Evaluating and Improving Navigability of Wikipedia: A Comparative Study of Eight Language Editions

Daniel Lamprecht
KTI, Graz University of Technology
Graz, Austria
daniel.lamprecht@tugraz.at

Dimitar Dimitrov
GESIS - Leibniz Institute for the Social Sciences
Cologne, Germany
dimitar.dimitrov@gesis.org

Denis Helic
KTI, Graz University of Technology
Graz, Austria
dhelic@tugraz.at

Markus Strohmaier
GESIS - Leibniz Institute for the Social Sciences
and University of Koblenz-Landau
Cologne and Koblenz, Germany
markus.strohmaier@gesis.org

ABSTRACT

Wikipedia supports its users to reach a wide variety of goals: looking up facts, researching a topic, making an edit or simply browsing to pass time. Some of these goals, such as the lookup of facts, can be effectively supported by search functions. However, for other use cases such as researching an unfamiliar topic, users need to rely on the links to connect articles. In this paper, we investigate the state of navigability in the article networks of eight language versions of Wikipedia. We find that, when taking all links of articles into account, all language versions enable mutual reachability for almost all articles. However, previous research has shown that visitors of Wikipedia focus most of their attention on the areas located close to the top. We therefore investigate different restricted navigational views that users could have when looking at articles. We find that restricting the view of articles strongly limits the navigability of the resulting networks and impedes navigation. Based on this analysis we then propose a link recommendation method to augment the link network to improve navigability in the network. Our approach selects links from a less restricted view of the article and proposes to move these links into more visible sections. The recommended links are therefore relevant for the article. Our results are relevant for researchers interested in the navigability of Wikipedia and open up new avenues for link recommendations in Wikipedia editing.

CCS Concepts

•Information systems → Web searching and information discovery; •Human-centered computing → Hypertext / hypermedia; Wikis;

OpenSym '16 August 17-19, 2016, Berlin, Germany

ACM ISBN 978-1-4503-4451-7/16/08.
DOI: http://dx.doi.org/10.1145/2957792.2957813

Keywords

Wikipedia, Navigability, Reachability, Bow Tie Model, Link Recommendations

1. INTRODUCTION

The multiple language editions of Wikipedia serve around 16 billion views per month as of 2016, with the English Wikipedia accounting for almost half of them [1]. Visitors of the free encyclopedia pursue a wide range of goals: looking up specific facts, learning about a topic of interest, making an edit, or simply browsing to pass time. Generally, the goals of users in retrieving information on the Web can be classified into three different ways [34]:

1. lookup of a specific object or fact,
2. search for something that cannot be explicitly described but will be recognized once retrieved, and
3. serendipitous (accidental) discovery.

Wikipedia supports these three ways of information retrieval by different means. The lookup of specific facts (or articles) is generally satisfiable with the Wikipedia-internal search engine or an external Web search engine. The other two ways, however, are not as well-supported by search engines. Users therefore need to rely on the hyperlinks that join the vast number of separate pieces of knowledge on Wikipedia in order to reach their goals. There exists several types of links to support users in connecting articles: links in the running text, links in tabular summaries such as infoboxes and links to groupings of articles such as categories or portals.

For the English Wikipedia, out of $3,279,134,602$ visits to articles collected in a clickstream in February 2015 [43], 30% were referred to from other Wikipedia articles, and 70% from external sources. This suggests that external links to Wikipedia, such as search engines, play an important role in referring visitors to the encyclopedia. However, it also shows that 30% of all clicks relied on links within Wikipedia and this implies that navigation plays a vital role for users.

On the Web, some users have been found to prefer the incremental process of navigation to direct retrieval, even

[1] http://stats.wikimedia.org/EN/TablesPageViewsMonthlyCombined.htm

Figure 1: **Navigational Views of a Wikipedia Article.** We investigate five different navigational views of Wikipedia articles: the unrestricted view of all links, the view of the links in the lead (shown with a blue border), the first lead paragraph (shown with a red border), the infobox (shown with a green border) and the very first link in the article (shown with brown background). These views enable us to understand the effects of limiting the number of links on the navigability of the article network.

when knowing what they are looking for [33]. Users also enjoy browsing without specific objectives, for example in recommender systems [12] or on entertainment websites such as YouTube [5]. Navigating articles via links is thus a vital component of Wikipedia even in the presence of powerful search engines. This is also manifested in the detailed description of the rules for placement and linking within Wikipedia's manual of style.

Problem. Recent research has shown the majority of the attention of Wikipedia users is focused on the areas located near the top, namely the lead section and the infobox [7, 19]. While previous work has already provided us with insights into the network structure of Wikipedia [4, 8], little is known about the effects of viewport restriction to, for example, the lead section. Understanding the impact of viewport restriction on the navigability of Wikipedia's article network would allow us to better understand and support the needs of users that only read parts of articles. We therefore investigate the following research questions.

RQ1. What is the state of navigability across multiple language versions of Wikipedia?

RQ2. How can this analysis be exploited to suggest improvements to navigability?

Approach. To assess navigability we make use of elements of a framework to evaluate navigability introduced in the context of recommender systems [18, 20]. We use this framework to study the navigability of eight language versions of Wikipedia based on the bow tie model, which partitions a network based on reachability criteria. We investigate five different navigational views constructed by taking into account a subset of all links in articles: (i) the unrestricted view of all links, (ii) the view of the links in the lead section, (iii) the view of the first lead paragraph, (iv) the view of the infobox and, finally, (v) the view of the very first link. Figure 1 shows an example of the areas used for these views.

Based on the results of this analysis, we then propose a method to improve navigability in Wikipedia by recommending specific links from articles to move into the more visible regions of an article or otherwise emphasize. We demonstrate our method by applying it to two Wikipedia versions and showing the effects.

Contributions. Our main contribution is a comparative study of navigability in terms of components and mutual reachability for the eight largest Wikipedia language versions. We show that for a restricted views of articles, only a relatively small share of the articles in Wikipedia is reachable by following links. Based on this analysis, we then propose a method to improve reachability in the article networks by suggesting specific links that could be emphasized, such as being moved to sections that users pay more attention to.

2. RELATED WORK

Navigation in Networks. Stanley Milgram's small world experiments [25, 36] were the first notable study of navigation in networks and investigated decentralized search in the social network of the United States. Participants were provided with a short description of a target person and were asked to forward the letter to a first-name acquaintances with the goal of reaching the target person. The striking result of the experiments was that participants were able to find very short paths to the target person across the social network of the entire United States.

As a result, the enabling property for efficient wayfinding in a network was named the small world property. Many networks that emerge in nature belong in fact to this class of networks. Watts and Strogatz proposed a generative model for small world networks based on rewiring of a ring lattice [39]. These models were subsequently extended based on networks with nodes organized in two-dimensional grid lattices and hierarchies [16, 38]. Kleinberg identified the properties that made these networks efficiently navigable with decentralized search algorithms [15].

The model of decentralized search was then applied to model human navigation in information networks such as Wikipedia, and in folksonomies [11, 10, 21, 35].

For Wikipedia, human navigation was extensively studied based on Wikipedia games. In these navigational games, players are challenged to reach a target article by following links in the text and not using the search function. Based on log files from these games, researchers have studied goal-directed navigation with explicitly specified target articles. Players are, in general, very efficient in finding targets on Wikipedia and use high-degree hubs as landmarks in navigation [40]. The resulting paths have been found useful to compute semantic relatedness of articles [30].

Table 1: **Datasets.** We used the eight Wikipedia language versions with the highest number of edits for this work.

Language	Articles	Edits
English	5,072,214	812,170,986
German	1,905,450	156,204,159
French	1,721,902	125,368,222
Spanish	1,232,123	94,262,365
Russian	1,287,687	88,544,149
Italian	1,251,650	83,993,233
Japanese	1,001,180	59,504,158
Dutch	1,854,708	46,955,102

Wikipedia Network Analysis. The article network of Wikipedia has been found to grow by preferential attachment, with most of the articles are contained in a (mutually reachable) strongly connected component [4]. In comparison to general Web pages, Wikipedia is more densely linked and contains a larger strongly connected component [14]. Authors contributing to different areas in the network have been found to support knowledge integration [8].

Ibrahim et al. have studied the structures and cycles emerging in the network of consisting only of the first links of the English Wikipedia [13]. The same first-link network has also been used to automatically categorize articles in Wikipedia based on its *is-a* descriptions and been shown to lead to categories with high precision when compared to human-curated categories [2].

Wikipedia Link Suggestions. The problem of improving the link structure of an information system has been studied in the context of social and information networks cast both as a link formation (i.e., link creation) [6, 22, 23, 32] and as a link removal problem [17, 29].

For Wikipedia, a number of approaches have been proposed: extracting potential links from text based on keyword extraction and word sense disambiguation [24], machine learning [26], factorization of the adjacency matrix [42], clustering documents [1] and mining navigational traces [41, 28]. In addition, there have been several approaches to suggest crosslingual links [27, 31, 37, 44].

3. MATERIALS AND METHODS

3.1 Datasets

For this work, we investigate the navigational properties of the top eight language editions of Wikipedia in terms of the number of edits. While there exist language versions containing a larger number of articles than the ones we investigate (e.g., the Swedish Wikipedia contains almost three million articles and the Cebuano Wikipedia contains almost two million), these versions have comparatively few edits and a large share of the articles are bot-created stubs. We hence restrict our analysis to Wikipedia language versions where the majority of edits is done by human editors. Table 1 shows the languages, numbers of articles and numbers of edits for the eight versions.

For all language versions, we used the articles in the version present at February 3, 2016, based on the Wikipedia IDs in the dump from that date. We obtained the HTML pages for the articles corresponding to all IDs from the Wikipedia API [2]. This had two distinct advantages over the XML dump containing Wiki Markup: First, it allowed us to view all templates in their resolved forms, which otherwise would be a very cumbersome to achieve. Second, it permitted us to resolve redirects by using the information contained in the API response, rather than relying on the (incomplete) redirect list that is part of the official Wikipedia dumps. In addition to the articles themselves, we used the page view count information for the entire month of January 2016 [3].

3.2 Navigational Views

After having downloaded all articles, we parsed the HTML files and extracted all links. We then constructed the article networks for all Wikipedia language versions we investigated. To this end, we regard all articles as nodes of a network and insert all links between pages as directed edges. We restrict the link analysis to those links occurring in the article itself (i.e., links in the text, in tables, divs, etc. within the content part) and exclude the remaining links, such as links to categories, links in the menu on the left side or links to external websites.

Previous research has shown that users of Wikipedia dedicate a large proportion of their attention to the top of articles (namely to the lead section and the infobox [7, 19, 28]). To better understand the implications of these behaviors, we investigate five distinct *navigational views* of Wikipedia:

1. **Entire Article.** This view represents the links from the entire article, including those in the lead section and any tables (such as infoboxes). This view shows how users navigate if they consider the links from the entire article.

2. **Entire Lead.** The links in the lead section receive a large share of user attention [19]. As the table of contents after the lead is by default expanded on Wikipedia, this presents an obstacle to users and frequently requires scrolling to read the first section.

3. **First Lead Paragraph.** This view comprises of all links in the first paragraph of the lead. This is similar to the excerpt shown by a Google search result for a search term. While the excerpt does not highlight the links themselves, the information contained in it represents what users can take away from it. If users are interested to learn more based on the excerpt, they might look into the Wikipedia articles for corresponding concepts.

4. **Infobox.** Infoboxes are tabular representation of the most important facts of an article and are present for 40% of articles on the English Wikipedia (and between 32 and 69% for the eight Wikipedia versions we investigated). Limiting the view to links contained in infoboxes represents users that look only at tabular information and the key facts of an article.

5. **First Link.** This view is restricted to the very first link that is not in parentheses, italics, or contained in a table. In a sense, the set of articles reachable this way represents the backbone concepts of a Wikipedia version.

[2] https://wikipedia.org/w/api.php

[3] https://dumps.wikimedia.org/other/pagecounts-raw/2016/2016-01

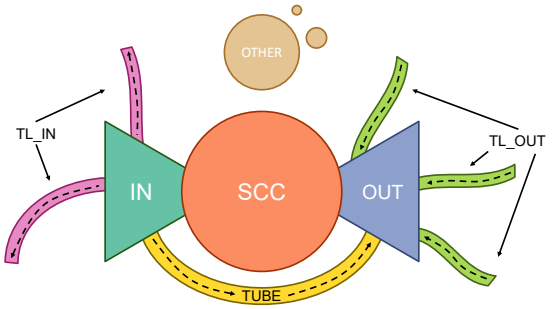

Figure 2: **Bow Tie Model.** For a directed network, the bow tie model [3] defines the largest strongly connected component (*SCC*) as the largest set of mutually reachable nodes. Nodes in the *IN* component all have an outgoing path leading to the *SCC*. Nodes in *OUT* are reachable from *SCC* but not the other way around. Note that *IN* and *OUT* are in general not strongly connected components themselves but consist of multiple components with the same reachability characteristics. In addition to these three main components, the bow tie model defines disconnected components (*OTHER*), *TUBES* (which connect *IN* to *OUT*) and tendrils leading away from *IN* or into *OUT*.

3.3 Bow Tie Analysis

To analyze the structure and connectivity of the article networks, we make use of the bow tie model. This model was proposed to study the structure of the Web graph [3] and took its name from the resemblance of a bow tie. Figure 2 shows the structure of this model. For a directed network, it defines the largest strongly connected component (*SCC*) as the largest set of mutually reachable nodes. Nodes in the *IN* component all have an outgoing path leading to the *SCC*. Nodes in *OUT* are reachable from *SCC* but not the other way around. Note that *IN* and *OUT* are in general not strongly connected components themselves but consist of multiple components with the same reachability characteristics. In addition to these three main components, the bow tie model defines disconnected components (*OTHER*), *TUBES* (which connect *IN* to *OUT*) and tendrils leading away from *IN* or into *OUT*.

Navigability of a Wikipedia article network measures the extent to which articles can be reached by following links. An important metric for navigability is hence the size of the largest strongly connected component (*SCC*), which measures the number of mutually reachable articles. We use the bow tie model to study navigability in the following way: Firstly, we study the membership of articles to partitions and the sizes of these partitions (in particular the size of the *SCC*). Secondly, we make use of the flow information contained in the model: For example, for all articles in *IN*, there exists a path to a node in the *SCC*. This allows us to investigate one-way reachability in the article network.

4. EVALUATING NAVIGABILITY

We evaluate the bow tie structure of the article networks and the navigational views with a membership change analysis, shown in Figure 3. The sizes of the components represent the percentage of articles contained in them, and the transition of node membership between navigational views is shown with connections between the partitions.

Table 2: **Correlation Analysis.** This table shows the correlation of the *SCC* sizes for all eight Wikpedia language versions with the number of edits and the median outdegree. The correlation was computed with Spearman's ρ, and ** denotes a p-value ≤ 0.05. The median outdegree strongly correlates with the *SCC* size when taking the entire articles into account. This implies that longer articles are correlated with a larger *SCC*. The number of edits correlates with the *SCC* size for the navigational view of the first lead paragraph, which suggests that a large number of edits introduces more navigable links into the lead section.

Navigational View	# Edits	Median Outdegree
First Lead Par.	0.74^{**}	0.48
Entire Lead	0.60	0.00
Infobox	0.12	-0.11
Entire Article	0.07	0.83^{**}

4.1 Entire Article

When taking the entirety of the links of Wikipedia articles into account, the *SCC* covers the vast majority of the articles. For the English Wikipedia, the *SCC* contains 94% of all articles, and for the remainder of language versions the coverage is between $87-97\%$, with the only exception being the Dutch version, for which only 63% of articles are contained in the *SCC*. A small share of articles for each investigated Wikipedia furthermore belongs to the *IN* component. These are articles that have outgoing paths into the *SCC*, but cannot be reached from it. Frequently, these are very short articles: For example, for the English Wikipedia the articles in *IN* have a median 6 links, while those in the *SCC* have a median 42. This indicates that articles in the *SCC* are substantially longer than those in *IN*.

Assuming that visitors carefully explore all links present at an article, these results imply that they could reach almost all of the articles on the encyclopedia by navigation. The coverage of articles by the *SCC* has notably increased since the earlier days of Wikipedia: A study of several language versions of Wikipedia from 2004 found that the *SCC* covered between 72 and 89% of articles for the Italian, Spanish, French, German and English Wikipedias and 67% for the Portuguese Wikipedia [4]. In addition, the number of Wikipedia articles has vastly increased since then (e.g., the English Wikipedia has grown from 340k to 5M articles).

4.2 Entire Lead, First Lead Paragraph, and Infobox

When restricting the navigational view to links occurring in the lead section, the sizes of the *SCC* drop to a range of 16% (Dutch) to 37% (Italian). This implies that for visitors not looking *below the fold* (which for Wikipedia mostly implies going further than the lead section and the table of contents), the share of mutually reachable articles in the network shrinks to $20-40\%$ of the *SCC* that is available for all links in the articles.

For the links in the first lead paragraph, the *SCC* sizes range between 3% (Dutch) and 7% (English). This implies that Wikipedia becomes effectively unnavigable for this view, except for very few concepts. The transitions of articles between the partitions of the bow tie model reveal that most of the articles that are in the *SCC* for the un-

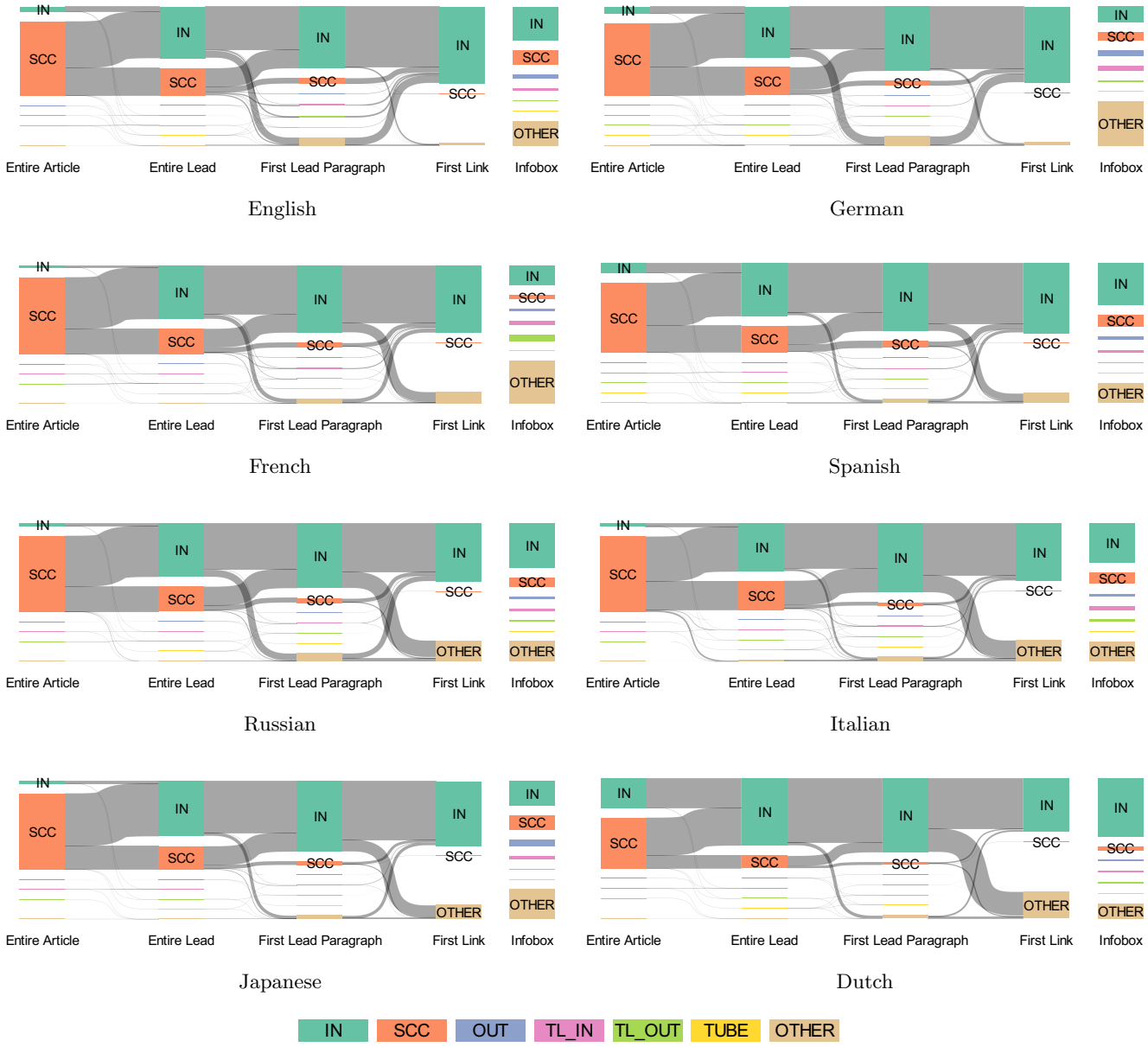

Figure 3: **Bow Tie Membership Change Analysis.** The figure shows the transitions from the unrestricted navigational view containing all articles to more restricted views. Colors and labels correspond to the ones used in Figure 2. The leftmost view (entire article) contains the second view (entire lead) and the membership transitions are shown between the states. For the unrestricted view of all links, the large majority of all articles are mutually reachable in the *SCC*. Restricting the navigational views to include fewer links decreases the size of the largest strongly connected component (*SCC*).

restricted navigational view of all links become part of the *IN* component when looking at links from the lead and the first lead paragraph. From a navigational perspective, this means that while fewer articles are mutually reachable, those articles still have outgoing paths leading to the *SCC*. Visitors looking at the lead section of these articles can therefore still reach the *SCC*. However, this navigation is necessarily one-way: Once in the *SCC*, the number of reachable articles is severely limited.

A possible explanation for this can be found in the guidelines for the lead section in Wikipedia. For example, for the English Wikipedia [4], these guidelines state that the lead section should summarize the most important aspects and provide links to more general articles. This likely restricts links from the less general articles to only point to more general ones. This in turn leads to the former becoming part of *IN* and the latter becoming part of the *SCC*.

A similar observation can be made for the infobox view. The *SCC* sizes for the corresponding navigational view range between 4% (Dutch) and 19% (Japanese). Like for the first lead paragraph, these reduced component sizes result in net-

[4] http://en.wikipedia.org/wiki/WP:Manual_of_Style/Lead_section

Table 3: **Cycles in the First-Link Networks of Wikipedias.** The first-link networks limit the navigational view to the first link that is not in a table, in parentheses or italics. The strongly connected components in this view are therefore cycles. The *IN* component shows the percentage of articles which eventually lead to a cycle when repeatedly following first links. The percentage after the first listed article states the size of the *IN* component for this article, excluding incoming links via the cycle. The articles in the cycles belong to very general topics and show the effect of the first sentence in articles frequently making use of an *is-a* relation. The article on philosophy is central to four of the eight investigated language versions.

Language	Size of *IN*	Article Titles (Translated)	Article Titles (Original)
English	97.0%	**Philosophy (92.1%)**, Existence, Ontology, Reality	Philosophy, Existence, Ontology, Reality
German	95.8%	**Philosophy (95.8%)**, World, Totality	Philosophie, Welt, Totalität
French	85.0%	**Philosophy (68.8%)**, Linguistics, Discipline (academia), Knowledge, Ancient Greek, Knowledge, Notion (philosophy), Greek language, Feature (linguistics), Isogloss, Centum and satem languages, Hellenic languages	Philosophie, Linguistique, Discipline (spécialité), Connaissance, Grec ancien, Savoir, Notion, Grec, Trait (linguistique), Isoglosse, Isoglosse centum-satem, Langues helléniques
Spanish	87.8%	**Psychology (87.7%)**, Profession, Activity, Specialization	Psicología, Profesión, Actividad, Especialización
Russian	73.7%	**Philosophy (58.9%)**, Mathematics, Science, Cognition, Object (philosophy), Set theory, Method (philosophy), Systematization, Objectivity (philosophy)	Философия, Математика, Наука, Познание, Объект (философия), Теория множеств, Метод, Систематизация, Объективность
Italian	73.2%	**Science (39.0%)**, Knowledge, Biology, Psychology, Psyche (psychology), Tissue (biology), Central nervous system, Brain, Nervous system, Awareness	Scienza, Conoscenza, Biologia, Psicologia, Psiche, Tessuto (biologia), Sistema nervoso centrale, Cervello, Sistema nervoso, Consapevolezza
Japanese	82.3%	**Person (82.3%)**, Interpersonal relationship	人間, 人間関係
Dutch	67.0%	**Knowledge (67.0%)**, Know-how	Kennis, Weten

works that allow only for a small fraction of mutually reachable articles. Again, there is a large number of articles in *IN* that can at least reach the *SCC*. The explanation for this is also similar as for the lead: as infoboxes state the key facts of articles, links to less general articles have a lower likelihood of being placed. There also exists a comparatively large fraction of articles in *OTHER*: These are the articles that do not possess an infobox and hence do not have any outlinks in this view.

In general, the difference in sizes of the *SCC* could be explained by the length of the article, as a longer text offers the possibility to include more links. To investigate this, we compute the correlation between the number of outlinks and the size of the *SCC* for all eight Wikipedia language version. We find that a significant correlation occurs only for the unrestricted view of the links in the entire article (see Table 2). A potential explanation for this is the restricted length of the lead section, the first paragraph and the infobox, which dampens the differences between long and short articles. We also find a strong correlation between the number of edits to a Wikipedia and the size of the *SCC* for the view of the first lead paragraph. This suggests that with an increasing number of edits, editors attach great importance to the links in the lead section, which as a result become more useful for navigational purposes.

4.3 First Link

The manual of style for the English Wikipedia states that the first sentence should give an easy-to-understand introduction, define the title, and put it in context [5] (other languages have similar guidelines). For example, in Figure 1, the first sentence defines Reykjavík to be the capital of Iceland and places the first link on the word for the country. In fact, many of the first links in the English Wikipedia are *is-a* relations. The tree structure created by these first links has been shown to lead to a category hierarchy with high precision compared to human-curated categories [2].

The navigational view of the first links is interesting from a theoretical perspective: According to popular belief, repeatedly following the first link in the English Wikipedia will eventually lead to the article on Philosophy [6]. To investigate the link structure of the first-link network, we look at its *SCC*. By definition, it consists of a cycle of articles (or a dead-end in the degenerate case). We then compute the number of articles in *IN*, which measures the number of articles from which navigation would end up at that cycle. For sake of clarity, we use the *SCC* with the largest corresponding *IN* component (and not the *SCC* containing the largest number of articles). Table 3 shows the results of this analysis. For all investigated Wikipedias, there is a single article in the cycle that accounts for more than half of the number of articles in its *IN* component, which confirms the navigational funnels identified by Ibrahim et al. [13]. For the English Wikipedia we indeed find that the vast majority of articles (97%) leads to the cycle containing the philosophy article. This finding also holds true for a large majority of articles in the German, French and Russian Wikipedias. For the Spanish and Italian Wikipedias, the dominant cycle contains the article on psychology, while for Dutch, the dominant cycle consists of the articles on knowledge and know-how. Interestingly, for the Japanese Wikipedia, the main cycle consists of the articles on person and interpersonal relationship. A possible explanation for this could be that this is an artifact of the importance of the status of relationships in the Japanese language, which uses an extensive range of honorifics for addressing conversational partners.

[5] https://en.wikipedia.org/wiki/WP:Redundancy

[6] https://en.wikipedia.org/wiki/WP:Getting_to_Philosophy

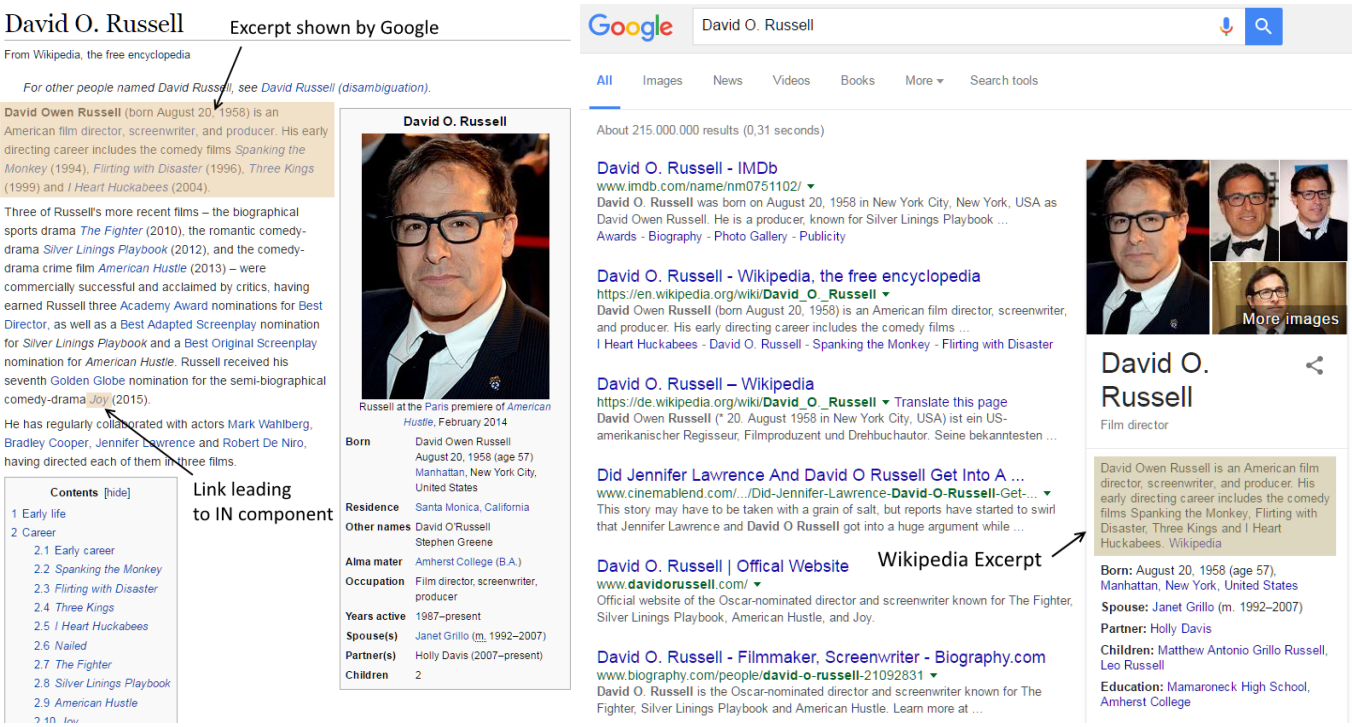

Figure 4: **Link Recommendation Example.** The English Wikipedia article on film director *David O. Russell* is in the largest strongly connected component (*SCC*) for the navigational view of the first lead paragraph. A Google search for the director shows a knowledge panel on the right side that shows the same first paragraph of the article as an excerpt. However, the first paragraph only includes information about *David O. Russell*'s early career. Moving the link to the film *Joy* to the first paragraph would not only better inform users of Google about the director but also add the article on *Joy*, which is in *IN*, to the *SCC* by closing a circle. Additionally, the article on *Joy Mangano*, which lies on a path from *Joy* to the *SCC* would also be added to the *SCC*. The two articles accounted for a total of 874,423 pageviews in January 2016.

5. IMPROVING NAVIGABILITY

The previous section has shown that restricted navigational views exert a strong influence on navigability. For visitors who only look at the links in a part of the article, navigability of the resulting article network is substantially reduced due to the reduced size of the set of articles in the *SCC*.

In this section, we propose an approach for improving navigability in an article network. Several methods already exist for augmenting the link structure of Wikipedia [24, 26, 41]. However, none of the approaches introduced so far considers restricted navigational views as criteria for navigability. The novel method we propose chooses additional links based on the bow tie structure of the network and takes the size of the *SCC* as well as the information about connectivity between the *IN*, *SCC*, and *OUT* components of the bow tie model into account.

Figure 4 shows an example of the application of our method. The main article shown is part of the *SCC*. The highlighted link points to an article in *IN*. Moving this link to the first lead paragraph would add it to the *SCC* by closing a circle to it, and also have the side effect of introducing a second article (that it links to) to the *SCC*. In what follows, we describe our method in more details.

5.1 Link Recommendation Approach

Our proposed approach of recommending links to increase navigability in the article networks makes use of the bow tie analysis. Generally, in order to add an article A to the mutually reachable set of articles of the *SCC*, two links are necessary: one link from A to an article of the *SCC* and a second link from an article in the *SCC* back to A. However, should one of the two links already be present, then adding the other type of link suffices. Moreover, there need not be a direct link from A to an *SCC* node: It is enough that there exist a path from A that reaches an *SCC* node.

The information about the links and paths is contained in the bow tie model: articles in *IN* have a path leading to the *SCC*, and articles in *OUT* are reachable from it. As the *IN* component largely dominates the *OUT* component in the Wikipedia article networks, we focus our attention on this component in what follows. However, the approach works for *OUT* in much the same way. In the following, we describe the steps necessary to compute the recommendations.

Computation of Link Candidates. The proposed method selects link recommendations from a given navigational view for inclusion in a more restricted view. For example, all links that are present in article in the entire lead but not in the first lead paragraph can potentially be recommended for inclusion in the first lead paragraph if that were to make the

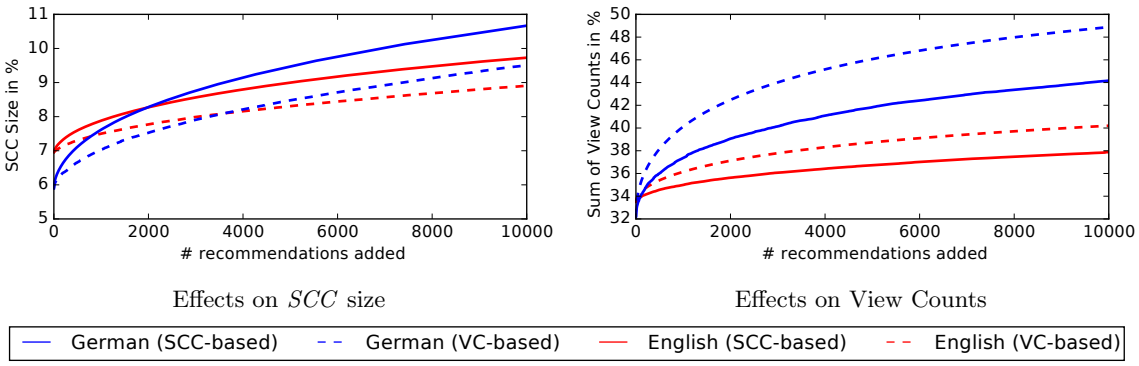

Figure 5: **Effects of Recommendations on the *SCC* Size and the Sum of View Counts**. For demonstration purposes, we compute the effects of adding the top 10,000 recommendations to the first paragraphs of the English and German Wikipedias. The y-axes show the fraction of articles in the *SCC* (left) and the fraction of the sum of view counts covered by articles in the *SCC*. The x-axes show the number of link recommendations added. The ranking of link candidates by *SCC* size (SCC-based) and by the sum of view counts introduced to the *SCC* (VC-based) shows a trade-off between these two effects.

network more navigable as a result. The recommendations are therefore links to articles that have already proven to be semantically relevant for the article by the community of Wikipedia editors. The recommended links could then be made more visible to users. This could be accomplished in several ways—for example, links could be emphasized by displaying them in italics or bold face. Perhaps an easier way, however, would be to move a link into a more visible section of the article, if this makes sense in the context of that section. As such, our proposed method would lend itself well to a link recommender that offers suggestions to Wikipedia editors wishing to make an article more navigable.

Ranking of Link Candidates. The computation of link candidates results in a large number of potential links. Next, we propose two methods to rank the links.

1. **Ranking by Number of Articles Added to the *SCC*.** Each newly introduced link increases size of the *SCC* by at least one article, but potentially several more. Specifically, consider an article that is located at the start of a path leading to the *SCC* via several hops. If that article receives a link from the *SCC*, all articles on the path become part of the *SCC* as well. A natural ranking method is therefore to rank link candidates by the number of articles that the link would add to the *SCC*.

2. **Ranking by Sum of Article View Counts Added to the *SCC*.** A second method to rank the link candidates is to take their importance in terms of view counts into account. We can hence rank link candidates by the sum of the view counts of all articles that a link adds to the *SCC*. This approach favors popular articles, which cannot be reached from the *SCC* without the added link and which can only be found via a search engine or a direct URL manipulation.

Introduction of Recommendations. Finally, the links can be moved or otherwise emphasized in the corresponding Wikipedia article. We propose that the computed information could be used as supplemental information for Wikipedia editors. By showing it alongside links in the edit view, the decision for what links to include would remain in the hands of the editors. In addition, the navigational effects of all other links could be made available to establish what effects the removal of a certain link would have.

5.2 Example of Link Recommendations

We now demonstrate the effects of making use of the recommended links. To this end, we compute both candidate ranking methods for the two largest Wikipedia language versions in our datasets (namely English and German) and incorporate the 10,000 top-ranked links in the network. For the example, we assume a navigational view of the links in the first lead paragraph. This is the view that users looking at results provided by the Google search engine showing excerpts of articles in the knowledge panel would have. We then select link candidates from the next-largest view, which are the links from the entire lead section.

Figure 5 shows the results for the exemplary application of this method to the English and German Wikipedias. The results show that both ranking methods lead to increases for the corresponding metric. However, it also shows that a trade-off exists between increasing the size of the *SCC* and increasing the number of page views covered by the *SCC*. Both effects could be made visible to Wikipedia editors, who could then make the editorial decision whether to emphasize a link.

6. DISCUSSION AND CONCLUSION

In this article we have assessed the navigability of eight large Wikipedia language versions and suggested a method to recommend links. Our research questions were as follows.

RQ1. What is the state of navigability across multiple language versions of Wikipedia?

When taking all links in the articles into account, the vast majority of all articles are contained in the largest strongly connected component and are mutually reachable. However, if we look at Wikipedia with a more restricted navigational view, navigability is substantially reduced. When looking

only at the links in the lead section, the fraction of mutually reachable nodes decreases to as little as 16–37%. When further restricting the view to the first lead paragraph, which mirrors the excerpt that could be shown as supplementary information by an external search engine, navigability further decreases.

RQ2. How can this analysis be exploited to suggest improvements to navigability?

To improve navigability, we have proposed a link recommendation algorithm based on the bow tie analysis of the article network. The algorithm selects links for a navigational view from a less restricted view. The suggested links are therefore semantically relevant and could be introduced by emphasizing them or moving them into a more visible region of the article. We have shown the effects of introducing link recommendations based on two examples. In a real-world setting, the decisions about link recommendations would be left to Wikipedia editors. Editors wishing to take navigability into account could be shown the additional information about the effects of specific links and receive suggestions to better connect articles.

Limitations and Future Work. The presented navigational views were selected based on evidence in previous studies suggesting that users dedicate a large portion of their attention to the area located close to the top. Due to the dynamic width of the Wikipedia in its Desktop view, the exact size of the viewport or the area above the fold is dynamic as well and no universally applicable method of establishing the number of links visible exists. The selected navigational views are therefore necessarily approximations to the true user viewports. However, the evaluation approach we presented is general and can be applied to any navigational view to analyze its effects on navigability. In future work, it could easily be adapted to test the effects of several specific screen resolutions.

The different language versions of Wikipedia generally have little overlap in their coverage of concepts [9]. In addition, every language edition of Wikipedia can establish its own policies and guidelines. Despite this, the eight language versions we investigated led to similar structure in terms of the bow tie model. While the navigational guidelines are likely to be influenced by the English language versions, it would be interesting to explore the differences and commonalities and their effects on navigability of the article networks in future work.

Finally, the use of the view counts to rank the importance of articles is a proxy measure that is subject to influence of crawlers, Wikipedia bots, traffic spikes due to external factors and periodic events such as holidays. For this work, we used the sum of all view counts within the most recent month before the page dump. While this is potentially subject to these limitations, it also carries with it the advantage of bringing to the attention articles that were popular but not reachable at that specific point in time.

We hope that our work stimulates discussion about the navigational effects of restricted views of Wikipedia articles and about methods to highlight the navigational effects of link editing to the Wikipedia community.

7. ACKNOWLEDGMENTS

This research was supported by a grant from the Austrian Science Fund (FWF) [P24866].

8. REFERENCES

[1] S. F. Adafre and M. de Rijke. Discovering missing links in wikipedia. In *Proceedings of the 3rd International Workshop on Link Discovery*, 2005.

[2] D. Brezhnev, S. Trusheim, and V. Yendluri. All paths lead to philosophy. part of the Stanford Network Analysis Project, 2013.

[3] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener. Graph structure in the web. *Computer Networks*, 33(1):309—-320, 2000.

[4] A. Capocci, V. D. Servedio, F. Colaiori, L. S. Buriol, D. Donato, S. Leonardi, and G. Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical Review E*, 74(3):03611:3–6, 2006.

[5] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. V. Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, B. Livingston, and D. Sampath. The youtube video recommendation system. In *Proceedings of the 4th ACM Conference on Recommender Systems*, 2010.

[6] D. Davis, R. Lichtenwalter, and N. V. Chawla. Multi-relational link prediction in heterogeneous information networks. In *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, 2011.

[7] D. Dimitrov, , P. Singer, F. Lemmerich, and M. Strohmaier. Visual positions of links and clicks on wikipedia. In *Proceedings of the 25th International Conference on World Wide Web*, 2016.

[8] I. Halatchliyski, J. Moskaliuk, J. Kimmerle, and U. Cress. Who integrates the networks of knowledge in wikipedia? In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, 2010.

[9] B. Hecht and D. Gergle. The tower of babel meets web 2.0: User-generated content and its applications in a multilingual context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing systems*, 2010.

[10] D. Helic, M. Strohmaier, M. Granitzer, and R. Scherer. Models of human navigation in information networks based on decentralized search. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media*, 2013.

[11] D. Helic, M. Strohmaier, C. Trattner, M. Muhr, and K. Lerman. Pragmatic evaluation of folksonomies. In *Proceedings of the 20th International Conference on World Wide Web*, 2011.

[12] J. L. Herlocker, J. A. Konstan, L. G. Terveen, and J. T. Riedl. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53, 2004.

[13] M. Ibrahim, C. M. Danforth, and P. S. Dodds. Connecting every bit of knowledge: The structure of wikipedia's first link network. *arXiv pre-print*, 2016. arXiv:1605.00309 [cs.SI].

[14] J. Kamps and M. Koolen. Is wikipedia link structure different? In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, 2009.

[15] J. M. Kleinberg. Navigation in a small world. *Nature*, 406(6798):845, August 2000.

[16] J. M. Kleinberg. Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems 14*, 2001.

[17] J. Kunegis, J. Preusse, and F. Schwagereit. What is the added value of negative links in online social networks? In *Proceedings of the 22nd International Conference on World Wide Web*, 2013.

[18] D. Lamprecht, F. Geigl, T. Karas, S. Walk, D. Helic, and M. Strohmaier. Improving recommender system navigability through diversification: A case study of IMDb. In *Proceedings of the 15th International Conference on Knowledge Management and Knowledge Technologies*, 2015.

[19] D. Lamprecht, K. Lerman, D. Helic, and M. Strohmaier. How the structure of wikipedia articles influences user navigation. *New Review of Hypermedia and Multimedia*, 2016. to appear; DOI 10.1080/13614568.2016.1179798.

[20] D. Lamprecht and D. H. M. Strohmaier. Improving reachability and navigability in recommender systems. *arXiv pre-print*, 2015. arXiv:1507.08120 [cs.IR].

[21] D. Lamprecht, M. Strohmaier, D. Helic, C. Nyulas, T. Tudorache, N. F. Noy, and M. A. Musen. Using ontologies to model human navigation behavior in information networks: A study based on wikipedia. *Semantic Web*, 6(4):403–422, 2015.

[22] J. Leskovec, D. Huttenlocher, and J. M. Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th International Conference on World Wide Web*, 2010.

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

[24] R. Mihalcea and A. Csomai. Wikify!: Linking documents to encyclopedic knowledge. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management*, 2007.

[25] S. Milgram. The small world problem. *Psychology Today*, 1(2):60–67, 1967.

[26] D. Milne and I. H. Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, 2008.

[27] J.-H. Oh, D. Kawahara, K. Uchimoto, J. Kazama, and K. Torisawa. Enriching multilingual language resources by discovering missing cross-language links in wikipedia. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, 2008.

[28] A. Paranjape, R. West, L. Zia, and J. Leskovec. Improving website hyperlink structure using server logs. In *Proceedings of the 9th International Conference on Web Search and Data Mining*, 2016.

[29] J. Preusse, J. Kunegis, M. Thimm, and S. Sizov. Decline - models for decay of links in networks. *arXiv pre-print*, 2014. arXiv:1403.4415 [cs.SI].

[30] P. Singer, T. Niebler, M. Strohmaier, and A. Hotho. Computing semantic relatedness from human navigational paths: A case study on wikipedia. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 9(4):41–70, 2013.

[31] P. Sorg and P. Cimiano. Enriching the crosslingual link structure of wikipedia—a classification-based approach. In *Proceedings of the AAAI 2008 Workshop on Wikipedia and Artifical Intelligence*, 2008.

[32] Y. Sun, J. Han, C. C. Aggarwal, and N. V. Chawla. When will it happen?—relationship prediction in heterogeneous information networks. In *Proceedings of the 5th ACM International Conference on Web Search and Data Mining*, 2012.

[33] J. Teevan, C. Alvarado, M. S. Ackerman, and D. R. Karger. The perfect search engine is not enough: A study of orienteering behavior in directed search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2004.

[34] E. G. Toms. Serendipitous information retrieval. In *Proceedings of the DELOS Workshop: Information Seeking, Searching and Querying in Digital Libraries*, 2000.

[35] C. Trattner, P. Singer, D. Helic, and M. Strohmaier. Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks. In *Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies*, 2012.

[36] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32(1):425–443, 1969.

[37] Z. Wang, J. Li, Z. Wang, and J. Tang. Cross-lingual knowledge linking across wiki knowledge bases. In *Proceedings of the 21st International Conference on World Wide Web*, 2012.

[38] D. J. Watts, P. S. Dodds, and M. Newman. Identity and search in social networks. *Science*, 296:1302–1305, 2002.

[39] D. J. Watts and S. H. Strogatz. Collective dynamics of small-world networks. *Nature*, 393(6684):440–442, 1998.

[40] R. West and J. Leskovec. Human wayfinding in information networks. In *Proceedings of the 21st International Conference on World Wide Web*, 2012.

[41] R. West, A. Paranjape, and J. Leskovec. Mining missing hyperlinks from human navigation traces: A case study of wikipedia. In *Proceedings of the 24th International Conference on World Wide Web*, 2015.

[42] R. West, D. Precup, and J. Pineau. Completing wikipedia's hyperlink structure through dimensionality reduction. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, 2009.

[43] E. Wulczyn and D. Taraborelli. Wikipedia clickstream, 2015. http://dx.doi.org/10.6084/m9.figshare.1305770, accessed Februar 17, 2016.

[44] E. Wulczyn, R. West, L. Zia, and J. Leskovec. Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, 2016.