# Structural Dynamics of Knowledge Networks

**Julia Preusse, Jérôme Kunegis, Matthias Thimm, Thomas Gottron and Steffen Staab**
Institute for Web Science and Technologies
University of Koblenz–Landau
{jpreusse, kunegis, thimm, gottron, staab}@uni-koblenz.de

**Abstract**

We investigate the structural patterns of the appearance and disappearance of links in dynamic knowledge networks. Human knowledge is nowadays increasingly created and curated online, in a collaborative and highly dynamic fashion. The knowledge thus created is interlinked in nature, and an important open task is to understand its temporal evolution. In this paper, we study the underlying mechanisms of changes in knowledge networks which are of structural nature, i.e., which are a direct result of a knowledge network's structure. Concretely, we ask whether the appearance and disappearance of interconnections between concepts (items of a knowledge base) can be predicted using information about the network formed by these interconnections. In contrast to related work on this problem, we take into account the disappearance of links in our study, to account for the fact that the evolution of collaborative knowledge bases includes a high proportion of removals and reverts. We perform an empirical study on the best-known and largest collaborative knowledge base, Wikipedia, and show that traditional indicators of structural change used in the link analysis literature can be classified into four classes, which we show to indicate growth, decay, stability and instability of links. We finally use these methods to identify the underlying reasons for individual additions and removals of knowledge links.

## 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, this paper 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. 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.

We begin in Section 2 by reviewing collaborative knowledge networks and giving an overview of related models and prediction methods. In Section 3, we state our model, and perform our experiments on Wikipedia datasets in Section 4.



[1] http://en.wikipedia.org/

[2] http://en.wikipedia.org/wiki/Wikipedia:Wikipedians

[3] http://www.diseasesdatabase.com/

[4] See Table 3 in this paper

Section 5 reviews works related to this one. We conclude in Section 6.

## 2 Background

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.

### 2.1 Link Analysis

The field of link analysis has primarily focused on social networks and has lead to a variety of social interaction theories. Balance theory (Heider 1958) states that people tend to align their preferences with others. Synthesizing this idea, Granovetter (1973) asserts the *strength of weak ties* which further develops the concept of *triadic closure*. 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. Exchange theory (Garlaschelli and Loffredo 2004) proposes that individuals choose to form the relationship they expect to profit from the most, or to have the lowest cost. According to this theory, individuals will stick to these relationships if they are rewarded and no other relationships provide better opportunities at lower costs. These theories suggest that individual relationships are driven by some amount of reciprocity, and thus unreciprocated edges dissolve more readily and are observed less often in the network. Lazarsfeld and Merton (1954) introduced the concept of *homophily* which states that individuals are likely to bond with others that are similar to themselves. Studies on social networks show that the disappearance of ties is influenced by several factors. In particular, homophily, reciprocity and the embeddedness of a tie in a larger group of well-connected people have positive effects on the persistence of a tie (Martin and Yeung 2006; Burt 2000), and the likeliness of decay goes down with the age of the tie and the age of actors, an effect coined *liability of newness*.

The field of network analysis is less researched for hyperlink networks than for social networks. The social sciences have considered hyperlink networks as a special case of social networks (Park 2003). From the perspective of computer science, the focus has been on applications for information retrieval, and particularly on ranking Web pages by popularity using Brin and Page's PageRank (1998) and Kleinberg's HITS (1999). These two algorithms have also been applied to the Wikipedia hyperlink network (Bellomi and Bonato 2005).

### 2.2 Network Evolution Models

In order to describe the effects that lead to the structure of real-world networks, different network evolution models have been proposed. Preferential attachment (Barabási and Albert 1999) and assortative mixing (Newman 2002) are structural theories about the way actors in a network pick other actors to bond with. Preferential attachment suggests that the likelihood of a node to form new links is proportional to its degree (the number of its neighbors), referred to as the "rich get richer" phenomenon. On the other hand, assortative mixing states that nodes are more likely to form links with nodes of similar degree. Several graph growth models include link disappearance in addition to link creation, for instance in a model to explain power laws (Akkermans 2012). Other examples can be found in (Eppstein and Wang 2002) and (Kleinberg et al. 1999), in which a model for growth of the Web is given in which edges are removed before others are added. While these methods succeed in predicting global characteristics of networks such as the degree distribution, they do not model the structure of the network, and thus cannot be used for predicting individual links.

### 2.3 Predicting Addition and Removal of Links

The problem of predicting the appearance of links in networks has received substantially more attention than the problems of predicting their disappearance. Recent surveys on the structural link addition prediction problem are provided by Liben-Nowell and Kleinberg (2007) 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. Other algorithms for links prediction include the index of Katz (1953), graph kernels (Ito et al. 2005) and diffusion models (Kondor and Lafferty 2002).

Work on the disappearance of links in networks has focused on social networks and on explaining why users on the social networking platforms Facebook and Twitter unfriend or unfollow each other. Kwak, Moon, and Lee (2012) have used structural features of the Twitter follower-followee relationship to ascertain when users decide to unfollow others.

Their findings suggest that ties persist when a user is acknowledged by its followee or when the users share followers and followees. An analysis of the unfriending behavior in Facebook by Quercia and colleagues (2012) found a correlation between personal traits and user information, and the likelihood to end a friendship on Facebook. The authors of that study found that friendships between neurotic or introverted users and others as well as between people who differ greatly in age are more likely to break, while well-embedded friendships or friendships sharing a common female friend are more robust.

As these studies use very specific user information, e.g., personality traits or 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. Furthermore, neither of both works use temporal features such as the age of a tie, due to the unavailability of this information.

# 3 Modeling Structural Changes in Knowledge Networks

Formally, 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.

Users in collaborative knowledge networks can edit the text inside knowledge items, as well as the links between them, by either removing or adding connections. We will assume that the semantic knowledge is captured in the links between knowledge items, and will thus only consider changes to the links, as well as the time of changes in the text, disregarding the actual changes in the content. Accordingly, we consider the following three types of events:

- **Add:** A link is added.
- **Delete:** An existing link is removed.
- **Update:** A knowledge item is changed textually.

We assume that a timestamp is given for each event.

## 3.1 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 1:

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

Table 1: 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.

- **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.

## 3.2 Features

A large number of features for predicting link appearance and disappearance can be found in the literature (Liben-Nowell and Kleinberg 2007; 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)-(f) cover known models from the literature. We introduce an additional hypothesis (e) exploiting additional information available for knowledge networks based on update and delete events. 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 (Burt 2000).
***Hypothesis:*** *The old age of an edge or a node are good indicators for link persistence.*

**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.

**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 $j$.

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

**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.

**Node freshness:** We define $nFresh(i)$ as the freshness of node $i$, denoting the last time that any event related to node $i$ occurred.

### (e) Instability

We consider a node as unstable if its content or its incident edges are frequently changed.
***Hypothesis:*** *The less stable nodes $i$ and $j$ are, the less stable the link $(i,j)$ is, whether present or not.*

**Update degree:** We define $d^U(i)$ to be the number of times that node $i$ has been updated, or equivalently, the number of update events for node $i$. Changes in the link structure are not counted.

**Node deletion coefficient:** We define node $i$'s add-degree $d^+(i)$ as the number of edges, going to or from node $i$, that have been added during the whole evolution of the network, and analogously the delete-degree $d^-(i)$. We further define the node deletion coefficient $ndc(i)$ as the regularized[5] ratio of the delete-degree and the add-degree of a node $ndc(i) = (d^-(i)+1)/(d^+(i)+1)$. This ratio summarizes how effective adds of node $i$ have been; if all links to or from $i$ were removed during the evolution we have $ndc(i) = 1$, and lower values indicate that a higher fraction of links persist.

We summarize the features and the expected behavior with respect to the predictability of new links and link removals in Table 2.

## 4 Experiments

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

### 4.1 Datasets

In our evaluation we use the largest dynamic knowledge network on the Web, Wikipedia. We use the directed article-hyperlink networks of four of the five largest[6] Wikipedias. We skip the largest one, the English Wikipedia, due to its size and limited computational resources.[7] 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. All datasets are available online as part of the Koblenz Network Collection (Kunegis 2013).[8]

For each of the four Wikipedias we consider all add, delete and update events until August 2011. An overview of the datasets is given in Table 3. The French Wikipedia is the biggest dataset by number of articles with around 1.8 million articles. During its whole evolution 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

[5] We regularize the ratio by adding one to the nominator and to the denominator.

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

[7] The evaluation of this dataset is ongoing.

[8] https://west.uni-koblenz.de/Research/DataSets/wikipedia-hyperlink/

| Model | Feature | | New links | Link removal | Expected state |
|---|---|---|---|---|---|
| Preferential attachment | Node degree | $d$ | $\nearrow$ | – | Growth / instability |
| Preferential attachment | Joint degree | $jd$ | $\nearrow$ | – | Growth / instability |
| Embedding | Common neighbors | $CN$ | $\nearrow$ | $\searrow$ | Growth |
| Embedding | Paths of length three | $P3$ | $\nearrow$ | $\searrow$ | Growth |
| Reciprocity | Back-links | $back$ | $\nearrow$ | $\searrow$ | Growth |
| Liability of newness | Edge age | $eAge$ | – | $\nearrow$ | Decay / instability |
| Liability of newness | Edge freshness | $eFresh$ | – | $\nearrow$ | Decay / instability |
| Liability of newness | Node age | $nAge$ | – | $\nearrow$ | Decay / instability |
| Liability of newness | Node freshness | $nFresh$ | – | $\nearrow$ | Decay / instability |
| Instability | Node deletion coefficient | $ndc$ | ($\nearrow$) | ($\nearrow$) | Instability |
| Instability | Update degree | $d^U$ | ($\nearrow$) | ($\nearrow$) | Instability |

Table 2: Summary of hypotheses about the ability of features to predict link addition and removal. "$\nearrow$" indicates a positive correlation; "$\searrow$" indicates a negative correlation. Novel hypotheses are shown in parentheses.

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

all link operations, thus accounting for a large part of structural changes. As shown in Figure 1, half of the edges are more than 23 months old.

### 4.2 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 the prediction of new edges $E^+$ and the prediction of deleted edges $E^-$.

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

such that

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

The sets of added and removed links are illustrated in Figure 2. The problem of predicting new links $E^+$ is called the link addition prediction problem, or simply the link prediction problem (Liben-Nowell and Kleinberg 2007). Typically, the link 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

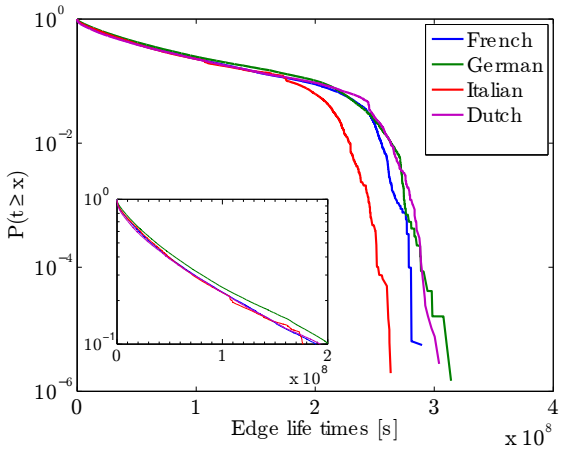



Figure 1: The decay of edges in the four studied Wikipedias.

define a 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 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 test set $E^-$, from those that are not removed, in the false test set $E^-_{false}$. The set $E^-_{false}$ is thus defined as

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

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 $t_2$

$$E^+_{false} \subset V \times V \setminus E_{t_1} \setminus E_{t_2},$$
$$|E^+_{false}| = |E^+|.$$

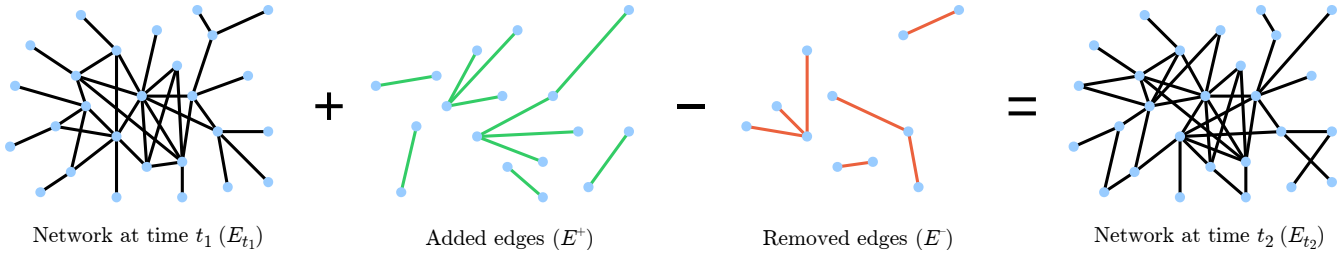



Figure 2: Schematic representation of the link addition and removal process. At time $t_1$, the network has the edge set $E_{t_1}$. After $t_1$, the set of edge $E^+$ is added and the set $E^-$ is removed, giving the set of edges $E_{t_2}$ at time $t_2$. Link directions are not indicated in the figure.

| **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 4: The size of our link addition and link removal test sets for the four Wikipedias we consider.

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 4 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 1. Link addition prediction functions (link removal prediction functions) can then be evaluated and compared.

## 4.3 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). In the following, we describe the ROC curve for link addition prediction; the definition is analogous for link removal prediction.

Let $f$ be a prediction function. All node pairs in the combined true and false test set $E^+ \cup E^+_{false}$ 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 is then defined as the area under the ROC curve, and is thus 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 us to build a measure of decay by negating a measure of growth.

## 4.4 Results

We compute all eleven features described in Section 3.2 and compute the AUC values of the link addition and removal prediction tasks. Figure 3 shows the performance of the features at the task of link addition and removal prediction for all studied datasets. Table 5 shows the top-three performing features for each of the four classes.

In the following, we compare our results with the projections of the hypotheses from Section 3.2.

### (a) Preferential Attachment

***Hypothesis:*** *The number of adjacent nodes is a good indicator for link addition.*

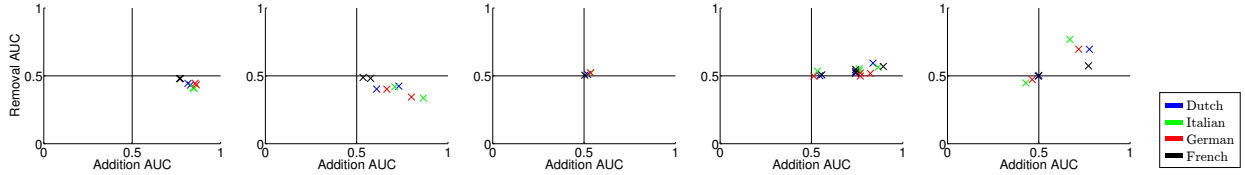
Following the hypothesis, we expect a good link addition prediction performance for features of preferential attachment. Figure 4(a) shows the AUC values for the two preferential attachment features. Our experiments show that preferential attachment features are indeed good indicators for the formation 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*

| **Decay** | | **AUC** | **Instability** | | **AUC** |
|---|---|---|---|---|---|
| Low node degree | $-d$ | 0.70 | High proportion of deletions | $ndc$ | 0.71 |
| Low joint degree | $-jd$ | 0.69 | Nodes have been changed recently | $nfresh$ | 0.71 |
| Few paths of length three | $-P3$ | 0.67 | Old edge | $eAge$ | 0.65 |
| **Stability** | | **AUC** | **Growth** | | **AUC** |
| Low proportion of deletions | $-ndc$ | 0.71 | High node degree | $d$ | 0.70 |
| Nodes have been unchanged for long | $-nfresh$ | 0.71 | High joint degree | $jd$ | 0.69 |
| Young edge | $-eAge$ | 0.65 | Many paths of length three | $P3$ | 0.67 |

Table 5: 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.




(a) Preferential attachment (b) Embedding (c) Reciprocity (d) Liability of newness (e) Instability

Figure 4: Link addition prediction and link removal 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 addition and removal 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.

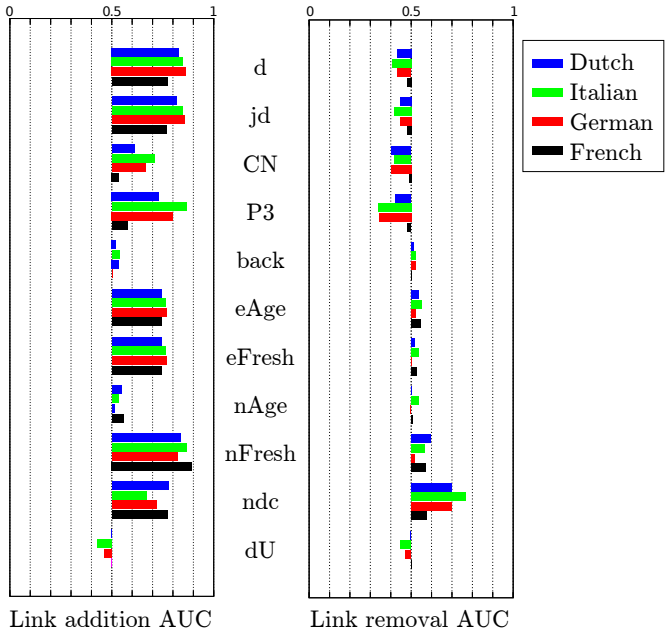



Figure 3: AUC values for the link addition prediction and link removal prediction tasks are shown for all features and all four datasets. Note that a below-random AUC value can be turned into an above-random one by the negation of the respective feature.

*existing links, i.e., it is an indicator for growth.*
Following the hypothesis, we expect a good link addition prediction performance and a bad link removal prediction performance for features of embeddedness. Figure 4(b) depicts the AUC values for this feature for link versus link removal 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 addition prediction performance and a bad link removal prediction performance. We depict the results for the binary feature of back-link *back* in Figure 4(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 old age of an edge and a node are good indicators for link persistence.*

Following the hypothesis, we expect a good link removal prediction performance for these features. Our findings shown in Figure 4(d) suggest that the age of a node is neither a good indicator for the formation of new links nor for the deletion of links. On the other hand, the three other features are good indicators for both the formation of new links and the removal of links. However, these features have a better performance for link addition prediction. Therefore the other three features are indicators of *instability*. In terms of knowledge networks, this implies that new knowledge is fragile, while established knowledge is more stable.

### (e) Instability

***Hypothesis:*** *The less stable nodes $i$ and $j$ are, the less stable the link $(i, j)$ is, whether present or not.*

Following the hypothesis, we expect a good link addition prediction performance and a good removal prediction performance for these features. Our findings shown in Figure 4(e) and Figure 3 suggest that the number of updates $d^U$ neither works well for link addition prediction nor for link removal prediction. A reason for the low predictive ability of $d^U$ may be the fact that the majority of updates do not change the content semantically but only typographically or orthographically, and thus the link structure remains unchanged. We have proposed the *node deletion coefficient* of a node, i.e., the ratio of the number of delete events to add events, as a second feature to ascertain the instability of a node. This feature is correlated positively with the prediction of new links as well as with the prediction of link disappearance (see Figure 3). If relatively few edges are removed, then a node is also unlikely to form new links. Therefore this feature is an indicator of instability. If we interpret a high number of edge deletions around a knowledge item as a repositioning of this item, then our results imply that relations to connected knowledge items will be affected as well.

### 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 addition prediction can be solved to a much higher accuracy ($\mathrm{AUC} \approx 0.90$) than the link removal prediction problem ($\mathrm{AUC} \approx 0.75$).

On the level of the four different classes of prediction problem which generalize the link addition and removal prediction problem. We observe that the problem of growth prediction can be solved well using embedding indicators (see Figure 4(b)), as can the instability prediction problem (see Figure 4(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 addition prediction, the features usually considered are not evaluated on the task of link removal prediction. Link removal 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 2007; Lü and Zhou 2011).

## 5 Related Work

In Section 2 we already revisited works that have been used as a foundation for our investigation on relationships of measures for link addition and link removal in knowledge networks. In the following, we discuss works on related *problem types* that are similar, but not identical to the prediction problem discussed in this paper.

**Link Decay** In many networks, links cannot be removed but are rather considered to become *inactive* or to *decay*. Research on predicting decay in mobile phone communication networks (Raeder et al. 2011; Hidalgo and Rodriguez-Sickert 2008) thus assumes that links decay if no communication was exchanged between the actors for a particularly chosen time period. Both works conclude that links are more likely to persist when the connection is reciprocated and when either both actors' degrees are lower or both high. Raeder and colleagues (2011) find that the "liability of newness" holds, i.e., the age of the tie is correlated with the persistence of the tie. However, this line of research deals with derived link removals as the datasets themselves do not contain explicit unlinks.

**Declining Participation** The decay of groups in social networks is studied in (Kairam, Wang, and Leskovec 2012), explaining it by interaction patterns. Another related phenomenon is called *churn*, describing the situation in which a user quits a social community. Churn can be modeled as the deletion of an edge between the user and the service, and thus corresponds to the deletion of edges in a bipartite graph (Karnstedt et al. 2010). Both problems are fundamentally bipartite, since they act on the network connecting users with items.

**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). Similarly, the disconnection of nodes has been predicted in mobile ad-hoc networks (De Rosa, Malizia, and Mecella 2005). These problems are structurally similar to the problem studied in this paper, but do not use features that are typical for link addition prediction such as the degree of nodes or the number of common neighbors.

**Citation Analysis** Another type of knowledge network is the citation network, i.e., scientific publications 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 that which adds 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 vertices 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.

## 6 Conclusion

In order to answer the question given at the beginning of this paper, 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. Since the methods presented in this paper work on a purely structural level, we propose that they can be used for analyzing the dynamics of other networks types, too.

## Acknowledgments

The research leading to these results has received funding from the European Community's Seventh Frame Programme under grant agreement n° 257859, `http://robust-project.eu/`ROBUST.

## References

Akkermans, H. 2012. Web dynamics as a random walk: How and why power laws occur. In *Proc. Web Science Conf.*, 1–10.

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

Bellomi, F., and Bonato, R. 2005. Network analysis for Wikipedia. In *Proc. Wikimania*.

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*.

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, S., and Page, L. 1998. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems* 30(1–7):107–117.

Burt, R. 2000. Decay functions. *Social Networks* 22(1):1–28.

De Rosa, F.; Malizia, A.; and Mecella, M. 2005. Disconnection prediction in mobile ad-hoc networks for supporting cooperative work. *IEEE Pervasive Computing* 4(3):62–70.

Eppstein, D., and Wang, J. 2002. A steady state model for graph power laws. In *Proc. Int. Workshop on Web Dynamics*.

Garlaschelli, D., and Loffredo, M. I. 2004. Patterns of Link Reciprocity in Directed Networks. *Phys. Rev. Lett.* 93:268701.

Granovetter, M. S. 1973. The strength of weak ties. *American Journal of Sociology* 78(6):1360–1380.

Guimerà, R., and Sales-Pardo, M. 2009. Missing and spurious interactions and the reconstruction of complex networks. *Proc. Natl. Acad. Sci. USA* 106(52):22073–22078.

Heider, F. 1958. *The Psychology of Interpersonal Relations*. New York: John Wiley & Sons.

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.

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*, 586–592.

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*, 673–682.

Karnstedt, M.; Hennessy, T.; Chan, J.; Basuchowdhuri, P.; Hayes, C.; and Strufe, T. 2010. Churn in social networks. In *Handbook of Social Network Technologies*. Springer. 185–220.

Katz, L. 1953. A new status index derived from sociometric analysis. *Psychometrika* 18(1):39–43.

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*, 1–17.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *J. ACM* 46(5):604–632.

Kondor, R., and Lafferty, J. 2002. Diffusion kernels on graphs and other discrete structures. In *Proc. Int. Conf. on Machine Learning*, 315–322.

Kunegis, J. 2013. KONECT – The Koblenz Network Collection. In *Proc. Int. Web Observatory Workshop*.

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*, 499–502.

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., eds., *Freedom and Control in Modern Society*. New York: Van Nostrand. 18–66.

Liben-Nowell, D., and Kleinberg, J. 2007. The link-prediction problem for social networks. *J. of the American Soc. for Information Science and Technology* 58(7):1019–1031.

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, J. L., and Yeung, K.-T. 2006. Persistence of close personal ties over a 12-year period. *Social Networks* 28(4):331–362.

Newman, M. E. J. 2002. Assortative mixing in networks. *Physical Review Letters* 89(20):208701.

Park, H. W. 2003. Hyperlink network analysis: A new method for the study of social structure on the Web. *Connections* 25(1):49–61.

Quercia, D.; Bodaghi, M.; and Crowcroft, J. 2012. Loosing 'friends' on Facebook. *Proc. Web Science Conf.* 251–254.

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.

Zeng, A., and Cimini, G. 2012. Removing spurious interactions in complex networks. *Phys. Rev. E* 85:036101.