*

Edge age: $eAge(i, j)$ is defined as the time passed since the first add-event, i.e., the first time that the edge (i, j) was added. Accordingly, links with a higher edge age are considered as older.

If an edge has never been present in the evolution of a network, the aforementioned feature is undefined. Thus, we define the following node feature.

Node age: We define $nAge(i)$ as the age of node i, i.e., the first time that any event related to node i occurred. Accordingly, nodes with a higher node age are considered as older.

We summarize the features and the expected behavior with respect to the predictability of new links and link removals in Table 3.5.

3.4.2 Methodology

In our evaluation we use again the largest dynamic knowledge network on the Web, Wikipedia.

Model	Feature		New links	Link removal	Expected state
Preferential attachment	Node degree	d	↗	–	Growth / instability
Preferential attachment	Joint degree	jd	↗	–	Growth / instability
Embedding	Common neighbors	*CN*	↗	↘	Growth
Embedding	Paths of length three	*P3*	↗	↘	Growth
Reciprocity	Back-links	*back*	↗	↘	Growth
Liability of newness	Edge freshness	*eFresh*	↗	↗	Instability
Liability of newness	Node freshness	*nFresh*	↗	↗	Instability
Stability of oldness	Edge age	*eAge*	↘	↘	Stability
Stability of oldness	Node age	*nAge*	↘	↘	Stability

Table 3.5: Summary of hypotheses about the ability of features to predict link addition and removal. "↗" indicates a positive correlation; "↘" indicates a negative correlation.

Wikipedia	Articles $[\times 10^6]$	Adds $[\times 10^6]$	Deletes $[\times 10^6]$
French	1.8	41.7	17.3
German	1.5	58.7	27.6
Italian	1.0	26.0	8.9
Dutch	0.8	15.3	4.7

Table 3.6: The datasets used in our evaluation. The number of articles also includes articles that were removed.

Datasets

We use the directed article-hyperlink networks of four of the five largest[7] Wikipedias. In the directed article-hyperlink network of Wikipedia, a link between two articles i and j is present if article i links to article j. We omit user pages and article discussion pages.

For each of the four Wikipedias we consider all add and delete events until August 2011. An overview of the datasets is given in Table 3.6.

Prediction Methodology

Given the set of links E_{t_1} present at a particular time t_1, how can the links E_{t_2} at time t_2 be predicted accurately? This problem involves again the prediction of new edges E^+ and the prediction of deleted edges E^- as defined in the preceding section

$$E^+ = E_{t_2} \setminus E_{t_1},$$
$$E^- = E_{t_1} \setminus E_{t_2}.$$

The problem of predicting new links E^+ is called the link addition prediction problem, or simply the link prediction problem [Liben-Nowell and Kleinberg, 2003]. Typically, the link

[7] `http://meta.wikimedia.org/wiki/List_of_Wikipedias`

Wikipedia	$\lvert E^+\rvert$ $[\times 10^6]$	$\lvert E^-\rvert$ $[\times 10^6]$
French	5.3	1.2
German	10.2	1.7
Italian	3.9	0.7
Dutch	2.3	0.5

Table 3.7: The size of our link addition and link removal test sets for the four Wikipedias we consider.

addition prediction problem is solved by *link addition prediction functions*, i.e., functions that map node pairs to numerical scores, based on the known edges in the set E_{t_1}. The problem of predicting the disappearance of edges can then be solved analogously by *link removal prediction functions*.

To compare the prediction accuracy of different link addition prediction and link removal prediction functions, we define a true test set and a false test set for each of the prediction problems. The test set contains the node pairs to be predicted; the false test set contains node pairs that must not be predicted.

For the link addition prediction problem, this means that node pairs in the true test set E^+ must be distinguished from those that where not added, i.e., those in the false test set E^+_{false}. Analogously, the prediction of link removal aims at distinguishing links that are removed, in the true test set E^-, from those that are not removed, in the false test set E^-_{false}. The definition of the two false test sets is analogous to Section 3.3.

To solve a prediction problem, one uses functions of the form

$$f : E_{t_1} \to \mathbb{R},$$

that take the structure of the network at time t_1 as input to compute scores for all node pairs in the test and false test sets.

When applied to the edge set E_{t_1}, f is a good link addition prediction function when it gives node pairs in E^+ higher values than node pairs in E^+_{false}. Analogously, f is a good link removal prediction function when it gives edges in E^- higher values than edges that are not removed, in E^-_{false}. In Table 3.7 we give an overview of the number of edge additions and removals in the test sets for our datasets.

The performance of a prediction function f at the two prediction problems can then be used to classify it into the four categories of growth, decay, stability and instability; see Table 3.4. Link addition prediction functions (link removal prediction functions) can then be evaluated and compared.

Consistently, we the experiments in Section 3.3, we again use the AUC-value to measure the predictive performance of a prediction function. As a reminder, the AUC-value of a predictor is ranged between 0 and 1, where the perfect predictor received an AUC-value of 1.0 and a random predictor receives and AUC-value of 0.5.

Decay		**AUC**	**Instability**		**AUC**
Low node degree	$-d$	0.70	Nodes have been changed recently	*nFresh*	0.71
Low joint degree	$-jd$	0.69	Old edge	*eAge*	0.65
Few paths of length three	$-P3$	0.67	Edge has been changed recently	*eFresh*	0.64
Stability		**AUC**	**Growth**		**AUC**
Nodes have been unchanged for long	$-nFresh$	0.71	High node degree	*d*	0.70
Young edge	$-eAge$	0.65	High joint degree	*jd*	0.69
Edge has been unchanged for long	$-eFresh$	0.64	Many paths of length three	*P3*	0.67

Table 3.8: The three best performing indicators for the four classes are shown along with their average AUC-values across the four datasets and the two prediction tasks.

3.4.3 Evaluation

In this section we report on experiments to determine which features are suitable signals for link addition and removal.

We compute all eleven features described in Section 3.4.1 and compute the AUC-values of the link addition and removal prediction tasks. Figure 3.8 shows the performance of the features at the task of link addition and removal prediction for all studied datasets. Table 3.8 shows the top-three performing features for each of the four classes.

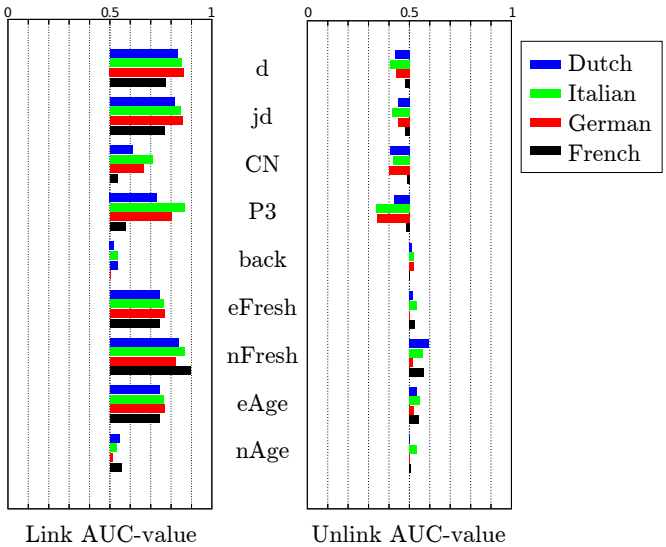
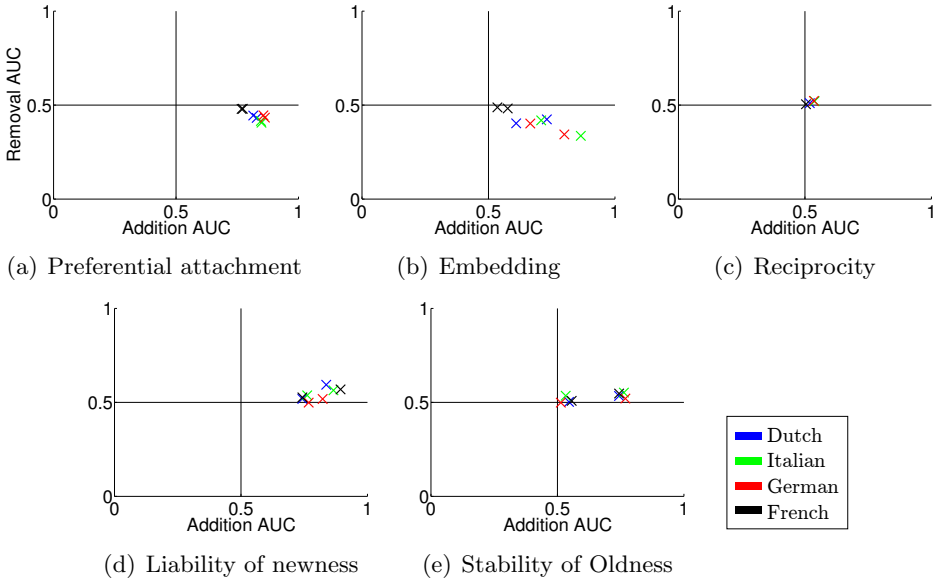


Figure 3.8: AUC-values for the link prediction and unlink prediction tasks are shown for all features and all four datasets. Note that a below-random AUC-value can be turned into an above-random one by the negation of the respective feature.

In the following, we compare our results with the projections of the hypotheses from Section 3.4.1.



(a) Preferential attachment (b) Embedding (c) Reciprocity

(d) Liability of newness (e) Stability of Oldness

Figure 3.9: Link prediction and unlink prediction AUC-values for the indicators based on the five models. The X and Y axes of each plot show the AUC-values of the link and unlink prediction tasks, respectively. The two lines showing an AUC-value of one half divide each plot into four quadrants, corresponding to the four classes of indicators.

(a) Preferential Attachment

Hypothesis: *The number of adjacent nodes is a good indicator for link addition.*
Following the hypothesis, we expect a good link addition prediction performance for features of preferential attachment. Figure 3.9(a) shows the AUC-values for the two preferential attachment features. Our experiments show that preferential attachment features are indeed good indicators for the formation of new links, as can be seen by the AUC-values above 0.5 for the two features. As all features scored below the AUC-value of 0.5 for the prediction task of link removal, we conclude that preferential attachment features are signals for *growth*. In terms of the knowledge networks, this implies that popular knowledge items tend to become integrated with more knowledge items.

(b) Embedding

Hypothesis: *The embeddedness of a link is suitable to predict the appearance of links and the non-disappearance of existing links, i.e., it is an indicator for growth.*
Following the hypothesis, we expect a good link prediction performance and a bad unlink prediction performance for features of embeddedness. Figure 3.9(b) depicts the AUC-values for this feature for link versus unlink prediction. For all four networks, this feature is situated in the lower right quadrant, implying that embedding is an indicator of *growth*. In terms of the knowledge networks, this implies that indirect relationships tend to be made explicit by direct knowledge connections.

(c) Reciprocity

Hypothesis: *The presence of a link makes the addition of a link in the opposite direction more likely and reciprocal links are likelier to persist. Thus, it is an indicator of growth.*
Following the hypothesis, we expect a good link prediction performance and a bad unlink prediction performance. We depict the results for the binary feature of back-link *back* in Figure 3.9(c). We observe a tendency of this feature to be correlated with the formation of new links, but the AUC-values are only marginally different from the random baseline. This confirms the fact that knowledge networks are inherently directed and that relationships between knowledge items are not necessarily symmetric as opposed to links in social networks. Therefore the feature of reciprocity does not fit into any of our four categories.

(d) Liability of Newness

Hypothesis: *The freshness of an edge or a node are good indicators for link change.*
Following the hypothesis, we expect a good link and unlink prediction performance for these features. In particular the node freshness is a good indicator for the formation of new links and the removal of links. The predictive performance of the edge freshness is good for the link prediction task but rather random for the unlink prediction task. Therefore, we can only confirm that the node freshness is an indicator of *instability*. In terms of knowledge networks, this implies that new a new knowledge item is fragile, the connections to other knowledge items will be changed more often.

(e) Stability of Oldness

Hypothesis: *The more stable nodes i and j are, the more stable the link (i, j) is, whether present or not.*
Following the hypothesis, we expect a bad link prediction performance and a bad unlink prediction performance for the age of a node and an edge. Our findings shown in Figure 3.9(e) and Figure 3.8 suggest that the age of an article is indeed correlated negatively with the likelihood of link addition and link removal. On the other hand, the age of an edge is an indicator of *growth* rather than of *instability* because it is correlated positively only with the formation of new links. Therefore we can confirm only that the node age is a good indicator of the *stability* of an edge. Old knowledge items in knowledge networks can thus be considered as more stable, the established connections are more likely to remain.

Comparison of Prediction Problems

We can use our evaluation to make a remark on the problems of link addition and link removal prediction. As a general rule, our results show that the problem of link prediction can be solved to a much higher accuracy (AUC ≈ 0.85) than the unlink prediction problem (AUC ≈ 0.60).

On the level of the four different classes of prediction problem which generalize the link and unlink prediction problem. We observe that the problem of growth prediction can be solved well using embedding indicators (see Figure 3.9(b)), as can the instability prediction problem (see Figure 3.9(e)). Since indicators for decay and stability can be derived from these two by inversions, it follows that all four types of prediction problems can be solved well.

Growth vs Instability

For the problem of link prediction, the features usually considered are not evaluated on the task of unlink prediction. Unlink prediction is however, even if it is rarely included in evaluation datasets, present in the majority of real-world networks. Thus, the distinction between indicators of growth, which correlate with the addition of edges and the non-removal of edges, and indicators of instability, which correlate with both the addition and removal of edges, should be made. As an example, a social recommender system ("you may also know these people") should use indicators of growth rather than indicators of instability. Even if an instable tie is likely to appear now, it is also likely to disappear later, and therefore should not be recommended. Our results thus show that preferential attachment-based and embedding-based indicators indicate growth and should thus be used for recommendation and other link prediction-type applications, while node and link age-based measures should not. This result is also in line with the link prediction literature, in which the best features are found to be based on preferential attachment and path counts [Liben-Nowell and Kleinberg, 2003, Lü and Zhou, 2011].

Link Versus Unlink Prediction

Note on Link Prediction False Test Set When looking at the auc-values at first glance on can easily get the idea that unlink prediction is a lot more difficult than link prediction. Whereas auc-values for Link Prediction are scored around 0.85, the best performing unlink prediction method can only reach 0.65. Note that these two values cannot be compared directly, since the set up for the two prediction problems is different. Whereas in the Link Removal Set Up, Links that remain are compared with links that are removed, the Link Prediction set Up compares links that are added versus random links that never appear in the network. The latter is by definition much easier.

Link Prediction Outperforms Unlink Prediction As we can see, the distributions of many link measures are highly skewed; many node pairs have the same small values whereas fewer node pairs have high values. This does not ressemble a problem for link prediction, because there the higher values are ranked better. Since the most succesful unlink prediction scores were obtained by negating link prediction scores, the low-scoring node pairs from the depicted distribution were shown. Since many node pairs have no common neighbors (60%) they will be ranked randomly. Therefore, the resulting AUC-value will not achieve values comparable with the link prediction problem.

3.4.4 Conclusion

Having performed experiments on four big Wikipedia datasets, we can state that indeed the appearance and disappearance of connections between items of knowledge in knowledge networks follow predictable patterns. As we showed, the patterns can be understood as an extension of link prediction models known in the literature, as well as of the much rarer link removal prediction problem. However, we found that to understand the dynamics of knowledge completely, a unified view of addition and removal must be adopted that distinguishes not two but four types of changes, namely growth, decay, stability and instability. We were able to verify empirically into which of these four categories the known prediction methods fit, showing that for all four,

suitable indicators exist. In particular, we were able to classify link prediction functions into those which actually indicate growth of the connectivity in a knowledge network, and those which indicate only instability. By reviewing known models of link-based network evolution, we were not only able to give a more detailed classification of known numerical indicators, but also to propose the novel indicator of the node deletion coefficient, which indicates instability, and is defined as the ratio of link deletion to link additions for a specific node.

3.5 Conclusions

In this chapter, we have studied the relationship between link and unlink prediction in two main experiments. Whereas the link prediction problem has been studied in various research, the unlink prediction problem has so far not been tackled in a general and structural manner. For the first experiment, we have considered two transformations of the link prediction problem into the unlink prediction problem. The *Complement Score Model* simply negates the link prediction score, therefore links that receive a low link prediction score are likely to be removed. The *Complement Network Model* computes the link prediction measures on the complement network of the original graph. Links that are likely to appear in the complement network can then be interpreted as potential unlinks in the original network. We have evaluated both models on five different Wikipedia datasets. The *Complement Score Model* turned out to be superior over the *Complement Network Model* for all considered prediction measures. We have demonstrated that unlink prediction is feasible with the classic structural indicators lent from link prediction. At the same time, the unlink prediction problem turned out to be more than a simple transformation of the link prediction problem. Both problems are distinct and only related prediction problems. In a second line of research, we have then analyzed the interplay of both prediction problems. As opposed to considering each problem as a separate prediction problem, we have provided a unified view which led us to categorize four different link change states of growth, decay, stability and instability. We have tested the predictive performance of different structural characteristics and categorized them into the four classes. Negated Structural measures of the embeddedness of a tie, i.e. the negated number of common neighbors ($-CN$) and the negated common neighbors of common neighbors ($-P3$) have shown to be superior over all other. Further we found that these two measures are also good indicators of the *growth* of a link – they are correlated positively with the likelihood of a new links and correlated negatively with the likelihood of link removal. On the contrary, the node freshness – the timestamp of the last edge addition for a node – was correlated positively for both prediction problems; therefore we consider it as a measure of *instability*. The work in this chapter has shown that unlink prediction is feasible with structural characteristics, but also that it lacks the strong predictive results of link prediction. The predictive performance of link prediction is significantly higher than for unlink prediction with the studied characteristics.

The work in this chapter was published in two papers:

- *Julia Preusse, Jérôme Kunegis, Matthias Thimm, and Sergej Sizov. DecLiNe - Models for Decay of Links in Networks. ArXiv e-prints, 2012.*
  Topic: Transformations of link to unlink prediction.
- *Julia Preusse, Jérôme Kunegis, Matthias Thimm, Thomas Gottron, and Steffen Staab. Structural dynamics of knowledge networks. In Proc. Int. Conf. on Weblogs and Social*

Media, pages 506–515, 2013.
Topic: Interplay of link and unlink prediction.

4 Temporal Models of Knowledge Networks

4.1 Introduction

A non-negligible part of social media is concerned, not with exchanging personal information, but with building knowledge bases. Such knowledge bases are for instance given by any part of the *semantic web*, in which knowledge is represented in a systematic way. Most prominently, the online encyclopedia Wikipedia represents one of the largest online communities dedicated to building a knowledge base spanning all areas of human knowledge – Wikipedia contains over four and a half million articles in the English language alone. While Wikipedia is a high-quality and up-to-date knowledge base with many practical uses, the question of Wikipedia's temporal stability remains unanswered. In contrast to traditional encyclopedias, Wikipedia is evolving very rapidly – articles are added, interlinked and revised constantly, reflecting the fast changes in many areas such as politics, art, and popular culture. Since change is an inherently temporal phenomenon, we may ask the question whether change in Wikipedia's hyperlink structure is mediated by temporal phenomena. To give pertinent answers to this question, we investigate three temporal working hypotheses: (1) temporal changes are mediated by the qualitative nature of past structure, (2) temporal changes are mediated by the recency of past edges and, (3) temporal changes are mediated by the temporal evolution of the neighborhood of individual nodes. To verify these hypotheses, we evaluate corresponding prediction algorithms, which must perform well under each hypothesis. In particular, we note that changes in a knowledge network can be of two fundamental types: the addition and the removal of edges. This leads to two separate prediction problems, the link prediction problem, and the unlink prediction problem.

Even though the links between articles of Wikipedia contain only a fractionally small part of the encyclopedia's total information, they can be used to get insight into the knowledge base with great accuracy. For instance, the addition of links to a network can be predicted, making it possible to suggest new connections between articles, and thus knowledge [West et al., 2009]. The accuracy of such link prediction algorithms is usually however not perfect. For instance, the imbalance between existing and non-existing edges needs special attention [Lichtenwalter et al., 2010], and typical accuracy values of link prediction algorithms attain values of 80% percents, as measured by the common measure AUC-value [Lü and Zhou, 2011].

As we have seen in the last chapter, the accuracy of observed structural methods are even as low as 60%. While this appears to be due to the fact that link prediction (and unlink prediction) algorithms only consider structural feature of a knowledge network, we show incidentally that this is not the case: The accuracy of link prediction can attain values as high as 95%, and that of unlink prediction 70%. As will be seen, the key to achieving these results lies in the temporal information. Whereas the last chapter has only considered measures of a snapshot of the data, this chapter uses the history of addition and removal events to improve the classification of links *and* unlinks.

Research Questions

This chapter will answer the following research questions.

RQ 2 *Does the exploitation of temporal data improve the classification of new links and unlinks?*

If the timestamps of individual addition and removal events are given for a dataset, how can this temporal information be exploited for the prediction of links and unlinks?

RQ 2.I *What strategies would be adequate to exploit temporal informations as to classify new links and unlinks?*

Information of addition and removal events can be leveraged on different levels; one could use the specific timestamp of an event, use only the ordering of events or exploit the qualitative information how often a link has been added or deleted.

RQ 2.II *Does the exploitation of temporal data improve the classification of new links and unlinks?*

Will the prediction results be significantly better than without temporal information? The snapshot representation of a dataset does not provide any evidence to whether links that are not in the snapshot have been present before. We hypothesize that information on unlinks, that can be extracted from temporal data, should improve the predictability of new unlinks.

The structure of this chapter is as follows. Section 4.2 will review related research that incorporates temporal information for link prediction tasks. Section 4.3 is dedicated to research question RQ 2.I and presents four models of the temporal evolution of knowledge networks. The evaluation set up for the proposed models is defined in Section 4.4. The added-value of temporal data, which corresponds to research question RQ 2.II, is then evaluated in Section 4.5. Finally, Section 4.6 concludes the chapter and discusses the results.

4.2 Related Work

4.2.1 Temporal Link Prediction

The temporal link prediction problem is defined as follows. Given linkage events of time $t_1 \dots t_N$, which links will appear at time t_{N+1}? Hence, the temporal link prediction problem uses temporal data for all events until t_N to predict links that will appear until t_{N+1}. In contrast, the problem of link prediction uses only a snapshot of links present at time t_N.

Related work on the temporal link prediction problem can be grouped into two categories – *weighted summary* and *temporal features*. The first category uses only the timestamps of edge events as input for the prediction algorithm. The second category of work deals with highly-parallel networks for which a weighted summary of the temporal history of each edge event is built. Both approaches measure characteristics of a single edge; they cannot be applied to paths, such as the number of common neighbors, or even to weight a node's neighborhood by time. We review the two categories in the following.

Temporal Features This line of work uses the first or last timestamp of an edge either as a single feature or plugged into an ensemble learning method to derive the likelihood of a particular edge. Tylenda et al. use the recency of a publication as an input feature for ensemble learning methods to predict links in publication networks [Tylenda et al., 2009]. Two studies on predicting decay in mobile phone communication networks define that links decay if no communication was exchanged between actors for a particularly chosen time period [Raeder et al., 2011, Hidalgo and Rodriguez-Sickert, 2008]. Preusse et al. have used a Wikipedia article's ratio of link removals and link additions to predict whether a link will be formed or will decay [Preusse et al., 2013].

Weighted Summary This line of work deals with networks with highly-parallel edges such as in communication networks. To infer the future number of interactions between two users, a temporally weighted summary of the interaction history of the user pair is implemented instead of just using the number of interactions between the two.

Targeting the problem to predict the next topic of an author's publication, Sharan and Neville implemented different kernel functions that weight the more recent topics on which an author has published higher [Sharan and Neville, 2008]. They have shown that this predictor achieves better results than a predictor that ignores the time that an item was published.

Spiegel et al. use tensor factorization to asses trends of data with parallel edges such as product ratings or interactions [Spiegel et al., 2012]. They apply exponential smoothing to the number of parallel edges in defined time segments to extrapolate the value to the target time segment.

A costly low-rank approximation of the adjacency matrix that captures latent relationships is presented in [Hayashi et al., 2009, Tong et al., 2008]. The dynamic representations are then used for Bayesian inference.

Koren analyzed the user rating behavior for movies in Netflix [Koren, 2010]. He proposed complex models that are based on observations such as that user preferences change from week to weekend and that the overall user taste and the movie rating shifts over time.

There are many works that use time-series analysis to predict the occurrence of the next event based on regularly re-occurring patterns. For instance in [Huang and Lin, 2009], methods of time-series analysis were combined with static structural approaches to predict new links in two highly-parallel event networks .

These methods cannot be applied to our status networks, because the history of an edge consists mostly of only as few as one or two events: the addition and the removal of an edge. As depicted in Figure 4.1 most edges in the knowledge network Wikipedia appear only once (between 62% and 88% for all networks) or are additionally removed once. In average for 90% of all edges only an addition and removal time stamp is given. Thus there are too few state changes to use the described methods.

In [Potgieter et al., 2007, Cooke et al., 2006], two temporal versions of classic link prediction characteristics, such as node degree, number of common neighbors and Adamic-Adar, to predict interactions in a social network are proposed. The link prediction measures are temporally weighted by the return, i.e. the ratio of the value at a defined start point and the value at a defined end point, and the average of a characteristic over some defined time points. The choice of the start point greatly influences the results of the weighted summary; finding an appropriate start point for all node pairs seems very tough.

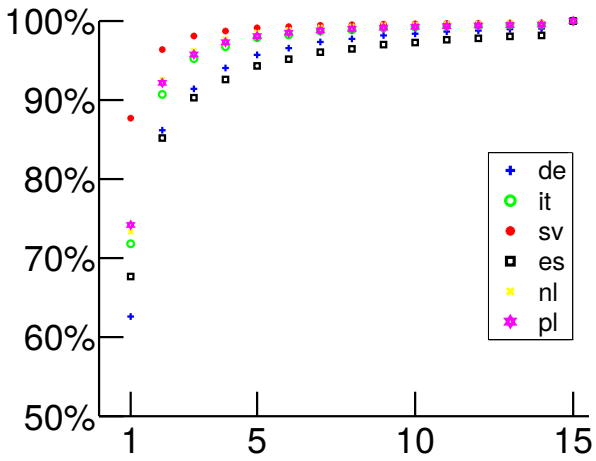


Figure 4.1: The distribution of the number of state changes per edge is depicted. The y-value corresponding to x=1 gives the number of edges that were added exactly once. We have cut the distribution at x=14 and added all remaining status changes to the corresponding y-value of x=15.

4.2.2 Related Problems

Link Decay In many networks, links cannot be removed but are rather considered to become *inactive* or to *decay*. Two studies that predict decay in mobile phone communication networks assume that links decay if no communication was exchanged between the actors for a particularly chosen time period [Raeder et al., 2011, Hidalgo and Rodriguez-Sickert, 2008]. Both works conclude that links are more likely to persist when the connection is reciprocated and when either both actors' degrees are low or both degrees are high. Raeder et al. observed that the "liability of newness" holds, i.e., the age of a tie is correlated with its persistence [Raeder et al., 2011]). Viswanath et al. found that the longer two users have engaged in wall-to-wall interactions on Facebook, the more likely they are to continue and thereby the less likely the interaction link in Facebook is to decay [Viswanath et al., 2009]. We cannot use features such as the interaction frequency, since links in a state network mostly change from present to removed and thus the history of a tie is not useful. Further, this line of research deals with derived link removals as the datasets themselves do not contain explicit unlinks.

Cross-Temporal Link Prediction Oyama et al. were the first to define the cross-temporal link prediction problem [Oyama et al., 2011]. In contrast to temporal link prediction, where links in a time slice T_{N+1} should be predicted, this problem aims at predicting links between different time slices. Applications of this problem are entity resolution, where events for entities are given for some time slices, but corresponding entities should be connected, and asynchronous communications, where messages should be linked to the messages they reply to, which is not given in many datasets or applications.

4.3 Temporal Models of Structural Change

The knowledge captured in knowledge networks is very dynamic. New pieces of knowledge are discovered and connections between existing knowledge items are made constantly. At the same time, connections between knowledge items are also removed if they are obsolete or wrong. We assume that the structural aspects of the knowledge are captured in the links between individual knowledge items, and will thus only consider changes to the links disregarding the actual changes in the content.

The structural evolution of knowledge follows predictable patterns, which fall into two main lines: time-agnostic ones that do not build upon temporal features to predict the network future and time-aware ones that exploit temporal trends to infer the future.

4.3.1 Hypotheses of Knowledge Evolution

In the following, we present the three hypotheses (H1-H3) stating which information in the network influences the future, and the baseline hypothesis that time does not play a role (H0).

Time-Agnostic Baseline Hypothesis

H0: *Future changes depend only on the network's current state.*

This hypothesis suggests that all information that is needed to infer the network's future is captured by the current state of the network. Existing approaches in the areas of link prediction and unlink prediction compute structural indicators for articles and article links and evaluate their performance on the actual future network. When temporal data is not present in a dataset, this assumption is implicitly made. A very large fraction of studies on link prediction consider only a static model of the network. As examples, the link prediction survey by [Liben-Nowell and Kleinberg, 2003] as well as that by [Lü and Zhou, 2011] only review methods based on this hypothesis.

Newer work on link prediction does exploit temporal information; this line of work is often referred to as temporal link prediction. Work on temporal link prediction can be broadly classified by the type of temporal information considered into three classes: works that take the qualitative nature of changes into account, works that consider the time of the last edge addition or removal event between any node pair, and works that consider the full add/remove history between all nodes. In the following, we derive three hypotheses from these three classes.

Qualitative Hypothesis

H1: *The appearance and disappearance of edges in the past influences the presence of edges in the future.*

Whereas the time-agnostic hypothesis uses only the current structure, this hypothesis states that the information on links and articles that have been removed and are not present in the current network, improves the predictability of the future network. As an example, an article's ratio of added and removed links was found to be a successful indicator for future link removals [Preusse et al., 2013].

Decay Hypothesis

H2: *The timepoint of the appearance and disappearance of an edge in the past influences the presence of an edge in the future.*

We refine this hypothesis by whether the weight of edges should increase or decrease by time, giving hypotheses of recency and longevity.

H2R: *Two articles with recent new links are more likely to link to each other.*

H2L: *Two articles with older new links are more likely to link to each other.*

Under the recency hypothesis, changes made long ago should affect the current network dynamics less than as recent changes. Measures such as the elapsed time since the last interaction were shown to be suitable indicators for future activity between two actors [Hidalgo and Rodriguez-Sickert, 2008, Raeder et al., 2011]. On the contrary, the *liability of newness* by [Burt, 2000] asserts that new links are very fragile and the effect of older established edges should be more indicative of future changes. Sharan and Neville built a weighted summary to predict the topics that an author will next publish on [Sharan and Neville, 2008]. For this application, the more recent publications appeared to be more important than the later onces.

Neighborhood Evolution Hypothesis

H3: *The temporal evolution of the neighborhood of an article influences its future neighborhood.*

Based on the evolution of their neighborhood, knowledge items can be classified as growing or stable with respect to new links or removed links. Trending items will change more as well as their interconnections with other items, whereas for older knowledge items most connections have been established. Thus, considering the local changes around a knowledge item, i.e. its neighborhood, should give a good indication of its future. For instance, extrapolating the event matrix to the future with exponential decaying weights of past events has been shown to perform well [Spiegel et al., 2012], and work in collaborative filtering demonstrates the usefulness of modeling time changing behavior throughout the lifespan of a bipartite rating graph [Koren, 2010].

Modeling Structural Change Having defined one time-agnostic and three time-aware hypotheses, we now present corresponding models of link change. We capture the different levels of temporal information from the hypotheses introduced in the previous section in each different network representation. Whereas the time-agnostic hypothesis (H0) is best modeled using the knowledge network's adjacency matrix, the qualitative and decay hypotheses (H1-H2) will be defined using weighted versions of it. The neighborhood evolution hypothesis (H3) is implemented by sequences of different neighborhood sizes which are extrapolated to estimate the future neighborhood of an article and consequently the network structure.

Link and Unlink Prediction Functions

The problem of link prediction is concerned with finding good indicators that predict whether an edge will appear in a network or not. A particular set of link prediction features has proven to be successful to predict the appearance of links in many different networks [Liben-Nowell and Kleinberg, 2003, Lü and Zhou, 2011]. These link prediction features are generally defined in a time-agnostic setting, i.e., without reference to appearance times of edges. In order to

implement link prediction methods corresponding to the hypotheses H1 and H2, we will extend these methods to take into account edge weights that are functions of an edge's history. We will restrict this study to the widely used (and supposedly best-performing) link and unlink prediction algorithms. Other, more complex algorithms include the index of Katz [Katz, 1953], graph kernels [Ito et al., 2005] and diffusion models [Kondor and Lafferty, 2002].

Work on the unlink prediction problem is much more sparse, and is mostly focused around specific social networks, for instance Twitter [Kwak et al., 2012] and Facebook [Quercia et al., 2012]. As these studies use very specific user information, e.g. personality traits or gender, they cannot be used to predict link removal in networks other than the chosen social networks. Another recent work classifies Wikipedia links into four categories: stable, instable, likely to appear and likely to be removed [Preusse et al., 2013]. Whereas suitable structural indicators for the formation of new links are found that reach an AUC-value of around 90% for some datasets, the best-performing features that are indicative of edge decay achieve an AUC-value of only 60%.

Definitions

Let a knowledge network be denoted as the directed graph $G = (V, E)$ consisting of a set of article nodes V and a set of directed hyperlinks between articles $E \subseteq V \times V$.

In order to analyze the detailed temporal evolution of a knowledge network we need to consider individual additions and removals of edges, which we both call *state changes*. Let $\mathcal{E} \subseteq V \times V \times \{+1, -1\} \times \mathbb{R}$ be the set of state changes, where each state change is either the addition of an edge or the removal of an edge. Each state change $e \in \mathcal{E}$ is of the form $e = (i, j, \pm 1, t)$, where i is the Wikipedia article containing the link, j is the linked-to article, t is the timestamp of the state change and the number $+1$ denotes an addition and the number -1 a removal of the link. We consider only simple links; multiple parallel links are coalesced into a single link in our treatment.

In order to take edge weights into account, it is useful to define the network's adjacency matrix. The asymmetric adjacency matrix A of the graph G is a $|V| \times |V|$ matrix with entries of $A_{ij} = 1$ denoting an edge from i to j, and entries $A_{ij} = 0$ denoting no edge. When a link or unlink prediction algorithm is defined in terms of the adjacency matrix, as many are, it can be extended to take edge weights into account by replacing the matrix A in its definition by a matrix W of the same size that contains edge weights.

Weighting of Edges

The regular (time-agnostic) link prediction scenario considers predictions measures based on the network at a particular time. All commonly used prediction functions can be defined using the adjacency matrix of the network as shown in Table 4.1, and are from two main categories: features based on preferential attachment, and features based on the embeddedness of a link. The principle of preferential attachment asserts that the number of new connections an article forms is proportional to the number of connections formed so far. Particular measures count the number of incoming links, the number of outgoing links, or their sum. On the other hand, the embeddedness of an article pair measures to what extent the articles are part of a larger cluster in the network. Whether two articles are well embedded is measured for instance by the number of common neighbors or by the number of common neighbors' neighbors.

Feature	Definition
Out-degree	$d_{\text{out}}(i) = \sum_{j \in V} W_{ij}$
In-degree	$d_{\text{in}}(i) = \sum_{j \in V} W_{ji}$
Common neighbors	$\text{CN}(i,j) = [(W + W^{\text{T}})^2]_{ij}$
Paths of length three	$\text{P3}(i,j) = [(W + W^{\text{T}})^3]_{ij}$

Table 4.1: Overview of prediction features and their definition in terms of the weighted adjacency matrix W. The definition of the corresponding usual time-agnostic prediction function is given by setting $W = A$.

In order to systematically investigate link and unlink prediction functions aligned with the hypotheses H0, H1 and H2, we extend the most common time-agnostic measures to allow weighted edges, with edge weights given by each model. We express these weights by replacing the 0/1 adjacency matrix A by a matrix of weights W in the definition of each prediction function. For hypothesis H3 (the neighborhood evolution hypothesis), we base the prediction functions on an extrapolation of the evolution of an article's neighborhood that does not fit the expressions in Table 4.1.

4.3.2 Time-Agnostic Model (M0)

The time-agnostic baseline hypothesis states that future edges are influenced only by current edges. In this model, the current state of the network can be represented by the adjacency matrix A, and all link and unlink prediction algorithms can then be defined in terms of this matrix. We define the features in the following.

Preferential Attachment Features Existing link prediction functions measure an article's in- and out-degree, which are defined as the number of articles an article links to and is linked to respectively. As a function to predict the directed edge $i \to j$, we thus use the out-degree of i and the in-degree of j.

Embedding Features The embeddedness of an article pair measures to what extent the two articles are part of a larger cluster. In analogy to other link prediction works, we quantify the embedding of an article pair by the number of common neighbors, i.e., the number of articles that both articles link to or are linked from. We also consider the number of paths of length three $\text{P3}(i,j)$ between articles i and j, measuring how many of i's neighbors are connected to j's neighbors. For these embedding methods, we always ignore edge directions, i.e., use the symmetrized matrix $A + A^{\text{T}}$.

We summarize all features in Table 4.1, where we also give their definition in terms of the weighted adjacency matrix W. Thus, the definitions of the time-agnostic prediction functions are given for $W = A$.

4.3.3 Qualitative Model (M1)

The previous representation of the network by its adjacency matrix contains no evidence of links that are not present anymore. Structural characteristics of this kind have been shown

to perform well on the link prediction task, but do not work well for unlink prediction. The latter is plausible because removed edges which may be the key to detect the removal of other edges – analogously to link prediction where added edges are used to infer which other edges will be added – are not contained in this network representation. Thus, hypothesis H1 states that the appearance and disappearance of edges in the past influences the presence of edges in the future. We propose a representation that captures links that have been removed: the removal-adjacency matrix A^- of a network. It is defined for an article pair i, j by

$$A^-_{ij} = 1 \Leftrightarrow \exists t : (i, j, -1, t) \in \mathcal{E} \wedge A_{ij} = 0,$$

i.e., it indicates which edges have been removed in the network and are not present in the current network.

We have defined a new weighting of the adjacency matrix, the removal adjacency matrix A^-, and can thus conclude the features summarized in Table 4.1 analogously.

Preferential Attachment Features The removal out-degree $d^-_{\text{out}}(i)$ and the removal in-degree $d^-_{\text{in}}(i)$ are defined as the number of outgoing and incoming removal edges of article i. Equivalently, they correspond to the definitions given in Table 4.1 when $W = A^-$. In Wikipedia, we have observed that some links are prone to be removed if already many links to this article have been removed, justifying the use of these degree measures for unlink prediction.

Embedding Features Exploiting the structure of the removal and addition network, we define new versions of the number of common neighbors and paths of length three. We define a new common neighbor feature that is given by the number of neighbors that one article links to and the other article is not linked with anymore. In the same way, the new paths-of-length-three feature of an article pair i, j is defined as the number of unlinked neighbors of i that are linked to neighbors of j ore vice versa.

We expect the path measures to be indicative for the removal of links. When the embedding of two articles is removed, the link between the two articles should be more likely to be removed, as well.

4.3.4 Decay Model (M2)

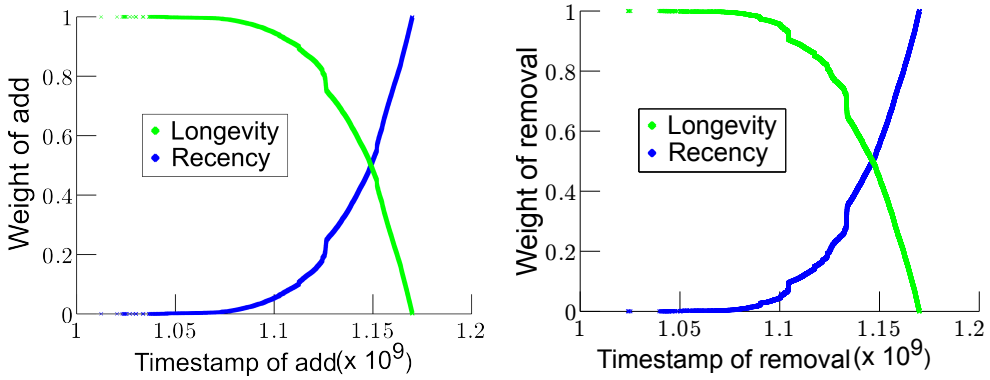
Whereas the qualitative model (M1) exploits information on whether a link has been added or removed, we hypothesize that exploiting the timestamps of these events improves the classification of links and unlinks even further. For that, we introduce two ways to weight edges by their timestamp: by recency and by longevity.

If edges are weighted by recency, the most recent edge receives the highest weight, whereas the oldest edge receives the highest longevity value. We define the recency and longevity adjacency matrices as follows. First, all known edges $(i, j) \in E$ are sorted by their timestamp in descending order for the recency and in ascending order for the longevity weights. Consequently, we obtain two orderings R and L for recency and longevity and define $R(i, j) \in [0, |R| - 1]$ as the position of an edge in the ranking $R(i, j)$ and $L(i, j)$ analogously. The recency and longevity weighting of the weighted adjacency matrix A_{R} and A_{L} for for all edges present in

the current network is then defined as

$$[A_{\mathrm{R}}]_{ij} = (|R| - R(i,j))/|R|,$$

and analogously for A_{L}. Thus, the most recent node pair has a recency value of one, whereas the oldest node pair has a recency value close to zero. In the same way, we define A_R^- and A_L^- as the recency removal matrix and longevity removal matrix to weight links that were removed by recency and longevity, as well. We depict the recency and longevity weight of all links in the Spanish Wikipedia as a function of their timestamps in Figure 4.2.



(a) Adds (b) Removals

Figure 4.2: The recency and longevity weights of all link events, i.e., edge additions and removals, are given as a function of their timestamp for the Spanish Wikipedia.

Hence, we have defined two models in accordance with the decay hypothesis: the *recency decay model* (M2R) and the *longevity decay model* (M2L), for which we derive the features summarized in Table 4.1 analogously.

Preferential Attachment Features The recency and longevity degree as well as their removal counterparts are then defined for the weight matrices $W = A_{\mathrm{R}}$ and $W = A_{\mathrm{L}}$. This way, we can distinguish between an article that has formed links to other articles recently or long ago.

Embedding Features For the paths features we will observe whether there is a difference between recent and older paths between two nodes. Consider the example illustrated in Figure 4.3. Should the more recent common neighbors formed within the last hours of the left example

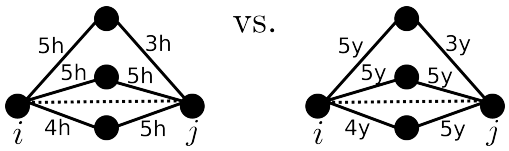


Figure 4.3: Two article pairs are contrasted: one article pair that has recent common neighbors formed within the last hours (left) and one article pair with common neighbors formed several years ago.

influence the likelihood of a new link (i, j) positively or are older links formed several years ago as depicted on the right more likely to trigger a new link? We weight all defined embedding characteristics of the qualitative model (M1) by recency and longevity, respectively.

4.3.5 Neighborhood Evolution Model (M3)

Preferential attachment states that the number of new links of an article is proportional to its degree and thus disregards its temporal evolution. An article's degree does not capture whether the article is growing or has stabilized.

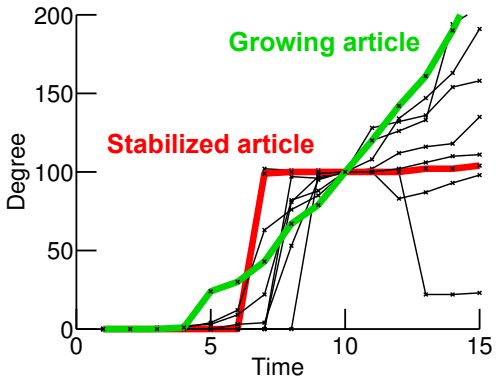


Figure 4.4: The evolution of the degree of nine sample articles is shown for a random subset of articles in the Spanish Wikipedia whose degree is 100 at time 10.

Consider the two highlighted articles in Figure 4.4. Both have the same degree d of 100 at time 10. When considering the degree evolution of both articles, the green article is growing, whereas the red article has stabilized. However, preferential attachment estimates the same number of new links for both articles. Thus, it fails to distinguish between growing and stable or even decaying articles. This brings us to our third and last hypothesis that using the temporal evolution of an article's degree improves the prediction.

Based on the temporal evolution of the article's degree, we estimate the number of new out- and in-links that will be added and removed. We will suppose that the add and remove events of the knowledge network's history are partitioned by time in multiple slices, where each slice represents the same time span.

Features We define the addition in- and out-degree $d^+_{in,k}(i)$, $d^+_{out,k}(i)$ of an article i in slice k as the number of new in- and out-links formed in slice k for article i. The removal in- and out-degree of a node pair in a slice is defined analogously. Given the sequence of, e.g., all addition in-degrees $d^+_{in,1}(i), \ldots, d^+_{in,T}(i)$ of article i for all slices $1, \ldots, T$, we employ exponential smoothing to estimate the future amount of additions to the in-degree of node i in the future slice $T + 1$ as

$$d^+_{in,T+1}(i) = \sum_{k=1}^{T} \alpha^+_{in} \cdot (1 - \alpha^+_{in})^{T-k} d^+_{in,k}(i), \tag{4.1}$$

where $\alpha_{in}^{+} \in (0, 1]$ is a smoothing factor. We also estimate removal in- and out-degrees with Equation (4.1). For $\alpha_{in/out}^{\pm} = 1$, only the last degree change is considered. Contrarily, for small $\alpha_{in/out}^{\pm}$ the influence of past degree changes increases and all degree changes are more weighted equally.

We expect the exponential smoothing to be a better approximation of an article's future degree than the classic preferential attachment measure.

4.4 Methodology

In order to test the statistical significance of our experiments, we will use multiple Wikipedias datasets in different languages. The list of language Wikipedias we use is given in Table 4.2. We exclude the three largest language Wikipedias due to their size, the English, French and German ones.

Language	**Articles** [$\times 10^6$]	**Adds** [$\times 10^6$]	**Deletes** [$\times 10^6$]	**Test slices**
Spanish	2.5	43.6	21.0	7
Swedish	1.9	21.7	3.8	3
Italian	1.2	26.0	8.9	7
Dutch	1.0	15.3	4.7	5
Polish	1.0	18.8	6.2	6

Table 4.2: The language Wikipedia datasets used in our evaluation. The number of articles includes articles that were removed, and is therefore higher than the value reported on the official Wikipedia statistics page.

In order to verify each hypothesis experimentally, we perform link prediction and unlink prediction experiments. To do that, we need to split each language Wikipedia dataset into a training set, which we use to compute the predictions, and a test set, which we use to evaluate the prediction methods. Let $\mathcal{E}$ be the set of events (link additions and removals) present in one language Wikipedia. We split the set of events into a training and equally-sized true and false test sets $\mathcal{E} = \mathcal{E}_{\text{training}} \cup \mathcal{E}_{\text{true}} \cup \mathcal{E}_{\text{false}}$.

Setup In order to measure the validity of a hypothesis, we aggregate the link prediction and unlink prediction methods suggested by that hypothesis into an ensemble link prediction and ensemble unlink prediction method. These ensemble algorithms are learned using regression methods on the training data alone, as described below, and their performance at the respective prediction task is thus an indication of the validity of their underlying hypothesis.

The big language Wikipedias have existed for over ten years, and thus the test set for each of them spans a range of at least five years. Thus, it is not a good benchmark to predict events at the end of the test set, when known edge additions and removal are almost 5 years in the past. Thus, we split the test set itself into slices spanning equal amounts of time each, and perform an experiment for each slice separately, in which the known events are all events preceding the slice. The number of test slices is different for each network and is chosen to ensure that each slice contains on average one addition and removal event per node. If the

slice size is chosen too small, then the slices are too sparse, i.e., there are no events for most nodes. We choose the number of slices as

$$\#slices = \left\lfloor \frac{|\mathcal{E}_{\text{test}}|}{2 \cdot |V|} \right\rfloor$$

and list the number of test slices for each dataset in Table 4.2. This gives a setup that corresponds more closely to the typical case of a recommender system that needs to predict changes in the network right now. For each slice of the test set, we then compute the accuracy of both link prediction and unlink prediction using the AUC (area under the curve), and finally use a statistical test to aggregate the AUC-values into a statement that tells us whether a single method is better than another one. This allows us to compare each hypothesis to the baseline method, as well as the individual hypotheses with each other.

Ensemble Prediction Methods Given the set of link prediction and unlink prediction methods defined for each model, we build ensemble link prediction and unlink prediction algorithms separately. Each ensemble algorithm is learned by logistic regression on the set of individual measures. In order to learn the regression weights, we again split the training set into a source set and a target set, such that the target set covers the same period of time as the individual test subsets described in the previous section.

If $f_1, f_2, \ldots, f_k$ are the individual prediction functions suggested by a hypothesis, then the ensemble prediction function is given by

$$f_* = L(b + a_1 f_1 + a_2 f_2 + \ldots + a_k f_k),$$

where b and a_i are the parameters of the ensemble method, which are learned by logistic regression, and $L(x) = 1/(1 + e^{-x})$ is the logistic function.

Evaluation Measure To measure the accuracy of a prediction function, we use the area under the curve (AUC), defined as the area under the receiver operating characteristic (ROC) curve [Bradley, 1997]. A random predictor yields an AUC-value of 0.5, a completely wrong predictor an AUC-value of 0 and a perfect predictor yields an AUC-value of 1.

4.5 Evaluation

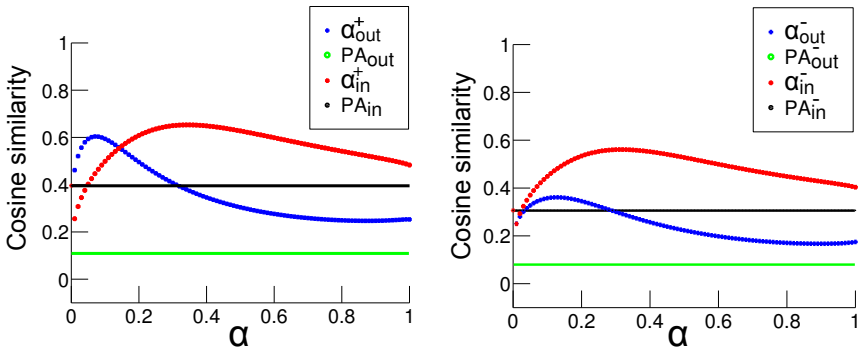
In order to verify each of the four hypotheses, we perform the following experiments. First, we investigate the influence of the exponential smoothing parameter α of the prediction methods behind model M3. In a second experiment, we compute the accuracy of prediction ensembles using methods based on each hypothesis, and compare the accuracies statistically in order to find out which hypothesis is correct, and which hypothesis gives a better model of change in a knowledge network. In a third experiment, we derive and compute an upper bound for prediction methods based on the neighborhood evolution model (M3).

4.5.1 Experiment 1: Fitting the Exponential Smoothing Factor α

In this experiment, we estimate the value of the exponential smoothing factors $\alpha^{\pm}_{out}$ and $\alpha^{\pm}_{in}$ that best predict the actual degrees, where $\alpha^{+}_{out/in}$ is the smoothing factor for weighting add-

degrees and $\alpha^{-}_{out/in}$ is the smoothing factor for remove-degrees.

In order to use the fitted $\alpha^{\pm}_{out}$ and $\alpha^{\pm}_{in}$ in subsequent prediction experiments, we use only the training data $\mathcal{E}_{\text{training}}$ to learn α. To learn the optimal α values, we split the training set into a source set and a target set, such that the target set covers the same period of time as the individual test subsets described in the previous section. Given the add- and remove-degrees of all additions and removals in all slices of the source set, we inserted all $\alpha^{\pm}_{out}, \alpha^{\pm}_{in}$ in the range $(0, 1]$ with a step width of 0.01 into Equation (4.1) to obtain an estimation for the degrees in the target slice. Having computed the degree estimation for overall 100 different values of $\alpha^{\pm}_{out}, \alpha^{\pm}_{in}$, we compute the cosine similarity between the estimated and actual degrees in the target slice. As an example, Figure 4.5 depicts the cosine similarity between the actual degrees in function of each α and their preferential attachment values for the Spanish Wikipedia.



(a) Additions (b) Removals

Figure 4.5: The cosine similarities of the degree estimation by exponential smoothing with varying coefficients $\alpha^{\pm}_{out/in}$ and the preferential attachment estimate is shown for the Spanish Wikipedia.

Results The cosine similarity for each variant reaches a maximum for $\alpha^{\pm}_{out}, \alpha^{\pm}_{out} \in [0.05, 0.3]$, and the optimal exponential smoothing estimation is more similar to the actual degree than its preferential attachment estimates for all degree variations. The small values of the exponential smoothing coefficient α suggest that older information of an article's degree should not be discarded and should be included into an estimate of the future degree. In particular, since all values of α are much smaller than 1.0, information on an article's recent changes is not sufficient to predict its future evolution. Figure 4.6 shows the resulting weights of each slice for the Spanish Wikipedia. For in-degrees, current events play a bigger role than older ones. We observe this effect consistently over all networks. The in-degree of an article can be interpreted as its popularity or importance. Thus, the importance of an article is not a global phenomenon, but can be better predicted by an article's recent evolution. On the other hand, the out-degree for additions and removals should be weighted more equally over time. This implies that an article's out-links are more stable over time.

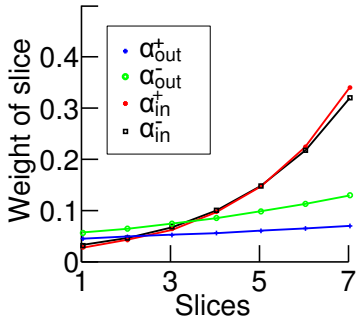


Figure 4.6: The weights of each slice is depicted for each optimal alpha for the Spanish Wikipedia.

4.5.2 Experiment 2: Comparison of Temporal Models

In order to validate each hypothesis, and to compare the hypotheses with each other, we implement the link and unlink prediction methods described in Section 4.3 and evaluate them on the training/test split of each language Wikipedia. We evaluate five ensemble methods using logistic regression of the logarithms of all features in each of the five models. The logarithm of features is used to achieve a multiplicative combination of features, in line with the observation that degree values are often combined multiplicatively to derive new prediction functions, such as the product and ratio of node degrees. Our experiments have shown that a logarithmic combination of features results in a better ensemble prediction for all cases. Since the neighborhood evolution model (M3) does not provide features for the embedding of a link, we also use the number of common neighbors and the number of paths of length three from the time-agnostic model (M0) to make the resulting model comparable with the others.

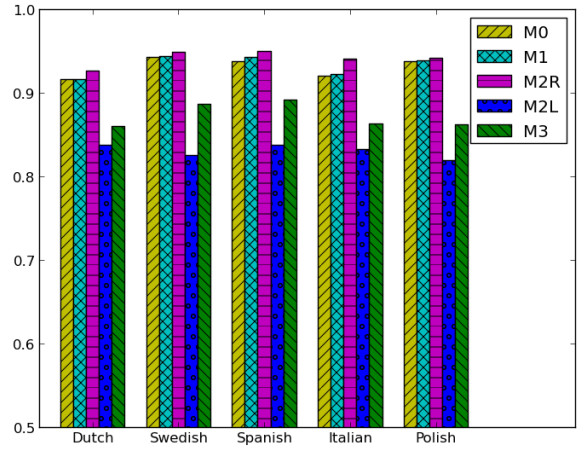
Results Figure 4.7 depicts the AUC-values of all ensemble methods for all language Wikipedias averaged over all test slices of each Wikipedia dataset. The regression weights of each feature are given in Table 4.3 for the Spanish Wikipedia.

Apparently, exploiting temporal information for link prediction does not lead to a big improvement for the prediction. On average, the recency decay model (M2R) performs best but improves the AUC-value on average only from 0.93 to 0.94. Given that the link prediction accuracy is already very high for the time-agnostic model (M0), an increase when exploiting temporal information is still notable.

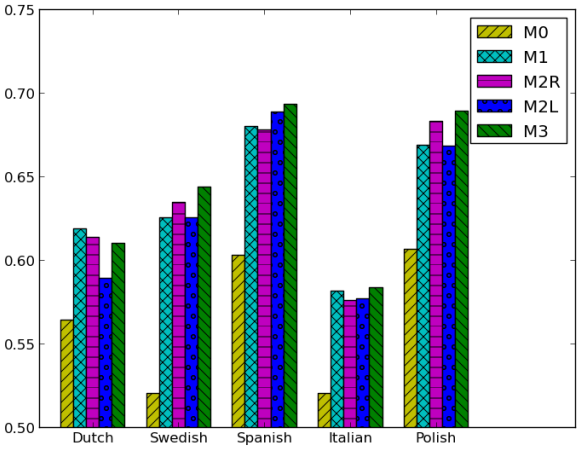
For the unlink prediction problem, a considerable improvement of prediction performance is apparent. Whereas the time-agnostic model (M0) reaches AUC-values between a very small 0.53 and 0.60 for the best-performing network, time-aware methods reach AUC-values between 0.63 and 0.70. These AUC-values for the unlink prediction problem improve the results of previous studies which had an AUC-value of 0.60 [Preusse et al., 2013].

To analyze whether the observed differences in AUC-values between methods are consistent for all test slices, we compute significance values for all differences. Figure 4.8 shows the differences between the AUC-values for link prediction (a) and unlink prediction (b) for each dataset, along with the significance (p-value) of each difference.

Whereas some differences are small, e.g. between the time-agnostic model (M0) and the



(a) Additions

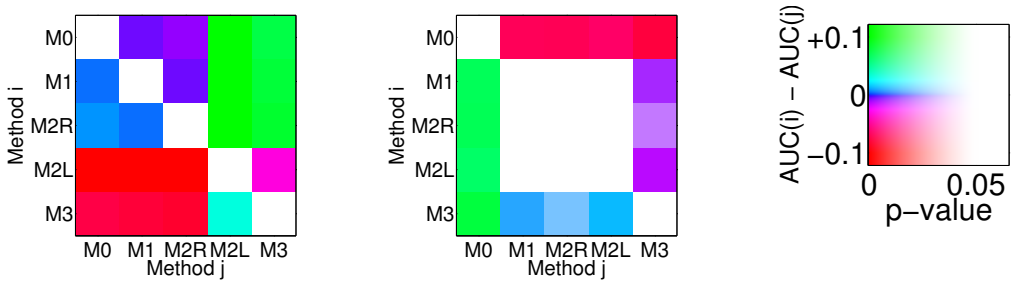


(b) Removals

Figure 4.7: The AUC-values of the ensemble methods on the link prediction and unlink prediction tasks.

qualitative model (M1) for link prediction, they are still significant, i.e., the qualitative model (M1) consistently performs slightly better than the time-agnostic model (M0). For the link prediction problem, the decay recency model (M2R) performs significantly best. The recency of edges is thus a bigger driving factor for the formation of new edges than the longevity.

For the unlink prediction problem, the neighborhood evolution model (M3) is significantly better than all other models. Furthermore, the time-agnostic model (M0) performs significantly worse than any time-aware model. The difference between the decay model (M2) and the qualitative model (M1) are very small and not significant. Notably, even if the temporal information in the qualitative model (M1) is very simple, i.e., it only captures whether an



(a) Link Prediction (b) Unlink Prediction

Figure 4.8: Pairwise comparisons between methods. For each pair of methods, the color of its cell denotes the comparison of both performances. The hue of the color indicates the difference in average AUC-values, and the saturation denotes the significance of the difference. Thus, significant differences are shown in green and red, and no significant differences in white.

edge has been removed or not, the improvement over the time-agnostic model (M0) is already apparent. Exploiting the temporal information further such as by weighting the evolution of an article's neighborhood, improves the prediction significantly but not by much in terms of its AUC-value. Interestingly, the differentiation between recent and longer-lived edge events is not important for unlink prediction. Thus, the information whether an edge has already been removed is sufficient and can only be improved when incorporating the whole neighborhood evolution.

4.5.3 Experiment 3: Upper Bound for the Neighborhood Model (M3)

Even with the best-performing method in our evaluation (M3), the problem of unlink prediction can still not be solved with a comparable AUC-value as link prediction. We may thus ask the question whether structural information is sufficient to predict unlinks after all. To answer this question, we go back to the neighborhood model (M3). What if the actual number of new out- and in-links for link additions and removals were known for all test slices? To find out, we employ one ensemble method that utilizes not estimated but *actual* degrees and analyze whether this will boost the performance of the unlink prediction problem significantly. Thus, we implement a variant of the neighborhood model (M3U) in which the algorithm has access to the actual degrees of all nodes in the test set. We stress that this is not an actual prediction algorithm – it is merely a way of deriving an upper bound on the AUC-values of algorithms that perform preferential attachment with estimated degrees of the nodes.

Results Figure 4.9 shows the AUC-values of the neighborhood model (M3) for each dataset along with the corresponding upper bounds (M3U). As we can see, unlink prediction using structural features could, in principle, attain AUC-values as high as 0.93. Thus, temporal structural information, such as the degree of a node in the current slice, may lead to a big improvement and demonstrates the theoretical feasibility of unlink prediction with structural informations only. If an article's current degree in a slice could be perfectly predicted, then the shown AUC-values could be reached. Thus, the problem of predicting an article's number

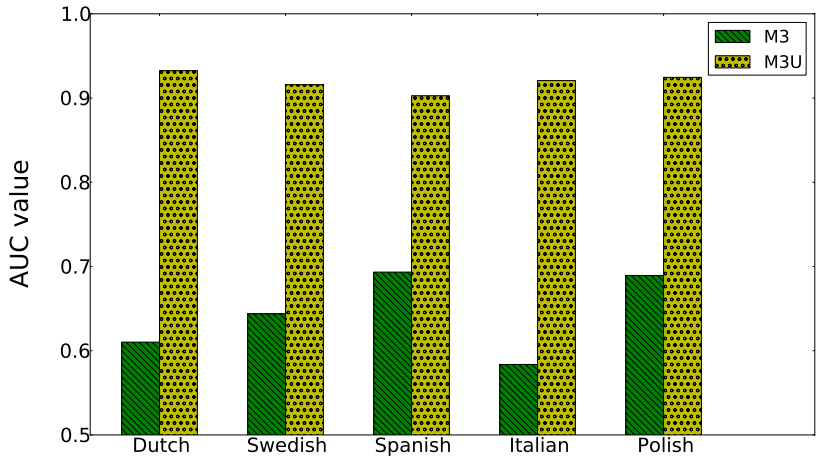


Figure 4.9: The AUC-values of the ensemble neighborhood methods (M3) and the upper bound derived from incorporating ground truth information from the test set into the them (M3U) for the unlink prediction task.

of out- and in-links is an important problem and worthwhile to consider further. If it could be solved better, the predictive performance of unlink prediction is likely to increase as well. At the same time, this also shows the limitation of our exponential smoothing approach, as the AUC-values of our proposed algorithm are far from the theoretical maximum.

4.6 Conclusion

Using temporal information in appearance and disappearance of links, we presented and implemented three models of temporal change. In contrast to the time-agnostic setting, the *qualitative model* captures which links have been removed. The *decay model* exploits the timestamps of state changes and the *neighborhood evolution model* uses the evolution of an article's neighborhood to reason about an article's future. We have shown that temporal information improves the classification of links and unlinks significantly. In particular, data on unlinks should not be discarded, but serves as valuable indicator for new links and unlinks. Further, we have demonstrated the theoretical feasibility of unlink prediction by using the actual neighborhood size as opposed to an estimation for the neighborhood evolution model.

In conclusion, we can state the the incorporation of temporal information increases the accuracy of both link and unlink prediction algorithms, and thus validates our three hypotheses H1-H3. Put in another way, the future evolution of a knowledge network does not only depend on its current state, but also on changes in its past. When exploiting temporal information, predictive performance is improved slightly but consistently in the case of link additions, and significantly for link removals. We presented two temporally weighted versions of the adjacency matrix that can also be applied to any other prediction measure that operates on the adjacency matrix.

For the link prediction problem, we improve the performance of state-of-the-art time-agnostic methods by employing a recency weighting of all edges. The significance experiment shows that the numerical differences are also significant across all datasets.

Problem	Model	$d_{\text{out}}(i)$	$d^{-}_{\text{out}}(i)$	$d_{\text{in}}(i)$	$d^{-}_{\text{in}}(i)$	$\text{CN}(i,j)$	$\text{CN}^{-}(i,j)$	$\text{P3}(i,j)$	$\text{P3}^{-}(i,j)$	Constant
Link	M0	0.1284	–	0.7005	–	1.3976	–	−0.0428	–	0.9712
	M1	−0.0075	0.1624	0.4906	0.2341	1.3323	0.9968	−0.1668	0.2139	3.6664
	M2R	−0.1603	0.2673	0.4057	0.2257	1.4510	1.3411	0.4859	−0.0349	6.0930
	M2L	−0.2284	0.2934	0.3317	0.3074	1.2445	1.3215	0.5466	0.0087	5.7731
	M3	0.1213	0.3612	0.6566	0.2416	1.4083	–	−0.1009	–	2.5656
Unlink	M0	0.2482	–	0.0190	–	−0.1545	–	−0.0736	–	−0.3833
	M1	0.0218	0.1312	−0.4149	0.3793	−0.1419	−0.0176	−0.0171	0.0345	0.7281
	M2R	0.0281	0.1325	−0.4939	0.4113	−0.1525	−0.0020	0.0828	0.0178	0.3849
	M2L	0.0206	0.1581	−0.3653	0.4059	−0.1346	−0.0446	−0.1159	0.0409	0.8775
	M3	0.1345	0.2178	−0.4253	0.5480	−0.1113	–	−0.1722	–	1.6555

Table 4.3: The regression weights of each feature for each model is given for the Spanish Wikipedia. The weighted sum of the logarithms of all features are inserted into the logistic function to compute the ensemble weight of each model. Each model has its specific definition of the given values, cf. Section 4.3. If a feature is not used in a specific model, this is indicated by '–'.

On the unlink prediction side, we improved the state of the art on the unlink prediction task by 0.083 in terms of the AUC-value, corresponding to an increase of eight percentage points. Further, we were able to show that in principle, temporal unlink prediction methods may achieve AUC-values as high as 0.9, justifying the temporal and structural approach. In particular, the prediction of the exact neighborhood size of an article could lead to tremendous improvements for the unlink prediction problem. Even if it is not realistic to expect 100% precision in the degree prediction task, we estimate that reasonable improvements may give AUC-values of up to 80% for the unlink prediction problem.

Not all tested hypothesis were equally valid, and it is not the case that the most complex hypothesis (M3) paint a more complete picture. For the unlink prediction task, M3 does give the best predictive accuracy, but for the link prediction task, the qualitative hypothesis (M1) and the recency decay hypothesis (H2R) seem to give better prediction methods. Comparing the recency-based and the longevity-based hypotheses (H2R and H2L), we note that the recency of in-links to an article is a larger driving factor of link changes than the recency of out-links.

The work in this chapter was published in one paper:

- *Julia Perl, Jérôme Kunegis, and Georg Ruß. If you want my future, don't forget my past: Temporal models of linked knowledge. Technical Report, 2014*

5 Predicting Link Additions and Removals in Social Networks

Whereas in Chapter 3, we have studied the formation and dissolution of links in knowledge networks, this chapter will analyze social theories and characteristics for changes of links states in social networks.

5.1 Introduction

The formation, maintenance and dissolution of social relationships has been widely studied in social networks ranging from married couples to criminal networks and high-school students [Parks, 2007]. Even if it is still not clear whether individuals behave similarly or differently in online and offline networks, the last century of sociological studies has developed several highly-interesting theories and discovered influence factors that are worthwhile to consider for online networks, too. To understand the evolution of a complex social system or a social network we need not only seek to understand the factors that drive the formation of new ties but also the factors that drive the maintenance or dissolution of existing ties and the interplay between them. Whereas a lot of research studied the formation of new links, unlinks have received much less attention, though they account for a high proportion of link changes [Myers and Leskovec, 2014, Preusse et al., 2013].

In this work we present a theory-driven computational approach that allows for exploring the different factors that explain the formation and dissolution of social ties in directed social networks where latent or explicit user groups may exist. We depart from sociological theories, describe how we operationalize these theories via quantifiable measures, and how we assess the utility of these measures within a link and an unlink prediction task. Our work builds upon a great body of previous research which mainly focused on the prediction of links in social networks (see e.g., [Liben-Nowell and Kleinberg, 2003, Lü and Zhou, 2011]). Though, this previous research has shown that simple principles like triadic closure are powerful predictors for link formations, some of these principles conceal the information about the directionality of links which is related to the potential motivation behind the formation of a link in directed social networks. While *attraction* and *support* are fundamental different motivations for the formation of social ties [Parks, 2007], both may have the same observable outcome, namely a closing triad. If user i connects to user j because j is supporting i's friends or members of i's group, that's not the same as if i connects to j because i was anyway already interested in many of j's friends or group members.

In this work we go beyond existing research by exploring theoretically motivated factors that may drive both, the formation and dissolution of social ties and analyze the interplay of these factors.

Consistently, with our results from knowledge networks, cf. Chapter 4, we observe that future changes in the structure of a social network are not only driven by the link network,

but are also largely influenced by past network connections. Especially the unlink prediction greatly benefits from the usage of unlink information.

The contribution of this chapter are twofold: First, we present an overview of social theories that aim to explain the formation and dissolution of social ties in a network; we present a computational approach for quantifying them and demonstrate the utility of our approach within an interesting case study about the evolution of the social network of German politicians on Twitter. Second, we present our empirical results on the impact of different theoretical influence factors on the formation and dissolution of ties in this social network. Though our empirical results are limited to one specific dataset, they clearly show interesting differences in the performance of the measures that operationalize the factors for both tasks.

We use our approach to conduct an empirical case study on the social network of German politicians on Twitter before, during and after the German federal election 2013. We use this dataset for two reasons: First, each politician belongs explicitly to one group – his or her political party – which allows us to compare group-specific and social-network-specific operationalizations of different factors. Second, the election was an external event that triggered many unlink-events and new links.

Research Questions

Concretely we address the following research questions:

RQ 3 *Which structural characteristics predict link formation and dissolution in directed social networks with latent or explicit groups?*

In the related work, some characteristics were shown to be indicative for the formation of a tie, while other characteristics were found to correlate with the dissolution of a tie. One can then asses which influence factors have the highest impact on the prediction of new links and unlinks.

RQ 3.I *Which influence do structural characteristics have on the prediction of new links and unlinks?*

While many datasets provide only a snapshot of the network that does not lent itself to derive unlinks [Kunegis, 2013], a dataset consisting of multiple snapshots can be used to derive links *and* unlinks.

RQ 3.II *What is the added value of unlink data for link and unlink prediction?*

This research question targets the question of how useful this additional unlink information is, i.e., how much the prediction of new links and unlinks is improved when unlink data is exploited.

This chapter is structured as follows: First, we related our work to existing research on link and unlink prediction. In Section 5.3, we present our approach to study the link evolution of directed social networks and the sociological background and theories which build the foundation of it. Next, we describe the prediction tasks which we used to assess the utility and added value of different factors and measures. We present our dataset and empirical results on the evolution of the social network of politicians in Section 5.4 and thereby answer research questions RQ 3.I and RQ 3.II. Finally, we conclude our work in Section 5.5.

5.2 Related Work

The problem of predicting the appearance of links in networks has received substantially more attention than the problems of predicting their removal (see e.g., [Liben-Nowell and Kleinberg, 2003] and [Lü and Zhou, 2011] for a good overview). Examples of well-known and well-performing structural indicators for link prediction are the number of common friends, the number of friends and the ratio of the number of common friends and the two persons' neighborhood sizes. Advanced machine learning algorithms for link prediction include the index of Katz [Katz, 1953], graph kernels [Ito et al., 2005] and diffusion models [Kondor and Lafferty, 2002].

Previous research on the dissolution of social ties has explored individual properties of users and the extent to which these properties correlate with the probability of a user to loose friends. For example, an analysis of the unfriending behavior in Facebook found that friendships which involve neurotic or introverted users and friendships between people who differ greatly in age are more likely to break [Quercia et al., 2012]. Further, the authors found that friendships which are well-embedded in the social network of both users and friendships where both users share a common female friend are more robust. [Incite, 2011] and [Sibona and Walczak, 2011] conducted surveys to reveal the motivation of individuals to dissolve social ties. [Incite, 2011] found the following three top reasons for Facebook users to remove friends: offensive comments (55%), don't know well (41%), trying to sell something (39%). [Sibona and Walczak, 2011] found that people who posted often about unimportant topics and people who often posted about controversial or inappropriate topics were more likely to loose friends [Sibona and Walczak, 2011]. Additionally, they observed that the initiator of a friendship request is unfriended much more often than expected and that it is likelier that the receiver of the friendship request ends the friendship than the other way around.

In contrast to the above mentioned research, we focus on sociological theories which manifest in the structure of a social network and may help to explain the formation and dissolution of social ties rather than studying individual properties of users. The advantage of this structural approach is that it is domain-independent and can thus be applied to any directed social network.

Most similar to our work, [Kwak et al., 2012] and [Kivran-Swaine et al., 2012] explored structural and interaction features of Twitter users to ascertain when users decide to unfollow others. The findings in [Kwak et al., 2012] suggest that ties persist when a user is acknowledged by the relational partner or when the users share followers and followees. The acknowledgment was measured by retweeting information. In [Kivran-Swaine et al., 2012] the authors observed that the more follower a user has, the more likely he or she will resolve an incoming tie. Further reciprocity of follower relationship is correlated positively with the persistence of a tie. Finally, the more common followers and followees two users have and the denser their surrounding network is, the higher the likelihood that the tie between them will persist.

Even if both aforementioned works have studied structural properties that characterize unlinks, they have not evaluated the properties in a prediction set up. Thus, it remains unclear how well these characteristics are suited to *predict* unlinks. This is probably related with the fact that in order to perform a predictive analysis, at least three consecutive snapshots of a network are needed, and in order to exploit past unlinking behavior, even four, as described in Section 5.3.4.

One recent work on structural characteristics of unlinks performed a data analysis for

friendship relationships between 32 freshmen over one year at seven different points in time [Snijders and Steglich, 2013]. This work trains a p* model to characterize unlinks and links in the dataset. Though that approach is very powerful, it does not scale well. The corresponding software package[1] is applicable to networks with 10 to 1,000 nodes and thus cannot even be applied to the dataset used in our study.

To summarize our work goes beyond previous research by presenting a theory-driven structural approach for link and unlink prediction in directed social networks which we evaluate within two predictions tasks: a link and an unlink prediction task.

5.3 Modeling the Formation and Dissolution of Ties

In the following section we present our approach that allows to analyze the formation and dissolution of ties in directed social networks. Firstly we review theories from sociology that aim to explain factors or phenomena that drive the formation or dissolution of ties in social networks. Finally, we describe how we conceptualize these theories in form of quantifiable measures and assess the utility of these measures within a link and an unlink prediction task.

5.3.1 Social Theories

According to the Dunbar number we are only capable of maintaining a certain number of relationships [Dunbar, 1992]. Therefore, we have to decide on a regular base which relationships to maintain and which to dissolve. Many of the studies that have targeted the understanding of personal relationships have focused on individual characteristics of actors rather than on describing actors as embedded in larger social networks. These two approaches correspond to *action theories* and *structuralist theories* [Parks, 2007].

Action Theories: Action theories emphasize the individual variability and choice of each actor in order to explain personal relationships. This theoretical branch assumes that individuals choose whom to interact with according to personal preferences to maximize their personal benefit. The *Social exchange theory* (originally proposed in [Homans, 1958]) is a prominent representative of this category. Social Exchange theory suggests that individuals choose to form the relationship they expect to profit from the most, or to have the lowest cost [Homans, 1958, Garlaschelli and Loffredo, 2004, Emerson, 1976]. According to this theory, individuals will stick to a relationship if they are rewarded and no other relationships provide better opportunities at lower costs.

Structuralist Theories: Structuralists explain individual behavior by the larger social structures that a person is embedded in. They see individual behavior not as the product of personal choice but rather as one's position held in a social network. People with the same position or function are assumed to behave similarly regardless of personal traits.

In this work we define changes of actors or entities in the network based on structuralist theories - i.e., we treat all users in the network the same and seek for general network mechanisms that explain the network evolution rather than focusing on individual differences. We

[1] `http://www.stats.ox.ac.uk/~snijders/siena/`

will use our approach later to explore the linking and unlinking behavior of politicians in Germany. Further, we choose a structural approach because social relations in political networks are very central since power is primarily defined in relational terms. Therefore, almost all political analysts are structuralists according to Knoke [Knoke, 1990].

5.3.2 Formalization

To analyze the dynamics of user relationships, either a dataset with users and their addition and removal events is directly given or the link changes have to be inferred by multiple snapshots of a dataset. In the following, we describe how to derive the formation of new links and unlinks for a dataset with multiple snapshots. Given the data for each snapshot $s = 1, 2, \ldots, n$, we define a network $N_s = (V, E_s)$ which contains a snapshot-*independent* set of users V. The snapshot-*dependent* edges E_s between users represent directed social relationships in each snapshot s

$$(i, j) \in E_s \Leftrightarrow i \text{ is a follower of } j \text{ in snapshot } s.$$

Note that there are no parallel edges in each network, since a user can only follow another user once. The group memberships is modeled as a function m which assigns a set of groups $\mathbf{G}$ to each user $v \in V$

$$m : V \to 2^{\mathbf{G}}.$$

Note that we assume that the group membership of a user remains consistent for all snapshots, thus we define $m(i)$ independent of the snapshot id s. In order to analyze the temporal evolution of social relationships in the given dataset, we need to extract individual additions and removals of edges, which we both call *state changes*. We define the set of state changes $\mathcal{E}$ as

$$\mathcal{E} \subseteq V \times V \times \{+1, -1\} \times \{1, 2, \ldots, 4\},$$

where each state change is either the addition of an edge or the removal of an edge in a snapshot $s \in \{1, 2, \ldots, n\}$. Let $N_0 = (V, \emptyset)$ be the empty network. We then define $\mathcal{E}$ as the union of the set of addition events $\mathcal{E}^+$ and the set of removal events $\mathcal{E}^-$ which are defined as

$$\mathcal{E}^+ = \bigcup_{s=1}^{n} \{(i, j, +1, s) : (i, j) \in E_s \wedge (i, j) \notin E_{s-1}\} \cup$$
$$\mathcal{E}^- = \bigcup_{s=2}^{n} \{(i, j, -1, s) : (i, j) \in E_{s-1} \wedge (i, j) \notin E_s\}.$$

Thus, $\mathcal{E}$ represents all edges as additions that were not present in the network before and all edges as removals that are not present in the current network anymore. Note in particular that edge removals can only be extracted from snapshot 2 onwards.

5.3.3 Conceptualization of Social Theories

In the following section we discuss structural theories mainly lent from [Parks, 2007][2] that aim to explain the life cycle (i.e., the formation, maintenance and dissolution) of social relationships in a social network and we explain how we transform the theories into factor categories (short factors). We categorize the measures in nine factor categories that are depicted in Figure 5.1. The *Activity* factor (out-going link from i in Figure 5.1) describes the general tendency of user i

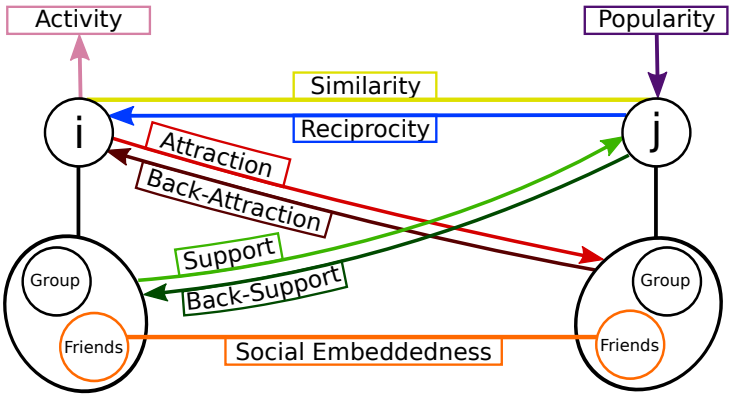


Figure 5.1: A visualization of network factors that can influence the formation of the tie (i, j)

to link/unlink and the *Popularity* factor (incoming link to j in Figure 5.1) depicts the tendency of user j to receive links or unlinks. These two factors are in line with the theory of preferential attachment [Barabási and Albert, 1999, Holland and Leinhardt, 1981] and only depend on one user rather than a user pair. Thus, the activity factor cannot distinguish between the formation of the tie (i, j) and the formation of the tie (i, k) as both ties start from user i. The popularity factor cannot distinguish between the formation of the tie (i, j) and the formation of the tie (k, j) as both ties have user j as target.

The *Similarity* between users i and j is a symmetric factor that depicts how similar the two users are. According to the theory of homophily [Lazarsfeld and Merton, 1954] individuals are likelier to bond if they are more similar and relationships between dissimilar individuals are likelier to dissolve [McPherson et al., 2001]. Based on empirical evidence [Parks, 2007], we further define the *Reciprocity* factor. This factor captures the tendency of individuals to reciprocate relationships, may it be linking or unlinking.

Triadic closure as a consequence of balance theory [Heider, 1958, Holland and Leinhardt, 1981] can be understood as the tendency of individuals to align their preferences with friends or relational partners. *Social Embeddedness* quantifies the embeddedness of a relationship in terms of the relational partner's friends. The more common friends or common friends of friends two users share, the likelier a tie will form [Parks, 2007]. On the other hand, the dissolution of a tie is likelier when the social embedding is low [Parks, 2007]. Social embeddedness is an undirected network concept - i.e., the social embedding of the relationship (i, j) is the same as of (j, i).

[2]Parks introduced the six effect categories of network distance, network overlap, cross-network contact, cross-network density, attraction to partner's network and support from partner's network.

To better explain the directed nature of the dataset at hand, in accordance with [Parks, 2007] we further introduce the concepts of *Attraction* (link from i to network of j) and *Support* (link from network of i to j) to assess the like or dislike for the partners network or respectively the amount of support that members of one's network have for the partner. Additionally, the factors of *Back-Support* (link from j to network of i) and *Back-Attraction* (link from network of j to i) quantify the inverse effects. Since the goal of this work is to quantify structural changes in *directed* social networks, the predictiveness of the back-measures might be radically different from the predictiveness of support and attraction. Only in networks with a very high amount of reciprocal connections, both the attraction or support measure and its respective back-counterpart would be highly correlated. For the latter four categories we use two different types of relationships between users (group-based and friendship-based relations) to quantify the amount of support or attraction. The group of a user i contains all users who are members of the same group as i, whereas the friends of i include all users who have a bidirectional relationship with i.

In the following we discuss each factor in detail and present measures that help to operationalize them. Without loss of generality, all measures below are defined for a user pair (i, j). Note that our network is directed and we aim to predict the formation and termination of directed links. Further note, that some powerful theories such as balance theory or homophily (and the related measures such as common neighbors and Jaccard similarity) are more suitable for undirected networks, since those theories may only indicate that it is likely that a new link will be created or removed between two users i and j but not if i will link to (or unlink from) j or the other way around.

Activity

The activity of the source user may impact the probability that a new link is created from the source user to any other user. The idea is simply that more active users are more likely to create new links, while users that often dissolve links are more likely to dissolve more links in the future. In the literature the activity factor is often also referred to as *productivity* [Holland and Leinhardt, 1981] of an actor or *out-degree activity* [Snijders and Steglich, 2013]. Again, this simple measure explains the change of links based on the structural properties of individual users rather than pairs of users. Thus, the activity factor cannot distinguish between the formation of the tie (i, j) and the formation of the tie (i, k) as both ties start from user i.

We measure the general willingness of user i to form new links by the number of users he linked to in the past (cf. A_+). Analogously, we quantify i's general willingness to resolve relations by the number of users he unlinked in the past (cf. A_-).

Popularity

The popularity factor (often also referred to as preferential attachment [Barabási and Albert, 1999] or attractiveness [Holland and Leinhardt, 1981]) states that the number of new links a user receives is proportional to its current link in-degree. Hence, high in-degree users are more likely to be picked by other users who want to create new links. In our work we make the following analog assumption for unlinks: The probability that a user j will be unlinked is proportional to the number of links which have been dissolved with j in the past (i.e., its unlink in-degree) We define the popularity of j by the number of all incoming links (cf. P_+). Analogously, we define the unpopularity (cf. P_-) by the number of unlinks to j from any user.

Similarity

According to the theory of homophily, user similarity is an important driving factor for the formation of new links and the dissolution of links. Homophily states that individuals are likely to bond with others that are similar to themselves ([Lazarsfeld and Merton, 1954, McPherson et al., 2001]). Consequently we hypothesize that dissimilarities may lead to the dissolution of ties.

We define two measures of similarity: the interest similarity and the perceived similarity of two users and visualize them in Figure 5.2. We take the set of users that a user links to as a proxy for his/her interest; thus the set of shared out-links is an indicator of the *interest similarity* (Sim_{int}). Note that we focus on the link network here rather than the unlink network, because in our dataset unlinks were too sparse for identifying users who were unlinked by the same set of users. The perceived similarity (Sim_{per}) is a measure of how similar other users perceive two users. Thus we take the shared in-link relationships of two users as a measure of their *perceived similarity*. Further, we define two normalizations where the proposed measures are either divided by the sum of incoming links of the two users or by the sum of the outgoing links of the two users (cf. Sim_{perN} and Sim_{intN}). Finally, we define $\text{Sim}_{\text{group}}$ as a binary indicator

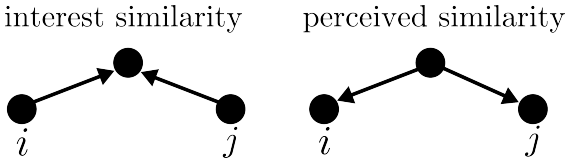


Figure 5.2: The measure for interest and perceived similarity are visualized. The interest similarity (Sim_{int}) measures the common out-links, whereas the perceived similarity (Sim_{per}) measures the common in-links.

of whether user i and j are members in the same group and may therefore be similar.

Reciprocity

Reciprocity describes the tendency of people to form symmetric relationships - i.e., if only an unidirectional relation is present, then it is likely that a new link is established to obtain a bidirectional relationship. On the other hand, unidirectional unlinks of a bidirectional link may indicate that the other direction should be resolved as well [Parks, 2007]. This factor is also in line with the balance theory [Heider, 1958, Holland and Leinhardt, 1981] which states that people tend to align their preferences with others.

We define two binary measures of reciprocity. One simply indicates if j links to i (R_+) and one indicates if j has unlinked i (R_-) .

Social Embeddedness

Several studies show that relationships are unlikelier to resolve when the cross-density of two partners' networks is high - i.e., if the partner's networks are well-connected [Parks, 2007, Milardo, 1987]. This indicates that the "social support" of the neighbors of the two users is important to determine if they will connect or disconnect. Theories such as the triadic closure [Simmel, 1950, Heider, 1958] state that a high network overlap between two actors

should increase the likelihood of the two actors getting connected. We use the network overlap between two users and the cross-density to estimate the social embeddedness of a tie between two users. Note that these measures are symmetrical and only help to predict that a link should be created or removed between two users i and j. However, one cannot infer the actual direction - i.e., whether i will link to j or the other way around.

To quantify the network overlap of two users i and j we define the following three measures on the friend network of i and j: the number of common neighbors (NCn), the Jaccard coefficient(NJa) and the Adamic-Adar characteristic (NAd) [Liben-Nowell and Kleinberg, 2003].

$$\mathrm{NCn}(i,j) = |CN(i,j)|, \tag{5.1}$$

$$\mathrm{NJa}(i,j) = \frac{\mathrm{NCn}(i,j)}{|\{k : k \in N(i) \vee k \in N(j)\}|}, \tag{5.2}$$

$$\mathrm{NAd}(i,j) = \sum_{\{k \in CN(i,j)\}} \frac{1}{\log |\{l : \{l,k\} \in E^F\}|}, \tag{5.3}$$

where $N(i)$ is defined as the set of all users that have reciprocal relationships with i and $CN(i,j)$ is the set of common neighbors of i and j. Note that we use the friendship graph - i.e., the reciprocal connections that both users have formed, as suggested by [Parks, 2007].

The cross-network density characterizes the amount of linkage between friends of user i and friends of user j. Thus, this characteristic counts the number of links that form paths of length three ($P3$) between i and j. The different kinds of paths of length three that are depicted in Figure 5.3:

- *Cross-network* links (CNe) between i's exclusive network and j's exclusive network, (depicted in red in Figure 5.3) indicate how friends of the users are clustered.

- *Cross-common-neighbor* links (CCn) between i's and j's common neighbors (depicted in green in Figure 5.3) reveal the amount of connectedness between the common friends. The more connections among the common neighbors, the more the common neighbors form one big cluster.

- *Cross-asymmetric* links (CAs) from common neighbors to either partner's exclusive network, (depicted in blue in Figure 5.3) indicate the number of asymmetric common friends which have connection to the network of only one of the two users.

Since the three defined link types provide interesting insights into the cross-network connections between two users, we consider all three link types and the classic $P3$ measure. For all cross-link types, we define the density - i.e., the number of actual links of the four types

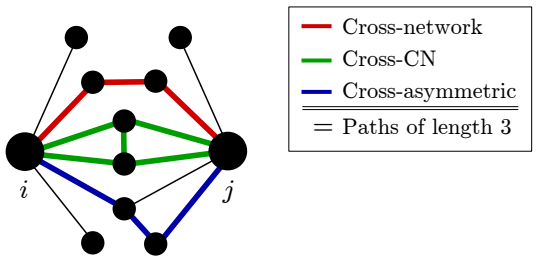


Figure 5.3: The three different types of paths of length three are depicted for a small toy network.

divided by the number of possible links, as follows.

$$\mathrm{CP3}(i,j) = \frac{|\{\{k,l\} \in E^F : \{i,k\},\{l,j\} \in E^F\}|}{|N(i)| \cdot |N(j)|} \tag{5.4}$$

$$\mathrm{CNe}(i,j) = \frac{|\{\{k,l\} \in E^F : \{i,k\},\{l,j\} \in E^F \wedge k,l \notin CN(i,j)\}|}{(|N(i)| - |CN(i,j)|) \cdot (|N(j)| - |CN(i,j)|)} \tag{5.5}$$

$$\mathrm{CCn}(i,j) = \frac{|\{\{k,l\} \in E^F : k,l \in CN(i,j)\}|}{\binom{|CN(i,j)|}{2}} \tag{5.6}$$

$$\mathrm{CAs}(i,j) = \frac{|\{\{k,l\} \in E^F : k \in CN(i,j) \wedge l \notin CN(i,j)\}|}{|CN(i,j)| \cdot (|N(i) \cup N(j)| - |CN(i,j)|)}. \tag{5.7}$$

The divisor of each measure is given by the number of possible cross-links for each of the four different link types. Similarly as for the network overlap, we cannot exploit reciprocated unlinks, to quantify the extent of cross-unlinkage.

Attraction

Attraction to the partner's network captures the amount of positive or negative affinity from user i to the network of j. Social science research suggests that disliking a person that is important to the partner, carries some potential for relationship dissolution [Cleek and Pearson, 1985]. Attraction is an asymmetric relation and can be positive (if i linked to many friends or members of j's group) or negative (if i unlinked many friends or members of j's group).

To quantify the positive or negative affinity of user i to the network of j, we use the link and unlink behavior of i as a proxy. We define the positive affinity of i to j by counting the number of j's friends who i links to ($\mathrm{Att}_{\mathrm{F}+}$) and by counting the number of members of j's group who i links to ($\mathrm{Att}_{\mathrm{G}+}$). Analogously, we measure the negative affinity between i and j by counting the number of j's friends who were unlinked by i ($\mathrm{Att}_{\mathrm{F}-}$) and by counting the number of members of j's group who were unlinked by i ($\mathrm{Att}_{\mathrm{G}-}$).

The back-attraction is the amount of attraction that i receives back from the friends or members of j's group. Therefore, it is the reciprocal effect of attraction. To quantify the back attraction we count the number of j's friends that link to or unlinked i ($\mathrm{BAtt}_{\mathrm{F}+}$ or $\mathrm{BAtt}_{\mathrm{F}-}$) and the the number of members of j's group that link to or unlinked i ($\mathrm{BAtt}_{\mathrm{G}+}$ or $\mathrm{BAtt}_{\mathrm{G}-}$).

Support

Support is defined as the amount of positive or negative affinity from user i's friends or members of j's group. Therefore, it characterizes the attitude of i's network towards j. Social science research suggests that relationships are likelier to be established with well-supported persons and that a missing support can lead to the dissolution of the relationship [Wellman et al., 1997]. Support is an asymmetric relation and can be positive (if e.g. many friends or members of i's group link to j) or negative (if e.g. many friends or members of i's group unlink j). We assume that link-relationships are an indicator for supporting behavior and unlinks are an indicator for nonsupporting behavior. We define the support user j receives from the perspective of user i as the number of i's friends who link to j (cf. $\text{Sup}_\text{F}+$) and the number of members of i's group who link to j (cf. $\text{Sup}_\text{G}+$). Analogously, we define the nonsupport user j receives from the perspective of user i as the number of i's friends who unlink j (cf. $\text{Sup}_\text{F}-$) and the number of members of i's group who unlink j (cf. $\text{Sup}_\text{G}-$).

The back-support is the amount of support that j gives back to the friends and members of i's group. Therefore, it is the reciprocal effect of support. To quantify the back-support, we count the number of i's friends that link to j or are unlinked by j ($\text{BSup}_{\text{F}+}$ or $\text{BSup}_{\text{F}-}$) and the number of members of i's group that are linked to i or unlinked by i ($\text{BSup}_{\text{G}+}$ or $\text{BSup}_{\text{G}-}$).

5.3.4 Prediction Methodology

In order to assess the utility of the defined measures, we perform link and unlink prediction experiments. The general methodology for prediction problems is as follows: given node pairs in the training set, the aim is to predict node pairs in the true test set against node pairs in the false test set. For the link prediction problem the new links that do appear (true test set) must be distinguished from node pairs that do not appear (false test set). In the unlink prediction scenario, the true test set contains links that are removed and the false test set consists of links that remain in the network. A good link or unlink prediction measure assigns higher values to node pairs in the true test set than to node pairs in the false test set.

We perform the link and unlink prediction experiment on network data consisting of 4 consecutive snapshots ($s = 4$). Note that our methodology can also be adapted to a dataset that consists of more than four snapshots. Figure 5.4 summarizes the three steps of our methodology. We split the dataset into training and test sets as follows: the training sets for both prediction problems consist of all edge events that are present at time $s = 3$. The true test set for link prediction consists of all links that are not present in the third snapshot and are present in the fourth snapshot. The corresponding false test set contains non-links, i.e. node pairs that are neither connected in the training set nor in the true test set. Note that we use a sample of non-links of the same size as the true test set which we randomly selected from the largest connected component of the network.

For unlink prediction, the true test set consists of all links that are present in the third snapshot and not present in the fourth snapshot. Conversely, the false test set contains a random sample of links that are present in the third and fourth snapshot. This random sample has again the same size as the corresponding true test set.

Assessment of Single Measures and Combinations of Measures: We assess the utility of individual measures and the combination of several measures by comparing the AUC-values

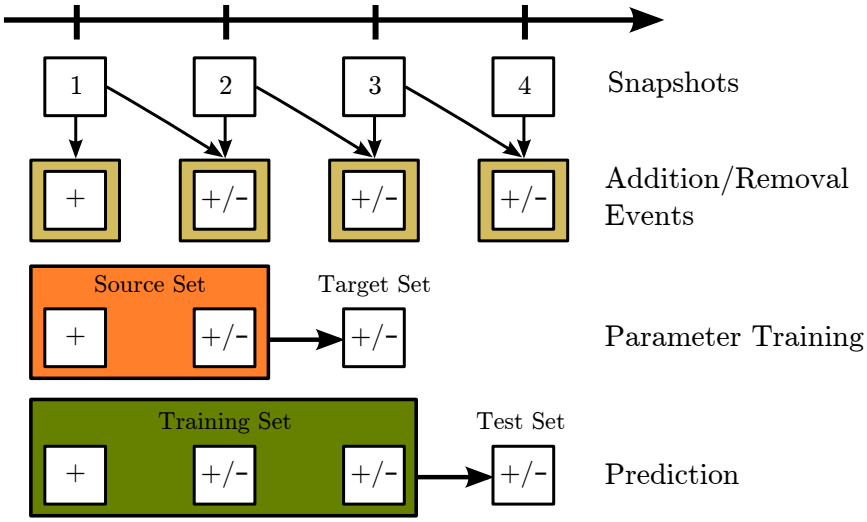


Figure 5.4: The three stages of our methodology are depicted. First, add and remove events are extracted from the snapshots. Second, the parameters of the logistic regression are trained. Third, the prediction is performed.

of the prediction functions including the corresponding measure(s). The AUC-value is defined as the Area Under the ROC-Curve and is ranged between 0 and 1, where 1 indicates a perfect prediction; a random baseline model would lead to an AUC-value of 0.5 [Bradley, 1997].

To combine several prediction measures, we train a logistic regression with multiple prediction measures. We perform the following greedy approach to select the best combination of different measures for the logistic regression. For each prediction problem, we first choose the single best performing measures. Iteratively, we test the combination with the greedy selected measure(s) to asses which additional measure leads to the largest performance gain. This measure selection procedure is performed until we have chosen *twenty* measures.

To train the parameters of the logistic regression, we split the dataset again into a source and a true and false target set. The source set consists of all edge events in the first two snapshots and the true target set contains the user pairs that are added in the third snapshot for link prediction and the user pairs that are removed in the third snapshot for unlink prediction. The respective false target sets are formed by random-non links for link prediction and remaining links for unlink prediction. Note that for our proposed methodology at least four snapshots are required to perform parameter training and prediction. The source set must consist of two snapshots, because the parameters need to be trained with unlink measures which are only observable after the second snapshot.

The true and false target set for the link prediction consist each of 814 node pairs; the true and false test set contain 749 node pairs. Thus, there are fewer new links in the fourth snapshot than in the third snapshot. For unlink prediction, each target set contains 551 links and each test set contains 552 unlinks. Hence, the number of unlinking events is stable for the third and fourth snapshot.

5.4 Empirical Study

We will use a Twitter dataset of German politicians from before, during and after the election to empirically evaluate our approach. First, we describe the Twitter follower network in general and the selected Twitter dataset in particular. Second, we discuss the empirical results we obtained when applying our approach to the dataset.

5.4.1 The Twitter Follower Network

Twitter is a microblogging service that is well-known for the fast propagation of news [Osborne et al., 2012]. It's Alexa-rank is 9 [3] and it has been extensively analyzed for its content, structure and structural changes. Research on Twitter content and interactions includes diverse topics such as real-world event identification [Becker et al., 2011], study of conversational practices [Boyd et al., 2010, Wu et al., 2011, Lietz et al., 2014], political polarization during the US election in 2010 [Conover et al., 2011], or driving factor for user's retweet behavior [Suh et al., 2010, Naveed et al., 2011].

Everyone can create a Twitter account and post *tweets* which can be up to 140 characters long and thus give Twitter the name of a *micro*blogging service. Registered users on Twitter can interact with each other by three means: they can *follow* other users, they can *retweet* the tweets of other users, or they can *mention* a user in a post. The follow relationship is not necessarily symmetric, user a can follower user b but b does not follow a. Then, a is said to be a follower of b and b is the followee of a. This relationship is illustrated in Figure 5.5. Users

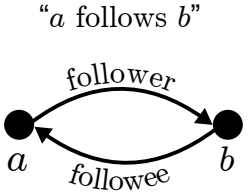


Figure 5.5: Visualization of the follower relationship from a to b.

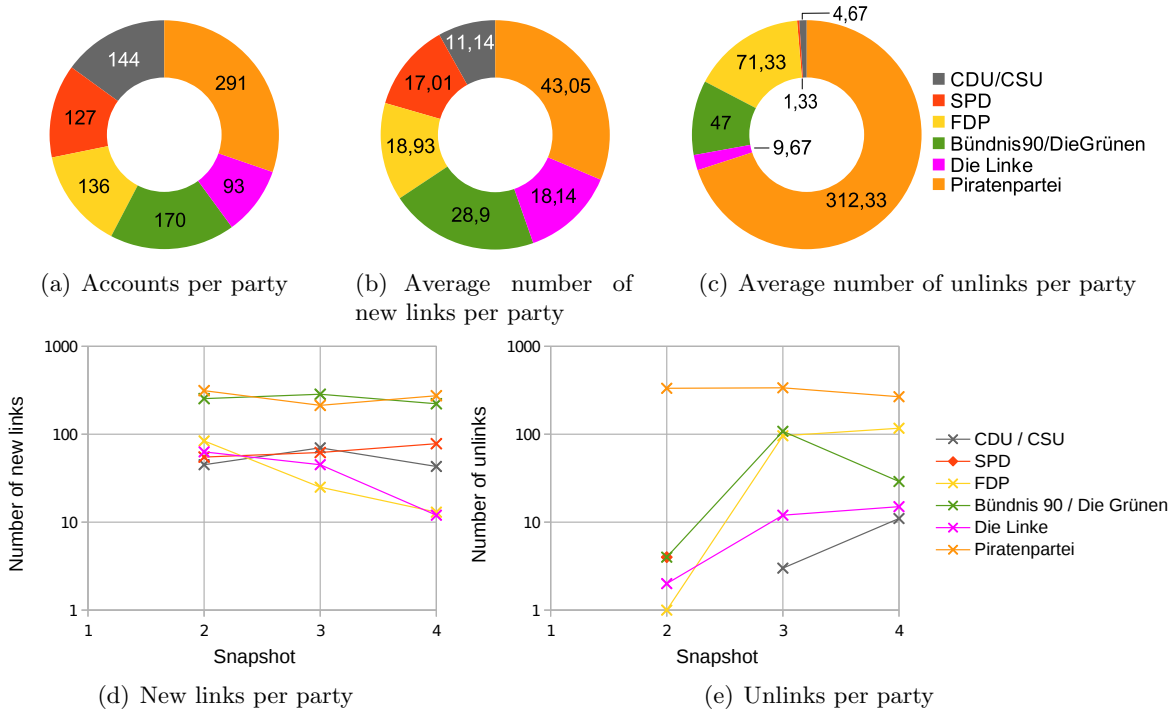
can also *unfollow* other users to not follow their latest tweets anymore. The underlying social network of Twitter consists of follower relationships and is, in contrast to many online social networks, not symmetric.

5.4.2 Dataset

GESIS[4] provides a Twitter dataset of the network of German politicians from before and after the German federal election in 2013 which was recently published in [Lietz et al., 2014]. The dataset consists of four snapshots that were taken in monthly intervals. Two snapshots were taken from before the election and two capture the network of politicians after the election. The snapshots were obtained by crawling all followers of 961 German candidates for the election at four timepoints [Kaczmirek et al., 2013]. Although *unlinks* and *links* are not explicitly captured in the dataset, they can be easily derived when comparing neighboring snapshots. Our dataset

[3] `http://www.alexa.com/siteinfo/twitter.com` on June, 2nd 2014
[4] http://gesis.org

contains Twitter users from six different political parties, see Figure 5.6(a). The Piratenpartei is the largest fraction and also the most active one followed by Bündnis90/DieGrünen. 77.17%



(a) Accounts per party

(b) Average number of new links per party

(c) Average number of unlinks per party

(d) New links per party

(e) Unlinks per party

Figure 5.6: The number of politicians per party that have a Twitter account, the average number of new links and unlinks per party and the temporal evolution of the number of new links and unlinks per party. The election was between the second and the third snapshot. One can see that the Piratenpartei is the largest party on Twitter. Bündnis90/DieGrünen and the Piratenpartei are also the most dynamic parties which create and resolve most links. From the dissolution of links over time one can easily see who were the big looser of this election - the FDP and Bündnis90/DieGrünen. Interestingly, the election triggered many unlinks, but did not (or only slightly) impact the formation of new links.

of the overall 2,799 new links in the dataset are between politicians of the same party. Figure 5.6(d) gives and overview of how the party-internal links are distributed per party. The link formation frequency of each party per snapshot is displayed in Figure 5.6(d). For Die Piratenpartei and Bündnis90/DieGrünen the number of new links per snapshot remains constantly high with around 200–300 new links per snapshot. On the other hand, the link formation activity for Die Linke and FDP goes down after the election (between snapshot ID 2 and 3).

The number of unlinks per party and snapshot are displayed in Figure 5.6(e). All parties except Die Piratenpartei show an increase in the number of unlinks after the election; in particular FDP and Bündnis90/DieGrünen. Interestingly, the election seems to have greater influence on the unlinking behavior of politicians than on their linking behavior. For the FDP and Die Piratenpartei, the number of unlinks even exceeds the number of new links after the election. It is also interesting to note that not only most links are created within a party but

#Unlinks	%	Name	Party
163	10.181%	tarzun	Piratenpartei
105	6.558%	Markus Kompa	Piratenpartei
73	4.560%	Vincent Thenhart	Piratenpartei
62	3.873%	Lars F. Lindemann	FDP
41	2.561%	Martin Lindner	FDP
36	2.249%	Dr.Toni Hofreiter	Bündnis90 / Die Grünen
36	2.249%	Le55ing	Piratenpartei
28	1.749%	René Rottmann	Piratenpartei
25	1.562%	Peter Meiwald	Bündnis90 / Die Grünen
22	1.374%	Miriam Seyffarth	Piratenpartei
591	36.914%		

Table 5.1: The ten politicians that received the most unlinks are given along with their political association, the number of unlinks they received and the relative number of unlinks with respect to all unlinks in the test set.

also most links are removed within a party (83.64% of the overall 1,601 unlinks in the dataset are party-internal).

Table 5.1 shows the ten politicians that received the most unlinks. The ten politicians receive 36.914% of all unlinks that overall all 961 politicians received.

5.4.3 Experiment 1: Predictive Performance of individual Measures

In the first experiment, we evaluate the performance of individual measures for the link and unlink prediction problem. All AUC-values are given in Table 5.2. Note that we have defined a measure to have a *positive* (marked by P) influence on the prediction if its AUC-value is bigger than 0.6; the influence is defined as *negative* (marked by N) if the AUC-value is smaller than 0.4. If the AUC-value is between 0.4 and 0.6 no influence can be identified since the measure performs similar to what we would expect from a random guesser which would have an AUC of 0.5.

Our results clearly show that most measures are useful for link predictions and can outperform a random baseline. However, for unlink predictions only few measures turn out to be useful and those measures tend to be based on unfollow information rather than on follow information. This indicates, that (1) information about past links are essential for predicting future unlink events while their benefit for predicting future links is marginal and (2) unlink prediction is a much more difficult problem than link prediction, partly because of how the false test set is constructed. While one chooses a random set of non-links as false test set in the link prediction task, one has to choose a random set of nodes which remain connected as false test set in the unlink prediction task. Differentiating between an actual new link and a completely random and unconnected node pair is much easier than differentiating between two pairs of nodes which were both connected in the past and where one link is removed while the other one remains. In other words, explaining why someone who works for party A in city X is not friend with a random politician of another party B at the other end of the country is easier than explaining why that person terminated his friendship with one of his party fellows in his city but not with the other one. In the following, we describe the best-performing measures

Category	Measure	Description	Link	influ.	Unlink	influ.
Activity	A_+	(links from i to any politician)	0.723	P	0.569	–
	A_-	(unlinks from i to any politician)	0.684	P	**0.705**	P
Popularity	P_+	(links to j from politicians)	0.746	P	0.422	–
	P_-	(unlinks to j from politicians)	0.614	P	0.500	–
Similarity	Sim_{int}	(users that i and j follows)	**0.907**	P	0.467	–
	Sim_{per}	(users that are followers of i and j)	0.796	P	**0.382**	N
	Sim_{group}	(binary: Are i and j in the same party?)	0.862	P	0.506	–
	Sim_{intN}	(normalized interest similarity)	**0.913**	P	0.457	–
	Sim_{perN}	(normalized perceived similarity)	0.783	P	0.436	–
Reciprocity	R_+	(Does j follow i?)	0.582	–	0.497	–
	R_-	(Has j unfollowed i?)	0.582	–	0.489	–
Social Embeddedness	NCn	(number of common neighbors)	0.825	P	0.432	–
	NJa	(CN weighted by neighborhood union)	0.828	P	0.447	–
	NAd	(CN weighted by degree of common neighbors)	0.827	P	0.435	–
	CP3	(linkage between neighborhoods)	0.800	P	0.416	–
	CNe	(linkage between non-shared neighborhood)	0.719	P	0.449	–
	CCn	(linkage between common neighborhood)	0.798	P	0.428	–
	CAs	(linkage between asymmetric neighborhood)	0.822	P	0.415	–
Attraction	Att_{G+}	(links from i to members of j's party)	**0.906**	P	0.560	–
	Att_{F+}	(links from i to j's friends)	0.837	P	0.449	–
	Att_{G-}	(unlinks from i to members of j's party)	0.743	P	**0.669**	P
	Att_{F-}	(unlinks from i to j's friends)	0.539	–	**0.613**	P
Support	Sup_{G+}	(links from i's party to j)	**0.917**	P	0.483	–
	Sup_{F+}	(links from i's friends to j)	0.847	P	**0.410**	N
	Sup_{G-}	(unlinks from i's party to j)	0.673	P	0.552	–
	Sup_{F-}	(unlinks from i's friends to j)	0.605	P	0.501	–
Back-Attraction	$BAtt_{G+}$	(members of j's party that follow i)	0.890	P	0.515	–
	$BAtt_{F+}$	(friends of j that follow i)	**0.897**	P	0.449	–
	$BAtt_{G-}$	(members of j's party that unfollow i)	0.640	P	0.584	–
	$BAtt_{F-}$	(friends of j that unfollow i)	0.584	–	0.524	–
Back-Support	$BSup_{G+}$	(members of i's party that j follows)	0.887	P	0.505	–
	$BSup_{F+}$	(friends of i that j follows)	0.893	P	0.470	–
	$BSup_{G-}$	(members of i's party that j unfollows)	0.726	P	0.572	–
	$BSup_{F-}$	(friends of i that j unfollows)	0.646	P	0.480	–

Table 5.2: The AUC-values of individual measure defined in Section 5.3.3 are given for the link and unlink prediction problem. The description of all measures relates to a tie (i, j) for which the likelihood to be added or removed should be characterized. The five highest AUC-values for each prediction problem are written in bold. The P symbol indicates that the predictive influence of the characteristic is positive, N indicates a negative influence and '–' indicates that the measure has no influence for the prediction.

for both problems in detail.

Link Prediction The number of links from members of i's party to j (Sup_{P+}), is the best-performing measure (AUC = 0.917) to predict new links (i, j). This means, i's probability of creating a tie with j increases with the number of i's party colleagues who established a tie with j. This suggests, that the follow behavior of individual politicians is in line with the follow behavior of their party. Since most new relations (85.71%) are created within a party,

we can conclude that politicians who are popular within the party are also more likely to be followed. If social ties are created across parties, individual politicians follow those politicians which are also supported by other members of their party.

The interest similarity (Sim_{int}) and the normalized similarity (Sim_{intN}) perform very well with AUC-values of 0.907 and 0.913 respectively. This indicates that users who are interested in the same users are more likely to follow each other. It is interesting to note that the perceived similarity and normalized perceived similarity perform much worse, while they are superior for predicting unlinks. This indicates, that the link formation behavior of politicians is more in line with what the people they observe do (i.e., their friends, party colleagues or users they follow) than their unlinking behavior. One potential explanation for this observations is social influence and group conformity - i.e., politicians potentially adapt their link creation behavior in order to fit within a group. An alternative explanation is homophily. That means, politicians potentially select friends or party colleagues who have similar interests and therefore create in part links with the same people. It is interesting to note that the unlink behavior of politicians does not show any evidence of the presence of group conformity or homophily. This suggests that social influence plays a smaller role in decisions about the dissolution of social ties than in decisions about the formation of social ties. However, further experiments are necessary to quantify the effect of social influence and group conformity on the social tie formation and dissolution behavior of people, since our observational data does not allow to encapsulate platform-specific effects such as friend recommendations which definitely impact the data we observe.

The attraction from i to members of j's party ($\text{Att}_{\text{P}+}$) performs also very well with an AUC-value of 0.906. This indicates that i's probability of establishing a relation with j increases with the number of links that i has already established to members of j's party. This indicates, that users are persistent in their cross-party-linking behavior. They either continue establishing links within their party or if they establish cross-party links then they continue focusing on the same parties as their party colleagues. The fifth best measure with an AUC-value of 0.897 is the Back Attraction Friend measure ($\text{BAtt}_{\text{F}+}$) which is defined as the number of friends of j that follow i. Around half of the links are reciprocal (51.12%). That means, that Back-Attraction and Attraction do not necessarily suggest the same links and unlinks. Hence, the attraction that i receives back from j's friends is a better indicator for the formation of new links than its counterpart, the attraction of i to j's friends (AUC = 0.837).

Unlink Prediction The number of politicians that i has unfollowed (A_-), is the best-performing measure (AUC = 0.705) to predict the dissolution of links (i, j). Politicians that have unlinked many politicians in the past are more likely to dissolve further links. Since this measures is independent of j, it cannot be used to predict which link i will dissolve, but rather expresses a general tendency of user i to unfollow other users. However, this also implies that all structural measures that take both users into consideration perform worse than the activity measures of individual nodes. This suggests, that no single structural power pattern (such as triadic closure in the link prediction task) exists for the unlink prediction task. Interestingly, the dissolution of ties can be better explained by user i dropping many relationships than by user j being unlinked by many users, as the AUC-value of A_- is 0.5 for the unlink prediction problem. In general, in-degree measures of j are bad predictors for unlinks. This also becomes apparent for the in- and out-degree unlinking distributions. Even though both distributions are skewed, the out-degree distribution appears to be more skewed. This means, while there are few users

who produce most unlinks, the amount of unlinks which individual users receive is more evenly distributed. This again confirms our observations that the unlinking behavior of users is less driven by social influence and group conformity since those factors would lead to situations where some users are unlinked by the majority of people.

The second and third best measures are attraction measures; the number of unlinks from i to members of j's party ($\text{Att}_{\text{P}-}$) performs well with an AUC-value of 0.669. Hence, when user i has already unfollowed many users from the same party as j, i is also likely to drop the follower relation to j. Further, the number of unlinks from i to friends of j ($\text{Att}_{\text{F}-}$) is also among the most predictive unlink measures with an AUC-value of 0.613. Thus, while the unfollow behavior of users seems to be driven by individual decisions, their decisions are far from being random since users show a persistent unlinking strategy.

As discussed before, contrarily to the link prediction task, the perceived similarity of two users (Sim_{per}) is ranked among the top five indicators of the unlink prediction task. An AUC-value of 0.382 states that the lower the number of shared followers, the likelier an unlink will occur; thus minus the number of shared followers achieves an AUC-value of 0.618.

5.4.4 Experiment 2: Predictive Performance of Combinations of Measures

In Experiment 1, we measured the predictive performance of *single* measures. The goal of this experiment is to find the best subset of measures for each task. We use a greedy approach to select the best combination of measures – i.e., we start with the best measure and extend this set by adding the measure which increases the AUC-value most.

	Link Prediction			Unlink Prediction		
Pos	**AUC**	**Meas.**		**AUC**	**Meas.**	
1	0.915	$\text{Sup}_{\text{G}+}$	links from i's party to j	0.710	A_{-}	unlinks from i to any politician
2	0.943	$\text{Att}_{\text{G}+}$	links from i to members of j's party	0.716	$\text{Sup}_{\text{F}+}$	links from i's friends to j
3	0.950	Sim_{intN}	normalized interest similarity	0.725	P_{+}	links from politicians to j
4	0.952	R_{+}	Does j follow i?	0.751	$\text{Att}_{\text{F-}}$	unlinks from i to j's friends
5	0.959	$\text{Sup}_{\text{F}+}$	links from i's friends to j	0.756	$\text{Sup}_{\text{G-}}$	unlinks from i's party to j
6	0.961	$\text{Att}_{\text{G-}}$	unlinks from i to members of j's party	0.760	R_{-}	Has j unfollowed i?
7	0.963	CNe	linkage between non-shared neighborhood	0.764	R_{+}	Does j follow i?
8	0.965	Sim_{per}	users that are followers of i and j	0.765	$\text{Att}_{\text{G}+}$	links from i to members of j's party
9	0.967	CCn	linkage between common neighborhoods	0.768	$\text{BSup}_{\text{F}+}$	friends of i that j follows
10	0.967	Sim_{int}	users that i and j follow	0.769	Sim_{perN}	normalized perceived similarity

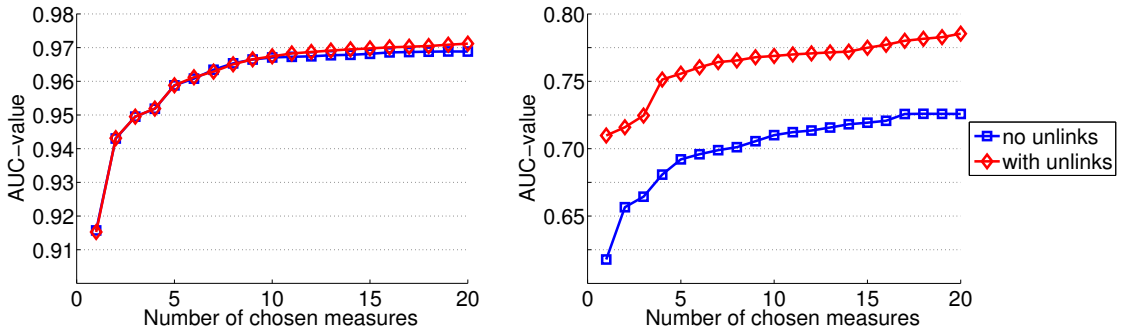
Table 5.3: The top ten measures in the selected subset for each prediction problem are given along with the respective AUC-values.

Table 5.3 shows the top ten measures and the corresponding AUC-values that can be achieved for each prediction problem. One can see that for the link prediction task the best performance that can be achieved when selecting a subset of 20 measures is 0.971, while for the unlink prediction task the best performance using a subset of 20 measures is 0.790. This shows that also for the unlink prediction task we can clearly outperform a random baseline (AUC=0.5). Our results clearly demonstrate that the unlink prediction task is much more difficult than the link prediction task. This can in part be explained by the fact that most research about the evolution of networks focused on the formation of new ties rather than the dissolution. Consequently, powerful structural patterns (such as triadic closure) have been

discovered for link networks and they work well for the link prediction task, but similarly strong structural patterns for the unlink prediction task are missing. However our results show that novel measures which are based on the unfollow-network rather than the follow network can help to address this problem and allow achieving a good performance. Nevertheless, the performance gap between link and unlink prediction still seems to be profound. But when comparing these two prediction tasks one also needs to take into account how the false test set is constructed (cf. Section 5.4.3). This explains why it is not fair to directly compare the performance of link and unlink predictions.

5.4.5 Experiment 3: Added value of unfollow information

In Experiment 3 we aim to quantify the added value of unlink information for the link and the unlink prediction task. Therefore, we ask the question of how well future changes can be predicted using measures based on information about currently existing links compared to measures that exploit information about past links (i.e., links which previously existed but were removed). Our results in Experiment 2 suggest that measures based on information about past links are especially useful for the unlink prediction task, where unlinks should also be trained on unlinking data[5]. As in Experiment 2 we select the best subset of measures for the



(a) Link Prediction (b) Unlink Prediction

Figure 5.7: The added value of unlinks for link and unlink prediction. A value pair of (X,Y) corresponds to the AUC-value (Y) of the top X measures. One can see that unlink measures help to increase the performance drastically in the unlink prediction task, but only have a marginal effect in the link prediction task.

link and the unlink prediction task. However, we change the base set from which the measures are selected as follows: once we only use measures based on currently existing links in the base set and once we also include measures based on information about past links. The results of the greedy measure selection are displayed in Figure 5.7. A value pair in this figure corresponds to the the AUC-values (y-axis) of the top N measures (x-axis). Our results show that for the link prediction task the added value of information about past links is marginal. However, for the unlink prediction task the performance can be significantly improved when including measures based on information about past links. This nicely shows that the future evolution of the network is not only influenced by present links, but also by past links.

[5]The alternative would be to train the classifier on link data only and assume that links for which we predict low ranks will be removed.

5.5 Conclusion

In this work we have analyzed theoretically-motivated factors to explain the formation and dissolution of ties in directed social networks. We departed from relational sociology and developed a computational approach that conceptualizes different factors and theories to quantify the likelihood of the creation and dissolution of social ties. Our approach exploits the structural information of social networks and allows to incorporate additional relations between users which might emerge when users form explicit or implicit groups. We tested the predictive performance of different factors by defining two prediction tasks: an unlink and a link prediction task. Finally, we demonstrated the utility of our approach in an empirical case study using consecutive snapshots of an interesting social network of German politicians on Twitter.

Consistently, with our results from knowledge networks [Preusse et al., 2013], we observe that future link changes are not only driven by the link network, but are also largely influenced by past network connections. Especially, the unlink prediction greatly benefits from the usage of unlink information.

Our empirical results clearly show that the follow behavior of individual politicians is in line with the follow behavior of their party. A politician is likelier to create a tie with j when more of his party colleagues also established a tie with j. This indicates, the usefulness of our approach to exploits group membership information such as the party affiliation especially for the prediction of the formation of new social ties.

We further find that while the same factor may drive the formation of links and unlinks, there are interesting differences in how the measures that operationalize these factors contribute to both tasks. For example, our results indicate that for predicting if a link between two politicians i and j will be established in the future, one should focus on the interest similarity between those politicians (i.e., how similar are the people they follow) while for predicting if they will unlink focusing on the similarity between users who follow them is more promising. This indicates, that the link formation behavior of politicians is more in line with what the people they observe do (i.e., their friends or users they follow) than their unlinking behavior. One potential explanation for this observations is social influence and group conformity - i.e., politicians potentially adapt their link creation behavior in order to fit within a group. An alternative explanation is homophily. That means, politicians potentially select friends or party colleagues who have similar interests and therefore create in part links with the same people. It is interesting to note that the unlink behavior of politicians does not show any evidence for the presence of social influence, group conformity and homophily. The unfollow behavior of politicians is not driven by what the people who they form a group with do, but seems to be more driven by individual decisions.

However, one needs to note that one potential explanation for that is that online social networks like Twitter influence the formation of new ties via friend recommendations which are based on the social ties that friends of a user created in the past. On the contrary, information about the removal of social ties is concealed. Therefore, users cannot easily observe who was unfollowed by most of his/her friends or group members, but they will likely see who was followed by many of their social contacts. The impact of the platform which generates the data is always a limiting factor of observational studies like ours and experiments are required in order to answer the question of whether the link and unlink behavior of users would still be driven by different factors if they would be handled in the same way by the platform.

Though our empirical results are limited to one specific platform and the biases which are introduced by the platform, our approach is general and can be applied to other observational and experimentally generated data depicting the evolution of social networks.

Though, our results suggest that unlink predictions are much more difficult than link predictions (amongst others due to the differences in the evaluation setup as we extensively discussed in Section 5.4.3), our results show that our approach allows to decrease the performance gap between link and unlink predictions when including novel measures which are based on the unlink-network. Finally, we hope that our approach helps to enhance our understanding about the hidden factors that drive the formation of links. Simple principles like triadic closure are powerful predictors for the link prediction task but conceal the information about the directionality of links which is related to potential theoretical explanations behind the formation (and also the dissolution) of new links.

Discussion

Different Paths of length three In former experiments, we have only measured the number of paths of length three and did not distinguish between the different kinds of paths that are visualized in Figure 5.3. For the Twitter data set we have computed all four paths of length three measures. In particular, there is no notable difference in the predictive performance of CCn, CNe and CP3. However, CAs performed slightly better than the other three measures for the link prediction problem. For the unlink prediction problem, the four measures perform very similar, thus for this data set and prediction problem, one doesn't need to compute all four paths measures and can only compute the classic path of length three feature.

Party-specific Predictors As [Lietz et al., 2014] discovered, the observed political parties behave very differently. Instead of training a joint predictor for all parties, one could train a single predictor for *each* party. In this research, we were not interested in deriving party-specific characteristics, but in general patterns that explain the formation and dissolution of ties. The dataset at hand can also not be used to train classifiers for each party, since there are too few data points for all but the pirate party.

Comparison with Link Changes in Knowledge Networks The symmetry for link changes in social networks is a bigger driving factor than for link changes in knowledge networks. This is in accordance with the intuition that knowledge is organized more hierarchical, whereas balance and thus symmetry is more important for the formation and dissolution of ties in social networks [Parks, 2007, Heider, 1958].

The work in this chapter was published in one paper:

- *Julia Perl, Claudia Wagner, Jérôme Kunegis, and Steffen Staab. A theory-driven approach for link and unlink predictions in directed social networks. Technical Report, 2014.*

6 Latent Negative Links in Social Networks

6.1 Introduction

Whereas chapters 3–5 have dealt with two transformations of non-edges, namely the additions of new links and the removal of links, this chapter will study the existence of particular non-edges. Online social networks allow people to connect with each other, forming a network. In most online social networks, only positive links between people are allowed such as *friendship*, *trust* and the *following* relationship. Relationships between people may also be of a negative type, for instance enmity as opposed to friendship, and distrust as opposed to trust. A very small number of online social networks actually do allow such negative links. Among them is Slashdot, a technology news website that lets its users tag other users as *friends* and *foes*, as well as the product review site Epinions that allows users to *trust* and *distrust* each other. In both cases, the negative link feature results in directed signed links between users that can be interpreted as approval and disapproval links, and that are used in the user interface of the two websites to decide which content is shown to users.

On Slashdot, the posts of users tagged as foes are given a lower score, and may thus be hidden. On Epinions, the trust and distrust information is used to determine the reviews shown, using an undisclosed algorithm. The negative links are thus used on both sites to enhance the site's content, and a *negative link* feature could similarly enhance he content shown on many websites.

Negative Links in Social Networks Social networks provide their users with a variety of functionality for connecting with other users. Examples of these features are friends on Facebook, circles on Google+ and followers on Twitter. Explicitly created connections in social networks can be displayed in the user profile and some users might want to boost their status by collecting as many visible contacts as possible. Besides consequences for the status of these users, these explicit social connections deeply influence the user experience within the social networking platform and the ability to interact with other users. The nature of an explicit link between two users is therefore dependent on its platform-specific implementation.

Here, we limit our investigation to links between users that are intended to more be permanent and therefore describe a long-lasting connection. This excludes links between users and other entities that form bipartite networks, e.g., ratings of movies, articles, comments, etc. Ratings of persons in dating sites [Kunegis et al., 2012] fall in this category too, since the rating and rated users have different roles. The same holds for one-time events such as elections, e.g., the elections of administrators in Wikipedia [Leskovec et al., 2010a].

Permanent social links between two users can be divided into two types according to their functionality, that can be described as positive and negative. It can be observed that large social networks such as Facebook and Google+ provide positively connotated linking functionality called *friend*, *contact*, or multiple *circles* with user-defined labels. These links are the defining concept for social networks, and they are crucial for them since they determine the visibility

of user-generated content for the creator and for potential readers. It is this functionality that makes the platform social since the user is supported in his interaction with selected other users. In the following, we define links that increase the visibility of users and content or which increase the ability to interact as positive links. Consequently, the links that decrease visibility of content or which decrease the ability to interact are called negative links. Negative links are associated with disapproval for another user. Labels for explicit negative links in social networks are for instance *enemy*, *foe*, *distrust*, *ignore*, *hide* and *block*. As negative aspects of a community are rarely advertised, these negative links are much less used and known than positive links. This might be one reason why only few social networks with negative links are publicly available for study and research.

Since many online social networks are however reluctant to implement a negative link feature, as shown by the very small number of sites featuring them, the question arises whether negative links have an added value for the network or whether their purpose can be replaced by a prediction algorithm that determines the negative social links automatically from the known, positive links. Such an algorithm could be applied to any online social network that does not want to allow explicit negative links, and would increase the accuracy of news streams, content filters and recommender systems embedded in these online social networking sites. However, two available social networks that contain positive and negative links are Slashdot and Epinions.

Research Questions

Based on these premises, this chapter investigates the following research question: *Can the negative links allowed in Slashdot and Epinions be inferred from the positive links only?* In particular, we study the following research questions:

RQ 4 Which structural characteristics are indicative for latent negative links in social networks?

For the first scenario, we assume that only the positive links, e.g. all friendships in a network, are given. The goal of this research is to find characteristic patterns for negative links in the network consisting of only positive relationships.

RQ 4.I *Which structural indicators infer negative links from only positive links?*

Some networks do not allow the user to label relationships as negative. Therefore we ask what the added value of the negative link feature for the prediction of negative links is. For that we compare two settings: How much easier is it to predict latent negative ties, when some negative information is used in contrast to the sole usage of only positive ties.

RQ 4.II *What is the added value of the negative link feature?*

The predictive performance of the two prediction settings will be compared to obtain the added value of the negative link feature.

This chapter is structured as follows. First, we review the related work on negative links in social networks in Section 6.2. To tackle our main research question, we will then introduce the *latent negative* link prediction problem and functions that solve the problem in Section 6.3. In Section 6.4, we present the evaluation methodology for the prediction problem. We will

evaluate the proposed prediction functions using Slashdot and Epinions data in Section 6.5 to answer research questions RQ 4.I and RQ 4.II. Finally, we conclude this chapter in Section 6.6.

6.2 Related Work

For the work of this chapter, we briefly review the relevant related work in the area of *link prediction* and *link sign prediction*.

Link Prediction Measures A major model of network analysis is preferential attachment, i.e., the rule that new edges are more likely to be attached to nodes with high degree [Barabási and Albert, 1999]. Another important model is that of a high clustering coefficient, i.e., the rule that typical networks contain a much higher number of triangles than predicted by a random graph model, and thus edges tend to connect nodes that have a high number of common neighbors [Watts and Strogatz, 1998]. A high clustering coefficient is one component of the *small-world* network model, and can be generalized to signed graphs to give balance theory, stating that triangles are likely to be balanced, i.e., to contain an even number of negative edges [Harary, 1953, Heider, 1958].

The preferential attachment model can be used to derive link prediction functions based on node-based centrality measures, such as the degree of nodes and PageRank [Brin and Page, 1998], whereas the clustering model leads to link prediction functions that compare the neighborhood of two nodes, such as the number of common neighbors and the cosine similarity.

Link Sign Prediction In the case where negative edges are allowed in a network, the problem of predicting the sign of new edges, given the known positive and negative edges is called the link sign prediction problem, and has been extensively studied [Kunegis et al., 2009, Leskovec et al., 2010b, Leskovec et al., 2010c]. In the link sign prediction problem, the known network contains both positive and negative edges, and thus sign information can be used for prediction. For instance, the multiplication rule lent from Balance Theory [Heider, 1958, Harary, 1953] stating that *the enemy of my enemy is my friend* can be used [Kunegis et al., 2009]. Leskovec et al. compare the number of triangles that are explained by balance theory and *Status Theory* to develop a better understanding of the underlying mechanisms that cause positive and negative links [Leskovec et al., 2010b]. Status theory states that a positive link from i to j indicates that j has a higher status than i and a negative link (i, j) indicates that j has a lower status than i

In a study on Epinions, user and interaction features (e.g. who replied to whom, who commented to which post, which comments are competing) were trained to predict trust among Epinions users. Hence, no trust or distrust information was used to predict the actual trust labels. A model of trust propagation that incorporates trust and distrust information is presented by [Guha et al., 2004]. These types of methods can however not be applied in the problem studied here, since in our case only positive edges are known.

A related problem is that of predicting the sign of new links, given both positive and negative links in a network [Yang et al., 2012]. In addition to the network itself, the method described in that work uses interaction information to achieve its prediction, as well as a small

sample of signed edges. Thus, the method cannot be applied to our scenario, since we assume no negative links are possible in the network.

6.3 Modeling Latent Negative Links

6.3.1 Slashdot & Epinions

Slashdot is a technology news platform where users can post and read other users' news articles and comments [Kunegis et al., 2009]. On Slashdot, users can create two types of explicit and directed social links between themselves and other users. These are labeled *friend* and *foe*. Both link types allow the user to change the visibility of the content the linked user has created. Although the effect of a link is not predetermined but user configurable the convention is that the *friend* link increases the content visibility, the *foe* link decreases content visibility of the target user. Therefore the *friend* link is a positive link, while the *foe* link is a negative link. The *friend* and *foe* link types are also called *fan* and *freak* from the point of view of the targeted user. The signed social network of Slashdot is called the *Slashdot Zoo* on Slashdot itself, and can be considered an extension to Slashdot's sophisticated moderation system [Lampe and Resnick, 2004].

Epinions is a website that collects community-created product reviews [Massa and Avesani, 2005]. Two types of links can be created by one user to a target user. One link is labeled *trust* the other link is labeled *block* (or formerly, *distrust*). These links influence the visibility of product reviews that are authored by the target user. The user who has created the *trust* link sees the reviews of the trusted user at a higher position in the list of all relevant reviews. Therefore this link is considered to be a positive link. Reviews by users that are *blocked* are not presented to the user, making it a negative link. The positive and negative links on Epinions are also used to predict a global trust score for individual users.

6.3.2 Definitions

Let $N = (V, E, w)$ be a social network (Slashdot or Epinions) with V the set of users, E the set of directed links between users, and $w : E \to \pm 1$ the edge sign function, with $w((i, j)) = +1$ denoting that user i approves of user j and $w((i, j)) = -1$ denoting that user i disapproves of user j. The fact that two nodes $i, j \in V$ are connected (in either direction) will be denoted by $i \sim j$, and the fact that i and j are connected by a directed edge (i, j) by $i \to j$. The degree of vertex $i \in V$, i.e., the number of vertices connected to i (in either direction) will be written as $d(i)$. The out-degree of node i, i.e., the number of nodes pointed to by i is denoted as $d_{out}(i)$.

At the task of ordinary link prediction, in which future links must be predicted from current links, both node-based and neighborhood-based measures are used. A link prediction function is defined to take as input a node pair (i, j), and returns a numerical score indicating how likely a new edge is to appear between i and j.

6.3.3 Link Prediction Functions

Link prediction functions can be divided into neighborhood-based and centrality-based functions, based on whether they include only vertex-based features of vertex-pair-based features. In the following, we list the link prediction functions used in our experiments, which correspond to the most common general link prediction functions used in the literature, and can

be found for instance in [Liben-Nowell and Kleinberg, 2003] and [Zhang et al., 2012]. The two nodes for which a link prediction score is to be computed will be called i and j.

Neighborhood-based Functions

These link prediction functions are based on comparing the neighboring nodes of i and j. In addition to the cosine similarity defined in Equation (6.5), we use the following neighborhood-based link prediction functions.

The number of common neighbors between i and j is defined as

$$f_{\mathrm{CN}}(i,j) = |\{k \mid i \sim k \wedge k \sim j\}|, \tag{6.1}$$

which equals the number of paths of length two between i and j.

Analogously, the number of paths of length three between i and j is defined as

$$f_{\mathrm{P3}}(i,j) = |\{(k,l) \mid i \sim k \wedge k \sim l \wedge l \sim j\}|, \tag{6.2}$$

where the sequence (i,k,l,j) forms a path of length three from node i to node j.

The Jaccard coefficient measures the amount of common neighbors divided by the number of neighbors of either vertex [Liben-Nowell and Kleinberg, 2003]:

$$f_{\mathrm{Jacc}}(i,j) = \frac{|\{k \mid k \sim i \wedge k \sim j\}|}{|\{k \mid k \sim i \vee k \sim j\}|} \tag{6.3}$$

The measure of Adamic and Adar counts the numbers of common neighbors, weighted by the inverse logarithm of each neighbor k's degree [Adamic and Adar, 2001]:

$$f_{\mathrm{Adad}}(i,j) = \sum_{k \sim i \wedge k \sim j} \frac{1}{\log(d(k))} \tag{6.4}$$

The cosine similarity is defined as the cosine between the two adjacency vectors of i and j, where the adjacency vector of a vertex is the 0/1 vertex-vector indicating to which vertices a given vertex is connected. The cosine similarity can be expressed in the following manner

$$f_{\cos}(i,j) = \frac{|\{k \mid i \sim k \wedge k \sim j\}|}{\sqrt{d(i)d(j)}}. \tag{6.5}$$

This measure thus weights the number of common neighbors that two nodes share by the weighted degree of both nodes.

The final two common proximity-based link prediction methods are graph kernels. They can be either defined as functions of the adjacency matrix A of the network, or as sums over all paths from i to j. The symmetric adjacency matrix A of the graph $G = (V,E)$ is defined as the $|V| \times |V|$ 0/1 matrix defined using $A_{ij} = 1$ when $i \sim j$ and $A_{ij} = 0$ otherwise. Both graph kernels have a parameter α, which we set to the value $0.85/\|A\|_2$, i.e., slightly less than the inverted spectral norm of the adjacency matrix.

The exponential graph kernel is defined as the exponential function of the adjacency ma-

trix [Kondor and Lafferty, 2002]

$$f_{\mathrm{EXP}}(i,j) = \left[e^{\alpha A}\right]_{ij} = \sum_{p \in P_*(i,j)} \frac{\alpha^{|p|}}{|p|!}. \tag{6.6}$$

The Neumann graph kernel is defined using matrix inversion [Kandola et al., 2002]

$$f_{\mathrm{NEU}}(i,j) = \left[(\mathbf{I} - \alpha A)^{-1}\right]_{ij} = \sum_{p \in P_*(i,j)} \alpha^{|p|}. \tag{6.7}$$

The basic idea behind these two measures is that longer paths between the node pair i and j should be weighted lower than shorter paths, because the closer neighborhood of two nodes is more indicative for the formation of a tie than nodes that are only reached via several hops. These expressions make use of the notation $P_*(i,j)$ for the (generally infinite) set of all paths in the network from node i to node j, and of the notation $|p|$ for the length of a path $p \in P_*(i,j)$.

Node-based Centrality Functions

Centrality-based link prediction functions are defined as products of centrality measures of the two vertices i and j; different choices of centrality measures lead to different link prediction functions. In addition to the PageRank product defined in Equation (6.9), we use the preferential attachment value.

The preferential attachment model states that the likelihood of a new node i to connect to node j is proportional to the degree of node j [Barabási and Albert, 1999]. Thus, the preferential attachment score is defined as

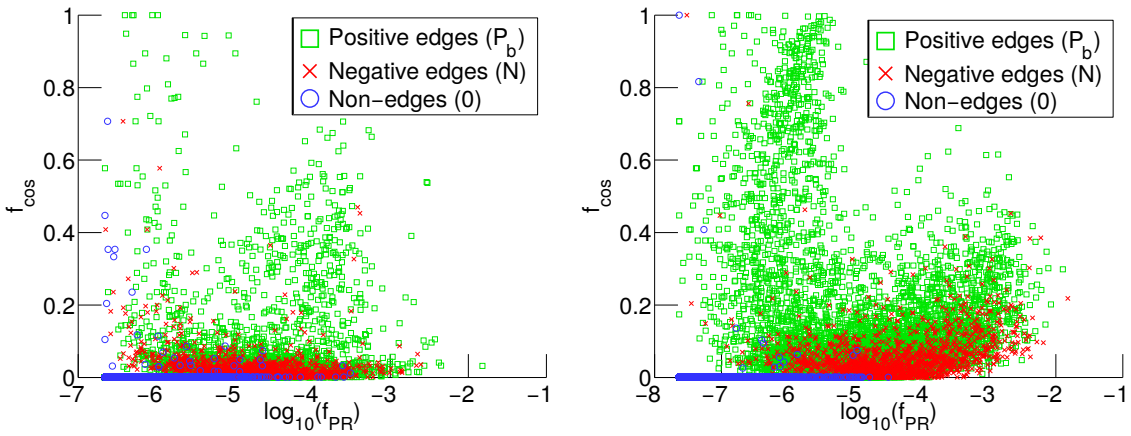
$$f_{\mathrm{PA}}(i,j) = d(i)d(j). \tag{6.8}$$

PageRank [Brin and Page, 1998] is a centrality measure in a directed network defined as the solution $\mathrm{PR}(i), i \in V$ of

$$\mathrm{PR}(i) = \frac{1-\alpha}{n} + \alpha \sum_{j \to i} \frac{\mathrm{PR}(j)}{d_{out}(j)}$$

where α is a parameter set to 0.85 [Langville and Meyer, 2006]. The general idea behind the PageRank measure is that the more incoming connections a node receives, the more popular it is which is then reflected in a high PageRank value. However, it depends on which nodes link to an article. If popular nodes link to a node i, then the popularity of i will be higher than the popularity of a node that only unpopular nodes connect to. The PageRank values are all positive by construction. The PageRank product link prediction function is then defined as the product of the two nodes' PageRanks

$$f_{\mathrm{PR}}(i,j) = \mathrm{PR}(i)\mathrm{PR}(j). \tag{6.9}$$



(a) Slashdot (b) Epinions

Figure 6.1: Scatter plots of the cosine similarity and the PageRank product with points colored according to their inclusion in the set of unknown positive edges P_b, the set of unknown negative edges N and the set of non-edges O.

6.3.4 Initial Analysis

Let P be the set of positive edges and N the set of negative edges, i.e., $E = P \cup N$ and $P \cap N = \emptyset$. To perform an initial analysis of the datasets, we split[1] the set of positive edges P randomly into two sets P_a and P_b such that $|P_\mathrm{a}| = 3|P_\mathrm{b}|$. We then consider P_a the set of known edges (all positive), P_b the set of unknown positive edges, N the set of negative edges to predict, and finally a randomly sampled set O of node pairs not in E with size $|O| = |P_\mathrm{b}|$.

We can now compute the PageRank product and the cosine similarity for all node pairs in the sets N, P_b and O, based on the known edges P_a. Figure 6.1 shows the scatter plot of the nodes pairs of the three unknown sets plotted in function of their PageRank product and cosine similarity values.

Two observations can be made:

- Most node pairs in the non-edge set O have a cosine similarity of zero, and a small value of the PageRank product.
- Node pairs in the positive edge set P_b have high cosine similarity and high PageRank product values (compared to non-edges).
- Node pairs in the negative edge set N have low but mostly nonzero cosine similarity values and high PageRank product values (compared to non-edges).

These observations are true for both the Slashdot and Epinions datasets. We conclude that negative edges can be identified from the cosine similarity and PageRank product in the following way:

- Negative edges connect nodes with high PageRank product values.

[1]The factor 3 is motivated by the fact that in the link prediction literature, the standard size of the test set is 25% of the total set of edges.

Table 6.1: The two signed social network datasets used in our evaluation. In both networks, all edges are directed.

Dataset	Vertices	Edges (pos. + neg.)
Slashdot Zoo [Kunegis et al., 2009]	79,120	515,581 (392,326 + 123,255)
Epinions trust [Massa and Avesani, 2005]	131,828	841,372 (717,667 + 123,705)

- Negative edges connect nodes with low but nonzero cosine similarity values.

Thus, we expect a combination of a positively weighted centrality measure with a negated neighborhood-based measure to solve our problem of predicting negative links, giving a combined prediction measure that takes into account both preferential attachment and balance theory.

6.4 Methodology

Datasets

Table 6.1 summarizes the two datasets. Both datasets are available on the Koblenz Network Collection site `http://konect.uni-koblenz.de`. Both networks have both positive and negative links between users, forming a directed, asymmetric signed network. Although the functionality that lies behind the link types is not fully identical between Slashdot and Epinions, it is very similar according to our definition of positive and negative links. Based on this similar functionality we assume similar properties of the two networks, and will use both datasets for our experiments.

Prediction Methodology

In this section, we describe our method for testing whether the negative links of a signed social network can be predicted from its positive links. We will review the link prediction problem itself, give suitable link prediction functions adapted to the problem at hand, and will describe two experiments, one for measuring the achievable predictive performance of the prediction problem, and one for computing an upper bound on that accuracy.

As defined in the previous section, the set of edges E can be divided into the set of positive edges P and the set of negative edges N. The problem can then be rephrased as the problem of evaluating whether the negative links N can be predicted from the positive links P. The general methodology we introduce for this kind of problem consists in predicting links of one type using only links of another type in the network. This problem extends the ordinary link prediction problem in which only a single link type is present.

The general methodology for link prediction is as follows. Given node pairs in the training set, predict node pairs in the true test set against node pairs in the false test set. We formalize the general prediction problem $\mathcal{P}$ as

$$\mathcal{P} : \text{Training Set} \rightarrow \text{True Test Set} \mid \text{False Test Set},$$

Table 6.2: The features used for learning a link prediction function.

Feature	Name	Ref.
$f_1 = \log(\text{CN})$	Common neighbors	Eq. (6.1)
$f_2 = \log(\text{P3})$	Paths of length three	Eq. (6.2)
$f_3 = \log(\cos)$	Cosine similarity	Eq. (6.5)
$f_4 = \log(\text{Jacc})$	Jaccard coefficient	Eq. (6.3)
$f_5 = \log(\text{Adad})$	Adamic–Adar	Eq. (6.4)
$f_6 = \log(\text{Exp})$	Exponential kernel	Eq. (6.6)
$f_7 = \log(\text{Neu})$	Neumann kernel	Eq. (6.7)
$f_8 = \log(\text{PA})$	Preferential attachment	Eq. (6.8)
$f_9 = \log(\text{PR})$	PageRank product	Eq. (6.9)
$f_{10} = \begin{cases} \log(\text{PR}) & \text{if } \cos = 0, \\ \min(\log(\text{PR})) & \text{otherwise.} \end{cases}$	Conditional PageRank	–

and search for prediction functions f that assign node pairs in the true test set higher values than node pairs in the false test set.

The result of a prediction function will be called the prediction score. Multiple prediction functions can then be compared to find a function that solves the prediction problem to a satisfying accuracy with the AUC-value, which is defined as the area under the ROC-curve. The AUC-value if ranged between 0 and 1, where 1 indicates that all node pairs in the true test set are ranked better than any node pair in the false test set and 0 indicates the opposite ranking. The AUC-value of a random predictor which produces a random ranking of node pairs in the true and false test set is 0.5.

Ensemble Link Prediction Functions

As shown in Section 6.3, neither centrality-based link prediction functions such as the PageRank product, nor neighborhood-based functions such as the cosine similarity are expected to predict negative links from positive links well. Instead, combinations of them are needed. Therefore, we propose a method for combining centrality-based and proximity-based link prediction functions into an ensemble.

To combine several link prediction functions, we use logistic regression applied to the logarithms of individual prediction functions. Some functions such as the number of common neighbors f_{CN} may be zero, and thus their logarithm is not defined; in this case we use the logarithm of the lowest possible value instead. Also, since the behavior of the PageRank product is different when the cosine similarity is exactly zero (as illustrated in Figure 6.1), we include as a feature the PageRank product multiplied by the indicator function of the cosine similarity being zero. We call this feature the conditional PageRank. The main rationality for introducing this feature is that negative edges can hardly be distinguished from non-edges and positive edges. Since most of the non-edges have a cosine value of 0, the PageRank feature appears to be a good characteristic to distinguish between positive and negative edges. If the cosine does not equal zero, the cosine value itself differentiates well enough between positive and negative links. Table 6.2 summarizes all features used in the evaluation.

We propose two ensemble link prediction functions, based on the basic link prediction

Table 6.3: The regression link prediction functions used in our evaluation.

Regression	Used features
f_{all}	$f_1, f_2, f_3, f_4, f_5, f_6, f_7, f_8, f_9$
$f_{\text{PR-cos}}$	f_3, f_{10}

functions of Table 6.2. Since the basic features $f_1, \ldots, f_9$ correlate among each other (for instance, the Adamic–Adar measure and the common neighbor count have Pearson correlation $p = 99\%$ for the Slashdot dataset), we restrict regression to the five best-performing individual link prediction functions.

- Logistic regression based on the five logarithmic features $f_1, \ldots, f_5$.
- Logistic regression based on the conditional PageRank f_{10} and the cosine $f_3 = \log(\cos)$.

The two logistic regression-based functions must be trained on the training set.

Given a set of features $f_1, f_2, \ldots, f_k$ then the ensemble prediction function is given by

$$f_* = L(b + a_1 f_1 + a_2 f_2 + \ldots + a_k f_k),$$

where b and a_i are the parameters of the ensemble method, which are learned by logistic regression, and $L(x) = \frac{1}{(1+e^{-x})}$ is the logistic function.

Table 6.3 shows the two regression features.

The ensemble link prediction methods are only used for Experiment 1, as using them in Experiment 2 to derive an upper bound on link prediction accuracy will skew the results.

6.5 Evaluation

In the following, we describe two experiments to measure how well the negative links can be predicted from the positive links in a signed social network. The purpose of the first experiment is to find good link prediction functions at that task, and to compute their accuracy. The second experiment consists in comparing this link prediction problem to the task of predicting negative links in networks where both positive and negative links are known. Since this task includes more information in the training set (i.e., negative links), the achieved accuracy of that problem is higher and gives an upper bound on the accuracy that can realistically be attained at the problem of predicting negative links when only positive links are known.

6.5.1 Experiment 1: Latent Negative Prediction

The goal of this experiment is to measure the accuracy of link prediction functions at the task of predicting negative links in social networks containing only positive links, and to observe which particular functions are well suited for that task.

In our scenario, we want to predict negative links from known positive links. Since we want to compare the scores of link predictions functions applied to node pairs connected by a negative link with the scores of node pairs that are unconnected or connected by a positive

Table 6.4: Learned weights of logistic regression. Weights marked as (–) denote functions that are not used in the respective regression type.

Dataset	Regression	log(CN)	log(P3)	log(cos)	log(PA)	log(PR)	f_{10}
Slashdot	f_{all}	−0.5411	−0.4866	−3.9434	0.2502	0.2321	–
	$f_{\text{PR-cos}}$	–	–	−6.113	–	–	0.2386
Epinions	f_{all}	−0.8587	−0.3827	−5.0360	−0.0105	0.8498	–
	$f_{\text{PR-cos}}$	–	–	−1.5103	–	–	0.5111

link, we split the set of positive edges P randomly into two sets P_{a} and P_{b}. We use the sizes $|P_{\text{a}}| = 3|P_{\text{b}}|$, corresponding to a training set containing 75% of all edges.

The training set is thus P_{a} and the true test set is N. The false test set can be chosen in three different ways to emphasize different features of the tested link prediction functions:

- (P_{b}) Other known positive links in the false test set force a good distinction capability between negative and positive links.
- (O) Only including non-edges in the false test set will emphasize the ability of a link prediction function to distinguish negative edges from non-edges.
- ($P_{\text{b}} \cup O$) Using both positive and non-edges in the false test set evaluates a link prediction function at the task of distinguishing negative edges from both positive edges and non-edges.

The three cases result in the following link prediction problems:

$$P_{\text{a}} \to N \mid P_{\text{b}} \tag{6.10}$$
$$P_{\text{a}} \to N \mid O \tag{6.11}$$
$$P_{\text{a}} \to N \mid P_{\text{b}} O \tag{6.12}$$

Although it may seem sufficient to use the third, combined false test set, our experiments will show that the relative accuracy of individual link prediction functions at the three problems may be radically different, and thus it is a requirement that a good link prediction method performs well at all three problems.

Results The AUC-values for all link prediction functions for all three link prediction problems are shown in Figure 6.2. The corresponding ROC curves for the link problem using $P_{\text{b}} \cup O$ as the false test set are shown in Figure 6.3. The weights learned for logistic regression are given in Table 6.4.

Observations A first observation from Figure 6.2 is that the easiest prediction problem is to predict negative links against non-links from positive links ($P_{\text{a}} \to N \mid O$) which reaches AUC-values as big as 0.93. Individual link prediction functions f_1 to f_9 perform well (AUC > 0.5) at the problem $P_{\text{a}} \to N \mid O$, while their inverses (AUC < 0.5) perform well at the task $P_{\text{a}} \to N \mid P_{\text{b}}$. Thus, none of these functions taken by itself is suited to solving our problem.

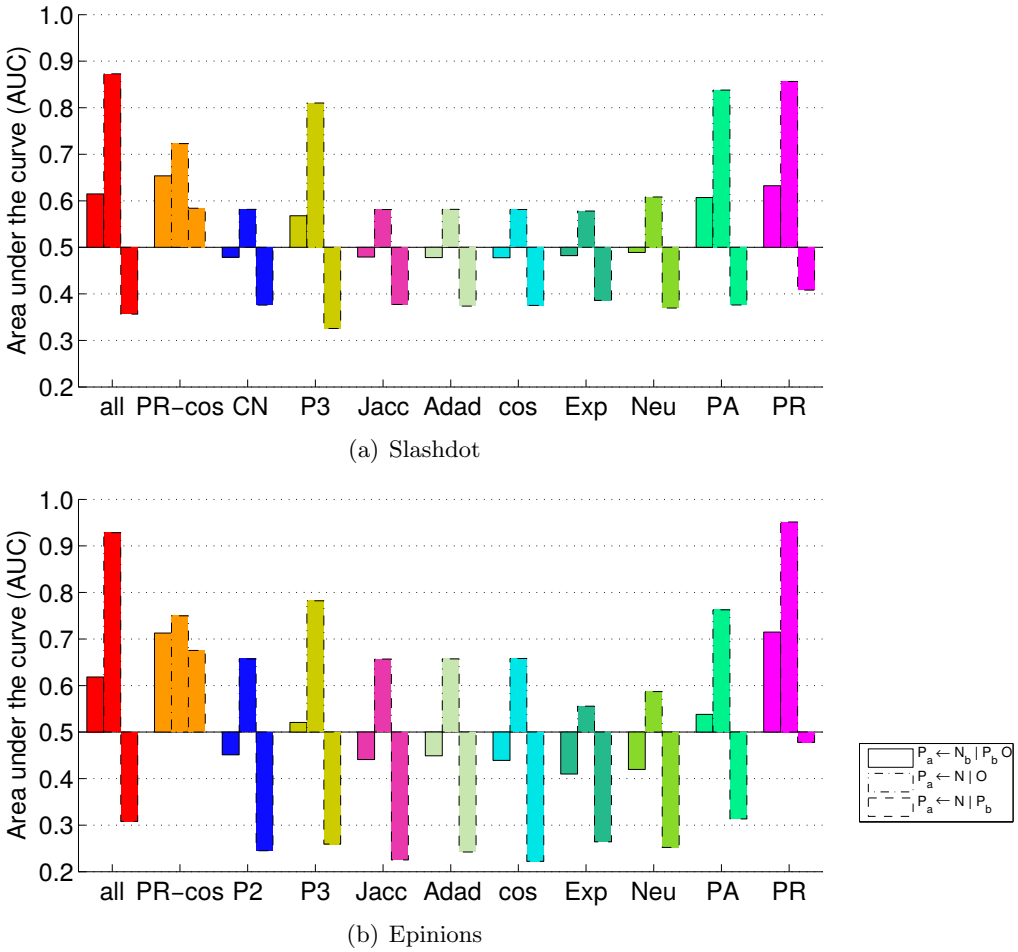
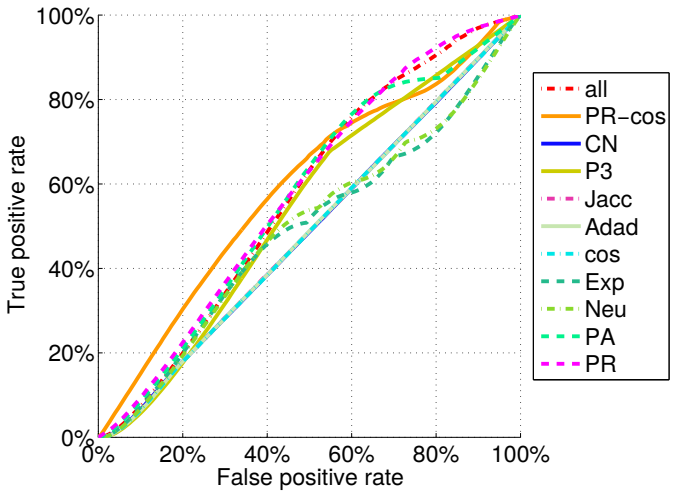


Figure 6.2: The AUC-value of the link prediction functions at the three link prediction problems of Experiment 1. The two leftmost functions are ensemble functions; the other functions are the basic link prediction functions. A suitable link prediction function at the task of predicting negative links must have an AUC-value larger than 0.5 for all three link prediction problem.

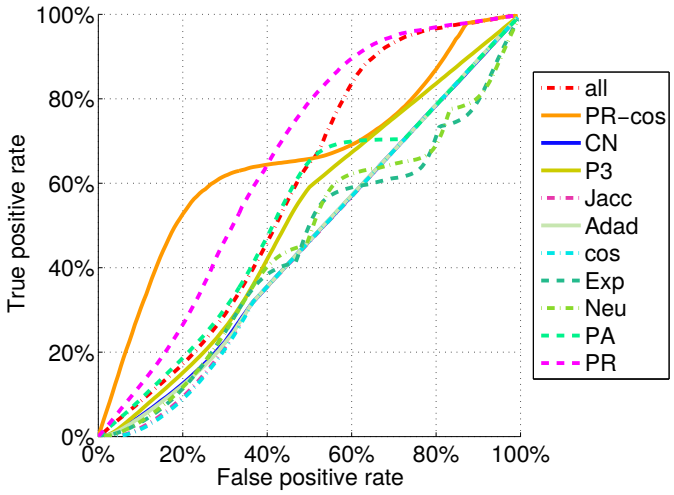
Instead, ensemble methods must be used. Our tests show that the only set of functions that perform well (AUC > 0.5) when combined include the conditional PageRank f_{10}, i.e., regression type $f_{\text{PR-cos}}$. Note that the regression weights in Table 6.4 cannot be interpreted individually. The regression weights learned for $f_{\text{PR-cos}}$ for both datasets have the same signed and relative weights, and suggest the prediction function

$$f = \alpha \left(\begin{cases} \log(\text{PR}) \text{ if } \cos = 0, \\ \min(\log(\text{PR})) \text{ otherwise.} \end{cases} \right) - \beta \log(\cos),$$

in which the weights $\alpha, \beta > 0$ must be determined experimentally. Figure 6.3 also shows that this method ($f_{\text{PR-cos}}$) also has the steepest ROC curve at the point $(0, 0)$, implying that this method is best at predicting the top-k unknown negative links for small k. This property is



(a) Slashdot



(b) Epinions

Figure 6.3: The ROC curves of all link prediction functions at the link prediction problem $P_{\mathrm{a}} \rightarrow N \mid P_{\mathrm{b}}O$ for both datasets. Well performing methods in this experiment have a ROC curve that is higher on the plot than other curves. A high steepness of the curve at the point $(0, 0)$ indicates a high precision for the top-k items, implying a good performance at recommendation tasks.

important for the application of recommender systems, in which only the top-k results are used and the rest ignored.

6.5.2 Experiment 2: Upper Bound

To assess whether the accuracy of link prediction achieved in Experiment 1 can be considered accurate enough to recommend against the introduction of explicit negative links in online social networks, we compare the results with the results of the related link prediction problem

in which negative links are known. This related link prediction function gives an upper bound for the accuracy attainable using the previous methods, and the difference in accuracy between both problems will thus characterize the added value that the *negative link* feature brings to a social networking platform.

We will assume that a part of the negative links in the social network are already known, and include them in the training set. We thus compare the two following link prediction problems:

$$P_\mathrm{a} \to N_\mathrm{b} \mid P_\mathrm{b}O \tag{6.13}$$

$$P_\mathrm{a}, N_\mathrm{a} \to N_\mathrm{b} \mid P_\mathrm{b}O \tag{6.14}$$

The set of negative edges N is thus split into the two sets $N = N_\mathrm{a} \cup N_\mathrm{b}$. The split of N is made in the same proportion as the split of P, i.e., $|N_\mathrm{a}| = 3|N_\mathrm{b}|$, which means that 75% of all negative links are used for the training set.

The first link prediction problem is the same as link prediction problem (6.10) in Experiment 1 up to the necessary replacement of N by N_b; the second one includes additional negative edges N_a in the training set. Note that any link prediction function that has a high accuracy in the first problem can be transformed into an accurate link prediction function for the second problem by simply ignoring the negative edges. Thus, the accuracy of link prediction functions at the second problem are upper bounds for the accuracy of link prediction functions at the first problem. The tightness of this bound can then be interpreted in terms of the added value of the negative edges. If the difference is high, negative edges contain information that is not recoverable using only the positive edges, and a *negative link* feature will increase the accuracy of news stream filters and recommender systems based on the social network. If the difference is small, negative links do not give such an added value.

For the second problem, the link prediction methods must be modified to work on signed edges. We follow the methods described in [Kunegis et al., 2009], which define the degree $d(i)$ as not depending on edge signs, and essentially replace the number of common neighbors

$$|\{k \mid i \sim k \wedge k \sim j\}|$$

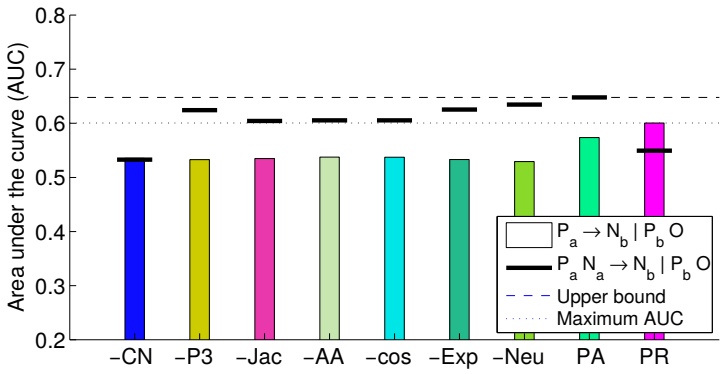
with the difference of positive and negative paths

$$\sum_{i \sim k, k \sim j} w(i,k)w(k,j),$$

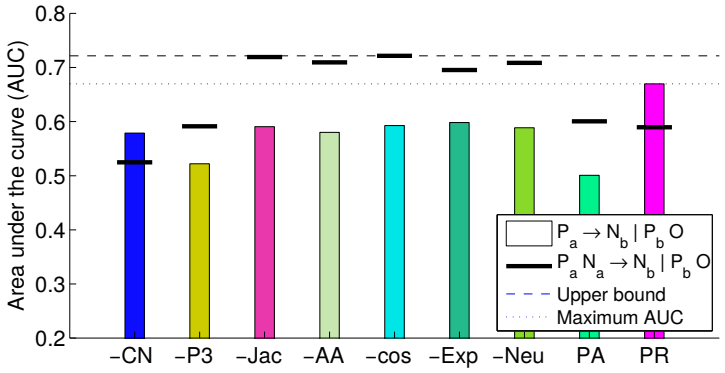
which reduces to the number of common neighbors in the unsigned case.

Results In this experiment, the performance of algorithms at the problem $P_\mathrm{a}N_\mathrm{a} \to N_\mathrm{b} \mid P_\mathrm{b}O$ serves as an upper bound for the performance of methods at the problem $P_\mathrm{a} \to N_\mathrm{b} \mid P_\mathrm{b}O$. Thus, the results of this experiment can be used to assess how much information is lost when negative links are not recorded in a social network. The results of the experiment for both datasets are shown in Figure 6.4.

The experimental results show that the best method when negative links are known performs by about 0.05 AUC points better than the best method when no negative links are known. Thus, allowing negative links in an online social network does have an added value for the network, although that added value is small, because the difference in AUC-values from one link prediction function to the next are larger than the observed difference of 0.05,



(a) Slashdot



(b) Epinions

Figure 6.4: Comparison of the accuracy of link prediction with and without N_a in the training set. The bars show the AUC-values of the link prediction problem in which no negative edges are known. The thick black lines represent the AUC-values at the task in which some negative links are known. For the neighborhood-based prediction functions, the plot shows the AUC-values of the inverted prediction functions, since these then have AUC-values of over 0.5.

suggesting that specific functions adapted to any dataset may be able to close that gap.

6.6 Conclusion

In this chapter, we have introduced the new problem of prediction latent negative links. We have defined the prediction set up to evaluate the predictive performance of several proposed measures for the latent negative prediction task. The experimental results derived in the two experiments show that the problem of predicting negative links in a social network, using only positive links is a variant of the link prediction problem that can only be solved by combining both centrality-based and neighborhood-based functions, using positive weights for centrality-based functions and negative weights for neighborhood-based functions. This result is congruent with the intuition that the existence of an edge (regardless of its sign) correlates positively with centrality-based functions, showing that models such as preferential

attachment, which predict a higher probability of edge attachment for nodes with high degree centrality, is valid independently of edge sign in networks where negative links are allowed. On the other hand, signed networks follow balance theory in that triangles in them tend to have an even number of negative edges, explaining why the neighborhood-based methods correlate negatively with the presence of negative edges.

We have shown that in the online social networks Slashdot and Epinions, the *foe* and *distrust* feature is used by users in a way that can be predicted to high accuracy from the *friend* and *trust* links. Thus, with regards maximizing the utility of news stream filtering and social recommendation, the negative link features of these two sites are redundant to a large extent. However, it does not follow that these features are useless. Quite the contrary is true; the *foe* feature of Slashdot is used as a personal organization tool (remembering who is considered a *troll*), or simply to let another user know one's disapproval of them. In Epinions, the *distrust* feature is likewise central to the Epinions's Web of Trust.

As a solution to the generic learning problem of predicting one link type from another one, we showed that the usual link prediction methodology can be applied, but only with the caveat that individual link prediction function may have inverted performance, e.g., the cosine similarity measure in the example of disapproval links.

Finally, as an application of our methods to online social networks that do not allow *foe* or *distrust* links, we propose that a link prediction function learned using regression with Slashdot and Epinions data may be applied. The only way however to ascertain the accuracy of these predictions is to perform the evaluation described in this work, which by nature of the problem is only possible when negative edges are known. Proxies for negative edges could be found in some platforms. For instance in Facebook, one can choose to not list the news of a friend in one's own news stream or one can deny a friend request.

The work in this chapter was published in one paper:

- *Jérôme Kunegis, Julia Preusse, and Felix Schwagereit. What is the added value of negative links in online social networks? In Proc. Int. World Wide Web Conf, 2013.*

7 Conclusions and Further Directions

7.1 Conclusions

In this thesis, we have demonstrated approaches how structural changes, such as the addition and removal of links, and the prediction of states, such as the presence of latent negative links, can be predicted only from the network structure.

In the first part of the thesis, we have investigated the relationship between link and unlink prediction. Sociological studies suggest that the corresponding problems of link and unlink prediction are highly related – many factors that are correlated positively with the formation of new ties were stated to correlate negatively with the dissolution of a tie. We evaluated the relatedness of the two prediction problems with two transformations from link to unlink prediction. These transformations indicated that indeed some link prediction characteristics are also suitable for unlink prediction. The two general prediction problems are however not the same; in particular the unlink prediction has proven to be much harder than link prediction. Since link and unlink prediction are not congruent, we have further analyzed their interplay. We have proposed a unified view that does not consider the performance of a characteristic at only one of the two prediction problems and instead considers the joint performance. This led us to define and evaluate the four states of growth, decay, stability and instability for several big Wikipedia datasets which are prominent representatives of knowledge networks. We have evaluated several characteristics and have demonstrated that structural characteristics for each category can be found. The contributions of this part are as follows.

- We have demonstrated that unlink prediction is more than a simple transformation of the link prediction problem.
- We have shown that link changes can be categorized into the four states of growth, decay, stability and instability.
- We found important indicators for the removal of a link in knowledge networks to be a small embedding of the relationship between two knowledge items and a low degree of both linked knowledge items.

In the second part, we utilized the temporal evolution of knowledge networks where time-stamps of each addition and removal event are known and proposed four temporal models that exploit the temporal information.We have evaluated the four models on the tasks of link and unlink prediction for several big Wikipedia knowledge networks and demonstrated that the link and unlink prediction problem greatly benefit from the incorporation of temporal information. Despite our observations that temporal information improves the predictability of unlinks, its prediction accuracy is still much smaller than for link prediction. In an upper bound experiment where we use exact link measures as opposed to link measures estimated from the temporal evolution, we demonstrated that it is in theory possible to predict addition

and removal events with very high accuracy using only the structure of the network. The contributions of this part are as follows.

- We have defined and evaluated four temporal models for a general dataset consisting of links, unlinks and timestamps for all link events at the task of link and unlink prediction.
- We have demonstrated that the exploitation of temporal information improves the prediction of links and unlinks.
- We performed an upper bound experiment to demonstrate that unlinks *could* be predicted with AUC-values up to 0.9.

In the third part, we have defined nine effect categories which describe reasons for the formation of new relationships and their dissolution in directed social networks where latent or explicit groups are given. We defined measures of each effect category and evaluated their individual and joint predictive ability for the task of link and unlink prediction. We evaluated our approach on a Twitter dataset of German politicians. Our results suggest that measures based on information about past links are extremely valuable for predicting the dissolution of social ties, while for the prediction of the formation of social ties measures based on the link network are sufficient. The contributions of this part are as follows.

- We have defined and evaluated a general framework that describes sociological effects for the formation and dissolution of ties in directed social networks where latent or explicit user groups are given. Our proposed framework exploits the group association information of users.
- We have demonstrated that information on *past* network connections boost the performance of link and in particular unlink prediction.

In the last part, we have investigated the new problem of how to predict *latent negative* links in a social network. These links are in contrast to positive links such as trust or friendship and are useful to hide content of distrusted users. Many online platforms prohibit the user to label relationships as negative; users can only be added as friends or followers. Although it is not possible in these platforms to explicitly label other users as foes or distrusted, users implicitly have negative relationships or opinions about other users. We have evaluated several measures on networks where positive and negative relationships were labeled. The contributions of this part are as follows.

- We have defined a new prediction problem: the prediction of latent negative links.
- We have proposed and evaluated structural characteristics for the latent negative prediction problem.
- We have demonstrated that the added value of the negative link feature in social network is rather small for the latent negative prediction problem. Hence, latent negative links can be successfully predicted from only positive links.

The structural consideration of the prediction problem has several advantages. Since the models that we presented in this thesis rely on the structure of a network and are thus in principle domain-independent, they can be applied to all networks. In particular, they can be combined

with domain-specific algorithms that make use of the content or the context information in a network. The structural viewpoint also allows us to compare the mechanisms of linking and unlinking across different networks and even network types. Even if we have shown that the structure of a network on its own may not be enough to solve the unlink prediction problem with a high accuracy, we were still able to beat state-of the art methods for unlink prediction.

Limitations For the analysis of which links are formed, removed or latent negative, the goal is to reveal the structural network characteristics that best describe the three types of links. Since we use actual data of link additions and removals, or friends and foes, these actions or relationship labels are greatly influenced by the respective platform policies in particular the recommendation system of the platform. We may not observe with whom users truly choose to connect or disconnect. Instead, we may observe which users or actions, that the recommender system has proposed, were chosen by a user. For instance, a user may only choose new followers that appear as recommendations in the interface of Twitter. Hence, the characteristics that we found to be indicative for the three link prediction problems could be the result of the technical system of the platform and not the user's natural behavior. To measure both effects, one would have to know what the technical system looks like, i.e. how Twitter's follower recommender works. Unfortunately, this information is not available for most platforms.

Furthermore, for most prediction tasks, one assumes that the interface or the platform policies don't change throughout the observation period. However, a site may employ a new recommender that changes the link addition behavior of its users, or it may introduce a new site element that changes the user's behavior. In general, research in link prediction disregards this technical changes and treats all user actions the same.

7.2 Future Directions

Some potential future research directions arise from the work done in this thesis.

False Test Set for Link Prediction To test the links that actually appear against the links that don't appear in a network, the false test set for the link prediction task is commonly chosen to contain random non-appearing links. In other words, the link prediction problem tries to distinguish between actually connected node pairs and random non-connected node pairs. Explaining why someone living in a village in Germany is not friend with a random person living in a town in Australia can easily be structurally distinguished from two people that are actually friends. As it turned out in our analyses, this choice greatly influences the precision of link prediction methods. Even very basic features perform surprisingly well, because their values in the true and false test set differ greatly. Thus, we ask a more practically appealing question: How can links that appear in a network be distinguished from links that *could* appear? This raises the question of defining when links *could* appear, which is in general tough to answer. In a given application however, links from the false test set might be naturally given. For instance in Facebook, new friends are suggested to users who mostly only add some of them; the other user were proposed but not chosen to befriend with. Thus, these kind of links could form a practically more appealing false test set.

This exact issue also gives the impression that unlink prediction is fundamentally more difficult than link prediction. When the false test set for link prediction is adjusted, we expect

the link prediction precision to drop significantly and to be more similar to the precision of unlink prediction methods.

Link Replacements In particular for Wikipedia, we have observed that sometimes a link (i, j) is removed and at the exact same time a link (i, k) is added. This gives the impression that the link (i, j) was replaced by the link (i, k). So far, we have considered the removal and addition processes separately, but we believe that concurrent link additions and removals, namely the *replacements* of links, is a new and interesting prediction problem. Replacements of relationships also appear in the context of job networks, where employees replace their employers by a new one, i.e., they quit their current job to start somewhere new. The replacement effect may also be observed in partnership or friendship networks, where the current romantic partner is directly replaced by a new one or the best friend is exchanged.

Short versus Long-term Link Removals In our studies we targeted the prediction of any unlink occurring in the network. If timestamps for all events, in particular removal events, are given in the dataset, one may ask whether structural indicators of short-term unlinks are different from indicators of long-term unlinks. Take for instance the social network Twitter. If a user unfollows a user within a minute after friending him, we can consider the previously established link as spurious. That is, maybe the user has just followed the wrong person or after the tweets of the followed person appeared in the tweet stream, he directly decided to unfollow this person because the posts are inappropriate or too many. Contrarily, if a user follows another user for a few years and then decides to unfollow this account, the reasons from the short-term unlinking should not apply. Therefore, we believe that predicting short-term and long-term links are related to very different aspects and it should thus be worthwhile to analyze them separately.

Structure versus Actions versus Content in Social Networks The scope of this thesis was to find *structural* predictors of link states and link state changes that use only the explicit structure of relationships in a network. In social networks, users cannot only establish social relationships such as friendship with other users, they can also interact with other users that are not among their friends, e.g., if two users reply to the same forum post. Further, a social network also contains the content of user posts or the content that is exchanged between users. We believe that the influence of these three feature categories – structure, interaction and content – may be very different for the evolution of relationships in a network. Whereas some platforms such as Twitter could be thought of as content-driven, the formation and dissolution of relationships in other platforms such as Facebook could be more driven by user interactions and the network structure. Whereas, measures for all three categories have been used individually or jointly in some works, e.g. [Aiello et al., 2012, Wagner et al., 2012, Raeder et al., 2011], a comparison between the influence of the three feature categories has, to the best of our knowledge, not been performed, yet. Juxtaposing the importance of the three factors across different platforms helps to understand the different motivations that users have for creating and removing links in different platforms.

Bibliography

[Adafre and de Rijke, 2005] Adafre, S. F. and de Rijke, M. (2005). Discovering missing links in wikipedia. In *Proceedings of the 3rd International Workshop on Link Discovery*, LinkKDD '05, pages 90–97, New York, NY, USA. ACM.

[Adamic, 1999] Adamic, L. (1999). The Small World Web. In Abiteboul, S. and Vercoustre, A.-M., editors, *Research and Advanced Technology for Digital Libraries*, volume 1696 of *Lecture Notes in Computer Science*, chapter 27, pages 852–852–852. Springer Berlin / Heidelberg.

[Adamic and Adar, 2001] Adamic, L. and Adar, E. (2001). Friends and neighbors on the Web. *Social Networks*, 25:211–230.

[Aiello et al., 2012] Aiello, L., Barrat, A., Cattuto, C., Schifanella, R., and Ruffo, G. (2012). Link creation and information spreading over social and communication ties in an interest-based online social network. *EPJ Data Science*, 1(1):12.

[Anderson et al., 1999] Anderson, C., Wasserman, S., and Crouch, B. (1999). A p* primer: logit models for social networks. *Social Networks*.

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

[Barabási et al., 1999] Barabási, A.-L., Albert, R., and Jeong, H. (1999). Mean-field theory for scale-free random networks. *Physica A*, 272:173–187.

[Becker et al., 2011] Becker, H., Naaman, M., and Gravano, L. (2011). Beyond trending topics: Real-world event identification on twitter. In *Fifth International AAAI Conference on Weblogs and Social Media*.

[Benczúr et al., 2005] Benczúr, A. A., Csalogány, K., Sarlós, T., and Uher, M. (2005). SpamRank – fully automatic link spam detection. In *Proc. Int. Workshop on Adversarial Information Retrieval on the Web*.

[Berg and McQuinn, 1986] Berg, J. and McQuinn, R. (1986). Attraction and exchange in continuing and noncontinuing dating relationships. *Journal of Personality and SocialPsychology*, 50:942–952.

[Bernheim, 1994] Bernheim, B. D. (1994). A Theory of Conformity. *The Journal of Political Economy*, 102(5):841–877.

[Boyd et al., 2010] Boyd, D., Golder, S., and Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10.

[Bradley, 1997] Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30:1145–1159.

[Brin and Page, 1998] Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the seventh international conference on World Wide Web 7*, pages 107–117.

[Burt, 2000] Burt, R. (2000). Decay functions. *Social Networks*, 22(1):1–28.

[Carrasco et al., 2011] Carrasco, B., Lu, Y., and da Trindade, J. M. F. (2011). Partitioning social networks for time-dependent queries. In *Proceedings of the 4th Workshop on Social Network Systems*, SNS '11, pages 2:1–2:6, New York, NY, USA. ACM.

[Clauset et al., 2009] Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-Law Distributions in Empirical Data. *SIAM Review*, 51(4):661–703.

[Cleek and Pearson, 1985] Cleek, M. G. and Pearson, T. A. (1985). Perceived causes of divorce: An analysis of interrelationships. *Journal of Marriage and Family*, 47(1):pp. 179–183.

[Conover et al., 2011] Conover, M., Ratkiewicz, J., Francisco, M., Goncalves, B., Menczer, F., and Flammini, A. (2011). Political polarization on twitter.

[Cooke et al., 2006] Cooke, R. J. E., Potgieter, A., and April, K. (2006). *Link prediction and link detection in sequences of large social networks using temporal and local metrics.* PhD thesis.

[Dunbar, 1992] Dunbar, R. (1992). Neocortex size as a constraint on group size in primates. *Journal of Human Evolution*, 22(6):469 – 493.

[Emerson, 1976] Emerson, R. M. (1976). Social Exchange Theory. *Annual Review of Sociology*, 2(1):335–362.

[Erdös and Rényi, 1959] Erdös, P. and Rényi, A. (1959). On random graphs i. *Publicationes Mathematicae Debrecen*, 6:290–297.

[Faust and Wasserman, 1992] Faust, K. and Wasserman, S. (1992). Blockmodels: Interpretation and evaluation. *Social Networks*, 14(1-2):5–61.

[Frank and Strauss, 1986] Frank, O. and Strauss, D. (1986). Markov Graphs. *Journal of the American Statistical Association*, 81(395):832–842.

[Garcia et al., 2013] Garcia, D., Mavrodiev, P., and Schweitzer, F. (2013). Social resilience in online communities: The autopsy of friendster. *CoRR*, abs/1302.6109.

[Garlaschelli and Loffredo, 2004] Garlaschelli, D. and Loffredo, M. I. (2004). Patterns of Link Reciprocity in Directed Networks. *Phys. Rev. Lett.*, 93:268701.

[Getoor and Diehl, 2005] Getoor, L. and Diehl, C. P. (2005). Link mining: a survey. *SIGKDD Explor. Newsl.*, 7(2):3–12.

[Granovetter, 1973] Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6):1360–1380.

[Guha et al., 2004] Guha, R., Kumar, R., Raghavan, P., and Tomkins, A. (2004). Propagation of trust and distrust. In *Proc. Int. World Wide Web Conf.*, pages 403–412.

[Guimerà and Sales-Pardo, 2009] Guimerà, R. and Sales-Pardo, M. (2009). Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078.

[Harary, 1953] Harary, F. (1953). On the notion of balance of a signed graph. *Michigan Math. J.*, 2(2):143–146.

[Hayashi et al., 2009] Hayashi, K., Hirayama, J.-I., and Ishii, S. (2009). Dynamic exponential family matrix factorization. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, PAKDD '09, pages 452–462, Berlin, Heidelberg. Springer-Verlag.

[Heider, 1958] Heider, F. (1958). *The Psychology of Interpersonal Relations*. John Wiley & Sons, New York.

[Hidalgo and Rodriguez-Sickert, 2008] Hidalgo, C. A. and Rodriguez-Sickert, C. (2008). The dynamics of a mobile phone network. *Physica A: Statistical Mechanics and its Applications*, 387(12):3017–3024.

[Holland and Leinhardt, 1981] Holland, P. W. and Leinhardt, S. (1981). An exponential family of probability distributions for directed graphs. *Journal of the American Statistical Association*, 76(373):33–50.

[Homans, 1958] Homans, G. C. (1958). Social behavior as exchange. *American Journal of Sociology*, 63(6):pp. 597–606.

[Huang and Lin, 2009] Huang, Z. and Lin, D. K. J. (2009). The time-series link prediction problem with applications in communication surveillance. *INFORMS J. on Computing*, 21(2):286–303.

[Incite, 2011] Incite, T. N. C. (2011). Friends & frenemies: Why we add and remove facebook friends. State of Social Media Survey.

[Ito et al., 2005] Ito, T., Shimbo, M., Kudo, T., and Matsumoto, Y. (2005). Application of kernels to link analysis. In *Proc. Int. Conf. on Knowledge Discovery in Data Mining*, pages 586–592.

[Järvelin and Kekäläinen, 2002] Järvelin, K. and Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446.

[Kaczmirek et al., 2013] Kaczmirek, L., Mayr, P., Vatrapu, R., Bleier, A., Blumenberg, M., Gummer, T., Hussain, A., Kinder-Kurlanda, K., Manshaei, K., Thamm, M., Weller, K., Wenz, A., and Wolf, C. (2013). Social media monitoring of the campaigns for the 2013 german bundestag elections on facebook and twitter. *CoRR*, abs/1312.4476.

[Kairam et al., 2012] Kairam, S., Wang, D. J., and Leskovec, J. (2012). The life and deatch of online groups: Predicting group growth and longevity. In *Proc. Int. Conf. on Web Search and Data Mining*, pages 673–682.

[Kandola et al., 2002] Kandola, J., Shawe-Taylor, J., and Cristianini, N. (2002). Learning semantic similarity. In *Advances in Neural Information Processing Systems*, pages 657–664.

[Karney and Bradbury, 1995] Karney, B. R. and Bradbury, T. N. (1995). The longitudinal course of marital quality and stability: A review of theory, methods, and research. *Psychological Bulletin*, 118(1):3–34.

[Karnstedt et al., 2010] Karnstedt, M., Hennessy, T., Chan, J., Basuchowdhuri, P., Hayes, C., and Strufe, T. (2010). Churn in social networks. In *Handbook of Social Network Technologies*, pages 185–220. Springer.

[Kashima and Abe, 2006] Kashima, H. and Abe, N. (2006). A parameterized probabilistic model of network evolution for supervised link prediction. In *Proceedings of the Sixth International Conference on Data Mining*, ICDM '06, pages 340–349. IEEE Computer Society.

[Katz, 1953] Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

[Kearns and Leonard, 2004] Kearns, J. N. and Leonard, K. E. (2004). Social networks, structural interdependence, and marital quality over the transition to marriage: a prospective analysis. *Journal of Family Psychology*, 18:383–395.

[Kirke, 2009] Kirke, D. M. (2009). Gender clustering in friendship networks: some sociological implications.

[Kivran-Swaine et al., 2012] Kivran-Swaine, F., Govindan, P., and Naaman, M. (2012). The impact of network structure on breaking ties in online social networks: Unfollowing on Twitter. In *Proc. Int. Conference on Weblogs and Social Media*, pages 1101–1104.

[Kleinberg et al., 1999] Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999). The Web as a graph: Measurements, models, and methods. In *Proc. Int. Conf. on Computing and Combinatorics*, pages 1–17.

[Knoke, 1990] Knoke, D. (1990). *Political Networks: A Structural Perspective*. Cambridge University Press, New York.

[Kondor and Lafferty, 2002] Kondor, R. and Lafferty, J. (2002). Diffusion kernels on graphs and other discrete structures. In *Proc. Int. Conf. on Machine Learning*, pages 315–322.

[Koren, 2008] Koren, Y. (2008). Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 426–434, New York, NY, USA. ACM.

[Koren, 2010] Koren, Y. (2010). Collaborative filtering with temporal dynamics. *Commun. ACM*, 53(4):89–97.

[Kossinets and Watts, 2006] Kossinets, G. and Watts, D. J. (2006). Empirical Analysis of an Evolving Social Network. *Science*, 311(5757):88–90.

[Kubica et al., 2003a] Kubica, J., Goldenberg, A., Komarek, P., Moore, A., and Schneider, J. (2003a). A comparison of statistical and machine learning algorithms on the task of link completion.

[Kubica et al., 2003b] Kubica, J., Moore, A., Cohn, D., and Schneider, J. (2003b). cgraph: A fast graph-based method for link analysis and queries. In *Proceedings of the 2003 IJCAI Text-Mining & Link-Analysis Workshop*.

[Kunegis, 2011] Kunegis, J. (2011). *On the Spectral Evolution of Large Networks*. PhD thesis, University of Koblenz–Landau.

[Kunegis, 2013] Kunegis, J. (2013). KONECT – The Koblenz Network Collection. In *Proc. Int. Web Observatory Workshop*.

[Kunegis et al., 2012] Kunegis, J., Gröner, G., and Gottron, T. (2012). Online dating recommender systems: The split-complex number approach. In *Proc. Workshop on Recommender Systems and the Social Web*, pages 37–44.

[Kunegis et al., 2009] Kunegis, J., Lommatzsch, A., and Bauckhage, C. (2009). The Slashdot Zoo: Mining a social network with negative edges. In *Proc. Int. World Wide Web Conf.*, pages 741–750.

[Kunegis and Preusse, 2012] Kunegis, J. and Preusse, J. (2012). Fairness on the web: Alternatives to the power law. In *Proc. Web Science Conf.*, pages 175–184.

[Kunegis et al., 2013] Kunegis, J., Preusse, J., and Schwagereit, F. (2013). What is the added value of negative links in online social networks? In *Proc. Int. World Wide Web Conf.*

[Kwak et al., 2011] Kwak, H., Chun, H., and Moon, S. (2011). Fragile online relationship: A first look at unfollow dynamics in Twitter. In *Proc. Conf. on Human Factors in Computing Systems*, pages 1091–1100.

[Kwak et al., 2012] Kwak, H., Moon, S., and Lee, W. (2012). More of a receiver than a giver: Why do people unfollow in Twitter? In *Proc. Int. Conference on Weblogs and Social Media*, pages 499–502.

[Lampe and Resnick, 2004] Lampe, C. and Resnick, P. (2004). Slash(dot) and burn: Distributed moderation in a large online conversation space. In *Proc. Int. Conf. on Human Factors in Computing Systems*, pages 543–550.

[Langville and Meyer, 2006] Langville, A. N. and Meyer, C. D. (2006). *Google's PageRank and Beyond*. Princeton University Press.

[Lazarsfeld and Merton, 1954] Lazarsfeld, P. F. and Merton, R. K. (1954). Friendship as a social process: A substantive and methodological analysis. In Berger, M., Abel, T., and Page, C., editors, *Freedom and Control in Modern Society*, pages 18–66. Van Nostrand, New York.

[Leskovec et al., 2008] Leskovec, J., Backstrom, L., Kumar, R., and Tomkins, A. (2008). Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 462–470, New York, NY, USA. ACM.

[Leskovec et al., 2010a] Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010a). Governance in social media: A case study of the Wikipedia promotion process. In *Proc. Int. Conf. on Weblogs and Social Media*, pages 98–105.

[Leskovec et al., 2010b] Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010b). Predicting positive and negative links in online social networks. In *Proc. Int. World Wide Web Conf.*, pages 641–650.

[Leskovec et al., 2010c] Leskovec, J., Huttenlocher, D., and Kleinberg, J. (2010c). Signed networks in social media. In *Proc. Int. Conf. on Human Factors in Computing Systems*, pages 1361–1370.

[Leskovec et al., 2005] Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 177–187, New York, NY, USA. ACM.

[Liben-Nowell and Kleinberg, 2003] Liben-Nowell, D. and Kleinberg, J. (2003). The link prediction problem for social networks. In *Proc. Int. Conf. on Information and Knowledge Management*, pages 556–559.

[Lichtenwalter et al., 2010] Lichtenwalter, R. N., Lussier, J. T., and Chawla, N. V. (2010). New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '10, pages 243–252, New York, NY, USA. ACM.

[Lietz et al., 2014] Lietz, H., Wagner, C., Bleier, A., and Strohmaier, M. (2014.). When politicians talk: Assessing online conversational practices of political parties on twitter. *International AAAI Conference on Weblogs and Social Media (ICWSM2014)*.

[Lima-Mendez and van Helden, 2009] Lima-Mendez, G. and van Helden, J. (2009). The powerful law of the power law and other myths in network biology. *Molecular bioSystems*, 5(12):1482–1493.

[Llullaku et al., 2009] Llullaku, S., Hyseni, N., Bytyçi, C., and Rexhepi, S. (2009). Evaluation of trauma care using triss method: the role of adjusted misclassification rate and adjusted w-statistic. *World Journal of Emergency Surgery*, 4(1).

[Lü and Zhou, 2011] Lü, L. and Zhou, T. (2011). Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and Its Applications*, 390(6):1150–1170.

[Martin and Yeung, 2006] Martin, J. L. and Yeung, K.-T. (2006). Persistence of close personal ties over a 12-year period. *Social Networks*, 28(4):331–362.

[Massa and Avesani, 2005] Massa, P. and Avesani, P. (2005). Controversial users demand local trust metrics: an experimental study on epinions.com community. In *Proc. American Association for Artificial Intelligence Conf.*, pages 121–126.

[McPherson et al., 2001] McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444.

[Mihalcea and Csomai, 2007] Mihalcea, R. and Csomai, A. (2007). Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, pages 233–242, New York, NY, USA. ACM.

[Milardo, 1987] Milardo, R. M. (1987). Changes in social networks of women and men following divorce: A review. *Journal of Family Issues*, 8(1):78–96.

[Milne and Witten, 2008] Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 509–518, New York, NY, USA. ACM.

[Mislove, 2009] Mislove, A. (2009). *Online Social Networks: Measurement, Analysis, and Applications to Distributed Information Systems*. PhD thesis, Rice University, Department of Computer Science.

[Mislove et al., 2013] Mislove, A., Koppula, H., Gummadi, K., Druschel, P., and Bhattacharjee, B. (2013). An empirical validation of growth models for complex networks. In *Dynamics On and Of Complex Networks, Volume 2*, Modeling and Simulation in Science, Engineering and Technology, pages 19–40. Springer New York.

[Myers and Leskovec, 2014] Myers, S. A. and Leskovec, J. (2014). The bursty dynamics of the twitter information network. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 913–924. International World Wide Web Conferences Steering Committee.

[Najork et al., 2007] Najork, M. A., Zaragoza, H., and Taylor, M. J. (2007). Hits on the Web: How does it compare? In *Proc. Int. Conf. on Research and Development in Information Retrieval*, pages 471–478.

[Naveed et al., 2011] Naveed, N., Gottron, T., Kunegis, J., and Alhadi, A. C. (2011). Bad news travel fast: A content-based analysis of interestingness on twitter. In *WebSci '11: Proceedings of the 3rd International Conference on WebScience*.

[Newman, 2002] Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, 89(20):208701.

[O'Madadhain et al., 2005] O'Madadhain, J., Hutchins, J., and Smyth, P. (2005). Prediction and ranking algorithms for event-based network data. *SIGKDD Explor. Newsl.*, 7(2):23–30.

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

[Osborne et al., 2012] Osborne, M., Petrovic, S., McCreadie, R., Macdonald, C., and Ounis, I. (2012). Bieber no more: First Story Detection using Twitter and Wikipedia. *SIGIR 2012 Workshop on Time-aware Information Access (#TAIA2012)*.

[Oyama et al., 2011] Oyama, S., Hayashi, K., and Kashima, H. (2011). Cross-temporal link prediction. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining*, ICDM '11, pages 1188–1193, Washington, DC, USA. IEEE Computer Society.

[Parks, 2007] Parks, M. R. (2007). *Personal Relationships and Personal Networks*. Lawrence Erlbaum Associates.

[Perkowitz and Etzioni, 1997] Perkowitz, M. and Etzioni, O. (1997). Adaptive web sites: an ai challenge. In *Proceedings of the 15th international joint conference on Artifical intelligence - Volume 1*, IJCAI'97, pages 16–21, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Perl et al., 2014a] Perl, J., Kunegis, J., and Ruß, G. (2014a). If you want my future, don't forget my past: Temporal models of linked knowledge. Technical Report.

[Perl et al., 2014b] Perl, J., Wagner, C., Kunegis, J., and Staab, S. (2014b). A theory-driven approach for link and unlink predictions in directed social networks. Technical report.

[Potgieter et al., 2007] Potgieter, A., April, K. A., Cooke, R. J. E., and Osunmakinde, I. O. (2007). Temporality in link prediction: Understanding social complexity. Sprouts: Working Papers on Information Systems.

[Preston and McDonald, 1979] Preston, S. H. and McDonald, J. (1979). The incidence of divorce within cohorts of american marriages contracted since the civil war. *Demography*, 16(1):pp. 1–25.

[Preusse et al., 2013] Preusse, J., Kunegis, J., Thimm, M., Gottron, T., and Staab, S. (2013). Structural dynamics of knowledge networks. In *Proc. Int. Conf. on Weblogs and Social Media*, pages 506–515.

[Preusse et al., 2012] Preusse, J., Kunegis, J., Thimm, M., and Sizov, S. (2012). DecLiNe - Models for Decay of Links in Networks. *ArXiv e-prints*.

[Quercia et al., 2012] Quercia, D., Bodaghi, M., and Crowcroft, J. (2012). Loosing 'friends' on Facebook. *Proc. Web Science Conf.*, pages 251–254.

[Raeder et al., 2011] Raeder, T., Lizardo, O., Hachen, D., and Chawla, N. V. (2011). Predictors of short-term decay of cell phone contacts in a large scale communication network. *Social Networks*, 33(4):245 – 257.

[Rowe, 2013] Rowe, M. (2013). Changing with time: Modelling and detecting user lifecycle periods in online community platforms. In *SocInfo*, pages 30–39.

[Rubin, 1986] Rubin, L. B. (1986). On men and friendship. *Psychoanal Rev*, 73:165–81.

[Saviotti, 2009] Saviotti, P. (2009). Knowledge networks: Structure and dynamics. In Pyka, A. and Scharnhorst, A., editors, *Innovation Networks*, Understanding Complex Systems, pages 19–41. Springer Berlin Heidelberg.

[Sharan and Neville, 2008] Sharan, U. and Neville, J. (2008). Temporal-relational classifiers for prediction in evolving domains. In *Data Mining, 2008. ICDM '08. Eighth IEEE International Conference on*, pages 540–549.

[Shvaiko and Euzenat, 2013] Shvaiko, P. and Euzenat, J. (2013). Ontology matching: State of the art and future challenges. *Knowledge and Data Engineering, IEEE Transactions on*, 25(1):158–176.

[Sibona and Walczak, 2011] Sibona, C. and Walczak, S. (2011). Unfriending on facebook: Friend request and online/offline behavior analysis. In *HICSS*, pages 1–10. IEEE Computer Society.

[Simmel, 1950] Simmel, G. (1950). *The sociology of georg simmel*, volume 92892. Simon and Schuster.

[Snijders and Steglich, 2013] Snijders, T. A. and Steglich, C. E. (2013). Representing micro–macro linkages by actor-based dynamic network models. *Sociological Methods & Research*.

[Spiegel et al., 2012] Spiegel, S., Clausen, J., Albayrak, S., and Kunegis, J. (2012). Link prediction on evolving data using tensor factorization. In *Proceedings of the 15th International Conference on New Frontiers in Applied Data Mining*, PAKDD'11, pages 100–110.

[Stager et al., 2006] Stager, M., Lukowicz, P., and Troster, G. (2006). Dealing with class skew in context recognition. *2012 32nd International Conference on Distributed Computing Systems Workshops*, 0:58.

[Suh et al., 2010] Suh, B., Hong, L., Pirolli, P., and Chi, E. H. (2010). Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *Proceedings of the 2010 IEEE Second International Conference on Social Computing*, SOCIALCOM '10, pages 177–184, Washington, DC, USA. IEEE Computer Society.

[Suitor and Keeton, 1997] Suitor, J. and Keeton, S. (1997). Once a friend, always a friend? effects of homophily on women's support networks across a decade. *Social Networks*, 19(1):51 – 62.

[Tong et al., 2008] Tong, H., Papadimitriou, S., Sun, J., Yu, P. S., and Faloutsos, C. (2008). Colibri: fast mining of large static and dynamic graphs. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 686–694, New York, NY, USA. ACM.

[Tylenda et al., 2009] Tylenda, T., Angelova, R., and Bedathur, S. (2009). Towards time-aware link prediction in evolving social networks. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis*, SNA-KDD '09, pages 1–10, New York, NY, USA. ACM.

[Viswanath et al., 2009] Viswanath, B., Mislove, A., Cha, M., and Gummadi, K. P. (2009). On the evolution of user interaction in facebook. In *Proceedings of the 2nd ACM workshop on Online social networks*, WOSN '09, pages 37–42, New York, NY, USA. ACM.

[Wagner et al., 2012] Wagner, C., Rowe, M., Strohmaier, M., and Alani, H. (2012). What Catches Your Attention? An Empirical Study of Attention Patterns in Community Forums. In *Proceedings of the International Conference on Weblogs and Social Media*. The AAAI Press.

[Watts and Strogatz, 1998] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of'small-world'networks. *Nature*, 393(6684):409–10.

[Wellman et al., 1997] Wellman, B., lin Wong, R. Y., Tindall, D., and Nazer, N. (1997). A decade of network change: Turnover, persistence and stability in personal communities. *Social Networks*, 19(1):27 – 50.

[West et al., 2009] West, R., Precup, D., and Pineau, J. (2009). Completing wikipedia's hyperlink structure through dimensionality reduction. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1097–1106, New York, NY, USA. ACM.

[White et al., 1976] White, H. C., Boorman, S. A., and Breiger, R. L. (1976). Social structure from multiple networks. i. blockmodels of roles and positions. *American Journal of Sociology*, 81(4):pp. 730–780.

[Wu et al., 2011] Wu, S., Hofman, J. M., Mason, W. A., and Watts, D. J. (2011). Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 705–714, New York, NY, USA. ACM.

[Yang et al., 2012] Yang, S.-H., Smola, A. J., Long, B., Zha, H., and Chang, Y. (2012). Friend or frenemy? predicting signed ties in social networks. In Hersh, W. R., Callan, J., Maarek, Y., and Sanderson, M., editors, *SIGIR*, pages 555–564. ACM.

[Zeng and Cimini, 2012] Zeng, A. and Cimini, G. (2012). Removing spurious interactions in complex networks. *Phys. Rev. E*, 85:036101.

[Zhang et al., 2012] Zhang, Q.-M., Lü, L., Wang, W.-Q., Zhu, Y.-X., and Zhou, T. (2012). Potential theory for directed networks. *CoRR*, abs/1202.2709.

Lebenslauf

Julia Perl, geb. Preusse

Kontaktinformation

Addresse	Institute for Web Science and Technologies Universität Koblenz-Landau Universitätsstr. 1 56070 Koblenz Deutschland
E-mail	jpreusse@uni-koblenz.de

Ausbildung

10/2005–12/2010	Universität Magdeburg	Diplom in Informatik, Nebenfach Mathematik
05/2005	Gymnasium Sonneberg (Thüringen)	Abitur

Arbeitserfahrung

01/2011–01/2015	Institute for Web Science and Technologies, Universität Koblenz-Landau	Wissenschaftliche Mitarbeiterin
08/2009–01/2010	Institut für Automation und Kommunikation, Magdeburg	Wissenschaftliche Hilfskraft
04/2009–07/2010	Institut für Wissensbasierte Systeme und Dokumentenverarbeitung, Universität Magdeburg	Wissenschaftliche Hilfskraft
09/2008–02/2009	Department of Mechanical Engineering, University of Melbourne, Australien	Forschungspraktikantin

Lehre

Semester	**Veranstaltung**	
SS 2014	Proseminar *Netzwerke und Dynamische Systeme*	Universität Koblenz-Landau
WS 2013/14	Übung zu *Grundlagen der Datenbanken*	Universität Koblenz-Landau
WS 2012/13	Übung zu *Grundlagen der Datenbanken*	Universität Koblenz-Landau
SS 2012	Übung zu *Social Networks and Dynamic Systems*	Universität Koblenz-Landau
SS 2008	Übung zu *Mathematik für Informatiker 2*	Universität Magdeburg
WS 2007/08	Übung zu *Mathematik für Informatiker 1*	Universität Magdeburg

Betreute Abschlussarbeiten

Bachelorarbeit	Miriam Kölle: *Supervised Link Prediction Approaches*
Masterarbeit	Io Taxidou: *Social Capital & Intellectual Capital in Online Communities*

Publikationen

- Julia Perl, Jérôme Kunegis, and Georg Ruß. If you want my future, don't forget my past: Temporal models of linked knowledge. Technical Report, 2014

- Julia Preusse, Jérôme Kunegis, Matthias Thimm, Thomas Gottron, and Steffen Staab. Structural dynamics of knowledge networks. In Proc. Int. Conf. on Weblogs and Social Media, pages 506–515, 2013.

- Jérôme Kunegis, Julia Preusse, and Felix Schwagereit. What is the added value of negative links in online social networks? In Proc. Int. World Wide Web Conf, 2013.

- Julia Preusse, Jérôme Kunegis, Matthias Thimm, and Sergej Sizov. DecLiNe - Models for Decay of Links in Networks. ArXiv e-prints, 2012.

- Jérôme Kunegis and Julia Preusse. Fairness on the web: Alternatives to the power law. In Proc. Web Science Conf., pages 175–184, 2012.

Glossary

A	Adjacency matrix
N	Network
V	node set
i, j, k, l	vertices
E	edge set
(i, j)	directed link
$\{i, j\}$	undirected link
$i \sim j$	the nodes i and j are adjacent
$N(i)$	neighborhood of a node
$w(i, j)$	edge weight of (i, j)
$d(i)$	node degree
$d_{in}(i)$	in-degree of a node
$d_{out}(i)$	out-degree of a node
$\mathrm{CN}(i, j)$	number of common neighbors between i and j
$\mathrm{P3}(i, j)$	number of paths of length three between i and j
$\mathrm{Jacc}(i, j)$	Jaccard coefficient
$\mathrm{Adad}(i, j)$	Adamic-Adar measure
$\cos(i, j)$	Cosine similarity
AUC-value	Prediction measure; area under the ROC curve
Training set	The set of node pairs that prediction measures are computed on
True test set	The set of node pairs that should be predicted with a high score
False test set	The set of node pairs that should not be predicted with a high score
Test set	All node pairs in true or false test set
Source set	The set of node pairs used to compute measures for parameter training
Target set	Set of target values to which the parameters are trained