Dipl.-Ing. Florian Geigl, BSc

Random Surfers as Models of Human Navigation on the Web

DOCTORAL THESIS

to achieve the university degree of
Doktor der technischen Wissenschaften
submitted to

Graz University of Technology

Supervisor
Assoc.Prof. Dipl-Ing. Dr.techn. Denis Helic

Institute of Interactive Systems and Data Science
Faculty of Computer Science and Biomedical Engineering
Graz University of Technology
Graz, Austria

Graz, April 2017

TO CAROLINA
You're my uppers and downers ♥

AFFIDAVIT

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly indicated all material which has been quoted either literally or by content from the sources used. The text document uploaded to TUGRAZonline is identical to the present doctoral thesis.

Date	Signature

Abstract

The Web is a central part of modern everyday life. Many people access it on a daily basis for a variety of reasons such as to retrieve news, watch videos, engage in social networks, buy goods in online shops or simply to procrastinate. Yet, we are still uncertain about how humans navigate the Web and the potential of factors influencing this process. To shed light on this topic, this thesis deals with modeling aspects of human navigation on the Web and the effects arising due to manipulations of this process. Mainly, this work provides a solid theoretical framework which allows to examine the potential effects of two different strategies aiming to guide visitors of a website. The framework builds upon the random surfer model, which is shown to be a sufficiently accurate model of human navigation on the Web in the first part of this work. In a next step, this thesis examines to which extent various click biases influence the typical whereabouts of the random surfer. Based on this analysis, this work demonstrates that exploiting common human cognitive biases exhibits a high potential of manipulating the frequencies with which the random surfer visits certain webpages. However, besides taking advantage of these biases, there exist further possibilities to steer users who navigate a website. Specifically, simply inserting new links to a webpage opens up new routes for visitors to explore a website. To investigate which of the two guiding strategies bears the higher potential, this work applies both of them to webgraphs of several websites and provides a detailed comparison of the emerging effects. The results presented in this thesis lead to actionable insights for website administrators and further broaden our understanding of how humans navigate the Web. Additionally, the presented model builds the foundation for further research in this field.

Kurzfassung

Das Web ist ein wesentlicher Punkt unserer modernen Gesellschaft. Menschen in der ganzen Welt verwenden es täglich aus einer Vielzahl von Gründen. Dazu gehört unter anderem das Ansehen von Videos in online Mediatheken, die Kommunikation in sozialen Netzwerken, das Kaufen von Produkten in online Shops, oder einfach nur als Zeitvertreib. Dennoch wissen wir nicht exakt, wie Menschen im Web navigieren und wie hoch das Manipulationspotential von Faktoren, welche diesen Prozess beeinflussen, ist. Um das Wissen in diesem Bereich zu verbessern, behandelt diese Doktorarbeit Aspekte der Modellierung von menschlicher Navigation im Web und die Effekte, welche durch Manipulation dieses Prozesses auftreten könnten. Vorwiegend präsentiert diese Arbeit ein solides theoretisches Konzept, welches es erlaubt, die potentiellen Effekte von zwei verschiedenen Manipulationsstrategien, beide mit dem Ziel Besucher einer Webseite zu steuern, zu untersuchen. Das präsentierte Konzept beruht auf dem „Random Surfer"-Model, welches im ersten Teil dieser Arbeit als ausreichend genaue Imitation von menschlicher Navigation im Web, überprüft wird. Im darauffolgenden Teil beschäftigt sich diese Doktorarbeit mit dem Ausmaß an Einflüssen von sogenannten „Klick-Biases" auf den Random Surfer. Basierend auf den dafür durchgeführten Analysen wird gezeigt, dass menschliche kognitive Biases ein hohes Potential aufweisen, welches die Frequenz mit welcher der Random Surfer bestimmte Webseiten besucht, zu manipulieren. Jedoch gibt es auch noch weitere Methoden um das Surfverhalten von Benutzern einer Website zu steuern. Im Speziellen können einfach neue Links in die Webseite eingebaut werden. Dies eröffnet den Benutzern neue Pfade über welche sie die Webseite erkunden können. Um herauszufinden welche der beiden Methoden das höhere Potential zur Manipulation von Surfverhalten besitzt, wurden beide auf empirische Webgraphen angewandt und die dabei aufgetretenen Effekte detailliert verglichen. Die daraus gewonnenen Einsichten in das menschliche Surfverhalten im Web führten zu praxisrelevanten Erkenntnissen, welche von besonderem Interesse für Webseiten-Administratoren sind. Zusätzlich bildet das in dieser Arbeit präsentierte Model die Grundlage für weitere Forschungsarbeiten auf diesem Gebiet der Wissenschaft.

Danksagung

In erster Linie möchte ich meinem Doktorvater Denis Helic für seine ausgezeichnete Führung und Betreuung durch mein Doktorat danken. Er hatte jederzeit ein offenes Ohr, war für neue Ideen zu begeistern und ging bei Problemen immer zielstrebig und mit gutem Beispiel voran—Denis, deine berühmten, seitenlangen, handschriftlichen Erklärungen, Skizzen, Ab- bzw. Herleitungen und die damit verbundenen exzellenten Erklärungen werde ich noch lange in guter Erinnerung behalten.

Ein weiterer Dank gilt Markus Strohmaier, welcher mich von Anfang an in seine wöchentlichen Donnerstagstreffen eingebunden hat. Markus, du hast es immer wieder geschafft intellektuell spannende Diskussionen zu kreieren, wodurch sich eine hervoragende Forschungsgruppe gebildet hat, die sich auf mehreren internationalen Konferenzen bewies.

Ein großer Dank gilt auch Simon Walk, welcher durch seine gesammelten Erfahrungen viele wertvolle Ratschläge einbrachte, oft mit gutem Rat voranging und mir wirklich jederzeit bei Problemen zur Seite stand—Simon, ich bin mir bis heute nicht sicher ob du auch mal schläfst oder gänzlich darauf verzichtest. Achja, irgendwo im Text habe ich für dich „cat videos" erwähnt.

Im selben Zuge möchte ich auch Daniel Lamprecht danken. Wir hatten unzählige Diskussionen in unserem Büro—zum Teil wissenschaftliche, zum Teil lustige (Duftlampe J/N, Fenster auf J/N und, mein Favorit, Pflanzen Pestizide Sprays...okay darüber wurde erst nach deren Verwendung diskutiert) aber zum Großteil über wichtige Filmzitate von Arnold Schwarzenegger. Du standest mir sprachlich immer zur Seite—sei es nun anfangs in Python oder später in Englisch.

Auch möchte ich mich bei Patrick Kasper und Thomas Wurmitzer, meinen Bürokollegen, herzlich bedanken. Patrick ohne dich wäre eine dermaßen gigantische Überraschungsei-Sammlung niemals möglich gewesen und Chomas (spanische Betonung), ohne dich hätte ich keine Zeit für Überraschungseier gehabt, da ich unendlich mehr Stunden gebrauchte hätte, um unseren Server am Laufen zu halten.

Natürlich möchte ich mich auch bei meiner Forschungsgruppe bedanken, welche im Laufe der Jahre viele neue Gesichter bekam. Namentlich möchte ich hier Dimitar Dimitrov, Lukas Eberhard, Rainer Hofmann-Wellenhof, Tomas Karas, Florian Klien, Philipp Koncar, Silvia Mitter, Lisa Posch, Philipp Singer, Tiago Teixeira dos Santos, Massimo Vitiello und Claudia Wagner nennen — danke für die vielen spannenden Diskussionen und das wertvolle Feedback zu meiner Arbeit.

In weitere Linie möchte ich meinen Co-Autoren Michael Goller, Ilire Hasani-Mavriqi, Stefan Hinteregger, Elisabeth Lex, Christine Moik und Subhash Pujari für die tolle Kooperation danken—wir hatten wertvolle und ideenreiche Meetings, woraus sich später tolle Publikationen entwickelt haben.

Im letzten Jahr meines Doktorats hatte ich die Ehre das Team des „Institut for Information Science“ der University of California kennen zu lernen. Hier möchte ich mich vor allem bei Kristina Lerman und José Luis Ambite für die hervorragende Zusammenarbeit und die Chance mein erlerntes Wissen und Geschick zu beweisen, bedanken. Mein besonderer Dank gilt hier auch Laura Alessandretti, Gully Burns, Emilio Ferrara, Lily Fierro (und ihrem Ehemann Generoso Fierro), Jonathan Gordon, Ulf Hermjakob und Farshad Kooti, welche mich herzlichst aufgenommen und mir die Stadt näher gebracht haben. Lily und Generoso, wir sehen uns spätestens in der Pension in Italien bei Wein und Pizza wieder.

Ein großer Dank gilt auch Detego—danke dass ihr mich alle so herzlich in eure große Familie aufgenommen und mir die nötige Unterstützung bei den letzten Schritten zum Doktorat gegeben habt. Im gleichen Zuge möchte ich mich auch für den Fehler auf dem Titelblatt entschuldigen—es heißt natürlich Florain G.

Auf den Konferenzreisen hatte ich das Glück viele interessante Menschen aus aller Welt kennen zu lernen, woraus sich auch anhaltende Freundschaften entwickelten. Hier möchte ich insbesondere Samuele Soraggi danken—wir hatten unglaublich viel Spaß und Gelati in Madrid, auch wenn wir deswegen mitten in der Nacht quer durch die Stadt laufen mussten.

Natürlich möchte ich mich auch bei all meinen Freunden für ihrer Treue und Nachsicht bei so manch verpassten „Festl“ bedanken. Da ich euch bis jetzt nicht losgeworden bin, werde ich euch wohl noch länger am Hals haben—und das ist auch gut so! Diana, dir gilt ein besonderer Dank für die professionelle, liebevolle und motivierende Korrektur meiner Arbeit—ohne dich wäre diese Arbeit grammatikalisch gesehen wahrscheinlich ein „total disaster“.

Abschließend möchte ich noch meinen größten Dank aussprechen, welcher natürlich nur meiner Familie, meiner Verlobten Carolina und ihrer Familie gelten kann. Ihr musstet in den letzten Jahren viel auf mich verzichten und wart mir dennoch nie böse—im Gegenteil, ihr seid mir Tag und Nacht immer zur Seite gestanden. Ohne eurer bedingungslose Liebe wären dies nur leere Seiten...

Institutional Acknowledgements

This thesis was partially funded by the FWF Austrian Science Fund Grant “Navigability of Decentralized Information Networks” (P 24866). Furthermore, I want to thank the Technical University of Graz for giving me the opportunity to conduct this thesis, and the Information Sciences Institute of the University of Southern California for my internship.

Contents

1. Introduction

1.1. Motivation

In 1989, Tim Berners-Lee, a scientist employed at CERN in Switzerland, devised an information system to facilitate communication among scientists at the institution. His proposed system made use of the recently established Internet to construct an information space, which allowed users to access various resources. Specifically, he built his system upon a technology called *hypertext*, which has displayed resources (i.e., text) on a screen, and further, allowed to directly connect, and thus relate, resources to each other through so-called *hyperlinks* [Berners-Lee et al., 1992, 2000]. He named his information system the *World Wide Web* or just *the Web*.

Over the last 30 years, engineers and scientists have improved many components of this technology and resolved some of its shortcomings, such as the lack of built-in search engines. However, the basic concepts have remained the same. In particular, we still employ web browsers to access the World Wide Web and still click on hyperlinks to navigate through the available resources (i.e., webpages). Today, we look back at the invention of the web browser as one of the most important steps towards making the Web as successful and popular as it is now. Nowadays, millions of people around the world not merely know the Web, but also utilize web browser to access it on a daily basis for various purposes, such as, communication with friends, online shopping, or sometimes just to watch videos of cats.

With an increasing user base, the Web has also experienced a vast growth in terms of numbers of online webpages. In 2005 there were over 11.5 billion indexable webpages online [Gulli and Signorini, 2005] and today, merely

ten years later, the Web is estimated to be over 40 billion pages in size[1]. An important reason for this enormous growth has been the simplicity with which everyone could add new resources (i.e., websites) to the system. As a consequence of the enormous growth and the decentralized structure of the system (i.e., no central index), it has become increasingly more challenging to retrieve resources and satisfy information needs. Hence, the need for web search engines became progressively more urgent, which is why many scientists and engineers have come up with various ways to tackle the problem of providing a searchable index which contains all currently online webpages. Over the years, web search engines have become so powerful that Google, the company offering the most popular search engine, has become one of the most valuable companies in the world[2].

Despite the increasing popularity of search tools, a large part of human navigation on the Web can be still attributed to following existing links [Gleich et al., 2010]. Moreover, it has been shown that humans are extremely efficient at navigating through large hypertext systems without even utilizing search engines [Helic, 2012; West and Leskovec, 2012a]. To examine and simulate human navigation on the Web, several models have been developed. Some of these models, such as the famous random surfer which follows links chosen uniformly at random, even influenced how search engines rank the results of search queries [Brin and Page, 1998].

Although the proposed models differ in various aspects, most of them have one characteristic in common: They mostly rely on the link structure of the Web, and not on the content of webpages. Specifically, they are based on the so-called webgraph, in which each page represents a node and each hyperlink is modeled as a directed edge (i.e., the source page contains a hyperlink towards the target page). These webgraphs should model the basic constraint of humans navigating the Web, that is, we can only click on existing links.

However, there are further properties, such as the layout of webpages, that influence how humans navigate the Web [Blunch, 1984]. For example, it has been shown that humans navigating the Web exhibit a strong bias

[1] http://www.worldwidewebsize.com/
[2] http://www.forbes.com/powerful-brands/list/

towards links that are located at the top of a page [Joachims et al., 2005; Murphy et al., 2006; Craswell et al., 2008; Yue et al., 2010; Lerman and Hogg, 2014; Dimitrov et al., 2016; Lamprecht et al., 2016]. In other words, links that are positioned at the top of a webpage will more likely be followed than those located at the bottom of the same webpage. Despite our knowledge of such human biases, to date there is no model which incorporates this kind of information.

However, a deeper understanding of such effects is crucial to efficiently guide users navigating the Web. Understanding those biases better would, for example, enable us to adjust the layout of webpages to facilitate navigation. Furthermore, it is currently not known whether human biases or the Web's link structure exhibit a larger potential to guide users on the Web.

In this thesis I am going to tackle those open questions and strive to get a deeper understanding of factors on which human navigation builds upon. Furthermore, I shed light onto the potential effects of exploiting human biases to guide them and how effects of other guiding techniques compare to them.

To that end, in Section 1.2 I provide a short overview of how humans typically navigate the Web, what their navigational traces look like, and how their biases influence their link selection process. Furthermore, I explain the basic concepts of models commonly used to simulate human navigation on the Web in Section 1.3. An overview of the problem statement, the objectives and main approach of this thesis is provided in Section 1.4. Subsequently, in Section 1.5 I list the specific research questions. A collection of the main publications of this cumulative thesis is shown in Section 1.6 and my contributions to each of them are emphasized in Section 3.1. In Section 1.7 further publications to which I have contributed during my time as a PhD student are listed. The contributions and implications of this thesis are summarized in Section 1.8. Finally, the entire thesis is outlined in Section 1.9.

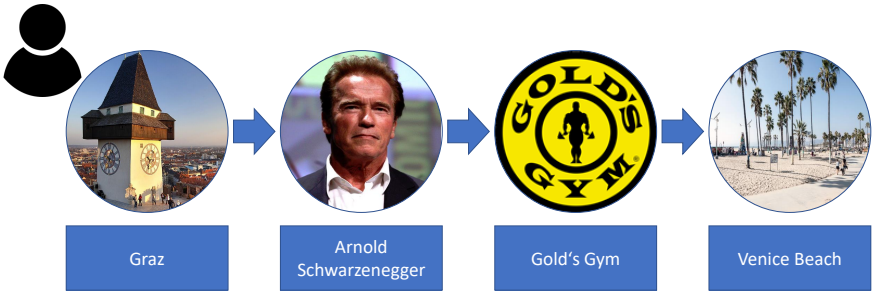


Figure 1.1.: **Illustrative Human Click Trail.** The figure depicts an exemplary click trail of a user navigating an online encyclopedia. In particular, the user starts to navigate at the article about *Graz*—the second largest city of Austria and wants to retrieve information about Venice Beach—a world-famous beach located on the coast of Los Angeles. In the first step, the user clicks on the link guiding her to the article about Arnold Schwarzenegger—a celebrity born near Graz who in his adulthood became the governor of California, USA. Subsequently, the user clicks on the link leading to the article about the famous Gold's Gym—often called "the mecca of bodybuilding"—which is located in Venice, California. In the next step the user clicks on the link towards the article about Venice Beach and, consequently, arrives at the target article. The chronologically sorted sequence of the pages visited by the user is called click trail. The short click trail shown here also includes a characteristic pattern typically observed in real user navigation. That is, in the first step the user navigates to a *popular* webpage—in that case an article about a famous person—followed by a click towards an article less famous but more *similar* to the target article—a Gym located in Venice. The former part of the pattern is often referred to as the "zoom-out" phase, whereas the latter is often called the "zoom-in" phase [Helic, 2012; West and Leskovec, 2012a].

1.2. Human Navigation on the Web

In this dissertation I analyze how humans navigate the Web and study the effects emerging through manipulations of this process. Humans navigating the Web leave behind *click trails*, which can be exploited to study their behavior. Typically, these trails are extracted from logfiles, or any other machine-readable formats, and allow to reproduce the chronologically sorted sequence of pages visited by a user. In other words, click trails are sequences of webpages a user consecutively visited in the past. An example of such a click trail produced by a human navigating Wikipedia—a large online encyclopedia—is depicted in Figure 1.1.

Most of the time, humans who navigate the Web possess various biases, such as the visual bias towards links on the top of a webpage. Furthermore, it has been discovered that it is possible to exploit these biases to manipulate a user's decision about which link to click on next [Lerman and Hogg, 2014]. As a consequence, steering users by means of their biases could potentially trigger changes in their typical whereabouts on the Web—which is why we want to guide (i.e., steer) them in the first place.

However, it is currently unknown how these changes emerge. In this thesis, I present a first stepping stone towards an answer to this question. To that end, I use a model-based approach, which can be summarized as follows. First, I utilize correlation analysis methods to evaluate existing models of human navigation on the Web (i.e., search algorithms). After determining that the random surfer is an appropriate model, I present an approach that allows me to intuitively incorporate biases into the model (i.e., proxies of known human biases). Subsequently, I conduct an in-depth analysis of the effects emerging due to the performed adjustments. Finally, I compare these effects to those triggered by structural modifications of the webgraph.

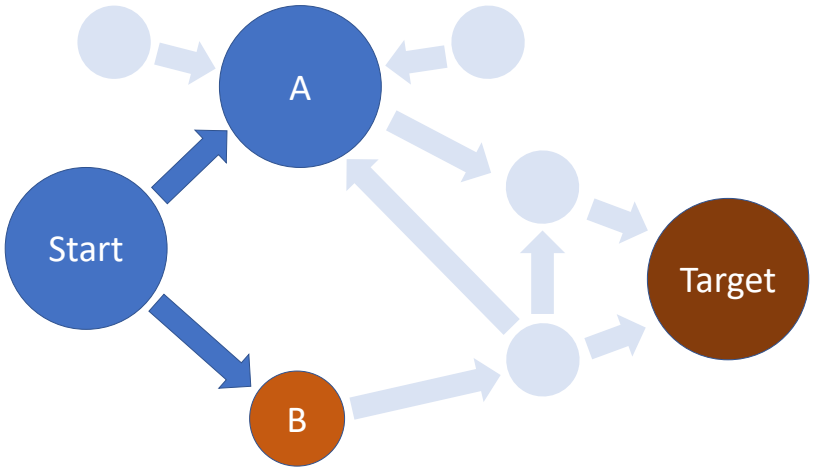


Figure 1.2.: **Illustrative Decentralized Search Example.** The figure depicts an exemplary step during a typical decentralized search process. In the presented scenario the algorithm—a search agent—aims to navigate from the "Start" node to the "Target" node in as few steps as possible. The difficulty of this task lies in the information available to the agent. In particular, it does not have access to global information about the structure of the system. Thus, only local information can be exploited. In the first step, the search agent has to decide whether to move forward to node "A" or "B", since these two nodes are the only ones accessible over a direct link from its current position. To decide which way to take, the agent exploits two different sources of available, local information: (i) the popularity of each neighboring node (represented by the node size) and (ii) the similarity of each neighboring node to the target (represented by node color). Hence, node "A" depicts a popular node, whereas node "B" is more similar to the target node. Depending on which of the two properties the search agent relies its decision on, it will greedily move forward to either "A" or "B". Because of this property, the algorithm is classified as a deterministic algorithm. Independent of the information source, the algorithm does not know which of the two options minimizes the distance to the target most. Consequently, the algorithm can not ensure that the decision it makes is optimal.

1.3. Models of Human Navigation on the Web

A well-known model of human navigation on the Web is *decentralized search* [Kleinberg, 2000a; Adamic et al., 2001; Helic et al., 2013]. This model builds upon the fact that during a navigation session, global information about the network is usually not available. Examples of such conditions include, but are not limited to, huge networks for which even state-of-the-art hardware is not able to provide global information, such as the shortest path between all pairs of nodes. Another example of such a situation is constituted by networks which exhibit a constantly changing structure, such as peer-to-peer networks or online social networks. In those case, it is impossible to provide accurate, global information about the entire structure of such networks. Consequently, there exists no possibility to create a centralized search engine which would enable immediate access to nodes. To tackle this problem, search algorithms implemented in such environments have to exploit local information to be able to perform search tasks efficiently.

In practice, humans face a similar problem while navigating the Web. In particular, the sheer number of webpages available on the Web and its constantly changing structure, make it impossible for them to memorize the entire webgraph. Thus, web users have to rely on information readily available to them, namely the information provided locally by the current webpage.

In general, humans and algorithms tackle such situations by utilizing a decentralized search approach. They first collect information about neighboring webpages and then use it to decide which link to click on next. Subsequently, they repeat this procedure till they either find the page they were looking for (i.e., target page) or decide to stop searching. Commonly exploited information in this process are (i) homophily (e.g., any kind of similarity to the target page [Pirolli, 1997; Kleinberg, 2000a]) and (ii) popularity (e.g., number of incoming and/or outgoing hyperlinks [Adamic et al., 2001; West and Leskovec, 2012a]). Figure 1.2 illustrates an exemplary step of such a process.

In the case of human navigation on the Web, available local information consists of a mix of background knowledge and intuitions associated with a link's anchor text (i.e., the text which represents a hyperlink on a webpage). By analyzing click trails, researchers discovered that humans who navigate through an online encyclopedia base their decisions about where to navigate next in the initial phase of a session mostly on the popularity of neighbors. After this so-called "zoom-out" phase humans start to rely more on the similarity between neighboring pages and the information they are searching for. This second phase is often referred to as the "zoom-in" phase [Helic, 2012; West and Leskovec, 2012a].

Another characteristic of human navigation is, that the human link selection process—the process in which they decide which link they will click on next—involves a large degree of randomness [Helic et al., 2013]. Consequently, the process is stochastic, meaning that the same situation can lead to different results for repeated runs. This is in stark contrast to decentralized search, where the same situation always leads to the same outcome (i.e., it is deterministic). To account for the randomness involved in human navigation on the Web, accurate models need to include stochastic procedures [Helic et al., 2013]. The simplest example of such a non-deterministic model is the so-called "random walk".

In a random walk model an agent—the "random surfer"—selects the next link to traverse uniformly at random out of all outgoing links. This step is repeated until the agent reaches the target node or it stops searching due to a predefined number of maximally allowed steps. Despite the model's simplicity, it provides a strong baseline for search algorithms. Furthermore, this model provides the basis for PageRank—the famous algorithm used by Google to rank the results of its search queries. Although it is often assumed that the random surfer is a well-fitting, or at least sufficient, model of human navigation on the Web, there is relatively little previous work that proves this with empirical data [Chierichetti et al., 2012; Singer et al., 2014b].

1.4. Problem Statement, Objectives and General Approach

Problem Statement. Human navigation on the Web includes many unknown variables, such as the background knowledge and intuition of an individual user and their effects on link selection. Consequently, predicting the specific link a user will click on next represents a difficult task. Having such a model of human navigation on the Web would equip website administrators with a valuable tool to better arrange important information on a page, or to display products in a way that enhances attention of their customers and subsequently increases sale. Moreover, a model of human navigation would be an important foundation to answer a multitude of related research questions.

While the task of accurately predicting the next click is complex, more simple models are valuable and sufficient for a range of applications. Website administrators are often interested in estimating the distribution of page views (i.e., the relative frequency with which a page gets visited by users), especially when adjusting parts of the site structure or the user interface. In practices, this is especially important in the case of unpublished websites for which no empirical data is available yet. Likewise, scientists can exploit these simpler models to analyze and compare page view distributions of various empirical, or even synthetic, websites without the need to gather real user data.

Feature selection for this modeling task is a critical part, as initially promising variables such as the history of previously clicked links can actually be neglected in modeling [Singer et al., 2014b]. Some human cognitive biases, on the other hand, have been shown to strongly affect link selection [Blunch, 1984; Joachims et al., 2005; Murphy et al., 2006; Craswell et al., 2008; Yue et al., 2010; Dimitrov et al., 2016; Lamprecht et al., 2016]. As an important example, the well-known position bias can be actively exploited to steer users [Lerman and Hogg, 2014].

Based on this insight, website owners could experiment with various layouts of their website to potentially redirect the typical whereabouts of users and thus use biases for their advantage (e.g., increase visits to a specific

product page). From a scientific point of view, investigating the potential of various biases and the effects triggered by them would be an important step towards a deeper understanding of human navigation on the Web. This is especially true in the case of Lamprecht et al. [2015a], where the authors present a system which recommends to reposition specific links to improve the navigability of Wikipedia. However, research so far has only demonstrated how we can actively manipulate the link selection process of users but has given little insight into the consequences emerging from this intervention. A model which incorporates various biases would thus be of great relevance and importance for website administrators and scientists alike.

Human cognitive biases are a prime example of an approach to influence human navigation on the Web. However, there exist further properties which can be capitalized on to this end. For example, it is common practice to introduce new links to a website, such in the case of interlinking related articles on Wikipedia or creating new friendships in online social networks. This action has a high potential to manipulate how users explore a website by providing new connections. Despite the frequent use of this method in practice, the potential effects of these modifications of the webgraph have not yet been investigated, and any potentially undesired ramifications have not been explored.

Furthermore, it is also not clear how the manipulation of the link selection process with biases (i.e., click biases) compares to the method of link insertion in terms of efficiency with which users are steered. The former action increases the probability with which users follow links towards selected pages, while the latter creates new paths to access these pages. Thus, both methods potentially increase user visits of the selected pages. Consequently, knowing more about these effects would help website administrators to decide better which of the two manipulation strategies—link insertion or click biases—possesses the higher efficiency to steer users towards specific pages. Moreover, the answer to this question lays out the basis for further research in the field of web science.

In summary, the effects of human biases and the comparison between link insertion and click biases represent a scientifically interesting prob-

lem with many real-world applications of utmost importance for website administrators.

Objectives. This thesis strives to shed light on the potential effects of click biases and to compare the consequences, implications and effectiveness of click biases and link insertion. As a first step towards this goal, an appropriate model of human navigation on the Web needs to be established. Based on such a model, this thesis aims to investigate global effects of click biases on the typical whereabouts of users on a website. Additionally, a major part of this thesis analyzes how click biases and inserting new links compare in terms of efficiency with which users are steered. A practically relevant goal of this thesis is to equip website administrators with a powerful tool that supports them in making decisions about potential modifications to steer their visitors.

General Approach. The general approach of this thesis is based on the random surfer model, which serves as a proxy for human navigation on the Web. In particular, I use the stationary distribution of the random surfer, which describes the probability of finding the random surfer on a specific node of the webgraph in the limit of infinitely large navigational sessions. In a first step, I validate that the selected model is a sufficiently good approximation of human navigation on the Web, by comparing it to empirical data gathered from an online encyclopedia. Second, I present an approach that allows me to incorporate click biases into the model. Based on this extend model, I utilize the models' stationary distribution as a proxy for the relative frequency with which a webpage gets visited by users. This allows me to provide an in-depth analysis of effects emerging due to various modeled biases. In the last part of this dissertation, I expand the model with the ability to simulate link insertion, which enables me to conduct several experiments comparing the effects of click biases to those of link insertion. An illustrative overview of the general approach is depicted by Figure 1.3.

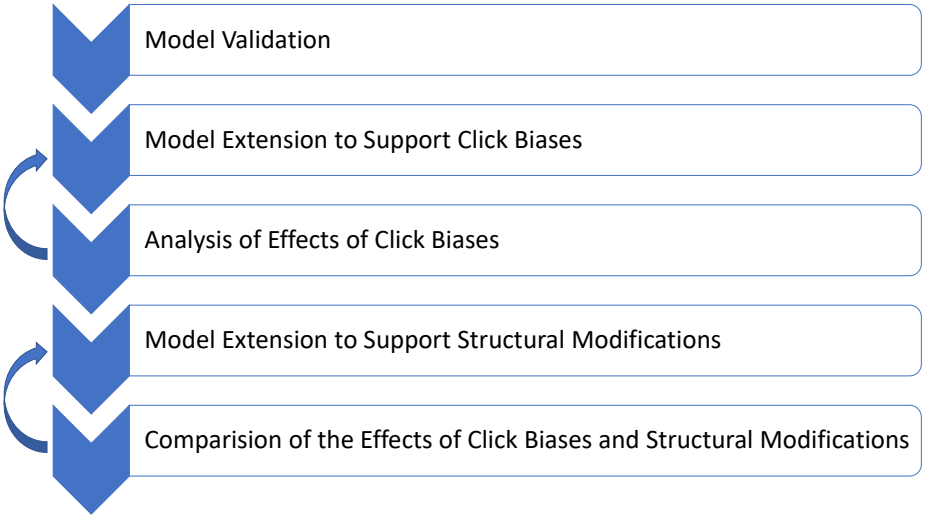


Figure 1.3.: **General Approach.** The figure outlines steps of the general approach of this thesis. First, I conduct experiments which validate the random surfer as an appropriate model of human navigation on the Web. Second, I extend the random surfer model in such a way, that it allows to incorporate click biases. Subsequently, I analyze the effects emerging due to various click biases. Based on insights gathered throughout this analysis, I further improve the model's support for click biases. After several iterations of this two steps, this thesis continues to extend the model further. Specifically, I enhance the model by introducing the support of targeted structural modifications of the webgraph. Again, I use an iterative approach to extend and analyze the model. However, in this case, the analysis aims to compare the efficiency of click biases and structural modification in terms of user guidance.

1.5. Research Questions

In this thesis, I utilize a model-based approach which allows me to shed light onto various effects of human biases on typical whereabouts of users on the Web. In the first part, I examine if the random surfer model is able to approximate user navigation on the Web (RQ1). Subsequently, I extend the random surfer model to examine the effects of induced biases (RQ2) and structural modifications (RQ3). In particular, based on the results presented in this thesis I answer the following research questions.

RQ1: Can we model human navigation using random surfers?

Problem. The Web has become an integral part of everyday life and serves a variety of purposes. A huge, diverse population of users around the globe uses the Web for a vast amount of reasons, such as to retrieve information, socialize, buy goods or simply to procrastinate. When using the Web, users base their decision about which link to follow next on assumptions originating from their constantly varying background knowledge and their current goals. All these variables and their complex interactions make the prediction of how a specific user will navigate the Web a hard task. What makes this task even more complicated is, that human navigation on the Web has been shown to include a large degree of uncertainty (i.e., randomness). However, for many tasks it is sufficient to know the relative frequency with which a page gets visited by users, such in the case of estimating which page will receive most attention of users.

Approach. To tackle this research question, it is necessary to compare empiric page views to those generated by the model under investigation. In particular, in Geigl et al. [2015] we incorporate empiric transition probabilities (i.e., relative frequencies with which humans follow outgoing links of a webpage) stemming from an online encyclopedia into the well-known random surfer model. Subsequently, we measure the correlation between the stationary distribution of the unmodified random surfer and the random surfer which acts according to the empiric data. Based on the

observed correlation, we then determine whether or not the random surfer is a valid model of human navigation on the Web.

Findings and Contributions. To answer the research question, we present an intuitive method to compare models of human navigation to actual humans browsing the Web [Geigl et al., 2015]. We find that on a macroscopic scale (i.e., distribution of page views) the random surfer exhibits strong similarities to empirical user data. However, these commonalities diminish as soon as humans use search engines while navigating the Web. The reason for this is that search engines allow users direct access to the page they are searching for. Consequently, pages on the top of a website's hierarchy (e.g., home page) receive fewer views, while specific pages, which are typically further down in the hierarchy, get visited more frequently.

RQ2: How can we model navigational biases of humans?

Problem. Although we know that the link selection process of humans is influenced by cognitive biases such as the position of links [Blunch, 1984; Joachims et al., 2005; Murphy et al., 2006; Craswell et al., 2008; Yue et al., 2010; Dimitrov et al., 2016; Lamprecht et al., 2016], we still do not know how such biases affect the typical whereabouts of users on the Web. In other words, we know that, for example, by altering link positions, we can actively manipulate a user's decision about which link to click on next [Lerman and Hogg, 2014]. However, the macroscopic effects caused by such modifications are still unknown.

Approach. In Geigl et al. [2015] we have shown that the random surfer model is, from a macroscopic point of view, capable of mimicking human navigation on the Web. To tackle this research question, we use a modification of the random surfer model to examine the potential effects of biases affecting the link selection process. In particular, we simulate biases towards popular, unpopular, similar, and dissimilar pages on empirical webgraphs. Subsequently, we provide and in-depth analysis of global effects triggered by these biases [Geigl et al., 2016b].

Findings and Contributions. The first contribution is the formalization of a solid theoretical framework which allows to analyze consequences of navigational biases on the browsing dynamics of humans. In particular, we investigate changes in visit probabilities of specific webpages of a website. Applying this approach to several empirical datasets, we find that, contrary to undirected networks [Sinatra et al., 2011], on (directed) webgraphs all biases under investigation increase the certainty of the random surfer when selecting a link. Additionally, we observe significant side effects of certain biases. These effects suggest that administrators should carefully decide whether or not to exploit a bias to actively steer users on their website. Furthermore, these severe side effects underline the value of our approach, which is, that it allows for an offline evaluation of several types of biases.

RQ3: How do navigational biases compare to structural modifications of networks?

Problem. In Geigl et al. [2016b] we have shown that manipulated transition probabilities of existing links (i.e., click biases) drastically influence the relative frequency with which humans visit pages of a website. In that case, we did not alter the underlying link structure of the website in any way. However, structural modifications of webgraphs is a common practice, such in the case of creating new friendships in online social networks or interlinking related Wikipedia articles. Yet, the emerging effects of this action have not been investigated until today.

Approach. To find an answer to this question, we base our approach on the insights obtained in Geigl et al. [2015]. Specifically, we utilize the random surfer as a valid model of human navigation. In a next step, we randomly pick subsets of pages (i.e., target sets) of several websites. Subsequently, our aim is to increase visit probabilities of the pages contained in these sets. Furthermore, we want to determine which of the two manipulation strategies under investigation—link insertion or click bias—should be preferred over the other in terms of efficiency of steering the random surfer. As a first stepping stone towards this

goal, we present a method that allows us to compare the performance of both strategies in a fair manner. Afterward, we apply this method to several empirical webgraphs to examine which of the two manipulation strategies is more efficient in increasing the visit probabilities of the target set. Additionally, we vary the size of the target set to investigate whether or not the performance of any of the two strategies is influenced by this variable [Geigl et al., 2016a].

Findings and Contributions. From a methodological point of view, we present a novel approach for measuring and fairly comparing the potential of click biases and link insertion. We find that, depending on the size of the target set, the optimal link modification strategy varies. In particular, for smaller sets of target pages, link insertion constantly outperforms click biases. This observation is especially prominent in the case in which the set of target pages consist mainly out of pages with almost no visits in the initial unmodified state. However, with an increasing number of target pages, click biases become progressively more efficient and start to outperform link insertion strategies. These findings can aid website administrators in their decision about which method to pick, based on their current situations and goals. Additionally, the generality of the presented framework makes it easy for website administrators to test their scenario in an offline simulation. The simulation tool can be adapted and extend by anyone as it is available as open-source on GitHub[3].

1.6. Main Publications

This cumulative thesis consists of the following three publications:

- **Article 1:** [Geigl et al., 2015] Geigl, F., Lamprecht, D., Hofmann-Wellenhof, R., Walk, S., Strohmaier, M. and Helic, D. (2015). Random Surfers on a Web Encyclopedia. *15th International Conference on Knowledge Technologies and Data-driven Business*

[3] https://github.com/floriangeigl/RandomSurfers

- **Article 2:** [Geigl et al., 2016a] Geigl, F., Lerman, K., Walk, S., Strohmaier, M. and Helic, D. (2016). Assessing the Navigational Effects of Click Biases and Link Insertion on the Web. *27th Conference on Hypertext and Social Media*

- **Article 3:** [Geigl et al., 2016b] Geigl, F., Walk, S., Strohmaier, M. and Helic, D. (2016). Steering the Random Surfer on Directed Webgraphs *International Conference on Web Intelligence*

1.7. Further Publications

Furthermore, I contributed to the following publications during my time as a PhD student:

- **Journal 1:** [Walk et al., 2016] Walk, S., Helic, D., Geigl, F. and Strohmaier M. (2016). Activity Dynamics in Collaboration Networks. *ACM Transactions on the Web*

- **Journal 2:** [Hasani-Mavriqi et al., 2016] Hasani-Mavriqi, I., Geigl, F., Pujari, S., Lex, E., and Helic, D. (2016). The Influence of Social Status and Network Structure on Consensus Building in Collaboration Networks. *Social Network Analysis and Mining*

- **Article 1:** [Hasani-Mavriqi et al., 2015] Hasani-Mavriqi, I., Geigl, F., Pujari, S., Lex, E., and Helic, D. (2015). The Influence of Social Status on Consensus Building in Collaboration Networks. *International Conference on Advances in Social Networks Analysis and Mining*

- **Article 2:** [Lamprecht et al., 2015a] Lamprecht, D., Geigl, F., Karas, T., Walk, S., Helic, D., and Strohmaier M. (2015). Improving Recommender System Navigability Through Diversification: A Case Study of IMDb. *15th International Conference on Knowledge Technologies and Data-driven Business*

- **Article 3:** [Helic and Geigl, 2015] Helic, D. and Geigl, F. (2015). Importance of Network Nodes for Navigation with Fractional Knowledge. *38th International Convention on Information and Communication Technology, Electronics and Microelectronics*

- **Workshop Article 1:** [Ambite et al., 2017] Ambite, J., Lerman, K., Fierro, L., Geigl, F., Gordon, J. and Burns, G. (2017). BD2K ERuDIte: The Educational Resource Discovery Index for Data Science *4th WWW Workshop on Big Scholarly Data*

- **Workshop Article 2:** [Geigl and Helic, 2014] Geigl, F. and Helic, D. (2014). The Role of Homophily. *2nd International Workshop on Dynamic Networks and Knowledge Discovery*

- **Poster 1:** [Geigl et al., 2017] Geigl, F., Moik, C., Hinteregger, S. and Goller, M. (2017). Using Machine Learning and RFID Localization for Advanced Logistic Applications. *11th Annual IEEE International Conference on RFID*

1.8. Contributions and Implications

The major scientific contribution of this thesis consists of an extension of the well-known random surfer model, enabling researchers to easily and intuitively formalize click biases and link insertion. From an empirical and practical point of view, I apply this method to several real-world scenarios and provide subsequent actionable insights for website administrators on how to efficiently guide their users.

Specifically, this thesis makes the following contributions:

- First, I use empirical data to validate that, on a macroscopic scale, the random surfer is a valid model of human navigation on the Web.

- Second, I extend the well-known random surfer model in such a way, that it is capable of mimicking the effects of (i) several types of biases, and (ii) structural changes made to the underlying network structure. Furthermore, the presented extension of the model enables

us to compare the two presented manipulation strategies in a fair manner.

- Third, I design a model that can be used for various scenarios. Specifically, I contribute a solid theoretical framework for further theoretical, empirical and practical analyses of human navigation on the Web.
- Fourth, from a practical point of view, I create an open-source tool that aids website administrators in the offline evaluation of several user-steering strategies. Furthermore, the presented tool can notify its users (e.g., website administrator) about probably unintended side effects potentially arising due to a specific manipulation strategy.

The results of this thesis indicate that exploiting human biases to manipulate the link selection process leads to drastic changes in the distribution of visits over the pages of a website. I find that due to some specific biases, such as the bias towards popular webpages, unforeseen and thus potentially unintended side effects may arise. Website administrators should be aware of these likely unwanted effects. Concerning the comparison between click biases and link insertion to efficiently steer humans on the Web, I find that, based on a few simple characteristics of a situation, there exists a rule of thumb about which manipulation strategy to prefer over the other.

1.9. Structure of this Thesis

The remainder of this thesis is structured as follows: In Chapter 2 I summarize the most relevant related work for this thesis . An important part of this is the concept of network search algorithms discussed in Section 2.1 and models of human navigation on the Web outlined in Section 2.2. A short overview of factors which are known to influence the link selection process of humans is given in Section 2.3. The central part of this cumulative thesis is Chapter 3, which includes all main publications as listed in Section 1.6. My personal contributions to each of those publications are described in Section 3.1, whereas detailed answers to the research questions tackled by this dissertation and the consequential findings are explained

in Section 4.1. Table 1.1 lists the main publications, their related research questions, topics, and main contributions. Additionally, Figure 1.4 provides the reader an illustrated structural overview of the research questions as well as the interdependencies between them. Implications and potential applications of the presented results are discusses in Section 4.2, whereas known limitations of this work are explained in Section 4.3. Finally, the last section of this thesis, Section 4.4, describes ideas and potential avenues for future work.

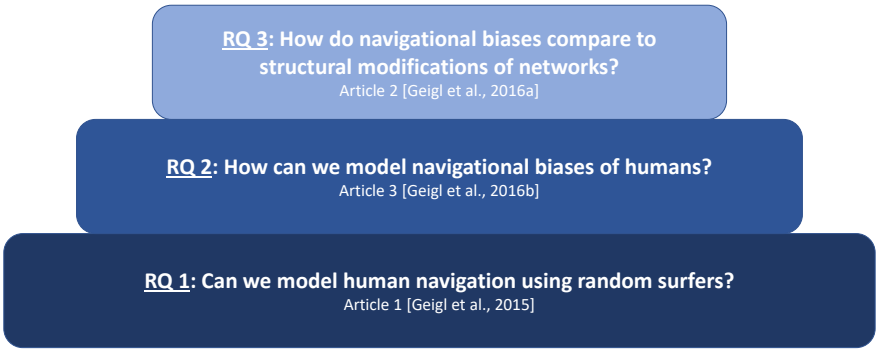


Figure 1.4.: **Structure of This Thesis.** This figure provides a structural overview of the research questions this thesis is tackling. It shows that the answer to the first research question—RQ 1—lays out the foundation for the other two research questions. In particular, it tackles the question whether or not the random surfer model can mimic human navigation on the Web. In a next step, to answer the question if and how we can model navigational biases of humans navigating the Web (RQ 2), I extend the random surfer model in such a way that it allows me to incorporate various biases. Subsequently, in RQ 3 I investigate and compare manipulation strategies regarding their efficiency in steering humans on the Web.

Table 1.1.: **Tabular Overview of the Main publications**. This table gives an overview of all publications, their scientific contributions and connection to the tackled research questions.

Article	RQ	Topic	Main Contributions
Article 1 [Geigl et al., 2015]	RQ 1	analyzing models of human navigation on the Web	Validating the random walk as an appropriate model of human navigation on the Web from a macroscopic point of view.
Article 2 [Geigl et al., 2016a]	RQ 2	exploring the effects of click biases onto the random surfers' typical whereabouts	Formalization of an approach to analyze consequences of navigational biases on the visit probabilities of specific pages of a website. Providing a solid theoretical model for further theoretical, empirical and practical analysis of human navigation on the Web. Applying the approach to empirical datasets to improve our understanding of the effects triggered by different biases.
Article 3 [Geigl et al., 2016b]	RQ 3	comparison of click bias and link insertion	Formalization of an approach to make click bias and link insertion comparable in an equitable manner. Utilizing the approach to evaluate both manipulation strategies on empiric datasets.

2. Related Work

In this chapter, I provide an overview of topics related to this thesis. A major part of the presented work is based on the concept of various search algorithms, which I briefly discuss in Section 2.1. In general, we categorize search algorithms into two groups. First, decentralized search—a deterministic model in which a search agent navigates through a network by greedily exploiting local information—and second, stochastic models—models that include any kind of randomness. Work related to the former model I discuss in Section 2.1.1, whereas models belonging to the latter are summarized in Section 2.1.2. Finally, in Section 2.2 I give a brief overview of how humans typically retrieve information from the Web, and how deterministic and stochastic navigation models have been used so far to imitate this process. Last but not least, in Section 2.3 I review literature dealing with biases that influence the link selection process of humans.

2.1. Network Search Algorithms

The goal of network search algorithms is to find a target node in a network by just traversing over existing edges. For evaluation of such algorithms pairs of randomly picked start and target nodes are passed to the algorithm. In the next step, the algorithm tries to find paths between each of the received pairs of nodes. Finally, the performance of the algorithm is measured as the so-called *delivery time*—that is the number of hops the algorithm needed to reach the target node.

2.1.1. Decentralized Search

Decentralized search algorithms try to solve this problem by greedily exploiting knowledge locally available to them. Hence, these algorithms can be categorized as deterministic models. Typical information locally available to these algorithms includes, for example, the neighboring nodes' popularity and/or similarity to the target node.

Kleinberg was the first scientists who conducted experiments investigating the importance of homophily (i.e., the probability that, based on the similarity between two nodes, there exists a direct link connecting them) for decentralized search [Kleinberg, 2000a,b, 2001]. In particular, he developed an efficient search algorithm for small-world networks that exploits information of the neighboring nodes' degree. Informally, the term "small-world networks" refers to networks in which the path lengths between all pairs of nodes is exponentially smaller than the number of nodes. In experiments conducted by Kleinberg [2000a] the author has used a slightly modified version of the well-known Watts and Strogatz model [Watts and Strogatz, 1998].

In particular, instead of starting out with a ring layout as in the original model, Kleinberg's method starts with a two-dimensional lattice. Additionally, the links within the model are directed, meaning that they can only be traversed in one direction. Moreover, Kleinberg introduced a new parameter p with which he manipulated the process of randomly inserting edges. Specifically, in the original Watts and Strogatz model nodes located next to each other on the underlying circular layout have been connected to each other. However, in Kleinberg's modified version we can connect some randomly picked nodes by setting p to a value higher than zero. For example, setting $p = 1$ results in a random graph in which each node becomes, independent of the ring layout, randomly connected to another node of the network.

To consider the distance between two nodes on the underlying lattice, Kleinberg's modified version of the Watts and Strogatz model includes another new model parameter named r. This parameter allows to manipulate the likelihood of connecting two nodes in such a way, that it

is proportional to the distance of these two nodes on the lattice. In his work, Kleinberg referred to the distance on the lattice as the *geographic distance* or *lattice distance*. Mathematically, he defined this distance $d(u, v)$ between two nodes u and v as the Manhattan distance between them on the underlying grid. Consequently, the probability of creating an edge from node u to v is proportional to $d(u, v)^{-r}$. To obtain a probability distribution Kleinberg normalized this value by the appropriate constant $\sum_v d(u, v)^{-r}$. Thus, setting r to zero results in the original Watts and Strogatz model, whereas with increasing value of r the insertion of links is biased towards pairs of nodes possessing a smaller lattice distance to each other. In other words, r controls how widely "networked" the underlying society of nodes is.

Utilizing this network generation method, Kleinberg investigated how well homophily, modeled as the lattice distance, can be exploited to efficiently solve search tasks in the network. Specifically, the information available to his search algorithm is the distance to the target node from (i) the current node, (ii) all its neighbors and (iii) all previously visited nodes. Based on this knowledge, the algorithm greedily navigates to the neighbor node that minimizes the lattice distance to the target. Exploring different values for r, Kleinberg found that with increasing values the algorithm is able to better exploit the geographic distance, while at the same time, long-range connections become less useful. Consequently, there exists a sweet spot for which the trade-off is optimal for the algorithm. In particular, Kleinberg has showed that this sweet spot is located at $r = 2$. Thus, Kleinberg stated the following Theorem:

Theorem 1 *There is a decentralized algorithm $\mathcal{A}$ and a constant α, independent of the number of nodes, so that when $r = 2$ and $p = q = 1$, the expected delivery time of $\mathcal{A}$ is at most $\alpha(\log n)^2$.*

Consequently, the presented algorithm based on homophily is particularly efficient in terms of delivery times in networks exhibiting a clustering exponent $r = 2$. In a next step, Kleinberg demonstrated that this property generalizes to networks generated on a d dimensional grid if the clustering exponent is set to $r = d$.

In the same year, Adamic et al. [2001] examined another important property for efficient decentralized search processes, namely the degree of neighbors. In particular, Adamic argued that in peer-to-peer file sharing networks (i.e., ad-hoc networks), the name of the file one is searching for is known. However, the location of the peer holding the file is unknown until we execute a real-time search. Consequently, during the search process, it is not possible for the algorithm to determine whether or not a certain step reduces the distance to the target peer. The naive approach implemented in such peer-to-peer networks was to perform a breadth-first search. Specifically, the algorithm asks all neighbors if they have the file, and if not, they should ask all their neighbors and so forth and so on. With the aim of improving this process, Adamic started out by comparing several search algorithms which made use of the knowledge that the degree distribution of such networks follows a power-law distribution [Kan, 2001].

Analyzing this property, she found, that despite the natural gravity of simple random walks towards high degree nodes, an explicit bias towards high degree nodes increases the performance drastically. In particular, Adamic numerically integrated the expected degree of the richest node among all neighbors and plotted the ratio between the degree of a node and the expected degree of the richest neighbor. For low degree nodes, the probability of having a connection to a node exhibiting a higher degree is very high. However, this probability drops when the node's degree increases. The exact point at which the probability of having high degree neighbor drops, is strongly dependent of the power-law exponent τ of the network's degree distribution. Thus, Adamic stated that to increase the efficiency of the search processes the algorithm should simply follow the degree sequence. In other words, the algorithm navigates towards the highest degree neighbors till it reaches the highest degree node of the network. Subsequently, by avoiding the highest degree node, it navigates towards nodes exhibiting approximately the second highest degree. Consequently, the algorithm quickly climbs towards the network's highest degree node followed by navigating down the degree sequence. However, Adamic also showed that this procedure only works if the network is sufficiently small, or if the power-law exponent of the degree distribution is close to 2 (i.e., $2 < \tau < 2.3$). Moreover, she proved that this is the most efficient way

to perform this type of sequential search [Adamic et al., 2001]. Precisely, she has found out that, the number of hops needed to find randomly picked targets in a power-law network scales sub-linearly with its number of nodes.

Based on the experiments conducted by Adamic and Kleinberg, the logical next step has been to combine both algorithms to create an even better one. Simsek and Jensen [2005] investigated this performance increasing opportunity. In their work, they formulate the search task as a decision-making task under uncertainty, in which the target is to minimize the expected path length l to the target. Furthermore, Simsek and Jensen define the expected path length l_{uv} from neighbor u to the target node v as the following series:

$$E(l_{uv}) = \sum_{\forall i} iP(l_{uv} = i) \tag{2.1}$$

Subsequently, they approximate the entire series by calculating just the first two terms under the assumption that the algorithm has access to the following information: (i) a list of already visited nodes, (ii) properties, such as degree and all attributes which are necessary to calculate the similarity between neighbors and the target node, and (iii) the relationship between the probability of observing a link and the similarity feature (i.e., homophily). In their paper they argue that the first two terms of the series denoted in Equation 2.1 capture enough information because they do not need to know the exact value of the estimation—only which of the neighbors possess the highest one. The estimation becomes zero if one of the neighbor nodes is the target node, whereas the second term of the series becomes zero if the neighbor has been visited previously. The latter is the case because previously visited nodes cannot have a link to the target—otherwise the algorithm would have found the target the first time it visited that node. In summary, we can explain the algorithm as following: If one of the current neighbors is the target node, navigate to it. Otherwise, navigate to the neighbor having the highest probability of being directly connected to the target node. Simsek and Jensen called this method *expected-value navigation*—or short *EVN*.

A nice property of EVN is, that if the algorithm has no information about the network's homophily, it reduces to the algorithm proposed by Adamic et al. [2001]. On the other hand, if no degree information is available EVN behaves equally to the algorithm introduced by Kleinberg [2000a]. Consequently, the algorithm performs at least equal to Adamic's as well as Kleinberg's algorithm. Please note that this algorithm is still deterministic as it does not involve any randomness.

2.1.2. Stochastic Search Models

Contrary to the algorithms reviewed in the last section, the following section deals with non-deterministic—stochastic—search models. The most simple and basic model belonging to this category is the random walk. In this model, an agent navigates the network by randomly traversing links until it finds the node it was looking for [Lovász, 1993; Woess, 1994]. Although that this is a very simplistic model, it proves to be a strong baseline for network search and lays out the basis for further research in the area of stochastic search models. Moreover, this model influenced many other areas of network science. For example, Blum et al. [2006] has based his webgraph generation model on random walks. Beside of that, in the area of community detection in networks, the random walk plays an important role [Pons and Latapy, 2005; Rosvall and Bergstrom, 2008; Zlatić et al., 2010]. In that case, scientists counted how often the random surfer traversed each link of the network. Many community detection algorithms are then based on the fact that nodes belonging to the same community are strongly interlinked to each other, whereas nodes outside of the community do not exhibit as many links to nodes of the community. In general, this characteristic represents the definition of communities in networks. Consequently, if the random surfer visits a node belonging to a certain community, it will most likely visit another node of the same community next, since most outgoing links point towards nodes belonging to the same community. Thus, frequently visited links indicate that the nodes connected by this link belong to the same community. Certainly, the opposite is also true.

Probably the most famous application of the random surfer has been presented in 1999, when Page and Brin [Brin and Page, 1998; Page et al., 1999] used the model to rank nodes of a network regarding their importance. The algorithm became famous under the name *PageRank* and was the foundation of Google, which is at the time of writing one of the largest and highest valued companies of the world. Page and Brin had the random surfer navigate the network for a very long time without a particular target. To avoid dead ends (i.e., nodes with incoming but not outgoing links) during the navigation, they introduced teleportation. This method allows the model to jump in each step with a small probability to a randomly picked node of the network. While the model was navigating the network, they counted how often each node gets visited by the random surfer. Based on these counts, Page and Brin calculated the probabilities for all nodes of being the one visited next by the random surfer. After letting the random surfer navigate long enough through the entire network, those probabilities converge to a state in which every node possesses a visit probability strictly higher than zero. By setting the model's teleportation probability to zero, this probability distribution over nodes becomes equal to the stationary distribution of a simple random walk. After applying this algorithm to webgraphs of the entire Web, Brin and Page used the derived probability distribution as a proxy for a website's popularity. Subsequently, they exploited this measure to rank the results of their web search engine. Indeed, this turned out to be an excellent idea.

At approximately the same time when Page and Brin invented PageRank, Kleinberg [1999] came up with a very similar idea of ranking nodes of a network. He called his algorithm *HITS*. Kleinberg's algorithm was able to rank nodes based on two properties: *hubs* and *authorities*. The former property is high if a node points towards many nodes that have a high authority. On the other hand, the authority value of a node is high, if many nodes that exhibit a high hub value link towards them. As a consequence of this definition, these two properties are directly dependent on each other. Thus, we need to calculate them in parallel. In particular, in the initial phase, the algorithm assigns each node a hub and an authority value of exactly 1. In the next step, the algorithm updates the authority values of all nodes by summing up the hub values of all neighbors pointing towards

them. After that, HITS calculates the hub values of all nodes by summing up the authority values of all outgoing neighbors (i.e., neighbors to which they have an outgoing link). Subsequently, the algorithm normalizes both measurements and starts over by calculating the authority values. This process HITS repeats until both measurements, namely the hubs and authorities vectors, converge. Although this algorithm seems more sophisticated than the one Page and Brin came up with, it never became as famous as PageRank did.

Overall, both methods were not intended to perform network search tasks efficiently, but rather to exploit the random surfer to generate rankings of nodes of a network. To examine the performance of stochastic search models we have evaluated such models in our own previous work [Geigl and Helic, 2014]. In particular, we conducted experiments in which we measured the performance of an algorithm, that bases its decision about where to navigate next on a convex mixture of homophily and popularity. Based on these mixtures, the algorithm weighted randomly picked the link to traverse next. While experimenting with various weightings of the two information sources used to calculate the convex mixture, we found that a ratio of $9 : 1$ of homophily and popularity respectively performs best. Our explanation for this finding is as following: At the beginning of each search task, it is unlikely that any neighbor exhibits a high similarity to the target node. Consequently, in that case, the homophily feature distribution over neighbors is similar to a uniform distribution. However, it seems that already minor information about popularity can help to steer the agent towards high degree nodes faster. Being on such a high degree node, it is very likely that at least one of the neighbors exhibits a high similarity to the target node. Thus, thenceforward homophily becomes more valuable than information about popularity. Based on this observation, we introduced a dynamical switch which changes the weights of the two features during the navigation. In particular, the agent starts by relying solely on the feature describing the popularity of neighboring nodes. However, as soon as it observes a small entropy in the homophily feature of neighbors (i.e., some neighbors are significantly more similar to the target than the others) it omits the influence of popularity and bases its decision solely on homophily. In experiments which we have conducted

on various empirical datasets, we observed that this method was able to outperform all the others examined in our work.

2.2. Search Algorithms Modeling Human Navigation

In the last section, we saw that we can utilize both, deterministic and stochastic, models to solve network search tasks efficiently. However, in practice, humans often perform such search tasks on their own. The following section reviews how humans usually navigate and retrieve information from the Web and by which search algorithms this process has been modeled so far. In the first part, I give a broad overview of the computational cognitive models used in the literature. Subsequently, I outline how deterministic (i.e., decentralized search) and stochastic (i.e., random surfer) models have been used to model this process.

2.2.1. Human Information Retrieval

A famous model which describes human information retrieval is called information foraging [Pirolli, 1997; Pirolli and Card, 1999]. Pirolli introduced this model in the late 90's. He based his idea on observations he made in nature. In particular, he took a look at how animals search for food. He stated, that animals that, for example, mostly consume cherries as food, have to decide for how long they stay at a certain tree eating its berries before they move on to the next one. The animals' decision whether or not to move on to the next tree is based on the number of remaining berries on the tree they are currently consuming food from. As the tree's berries get fewer and fewer—because of the animals eating them—there comes the point where there are still some berries left on the tree but it might take too long to find them. Hence, it makes sense to move on to the next tree which carries berries to abound. From the point of view of process optimization this means, that animals are minimizing the time it takes them to eat as many berries as possible, which, of course, makes sense as they would die without eating enough.

We can transfer the concept behind this process directly to information retrieval of humans on the Web. In particular, in 2001 Chi et al. [2001] assumed that humans are guided by the so-called information scent. This means that they have an expectation of which information will be available on a certain webpage, based on information available to them prior to visiting the site (e.g., title of a webpage or the anchor text of a link towards the page). Subsequently, humans decide whether or not it makes sense for them to forage this particular webpage for information. In the case they decide to not follow the link, they continue searching for another webpage having a high probability of containing information they are looking for, based on the information scent. On the other hand, if they visit the page, they start to collect information. Subsequently, as the type of information satisfying their needs becomes sparse, there comes the point at which they decide to move on to the next webpage.

Later, Fu and Pirolli [2007] successfully have used the idea of information scent to create the so-called *SNIF-ACT*. The abbreviation stands for "Scent-based Navigation and Information Foraging in the ACT cognitive architecture". The aim they pursued with this model was to predict as accurately as possible the outcome of the link selection process of humans. In other words, they wanted to forecast which links humans click on next. However, the SNIF-ACT models navigation between webpages, which is why other researchers started to examine models of inter page navigation. Soon, Kitajima et al. [2000] invented *CoLiDes*—Comprehension-Based Linked Model of Deliberate Search—which accurately modeled the inter page navigation of humans. Later, researchers combined both models and further included information about webpages visited previously by users with the aim to increase the accuracy with which they could predict the next click of a human surfing the Web [Juvina et al., 2005; Kitajima et al., 2007].

Beside of information foraging, scientists also examined other phenomena observed in nature to model human information retrieval. In particular, as berries do not only grow in bunches (e.g., on trees) there also exists a model which assumes that they grow on bushes. The important difference is that on bushes there is not a single particular place where one finds a bunch of berries, but rather they can be found distributed over larger areas.

Consequently, one has to constantly and dynamically adapt where to look next to find more berries. The cognitive model describing this process regarding human information retrieval is suitably called berrypicking [Bates, 1989].

Another model of human navigation on the Web is called orienteering [O'Day and Jeffries, 1993]. The model itself is to some extent similar to berrypicking. It describes the process of humans looking for information on the Web as a session dependent process. In other words, they pick only the most relevant information before they move on to the next webpage which probably looks at the same topic from a different angle.

Nevertheless, all those models are complex and thus computationally expensive. Furthermore, to be able to conduct experiments with them, one needs to have access to a vast amount of data (e.g., user sessions, webpage content). The complexity of these models, which further makes them hard to interpret and extend, and the need for appropriate empirical data were the reasons why I did not consider them in my thesis as models for human navigation. Moreover, it turned out that simpler models mimic human navigation on the Web accurately enough for the experiments conducted for this thesis [Geigl et al., 2015].

2.2.2. Decentralized Search

The idea to model human navigation behavior by using decentralized search algorithms started out with a nowadays famous experiment conducted by Stanley Milgram in 1969 [Travers and Milgram, 1969]. He came up with the idea of letting people forward letters through the United States of America with the purpose of investigating how humans perform on this task. In particular, he asked a few randomly picked people living in Omaha, Nebraska and Wichita, Kansas to forward a letter to a stockbroker living in a suburb of Boston, Massachusetts. Those cities he chose purposefully because they were considered especially distant regarding their geography as well as socially. Each of the randomly picked persons received a letter asking them to participate in the study. Furthermore, they were asked whether they knew the stockbroker by first-name and if so they could

directly forward the letter to him. The rest, all participants who did not know the stockbroker personally, were asked to send the letter to one of their friends of whom they tough that she or he might know the stockbroker by first-name, or at least has a friend who does know him personally. Conceivably, it took a while until the first letters finally made it to the target person living in Boston. The remarkable result of the experiment was that the average number of forwards till the letter finally reached the stockbroker was less than six. Based on this result the famous "Six-degrees of separation" phenomena was discovered. It states that every person living in the United States of America is, on average, separated by just 6 persons to everyone else. The experiment also emphasized how efficient humans are in routing information (e.g., in the form of a letter) across their social network of which they only possess local and thus, compared to the entire network, very limited knowledge. These findings have been validated again in a study conducted in 2003 by Dodds et al. [2003].

Furthermore, Leskovec and Horvitz [2008] examined this phenomenon in a slightly more modern experiment in which they measured the average length of connection chains between users of an online messaging network. They found out that in this huge network, which consists of millions of users, the average connection chain between two randomly picked users was only 6.6.

In 2012, Helic [2012] analyzed click trails of users playing a game on Wikipedia in which they were challenged to navigate from a randomly picked article to another randomly selected article with as few clicks as possible. Please note that this is the same target function which the decentralized search algorithms presented in Section 2.1.1 are trying to minimize. Back then the English Wikipedia consisted of around 10 million articles connected through roughly 250 million links. Helic found that, despite the vast amount of pages available on Wikipedia, human navigation patterns were similar to those observed in the experiment conducted by Milgram. Remarkably, the average path length was just 6.27—suggesting that certain commonalities exist between navigation of humans in social and information networks. Furthermore, this result again underlines how efficient humans can navigate through network-based systems.

Based on all those fascinating insights scientists began to examine characteristics of humans navigating the Web. One direction in this field of research has been to estimate the teleportation probability of humans while they were surfing the Web. In this case, teleportation represents the action of manually typing in a web address or clicking on a previously stored bookmark. To juxtapose, we call it navigation, if a user clicks on any link of a webpage. In 2010, Gleich et al. [2010] empirically measured this probability by analyzing click trails of humans. They reported an estimated damping factor (i.e., how likely a human keeps clicking on links available on the current webpage) between 0.6 and 0.72 for the entire Web. In other words, on the Web humans tend to click on links available on a webpage with a probability between 60% and 72%, whereas they teleport to any other page with a probability between 28% and 40%. Gleich et al. [2010] furthermore measured the damping factor of humans on Wikipedia and found that it was drastically less than the one observed on the entire Web (i.e., 0.33 and 0.43). One reason for this observation might be the way how humans access Wikipedia. Specifically, users often utilize search engines to navigate directly to the article of interest on Wikipedia. Consequently, there is no need to further navigate on Wikipedia, since users mostly find all required information on the page suggested by the search engine.

The results reported by Gleich et al. [2010] suggest that still a significant amount of all clicks made by humans on the entire Web can be classified as navigation. Hence, researchers directed their focus on the link selection process of humans. Specifically, they were interested which factors potentially influence this process. In 2012, West and Leskovec [2012a] manually designed various features they thought would play an important role in the link selection process of users. In a next step, they used machine learning techniques to find out the impact of each feature on the link selection process by fitting models to click trails of humans. The click trails analyzed by them originated from the same Wikipedia navigation game used in Helic [2012]. To prevent users from leaving Wikipedia or using search engines the game embedded all articles into its own interface which disabled all those disallowed options. By analyzing the feature importance of their machine learning method West and Leskovec [2012a]

found that the link selection process of humans could mostly be explained by popularity and similarity properties of neighboring nodes. Interestingly, those are the same characteristics used by Kleinberg [2001] and Adamic et al. [2001] in the node finding task to outperform the simple random surfer model.

Another study shedding light onto the process of humans navigating complex networks was conducted in the same year by Sudarshan Iyengar et al. [2012]. By utilizing network analysis methods, they found convincing evidence of why humans were able to perform navigation in such networks sufficiently efficient (but not necessarily optimal) for them. In particular, the authors showed that after the identification of a set of landmarks in the network, the performance in exploring it increased drastically. Furthermore, the landmarks were identified to mostly be popular nodes. This suggests that humans use a potentially similar method to the one utilized by Adamic et al. [2001] to generate an efficient search algorithm. Further evidence for this can be found in the work of Helic [2012] and West and Leskovec [2012a]. Both studies highlighted that empirical click trails of humans tend to include a so-called *zoom-out* phase in which they mostly navigate towards popular nodes of the network.

In 2013, Helic et al. [2013] applied stochastically biased random surfers with the purpose of modeling human navigation in information networks. They applied well-established decentralized search algorithms which initially were developed for social networks to information networks. Specifically, they compared paths generated by the models to empirically observed click trails of humans. In particular, they examined *greedy*, ϵ-*greedy*, *softmax rule* and *inverse distance rule* methods. The presented greedy algorithm consists of a search agent which strictly navigates to the neighboring node with the shortest distance to the target. In contrast to Kleinberg [2000b] the authors based the distance between nodes on information derived from hierarchies extracted out of the network. They performed the extraction of those hierarchies using the algorithm introduced in Muchnik et al. [2007]. The presented method is deterministic under the assumption that the hierarchy is fixed and that a tie is always processed in the same way (e.g., fixed enumeration of neighbors and always favoring the first one exhibiting the smallest distance). However, as human navigation potentially involves

randomness, they presented an ϵ-greedy algorithm which introduces some variance into the process. Specifically, in this method the agent reacts greedy with a probability of $1 - \epsilon$, whereas in the remaining cases it simply performs a random walk. Helic et al. [2013] performed a parameter search over a wide range of values for ϵ and determined that the one producing most human-like paths is 0.15. In other words, the model producing paths which exhibit the highest similarity to those of humans, performed in 15% of all hops a random walk. While going into further detail, the authors found that there was still space for improvement. Specifically, they found that the link selection process of humans is especially random at the beginning of their navigation. They conclude that users might need some time to gain orientation in the network before they can efficiently explore it—this observation is in line with the results presented in Sudarshan Iyengar et al. [2012].

Nevertheless, some findings presented in Helic et al. [2013] might also be caused by an artifact of the investigated empirical dataset (i.e., click trails of the same Wikipedia navigation game used by West and Leskovec [2012a]). In the game users started at an article selected uniform at random out of a vast amount of articles available in the dataset. Hence, it is very likely that the user has never seen the article (and the contained links) before. Consequently, randomly clicking on any of the links might be the action taken by the user. To account for this situation, Helic et al. invented a new method called *decaying ϵ-greedy* which adapted ϵ based on the path length. For example, the algorithm started out with $\epsilon = 0.8$ and after one hop it divided ϵ by 2. In that case ϵ would evolve with the path length as follows: 0.8, 0.4, 0.2, 0.1 and so forth and so on. Using this algorithm they were able to further increase the similarity of the generated paths to those of humans. Finally, they concluded that their method produced paths possessing characteristics which are similar to those generated by humans. This result highlights that the process of humans navigating the Web includes much randomness—especially if users explore new areas of the Web (i.e., webpages they have not visited in the past). A study conducted by Lamprecht et al. [2015b] further underlines the value of these results. In particular, they have shown that the idea to model the background knowledge of humans using hierarchies (i.e., ontologies) is a valid method

to generate human-like paths on Wikipedia. Specifically, they used various biomedical ontologies, such as the ICD-10, as the information source for a decentralized search algorithm. The informed search agent then had to solve search tasks on a subset of Wikipedia (i.e., all articles about biomedicine related topics). Their findings have shown that this approach was able to outperform simple random surfers, and more importantly, that the paths generated by the informed search agent exhibited characteristics similar to the one observed in empirical data.

Later, Singer et al. [2014b] were interested to what extent human navigation is influenced by previously visited pages. Technically speaking, they investigated whether human navigation is Markovian, meaning that the next click of a user is only dependent on the most recent click. In their experiments, they used several methods for model selection (i.e., to find the appropriate order of the Markov chain). Specifically, they applied Maximum Likelihood, Bayesian Inference, Akaike information criterion, Bayesian Information Criterion and simple Cross Validation. In their experiments, all methods showed a similar picture, namely that one specific model explains the observed empirical data best. In particular, the authors found that a Markov chain of first order fits the data best. Consequently, the probabilities on which link users might click next are independent of the links they had previously clicked on. This insight indicates that stateless models (i.e., memoryless) are not only sufficient but rather the best to go with for modeling human navigation on the Web.

To be able to explore even more hypotheses about human behavior Singer et al. [2014a] presented a Bayesian approach capable of testing various hypotheses against each other. The method presented in the article allows to investigate whether specific factors may have had an influence on humans navigating through different systems. The system to which the authors applied this method to included, but were not limited to, yelp—a online restaurant recommender—and last.fm—an online music platform. In their experiments they formulated domain-specific hypotheses which they subsequently compared to a uniform (i.e., random surfer) hypothesis. They found that for click trails stemming from the Wikipedia navigation game (the same data analyzed by West and Leskovec [2012a] and Helic et al. [2013]) one such domain-specific hypothesis is more plausible

than the one suggesting a random-walk-like behavior. In particular, the hypothesis describing a preferred navigation over semantically related topics was the most plausible one. This is in line with previous research in this field [White and Huang, 2010; West and Leskovec, 2012a; Singer et al., 2013]. Furthermore, it reflects the idea of exploiting homophily for decentralized search proposed by Kleinberg [2000b]. Nevertheless, for this thesis the model itself cannot be used as it only allows to investigate the plausibility of various hypothesis, but it cannot be utilized to generate new synthetic data.

2.2.3. Random Surfers as Model of Human Navigation

In the literature, the random surfer model has often been assumed to mimic human navigation on the Web accurately. Even the famous PageRank algorithm introduced by Brin and Page [1998] builds its basic idea upon this assumption. Nevertheless, only a few studies tried to validate this assumption on empirical data [Chierichetti et al., 2012; Singer et al., 2014b]. This might mainly be due to the fact, that in the past user data was not available and also not processable on a large scale. As both became available in recent years, researchers seem to have skipped this step in favor of more sophisticated models powered by machine learning, or similarly complex approaches [Kitajima et al., 2000; Chi et al., 2001; Juvina et al., 2005; Kitajima et al., 2007; West and Leskovec, 2012a]. Nevertheless, in other research areas which investigate human behavior, such as human traveling behavior, slightly modified random walks seem to be capable of accurately modeling human behavior [Brockmann et al., 2006].

2.3. Influencing Factors in Human Navigation

Researchers have demonstrated that there are a few factors, such as similarity or popularity, that play an important role in the link selection process of humans while they are navigating the Web [West and Leskovec, 2012a; Helic, 2012; Singer et al., 2014a]. However, many of these factors are not easy to shape if the aim is to manipulate a user's link selection process

actively. For example, one cannot alter the entire content of an article on Wikipedia just to increase its semantic similarity to other articles. If we attempt to exploit the popularity of articles to steer visitors, we might encounter a similar problem, that is, potentially connecting unrelated articles on Wikipedia. While the latter problem might be tackled by implementing boxes on the webpage such as *article of the month*, the former problem is not easy to solve. Naturally, the question arises if other extrinsic factors can be exploited to actively steer users navigating the Web. In the past, scientists invested a lot of effort into answering this question. One of the most prominent factors known to influence user navigation is the so-called *position bias* of humans which has been subject of many scientific articles in the past [Blunch, 1984; Joachims et al., 2005; Craswell et al., 2008; Buscher et al., 2009; Yue et al., 2010; Lamprecht et al., 2016].

The first study investigating the effect of the human position bias onto their behavior was conducted in 1984. In the experiments, Blunch examined how a user's decision about which answer to select in a multiple-choice questionnaire is influenced by the sorting (i.e., position) of the answers. He concluded that the leverage of an answers position is quite strong. Many years later Joachims et al. [2005] took advantage of modern technical equipment to further analyze these findings. Specifically, they utilized eye-tracking techniques to analyzes the sequence in which users look over a webpage. Even though the experimental setting was different to those utilized by Blunch, the results were in line with the findings of Blunch. The authors concluded that users typically read pages from top to bottom, resulting in more clicks onto links positioned at the top of a page. Buscher et al. [2009] conducted a similar study in which 20 participants had to engage in information foraging and page recognition tasks on 361 webpages while their visual focus was logged using an eye-tracking system. They found out that users start skimming the webpage in the left upper corner before looking at content located towards the bottom of the page. Further, Craswell et al. [2008] compared click position bias models to each other and analyzed which of them could best explain empirically observed data gathered from various search engines. Their conclusion was that not only the position but also the link's information sent plays an important role. In

other words, users start reading from the top and consequently prefer links on the top over the those positioned further down on the page. However, if the anchor text of a link does not look promising to lead them to a page containing the information they are looking for, they consider the next click in the same manner. Thus, there is a trade-off between the position of the link and its information scent. In general, strong evidence suggests that the position of a link drastically influence whether or not users click on them. Furthermore, the major advantage of this bias is, that it only requires altering the positions of already existing links on a webpage. This makes it especially easy to be implement in already existing webpages.

In further studies, the position bias was exploited to manipulate human behavior actively. In particular, Lerman and Hogg [2014] showed that by altering the position of items they were able to change the way peer recommendation worked. Exploiting this effect the authors steered user attention so as to improve the outcomes of peer recommendations. This insight is of utmost relevance for this thesis as it proves that there are ways to manipulate the link selection process. Furthermore, Lamprecht et al. [2016] already presented a recommendation system that provides suggestions about the repositioning of links. In particular, the authors of this study explored what happens if users click on only a few links on the top of articles on Wikipedia. They concluded that by doing so, users get stuck on just a few articles of the entire encyclopedia. Subsequently, they presented a system that recommends specific repositioning of links within an article to improve the system navigability for humans.

In summary, scientist found that we can exploit various biases to actively steer user navigation and the community already started to introduce frameworks capable of exploiting these findings to reach a certain goal (e.g., increased navigability). However, the effects thereof have not been investigated until today. Consequently, the examined effects arising due to various induced biases in experiments of this thesis are of practical relevance for website administrators considering to actively steer their visitors. Furthermore, scientists investigating human navigation behavior on the Web should keep in mind the existence of such effects.

3. Publications

3.1. Contributions to the Main Publications

The following section lists all of my contributions to the main publications of this cumulative thesis.

- **Article 1:** [Geigl et al., 2015] Geigl, F., Lamprecht, D., Hofmann-Wellenhof, R., Walk, S., Strohmaier, M. and Helic, D. (2015). Random Surfers on a Web Encyclopedia. *15th International Conference on Knowledge Technologies and Data-driven Business*

In this article I was responsible for designing the approach. Specifically, my task was to provide a framework for all experiments, execute them and examine all results. The framework was written by myself in Python. Furthermore, with the purpose of making the presented results reproducible for everyone, I made the framework available as open-source on Github[1].

The ideas for the experiments and employed methods were developed during discussion between Daniel Lamprecht, Simon Walk, Markus Strohmaier, Denis Helic and myself. Rainer Hofmann-Wellenhofer was responsible for preprocessing the data used for all experiments. He also produced the basic statistics of the dataset as seen in the article. All authors contributed to the writing of the article itself.

- **Article 2:** [Geigl et al., 2016a] Geigl, F., Lerman, K., Walk, S., Strohmaier, M. and Helic, D. (2016). Assessing the Navigational Effects of Click Biases and Link Insertion on the Web. *27th Conference on Hypertext and Social Media*

[1] https://github.com/floriangeigl/RandomSurfers

In this work, I was responsible for the design of the entire approach. This included the extension of my Python framework used in Geigl et al. [2015]. This extension is publicly available as open-source on Github[2]. Additionally, I carried out all experiments and produced the visual representations of the corresponding results.

The ideas for this paper and decisions about the applied methods originated in discussions held in conjunction with Kristina Lerman, Simon Walk, Markus Strohmaier, Denis Helic. All authors contributed to writing the paper.

- **Article 3:** [Geigl et al., 2016b] Geigl, F., Walk, S., Strohmaier, M. and Helic, D. (2016). Steering the Random Surfer on Directed Webgraphs *International Conference on Web Intelligence*

My contribution to this article was the design of the approach as well as the execution of the experiments and visual representation of the obtained results. Part of this work was to expand my existing Python framework to handle and process the new biasing methods introduced in this article. The updated version of the framework is publicly available as open-source on Github[3].

The concept for the paper and the development of the employed methods stem from discussions between Simon Walk, Markus Strohmaier, Denis Helic and myself. In the writing process of the paper itself all authors were involved.

3.2. Contributions to Further Publications

- **Journal 1:** [Walk et al., 2016] Walk, S., Helic, D., Geigl, F. and Strohmaier M. (2016). Activity Dynamics in Collaboration Networks. *ACM Transactions on the Web*

To this journal I mostly contributed by providing an efficient data preprocessing and visualization pipeline written in Python. This was particularly

[2] https://github.com/floriangeigl/RandomSurfers
[3] https://github.com/floriangeigl/RandomSurfers

relevant for the design of the model presented in this journal as we needed to get deeper insights into empirical data. Applying this onto large Stackexchange[4] datasets enabled us to integrate knowledge derived thereof into our model. Furthermore, I have been involved in discussions about the concept of the paper and the development of the model.

- **Journal 2:** [Hasani-Mavriqi et al., 2016] Hasani-Mavriqi, I., Geigl, F., Pujari, S., Lex, E., and Helic, D. (2016). The Influence of Social Status and Network Structure on Consensus Building in Collaboration Networks. *Social Network Analysis and Mining*

My major contribution to this paper was the development of the framework used for the experiments. In particular, this involved the basic architecture of the software. Moreover, a considerable part of my work for this paper was devoted to improving the framework to be capable of executing complex simulations within a reasonably short period of time. Besides, I was actively involved in the iterative process of developing the presented model and dealing with the interpretations of the results in the discussions.

- **Article 1:** [Hasani-Mavriqi et al., 2015] Hasani-Mavriqi, I., Geigl, F., Pujari, S., Lex, E., and Helic, D. (2015). The Influence of Social Status on Consensus Building in Collaboration Networks. *International Conference on Advances in Social Networks Analysis and Mining*

As the Journal [Hasani-Mavriqi et al., 2016] was an extension of this paper, my contributions to this paper are the same as the one listed above.

- **Article 2:** [Lamprecht et al., 2015a] Lamprecht, D., Geigl, F., Karas, T., Walk, S., Helic, D., and Strohmaier M. (2015). Improving Recommender System Navigability Through Diversification: A Case Study of IMDb. *15th International Conference on Knowledge Technologies and Data-driven Business*

I contributed to this article mostly by providing valuable feedback during discussions held between Daniel Lamprecht, Tomas Karas, Simon Walk, Denis Helic, Markus Strohmaier and me. Furthermore, I was involved in the writing process of the article and all discussions held about the development of the utilized framework.

[4] https://archive.org/details/stackexchange

- **Article 3:** [Helic and Geigl, 2015] Helic, D. and Geigl, F. (2015). Importance of Network Nodes for Navigation with Fractional Knowledge. *38th International Convention on Information and Communication Technology, Electronics and Microelectronics*

The ideas for this work stem from discussions between Denis Helic and myself. Initial experiments and results executed by me were important stepping stones for the presented article. Furthermore, I have been involved in the final writing process of the article itself.

- **Workshop Article 1:** [Ambite et al., 2017] Ambite, J., Lerman, K., Fierro, L., Geigl, F., Gordon, J. and Burns, G. (2017). BD2K ERuDIte: The Educational Resource Discovery Index for Data Science *4th WWW Workshop on Big Scholarly Data*

In this workshop article, I was responsible for the execution of a large part of the presented experiments. A major part of this was to find a reasonable good machine learning method for automated tagging of learning resources dealing with data science topics. Most of the experiments I executed during my time as a visiting PhD-Student at the Information Science Institute of the University of Southern California. During this time I was also an active part of all discussions dealing with this topic. Moreover, I developed the foundations for the portable and easy to extend automated crawling system needed to acquire new resources from the Web. I applied this framework on two different web portals to increase our dataset. Last but not least, I investigated experimental methods to automatically detect prerequisites between resources.

- **Workshop Article 2:** [Geigl and Helic, 2014] Geigl, F. and Helic, D. (2014). The Role of Homophily. *2nd International Workshop on Dynamic Networks and Knowledge Discovery*

First and foremost, I developed the framework in Python which allowed us to conduct all necessary experiments. Furthermore, I was responsible for gathering and preprocessing the data used in our experiments. Also, I ran all the experiments and produced all the numerically and visually presented results. The ideas for this work, which includes the used methods, originated in discussions with Denis Helic. The interpretation of the

results, their discussion and the writing of the article itself stem from the collaboration with Denis Helic.

- **Poster 1:** [Geigl et al., 2017] Geigl, F., Moik, C., Hinteregger, S. and Goller, M. (2017). Using Machine Learning and RFID Localization for Advanced Logistic Applications. *11th Annual IEEE International Conference on RFID*

In this poster, I was responsible for executing and visualizing all presented experiments. The synthetic data used in this article were created by Stefan Hinteregger, who also published an own article describing this procedure in detail [Hinteregger et al., 2017]. The ideas for the article origin from discussions held between all authors. Furthermore, all authors contributed to writing the article.

3.3. Random Surfers on a Web Encyclopedia

The following article tackles the first research question of this thesis by comparing empirical click trails to those produced by a random surfer. Moreover, it provides an in-depth description of the process used to analyze differences and commonalities between synthetic and empirical data. The article starts with a detailed explanation of the preprocessing steps applied onto logfiles of a large online encyclopedia to retrieve click trails of humans as accurately as possible. This process includes, for example, a detailed description of the rules determining wheter or not to merge consecutive sessions of a user into a single session. Furthermore, we provide a statistical overview of the processed click trails used for the experiments.

With the experiments conducted in preparation of the article, co-authors and I were specifically interested in determining whether or not the random surfer is capable of producing human-like navigation patterns—postulating that researchers can use the model to conduct further experiments about human navigation on the Web.

Specifically, this article presents a method that allows to incorporate empirical transition probabilities derived from click trails into the random surfer model—resulting in a biased random surfer. To account for lateral access, that is users arriving directly from webpages that are not part of the website under investigation (e.g., search engine), we additionally investigate page views gathered by the website. To measure how good the random surfer can simulate empirical data we correlate the stationary distribution of the random surfer with (i) the random surfer made biased with empirical data, and (ii) the normalized page views.

The results presented in the article indicate that the random surfer is a valid model of human navigation on the Web. However, as soon as users utilized search engines (i.e., lateral access) this was not true anymore. Specifically, my co-authors and I found that due to the lateral access the skewness of the stationary distribution decreases. In other words, search engines allow users to directly access very specific pages of a website without the need to navigate the website's hierarchical structure from top to bottom.

3.3.1. Abstract

The random surfer model is a frequently used model for simulating user navigation behavior on the Web. Various algorithms, such as PageRank, are based on the assumption that the model represents a good approximation of users browsing a website. However, the way users browse the Web has been drastically altered over the last decade due to the rise of search engines. Hence, new adaptations for the established random surfer model might be required, which better capture and simulate this change in navigation behavior. In this article we compare the classical uniform random surfer to empirical navigation and page access data in a web encyclopedia. Our high level contributions are (i) a comparison of stationary distributions of different types of the random surfer to quantify the similarities and differences between those models as well as (ii) new insights into the impact of search engines on traditional user navigation. Our results suggest that the behavior of the random surfer is almost similar to those of users—as long as users do not use search engines. We also find that classical website navigation structures, such as navigation hierarchies or breadcrumbs, only exercise limited influence on user navigation anymore. Rather, a new kind of navigational tools (e.g., recommendation systems) might be needed to better reflect the changes in browsing behavior of existing users.

3.3.2. Introduction

The last decades have seen immense growth of the Web, which now has an approximate size of over a billion webpages[5]. The Web provides people around the world with access to a host of information resources and serves uncountable use cases, such as gathering information, studying, making financial transactions, shopping, or booking hotels. To find relevant information in this huge information system, web users apply various information retrieval techniques. A very common—and probably the most basic and straight-forward—strategy consists of simply navigating between webpages by traversing the provided hyperlinks from one webpage to

[5] http://www.internetlivestats.com/

another. In many cases, users also jump directly to other webpages by typing the *URL* of the new target page in the browser address bar or by using a search engine and following one of the search results. These cases are typically referred to as *teleportation*, as users "teleport" from the current webpage to another one [Brin and Page, 1998].

The importance of web navigation is even further amplified by an alternative informational retrieval strategy—web search. Ranking algorithms used by search engines are based on variants of PageRank, which assigns weights based on hyperlinks [Brin and Page, 1998]. These ranking approaches assume a so-called random surfe—a model of a user who traverses the Web by following hyperlinks uniformly at random with a small chance of teleporting at each navigation step. In their original paper, [Brin and Page, 1998] suggested a damping factor of 0.85, meaning that, for each step, users traverse hyperlinks with a probability of 85%, while exhibiting a probability of 15% of teleporting to a page selected uniformly at random. The number of visits of an indefinitely navigating random surfer to each particular page is then a direct measure of page importance for web navigation and is used to rank search results.

Problem. Although the random surfer model has proven to be extremely useful in practice, only a few studies have analyzed the capabilities of this model to imitate real user behavior in different contexts. Moreover, most of these studies concentrated on empirically analyzing the damping or teleportation factor (such as Gleich et al. [2010]). In this work, we compare *clickstream data of real users* with the *random surfer model*. In particular, we are interested in analyzing how real users assess the importance of webpages for navigation and how that assessment compares to that of the random surfer. Moreover, we also study to what extent the navigation of human users is influenced by the modern search engines. To this end, we analyze page view counts, which also account for landing pages from search engines.

In particular, we are interested in answering the following research questions:

RQ1 Comparison of a random surfer with real users. To what extent does a random surfer with teleportation imitate user navigation behavior?

RQ2 Influence of search engines. How do search engines affect how users access and navigate websites?

Approach & Methods. For our analysis, we first calculate the stationary distributions of a *uniform* random surfer, traversing the information network uniformly at random with a teleportation probability of 15%. We then compare this stationary distribution with the stationary distribution of a *pragmatic* random surfer, who selects the links with a probability that is proportional to the transition counts from empirical data (human users). For the pragmatic random surfer we again use 15% teleportation probability. Finally, we compare stationary distributions of both uniform and pragmatic random surfer with the stationary distribution (normalized page view count distribution) of a *lateral* random surfer, which accounts for the lateral access from a search engine to a given website.

For the distribution comparison we calculate linear correlation factors and Gini coefficients to investigate the alignment of distributions, and the distributions' inequality, respectively.

Contributions. Our high-level contribution is a better understanding of human navigation behavior and how it compares to a navigational model such as the random surfer model.

Methodologically, we compute and analyze stationary distributions using a set of standard measures with a clear interpretation in the context of web navigation.

Empirically, we provide evidence that, despite its simplicity, a random surfer model is a very accurate model of basic human navigation behavior in our dataset. Our results suggest that the general navigation behavior of users is very much in line with the random surfer model—both assess the navigational page importance in a similar and highly skewed way, meaning that just a few pages are extremely important. These results also hold for cases where website operators decide to provide specific navigational structures (as in our dataset) such as navigational hierarchies. Users, as

well as the random surfer, do not make any particular distinction between different types of links present on the website. However, the lateral access from search engines reduces the imbalances, at least for human users, and need therefore to be taken into account when modeling user navigational behavior.

3.3.3. Related Work

Our work relies heavily on the random surfer model, which is a simple but well-studied model for modelling navigation on the Web [Lovász, 1993; Woess, 1994]. Apart from navigation, the random surfer model has also been applied to a variety of different problems such as graph generation and graph analysis. In particular, [Blum et al., 2006] used the model for the creation of webgraphs while Pons and Latapy [2005], Rosvall and Bergstrom [2008] and Zlatić et al. [2010] have applied the model to detect community structures in networks.

Algorithms such as PageRank [Brin and Page, 1998; Page et al., 1999] or HITS [Kleinberg, 1999], use the random surfer as the basis for calculating node centralities in networks. PageRank includes a parameter to define the probability of teleportation for the random surfer. This parameter is often referred to as the *damping-factor* α, representing the probability that the random surfer traverses one of the links pointing away from the current node. With probability $1 - \alpha$ it jumps to a network node chosen uniformly at randomly and continues surfing from there. In 2010, researchers have empirically measured this factor by analyzing clicktrails of humans and reported an estimated damping factor between 0.6 and 0.72 for the entire Web [Gleich et al., 2010]. In contrast, the damping factor for Wikipedia has been determined to be between 0.33 and 0.43. This difference in damping factors might be caused by the way users access Wikipedia—they use search engines that point them directly to the article of interest, rendering additional navigational efforts unnecessary. Researchers additionally investigated the connection between the damping factor and the convergence rate of the PageRank algorithm and found that it converges very fast for a value of 0.85 [Haveliwala and Kamvar, 2003; Kamvar et al., 2003]. However, in this paper we investigate the

influence of the damping factor onto the stationary distribution of the random surfers.

Qiu and Cho [2006] presented a framework that was able to personalize PageRank on a very small set of user-based clickdata for websites. Additionally, Al-Saffar and Heileman [2007] compared these personalized and topic-sensitive PageRank results with results from the unbiased (original) PageRank and came to the conclusion, that both ways of personalizing the PageRank produce a considerable level of overlap in the top results. In particular, the authors conclude that biases, which do not rely on the underlying link structure of the network under investigation, are needed to further improve the personalization of PageRank. In this paper we are interested in the stationary distribution of PageRank personalized by observed user transitions.

Researchers also looked closely into modeling human navigation behavior, using this biased random surfer model. For example, West and Leskovec [2012b] investigated human click trails of a navigation game played by humans on Wikipedia. Participants were asked to navigate from a given start article in Wikipedia to a specific target article, using as few clicks as possible. Using the results of this study, West and Leskovec [2012a] designed different features for steering a probabilistic random surfer. They also compared paths produced by the biased random surfer with those of humans and found that navigation of humans was based mostly on popularity and similarity biases. Helic et al. [2013] compared click trail characteristics of stochastically biased random surfers with those of humans. They concluded that biased random surfers can serve as valid models of human navigation. Furthermore, Singer et al. [2014b] conducted experiments to find out whether human navigation is Markovian, meaning that the next click of a user is only dependent on the most recent click. They showed that on a page level, human navigation can be best explained by first-order Markov chains. This finding is particularly relevant for us, as it allows us to use simple biases which do not consider previously visited nodes of the random surfer for our experiments.

3.3.4. Materials & Methods

Dataset

Austria-Forum. In this paper we use change and click data from Austria-Forum[6], an Austrian oline encyclopedia which was initially created more than two decades ago and restructured in 2009. Austria-Forum tries do distinguish itself from other well-established web encyclopedias by providing mechanisms to counteract some specific drawbacks: For instance, Austria-Forum tries to fight against the apparent (personal) biases of anonymous contributions by having (and enforcing) approved and named authors as the only contributors to the knowledge base. Authors are mostly academics well-established in their field, which has the positive aspect of thoroughness since they exhibit a personal interest not to produce literature of low quality. As the name suggests, the information published is geographically limited to all things concerning the country of Austria. Compared to other resources on the Web, Austria-Forum tries to transmit the knowledge on a more granular level. Not only does it provide users with several differently scoped articles, but also with entire digitized books as web books on a variety of different cultural and historical aspects of Austria. In order to increase the amount of displayed content, Austria-Forum added the capability of including entire pages from different external domains into their Wiki (e.g., of the German Wikipedia).

Most of the interactions of a user with an encyclopedia are limited to single page views, usually generated by direct requests via a search engine. For other users, who are interested in browsing the website and learning more about Austria, Austria-Forum has divided its content into several different categories, such as culture, people, scenery, nature and more, with the ultimate goal of keeping users engaged and increasing their session lengths as well as clicks on the website. The link structure of Austria-Forum mostly forms a huge hierarchy. Arriving at the main page users can choose one of 22 main categories and start navigating the hierarchy downwards to a specific topic (e.g., *main page/nature/fossils/amber*). Overall, nearly 90

[6] http://www.austria-forum.org

percent of all links within Austria-Forum can be categorized as hierarchical links.

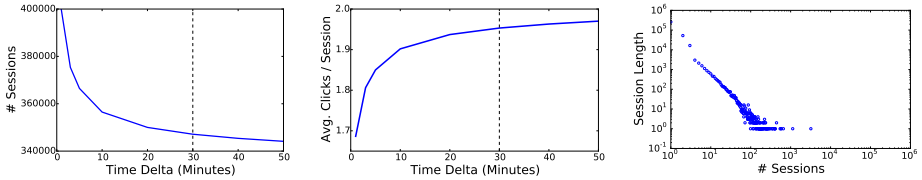
Log Data. For our analysis we use data that was gathered by logging *HTTP-Requests* on http://www.austria-forum.org, as well as other domains—such as the outdated http://www.austria-lexikon.at—which link to it. The observation period of our logs consisted of 59 days in April, May, and June of 2015.

Table 3.1 lists the parts of the *HTTP-Requests*, which were logged and provides a typical example *HTTP-Request* of a successful access request to Austria-Forum.

As we are mainly interested in user navigational behavior, we have extensively filtered the logs. First, we filtered the *Content-Type* to only include human-readable *HTML* pages, eliminating *XML*, *templates* and *attachments*. Second, *Referrers* and *Targets* indicating admin or irregular user behavior, were removed. The removed logs included previewing an edit for a page, pressing the upload button to attach files to articles, or *RSS-Feed-Requests*. Third, we have only kept *Requests* which successfully transmitted a page to the user, indicated by the *Response Code*. Therefore, we have removed all *Requests* with *Response Codes* other than 200 (OK).

Table 3.1.: **HTTP-Request Log Entry.** The table shows the HTTP parameters which were logged and an example query entry where the user came from Google and visited the page of *Waltraud Klasnic* which was successfully transmitted.

Date	2015-04-12 23:22:13,893
Method	GET
Response Code	200
Server Name	austria-forum.org
Target	[...]/Biographien/Klasnic,_Waltraud
Request-Query	None
Content-Type	text/html;charset=UTF-8
Session-ID	DC8F6B58BE968C906740853F4E6D4F41
Remote-IP	1.1.1.1 (for anonymity)
User-Name-Hash	None
Referrer	https://www.google.at/
User-Agent	Mozilla/5.0 (iPad; CPU OS 8_2 like Mac OS [...]



(a) Number of Sessions (b) Session Length (c) Session Distribution

Figure 3.1.: **Dataset Description.** The figures depict characteristics of our dataset as well as the highly skewed heterogeneous distribution of the resulting sessions. The y-axis of Figure 3.1a represents the number of sessions, while the x-axis represents the *Time Delta*—the maximum time a user can spend between two clicks without creating a new session. We identified 30 minutes to be a good compromise between numbers of sessions and session lengths. Figure 3.1b depicts the *average clicks* a user makes per session (y-axis) over different *Time Deltas* (x-axis). We highlighted the chosen Time Delta of 30 Minutes in both figures (Figure 3.1a and Figure 3.1b). As can be seen, increasing the Time Delta would only result in a very small increase of session lengths. Figure 3.1c visualizes the session lengths (y-axis) over the total number of observed sessions of specific length (x-axis). In our dataset we have many sessions of short lengths. With increasing session lengths, the number of observed sessions decreases, following a power-law distribution with $alpha = 1.52$ [Alstott et al., 2014].

In order to be able to identify pages with multiple *URLs*, *Requests* were normalized by removing the "*www.*" prefix as well as trailing slashes "/" when applicable. We stripped the data of all entries created by well-known *User-Agents* of crawlers, such as GoogleBot, or whenever the *User-Agent* contained a specific substring, such as *crawl*, *slurp*, *spider* or *bot*, which suggested bot activities. Furthermore, to identify bots which do not want to be recognized as such, we removed all entries which had the same *Target* as *Referrer*, which is abnormal behavior as standard page-refreshes usually retains the last *Referrer*. As many bots leave the *Referrer* in their *Requests* empty, all sessions with 4 clicks and more (47, 312) that had more than half of its *Referrers* missing were removed. Using this procedure, we removed a little over half (24, 293) of those sessions.

The specific method that was used on the server to generate *Session-IDs* is unknown to us. As we assume that the *Remote-IP* as well as cookies are likely considered for generating sessions, it is no simple task to combine, split, recreate and aggregate *HTTP-Requests* into navigational sessions. The number of *Session-IDs* exceeds the number of *Remote-IPs* by a large margin, which we presume is due to static *IPs* of some users such as schools using the same *IP* for all students, and users with browser add-ons to increase anonymity (so that no *Session-ID* can be mapped to that specific user). To make sure that sessions by the same user in different periods could be recognized as such, we introduced a time delta which—if exceeded between two requests—indicates the start of a new session. Hence, a smaller delta increases the number of sessions (Figure 3.1a). Decreasing delta too far would split sessions at pages where users spent a lot of time, even tough in reality the users were still active in their sessions.

Meiss et al. [2009] showed that separating *HTTP-Requests* (which they gathered on the entire Web) into sessions, can not be done in a clean way solely based on timeouts. Hence, they introduced the concept of logical sessions. In particular, users can have multiple logical session at the same time. For example: browsing domains consisting of mostly images in one tab while navigating on encyclopedias in others. Depending on the domain, average time spent per page varies greatly, as images can be consumed much faster than textual content. In their research they identified a timeout of 15 minutes as a good approximation of a logical user session. Since users tend to browse Austria-Forum for research, information, self-improvement, or just to educate themselves further, their sessions can be seen as logical as long as the time between two requests is not exceedingly long. It can be assumed that the time users spend on a page in an encyclopedia can be substantially longer than on an average webpage, due to long (and possibly) complex articles. Taking these factors into consideration, we found that setting our delta to 30 minutes still split several sessions while granting our users enough time for longer page visits. With delta set to 30 Minutes, the average session was 1.95 clicks long (Figure 3.1b).

The distribution of sessions can be seen in Figure 3.1c. It is apparent that the distribution is highly skewed and heterogeneous, indicating many short

sessions of few clicks (portrayed by many sessions which are situated low on the y-axis) and a few very long sessions (represented by a few sessions in the upper left corner). The short sessions are mostly users who were referred to Austria-Forum by a search engine and either instantly found the information they needed or ceased looking for the needed information on Austria-Forum.

Crawling the Link Structure. To compare the navigation behavior of website visitors to the random surfer, we have crawled the whole link structure of Austria-Forum. To this end, we have developed a simple web crawler that we pointed towards the main page of the website, and which then recursively crawled and followed all encountered (internal) links by pursuing a breadth-first strategy. Some of the encountered links were removed, such as all requests to display the raw Wiki sources for each page that are easily identified by the *skin=raw* parameter in the *URLs*. Further, links to binary files, such as *.mp3*, *.mp4*, *.jpg*, and many more, have been removed as well, as we are only interested in the navigation behavior of users while browsing and exploring the underlying website.

Limitations. We were not able to include the clicks of users within the web books of Austria-Forum in our study. Further, to simplify the data preprocessing, we cut off active sessions at midnight.

Random Surfer

Preliminaries. Mathematically, a random surfer is represented by a random walk on a weighted directed graph. Thus, we start by introducing some basic notion for such random walks.

Let $\boldsymbol{A}$ be the weighted adjacency matrix of a directed and weighted graph G with $A_{ij} > 0$ if node j points to node i and 0 otherwise. The value of A_{ij} represents the weight of the link from j to i. The weighted out-degree k_i^+ of a node i is defined as the sum over the weights of outgoing links:

$$k_i^+ = \sum_{j=1}^{n} A_{ji}. \tag{3.1}$$

Let $\boldsymbol{D}$ be a diagonal matrix of weighted out-degrees, so that $d_{ii} = k_i^+$ if $k_i^+ > 0$, otherwise we set $d_{ii} = 1$. The matrix $\boldsymbol{P}$, defined as

$$\boldsymbol{P} = \boldsymbol{A}\boldsymbol{D}^{-1}, \tag{3.2}$$

is than a transition matrix of a random walk on the weighted directed graph G. An element P_{ij} of the matrix defines the probability of a random surfer moving from node j to node i.

A stationary distribution of a random walk is defined as a probability of finding a random walker at a particular page in the limit of infinitely many steps. Algebraically, the stationary distribution is equal to the right eigenvector corresponding to the largest eigenvalue of the transition matrix $\boldsymbol{P}$. If the graph G is strongly connected and the transition matrix does not allow only periodic returns to a given state, then the largest eigenvalue of the matrix $\boldsymbol{P}$ is 1, and the stationary distribution is unique. In the case of a graph G that is not strongly connected, teleportation represents a simple technical solution as it connects each page to every other page with small weight. Teleportation also guarantees that there are not exclusively periodic returns to any given state in the network since there is a constant small probability to remain at the current page after teleporting the surfer to exactly that page. Thus, we therefore include teleportation in our calculations and calculate PageRank vectors of pages from G.

The calculation of the PageRank vector of the weighted adjacency matrix simplifies to (details are given in e.g., Newman [2010]):

$$\boldsymbol{\pi} = \boldsymbol{D}(\boldsymbol{D} - \alpha\boldsymbol{A})^{-1}\mathbf{1}, \tag{3.3}$$

where $\alpha \in [0, 1]$ is the damping factor.

Uniform Random Surfer. For the uniform random surfer we use the graph G, that we crawled from Austria-Forum. We do not set weights to hyperlinks for the uniform random surfer, thus we set $A_{ij} = 1$ if node j points to node i and 0 otherwise.

Pragmatic Random Surfer. To create a weighted adjacency matrix containing information of user transitions we first filter out teleportations,

meaning transitions which are not present in the adjacency matrix of the network. Afterwards we account for user transitions that we observed in the network adjacency matrix. For that purpose, we apply sublinear scaling to the transition counts, which is a common scaling technique in the field of information retrieval—a word which occurs, for example, 20 times in an document is not assumed to be 20 times more significant than a word occurring only once. For navigation we can make an analogous assumption, meaning that 20 observed transitions from page A to page B does not make this transition 20 times more significant than a single transition from, for example, page A to page C. In many cases there are several links between any two pages and some of these links are prominently presented in the user interface (e.g., in the navigation bar) inducing bias to the link selection process by users.

Therefore, sublinear scaling seems to be an appropriate approach to account for such situations. We scale the transition counts in the following way. Let t_{ij} be the number of transitions between pages j and i. We then calculate scaled transition count c_{ij} as:

$$c_{i,j} = \begin{cases} 1 + \ln t_{i,j} & \text{if } t_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{3.4}$$

After scaling down the transition counts we calculate the weighted adjacency matrix for the pragmatic random surfer in the following way. Let $\boldsymbol{C}$ be a matrix containing scaled transition counts, with C_{ij} being the scaled number of transitions between pages j and i. Further, we define a vector $\boldsymbol{v}$ which is a binary vector with $v_i = 1$ if the page i has been visited at least once by any of the users. Otherwise we set $v_i = 0$. Finally, let $\boldsymbol{V}$ be a diagonal matrix with vector v on the diagonal. Then the adjacency matrix of a directed network weighted with the scaled user transition counts can be calculated as follows:

$$\boldsymbol{A} = \boldsymbol{V}(\boldsymbol{A}_u + \boldsymbol{C})\boldsymbol{V}, \tag{3.5}$$

where $\boldsymbol{A}_u$ is the adjacency matrix of the unweighted graph as used for the uniform random surfer. After removing all rows and columns consisting

of only zeros this results in the adjacency matrix of the induced sub graph, which only includes nodes visited at least once by any user and all edges between those nodes (independent if traversed by any user or not). Now, the stationary distribution $\boldsymbol{\pi}$ may be calculated as given by Equation 3.3.

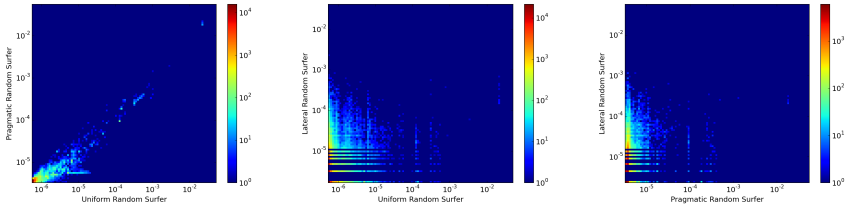
Lateral random surfer. We represent the lateral random surfer only through its stationary distribution. The stationary distribution of the lateral random surfer we calculate by simply normalizing page views we directly obtained from the server access logs. Specifically, we do not have a random surfer in this case, but observe the resulting stationary distribution of an underlying random navigation process.

Gini coefficient

The Gini coefficient is a metric for measuring inequality of a distribution. It computes the area between the Lorenz curve [Gastwirth, 1971] and the uniform distribution. Higher values indicate a larger difference and higher inequality. For our analyses, we calculate the Gini coefficient for the stationary distributions of all three random surfer types.

3.3.5. Results & Discussion

In our experiments we are interested in comparing and analyzing the differences and commonalities between the uniform random surfer model, the pragmatic random surfer model and the lateral random surfer model (cf. Section 3.3.4). We use the power iteration method to calculated the PageRank vector [Brin and Page, 1998]. In the first experiments we set α to a fixed value of 0.85. This correspond to teleportation probability of 15%, analogously to the original PageRank algorithm [Brin and Page, 1998]. Hence, the damping factor corresponds to the probability of a user to keep navigating over adjacent pages at each step. In later experiments we analyze the influence of various values for α. Figure 3.2 depicts the different correlations between the stationary distributions of all three random surfer models. In particular, the Pearson correlation coefficient

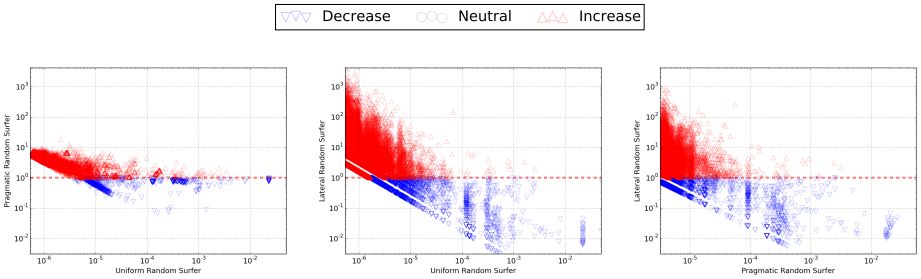


(a) Uniform - Pragmatic (b) Uniform - Lateral (c) Pragmatic - Lateral

Figure 3.2.: **Correlation Scatter.** This figure depicts the correlation of the stationary distributions of all three random surfer models on a log-log scale. It shows binned elements of a scatter plot using a heat map. Colors refer to the amount of elements falling into a bin. Note that the color range is also on a log scale. We identified the strongest correlation between the uniform and pragmatic random surfer (Figure 3.2a) with a Pearson correlation coefficient of ρ =0.98. In contrast, the correlation between the uniform and lateral random surfers (Figure 3.2b) is rather low with ρ =0.38. Figure 3.2c depicts the correlation of pragmatic and lateral random surfer with a Pearson correlation of ρ =0.47.

between the uniform and pragmatic random surfer of ρ =0.98 indicates nearly perfect positive correlation. Thus, this correlation analysis shows that there is a considerable overlap between the behaviors of the uniform and pragmatic random surfer models. In conclusion, the uniform random surfer model appears to be a very good approximation of the pragmatic random surfer—which in our case represents a proxy for user behavior—on Austria-Forum.

On the other hand, the uniform (ρ =0.38) and pragmatic (ρ =0.47) random surfer models exhibit only weak levels of correlation to the lateral random surfer. Further, the heat maps depicted in Figure 3.2 strengthen our findings, as the lateral random surfer, representing users entering the website from for instance search engines, exhibits higher probabilities to visit pages which are rated as unimportant by the uniform or the pragmatic random surfer. In other words, they are pointed directly to specific pages without the need to navigate the hierarchy of the website. Thus, search engines appear to reduce the need for users to navigate (hierarchical)



(a) Uniform - Pragmatic (b) Uniform - Lateral (c) Pragmatic - Lateral

Figure 3.3.: **Ratio of Stationary Probabilities.** The figures depict the ratio between stationary probabilities of pages for uniform, pragmatic and lateral random surfer. It contains basically the same information as Figure 3.2 transformed to ratios between values of the two stationary distribution under investigation. Figure 3.3a shows the ratio between the uniform random surfer (as baseline) and pragmatic random surfer. Pages that are important for the uniform random surfer appear to be less important for the pragmatic random surfer. However, this difference is not significant (corroborated by a high correlation between those two random surfers), meaning that both surfers rate (nearly all of) the same pages as the most important ones. The ratio between the uniform random surfer and lateral random surfer (3.3b) shows that the latter strongly emphasizes pages with low stationary distribution values of the uniform random surfer. Thus, users have a higher tendency to visit just one page—nested deeper in the hierarchical network structure—of the Austria-Forum. Similar observations can be made for the pragmatic and lateral random surfers (3.3c).

website structures and therefore are an important factor to include in (future) analyses of user navigation behavior.

Finding 1: Uniform random surfer is a very good model of user navigational behavior in our dataset. It correlates almost perfectly with the pragmatic random surfer constructed from the clickstream data. On the other hand, both uniform and pragmatic random surfer significantly differ from the lateral random surfer, which also reflects user visits from search engines.

In further experiments we varied α (damping factor of PageRank) and found that with lower values of α (e.g, $\alpha = 0.2$) the correlation between uniform and lateral random surfer increases from ρ =0.38 to $\rho = 0.49$, which suggests that higher teleportation probabilities better capture the lateral user access from search engines. However, at the same time the correlation between the pragmatic and the lateral random surfer decreases from ρ =0.47 to $\rho = 0.29$ for $\alpha = 0.2$ while the correlation between the uniform and the pragmatic remains stable and above 0.9. This result suggests that the lateral access to a website can not be solely captured by a random surfer with teleportation. Rather we need to extend this basic model. For example, we could use the basic model to also model navigational sessions. In this model teleportation probability increases with every new click to account for an increased likelihood of switching to a new session as the user makes progress in the current session.

Finding 2: To capture the lateral access to a website from a search engine we need a new kind of random surfer model.

Furthermore, we calculated and compared the ratios of stationary probabilities for each page and between all combination of three random surfer models to investigate commonalities and differences between them (see Figure 3.3). Although the uniform and pragmatic random surfer models exhibit a Pearson correlation coefficient of almost $\rho = 1$, there are a few pages with a ratio of 10 or 0.1. This means that those pages are 10 times more (less) important for the pragmatic random surfer than for the uniform random surfer. Figure 3.3a depicts a specific trend showing that pages with a low value in the stationary distribution of the uniform random surfer often obtain much higher values with the pragmatic random surfer. This difference is compensated by somewhat smaller importance for the pragmatic random surfer of the mid and high important pages for the uniform random surfer.

When comparing the ratios of the uniform and lateral random surfer models, we can see even stronger tendencies than in our previous analysis. The general shape of the differences remains the same, meaning less important

pages for the uniform random surfer become more important for the lateral one, but the magnitude of the differences is larger now and goes in some cases up to 100. Similar observation can be made for the most important pages for the uniform random surfer, which now become less important also in some cases by a factor of 100 (see Figure 3.3b). Finally, Figure 3.3c depicts the ratios of the pragmatic random surfer compared to the lateral random surfer. Again, we make a very similar observation as in the case of differences between the uniform and the lateral random surfer.

Finding 3: Although the assessment of individual page importance between the uniform random surfer and the pragmatic random surfer differs in some cases by a factor of 10, the assessments are generally very well aligned. The differences in assessments between the uniform and the pragmatic on the one side, and the lateral random surfer on the other side are often very large (factor of 100). The general alignment in the assessment between the lateral and other two models is not given in our dataset.

The Lorenz curves of the stationary distribution of all three random surfers are shown in Figure 3.4. The uniform random surfer achieves a Gini coefficient of 0.96. With a value of 0.83, the pragmatic random resulted in a lower coefficient. This means that the inequality in the stationary distribution of the pragmatic random surfer is lower than that of the uniform random surfer. In other words, the imbalances in the individual page importance are reduced as low importance page become more important, and vice versa highly important pages are less important for the pragmatic random surfer. Finally, the lateral random surfer exhibits the comparatively lowest Gini coefficient of 0.7. Due to the bias towards more specific pages located in lower levels of the website hierarchy in the lateral random surfer, this type of the random surfer is less likely to be directed towards highly popular pages as compared to the uniform random surfer.

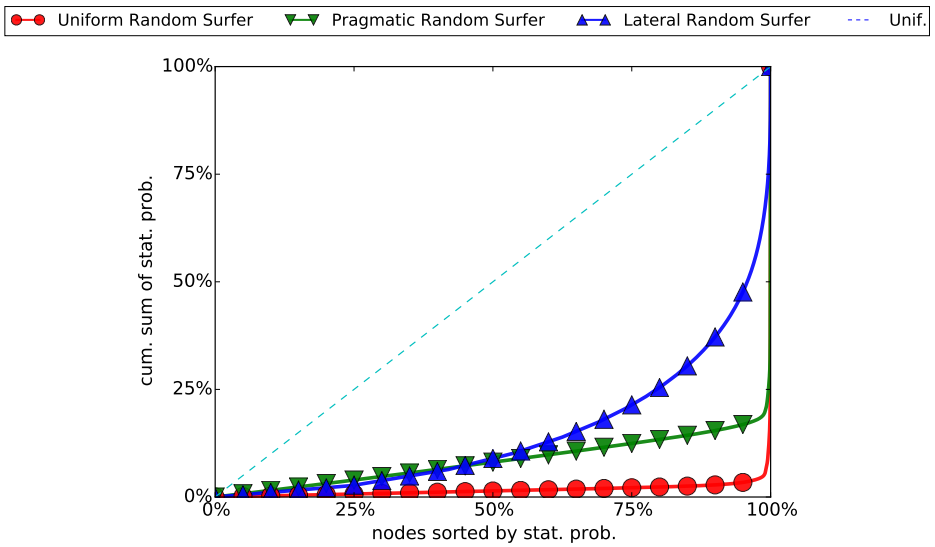


Figure 3.4.: **Lorenz-curves.** The plot depicts the Lorenz-curves of all three stationary distributions. We obtained the highest Gini coefficient of 0.96 for the uniform random surfer, followed by the pragmatic random surfer with 0.83. The lateral random surfer achieved the lowest Gini coefficient (0.7). Thus, search engines (or other in-going links from external pages) likely point users to very specific pages of the Austria-Forum, tackling the problem of directing users to high importance pages, helping to mitigate the influence of popular websites on navigation behavior.

Finding 4: The imbalances in the relative page importances are reduced for the pragmatic random surfer (only slightly) and for the lateral random surfer (significantly) as compared to the uniform random surfer. Direct lateral access from search engines towards more specific pages reduces the degree to which a random surfer is directed towards high importance pages.

3.3.6. Conclusions & Future Work

In this paper we presented new insights into the commonalities and differences between a uniform random surfer, a user clickstream biased (pragmatic) random surfer and a page visits biased (lateral) random surfer. We compared the navigation behavior of these three different random surfer models in an online encyclopedia, namely Austria-Forum. Using empirical user data we showed that the random surfer represents a good approximation of navigational user behavior for the investigated website—allowing researches to conduct user navigation experiments using a simple random surfer without the need to collect user clickstreams. Due to the low correlation between uniform and lateral random surfer we conclude, that the hierarchical structure of a website does not play such an important in role in terms of user navigation as it did before the rise of search engines. The majority of users enter the website using a search engine and leave after consuming the landing page. Hence, the uniform random surfer model is a good approximation of user navigation as long as no search engines are involved. However, hierarchical structures are needed for most search engines to rank the results of search queries. Nevertheless, the observed behavior leads to the question if website administrators should additionally provide page recommendations to keep users navigating their page.

Further experiments with varying teleportation probabilities (i.e., lower α) for the random surfer show that we can increase the correlation of stationary distributions between the uniform and lateral random surfer, but at the same time decrease the correlation between the pragmatic and the lateral random surfer. These differences in modeling navigational user behavior with and without search engines represent the directions for future work for modeling and hence optimizing navigational potential of a website.

Our results represent important insights for website administrators, search engine providers and researchers who want to broaden their understanding of user navigation and the models thereof. The contributions of this paper may serve as an interesting input to modify the models and for example link recommendation algorithms to influence navigational behavior of users.

With this work we contribute to the analysis of user navigational behavior by (i) providing a comparison of random surfer model data with clickstream data, (ii) a thorough analysis of the differences between these random surfer models on a web encyclopedia and (iii) presenting a methodology that allows us to estimate the optimization potential of a website in terms of keeping users navigating on the website as long as possible.

Future Work. In future work, we plan to verify our results on other websites where user clickstreams are available (e.g., the English Wikipedia). Furthermore, we want to use our model to test different types of biases introduced into the front end (e.g., recommendations of other pages) of a website to analyze to which extent such biases are able to influence users in their navigation. Another idea is to modify the order of recommendations in a recommendation network and analyze—based on the assumption that recommendations on the top are clicked more often [Blunch, 1984; Joachims et al., 2005; Murphy et al., 2006; Craswell et al., 2008; Yue et al., 2010; Lerman and Hogg, 2014; Dimitrov et al., 2016]—the influence thereof.

3.3.7. Acknowledgments

This research was in part funded by the FWF Austrian Science Fund research project "Navigability of Decentralized Information Networks" (P 24866). We thank Gerhard Wurzinger for providing access to Austria-Forum server logs.

3.4. Steering the Random Surfer on Directed Webgraphs

The following article strives to give an answer to the second research question. To that end, the article presents a method to explore the effects arising due to various human biases which influence the link selection process. We use the random surfer as a valid model of human navigation on the Web [Geigl et al., 2015] to present a novel method which allows to encode biases into the random surfer model intuitively. In the remainder of the article, we use this method to explore the effects of typical biases, such as the bias towards similar webpages.

First and foremost, the results gathered in this work indicate that all investigated biases increase the certainty of the link selection process. Second, we find that, from a macroscopic point of view, click biases are capable of drastically altering the typical whereabouts of users. Last but not least, the conducted in-depth analysis shows that in some extreme cases strong side effects emerge and spread throughout the entire website. Specifically, my co-authors and I found that some biases lead to a situation in which large fraction of webpages enter a so-called "ground-state". In this state, the probability that these sites are visited by the random surfer is essentially zero.

The insights described in this article suggest that website administrator should carefully decide whether or not to exploit such biases on their websites. To further aid them in their decision-making process we open-source the presented framework so that they can examine the potential effects of biases in an offline setting.

3.4.1. Abstract

Ever since the inception of the Web website administrators have tried to steer user browsing behavior for a variety of reasons. For example, to be able to provide the most relevant information, for offering specific products, or to increase revenue from advertisements. One common approach to steer or bias the browsing behavior of users is to influence the link selection process by, for example, highlighting or repositioning links on a website. In this paper, we present a methodology for (i) expressing such *navigational biases* based on the *random surfer model*, and for (ii) measuring the consequences of the implemented biases. By adopting a model-based approach we are able to perform a wide range of experiments on seven empirical datasets. Our analyses allows us to gain novel insights into the consequences of navigational biases. Further, we unveil that navigational biases may have significant effects on the browsing processes of users and their typical whereabouts on a website. The first contribution of our work is the formalization of an approach to analyze consequences of navigational biases on the browsing dynamics and visit probabilities of specific pages of a website. Second, we apply this approach to analyze several empirical datasets and improve our understanding of the effects of different biases on real-world websites. In particular, we find that on webgraphs—contrary to undirected networks—typical biases always increase the certainty of the random surfer when selecting a link. Further, we observe significant side effects of biases, which indicate that for practical settings website administrators might need to carefully balance the desired outcomes against undesirable side effects.

3.4.2. Introduction

Millions of people access the Web on a daily basis to conduct a variety of different tasks, such as maintaining social contacts, buying products in webshops, gathering information, or just passing time. While surfing the Web, users usually either traverse static (e.g., breadcrumb navigation) or dynamic (e.g., personalized recommendations) links, type in the URL of a website, or use a search engine to find their desired resource. Previous

research has already established that users exhibit a 65% probability of exploring websites through static links [Gleich et al., 2010]. Many researchers already directed their efforts towards these 65% of clicks, analyzing different aspects of the navigational behavior of users, such as estimating the probability of a user to traverse a given link by analyzing user click and interaction trails [West and Leskovec, 2012a,b; Helic et al., 2013; Singer et al., 2014b; Walk et al., 2014, 2015, 2016]. Granka et al. [2004] demonstrated how specific user behaviors directly influence which links are selected for browsing a website. Furthermore, Lerman and Hogg [2014] showed that users can be steered towards certain links by manipulating the interface (e.g., the position of links). In practice, website administrators often modify the interface to steer visitors towards certain pages. For example, owners of online shops might want to steer visitors towards best-sellers to increase revenue by modifying the probability that those pages are visited (e.g., by highlighting or repositioning links towards them).

Problem & Approach. Website administrators are typically not aware of the exact effects and implications of a particular modification. Moreover, such modifications may also affect the selection of other links and may trigger unpredictable and complex side effects. In fact, we still know very little about the (potentially) complex impacts of modifications and manipulations of linking structures on websites. In this paper we set out to close this knowledge gap. Specifically, we aim at assisting website administrators in estimating the consequences of inducing specific biases on their website. In addition, we seek to increase our understanding of the emerging effects through biased link selection processes.

To this end, we present an approach for assessing the impact of different navigational biases on visit probabilities and browsing dynamics on directed webgraphs. We adopt a model-driven approach, based on the well-established random surfer model [Brin and Page, 1998], to simulate users browsing a website. Although the model itself is very simple and straightforward, it provides a good approximation of actual user browsing behavior [West and Leskovec, 2012a,b; Helic et al., 2013; Geigl et al., 2015].

In particular, we are interested in answering the following research questions:

Website Coverage. Can certain biases increase the effective number of pages visited by the random surfer or do they trap the surfer within specific (small) parts of a website?

Surfer Guidance. Given a specific bias, what is the degree of guidance (i.e., certainty) induced by that bias? How many options (on average) are random surfers confronted with when they select the next link to follow? In other words, to what extent are browsing decisions purely random and to what extent do they adhere to a certain structure?

Webpage Response. How do visit probabilities of webpages respond to a given bias and how do such responses propagate through a network? For example, are those responses coupled and how? Specifically, what is the coupling between neighboring pages?

Contributions. In this paper we extend our framework[7] for simulating biased random surfers on networks [Geigl et al., 2016a] by analyzing, comparing and modeling the impact of unbiased and biased random surfers on directed webgraphs. In particular, we study real-world, empirical networks to obtain new insights into the global and local effects of different biases on the random surfer. Our results suggest that we can strongly and specifically influence the effective website coverage by using certain biases. Furthermore, we show that typical biases, such as popularity biases, always increase the certainty of the link selection process (i.e., provide a better guidance for the random surfer). Finally, we analyze potentially unwanted side effects that occur when inducing different biases, which affect a large proportion of all webpages of a website.

3.4.3. Related Work

The random surfer is a simple but well-established model, which has already been extensively investigated by researchers in the past [Lovász, 1993; Woess, 1994]. It also represents the basis for the calculation of more

[7]The framework is available as open-source software at https://github.com/floriangeigl/RandomSurfers

complex node properties such as PageRank [Brin and Page, 1998; Page et al., 1999] or HITS [Kleinberg, 1999]. The PageRank model includes a parameter for the probability of the random surfer to teleport to a different node. This parameter is also often referred to as the damping factor α, which is the probability that the random surfer continues to follow links at the current node. Conversely, with probability $1 - \alpha$ the random surfer "jumps" to a randomly selected node and continues traversing links from there. Gleich et al. [2010] empirically analyzed human click trails and estimated that the damping factor is in range between 0.6 and 0.725 for the Web.

Researchers have also manipulated the random surfer by applying different biases on the model to influence the link selection process [Goldhirsch and Gefen, 1987; Hill and Häder, 1997; Qiu and Cho, 2006; Fronczak and Fronczak, 2009]. In such cases, the links are weighted and the link selection is not uniformly at random anymore. Instead, the link selection probability is proportional to the link weights. Richardson and Domingos [2001] used biased random surfers in the field of information retrieval. In their work they were able to outperform PageRank in terms of quality of the results. Despite an increase in computational costs and memory requirements, the authors argue that the algorithm is still reasonably feasible for large-scale search engines.

Al-Saffar and Heileman [2007] later compared personalized and topic-sensitive PageRank with the original formula and came to the conclusion that both ways of personalization produce a considerable level of overlap in the top results. The authors conclude that new biases, which should not rely on the underlying link structure, are needed to improve the personalization of modified PageRank algorithms. In this paper we are not interested in improving the personalization of a node ranking algorithm. Instead, we want to broaden our understanding of the effects of different biases on the stationary distribution of a random surfer.

West and Leskovec [2012b] investigated human click trails from a Wikipedia navigation game. Based on the results of this study, they designed different features for steering a probabilistic random surfer [West and Leskovec, 2012a]. In their work they compared paths produced by the biased random

surfer with those of humans. They found that human navigation was mostly based on popularity and similarity of articles. To further investigate this, we focus in this paper on the effects of popularity biases.

Helic et al. [2013] compared click trail characteristics of stochastically biased random surfers with those of humans. Their conclusion was that biased random surfers can serve as valid models of human navigation. In our previous work, we validated this finding by showing that the result vector of PageRank and click data biased PageRank have a strong correlation for the example of an online encyclopedia [Geigl et al., 2015].

Regarding the number of pages which are effectively visited by random surfers, Hwang et al. [2012] investigated the probability of returning to the start node of random surfers in scale-free networks. They found that this probability depends on the degree of the starting node, and thus the total distribution is similar to a power-law distribution. By investigating the stationary distribution of the random surfer, we circumvent this problem as the distribution is independent of the starting point.

In previous work we have investigated how biases towards different subgroups of nodes influence the visit probability of the random surfer and how such biases compete with link insertion [Geigl et al., 2016a]. In this paper we extend our methodology to allow for the simulation of biases based on structural properties of nodes, expanding the arsenal of tools to analyze the effects of biases on random surfers.

3.4.4. Methodology

First, we introduce a basic notion for random surfers on a directed graph. Let $\boldsymbol{W} \in \mathbb{R}^{n \times n}$ be the adjacency matrix of a directed graph with $W_{ij} = l$ where l is the number of links that point from node j to node i (i.e., 0 if there are no links). The out-degree k_i^+ of a node i is defined as the number of outgoing links, that is $k_i^+ = \sum_{j=1}^{n} W_{ji}$. Further, let $\boldsymbol{D} \in \mathbb{R}^{n \times n}$ be a diagonal matrix of weighted out-degrees (i.e., $d_{ii} = k_i^+$). Then the equation

$$\boldsymbol{P} = \boldsymbol{W}\boldsymbol{D}^{-1} \tag{3.6}$$

defines the transition matrix $\boldsymbol{P}$ with elements P_{ij} equal to the probability of a random surfer moving from node j to node i.

If we think about nodes as states and links as transitions between states, the transition matrix $\boldsymbol{P}$ defines a first-order Markov chain. If a Markov chain is irreducible (i.e., any state can be reached from any other state with a non-zero probability) and aperiodic (i.e., returns to all states occur at irregular times), the chain has a unique stationary distribution $\boldsymbol{\pi}$. This distribution represents the probability of finding a random surfer on a given node in the limit of large number of steps. To ensure that the Markov chain $\boldsymbol{P}$ is irreducible we only use the largest strongly connected component from our datasets. On the other hand, a random walk on a connected directed graph is aperiodic if and only if there is no integer greater than 1 that divides the length of every cycle in the graph. Thus, it suffices to show that there is at least one cycle of length 2 and one cycle of length 3 in a directed graph for it to be aperiodic. We find that in all our datasets.

An algebraic solution for the stationary distribution yields $\boldsymbol{\pi} = \boldsymbol{P}\boldsymbol{\pi}$. Thus, the stationary distribution is an eigenvector of the transition matrix $\boldsymbol{P}$, corresponding to the largest eigenvalue 1. In related literature, the stationary probability of a node is often referred to as the *energy* of a node [Langville and Meyer, 2004; Bianchini et al., 2005; Geigl et al., 2016a]. As the random surfer is a conservative process [Ghosh and Lerman, 2012], the system total energy is constant and equals 1. However, the distribution of energy over nodes is dependent on the link selection process of the random surfer under investigation.

Inducing Bias. In practice, we can influence the link selection process of users by, for example, repositioning links [Lerman and Hogg, 2014]. In our analysis, we bias the random surfer by weighting links in a given network to achieve similar effects. To that end, we investigate different structural properties of nodes and weight all links pointing towards nodes proportional to a given structural property. For example, to induce a *popularity* bias we weight links according to the popularity (i.e., degree) of the target node.

Algebraically, we represent a bias as a diagonal matrix $\boldsymbol{B} \in \mathbb{R}^{n \times n}$ with node weights $\boldsymbol{b} \in \mathbb{R}^n$ on its diagonal. Matrix $\boldsymbol{W'}$ is then the weighted adjacency matrix of the biased network, which we calculate as the product of $\boldsymbol{B}$ and $\boldsymbol{W}$:

$$\boldsymbol{W'} = \boldsymbol{BW} \,. \tag{3.7}$$

Using the weighted out-degree diagonal matrix $\boldsymbol{D'}$ of $\boldsymbol{W'}$ we calculate the corresponding biased transition matrix $\boldsymbol{P'}$ as:

$$\boldsymbol{P'} = \boldsymbol{W'}\boldsymbol{D'}^{-1} \,. \tag{3.8}$$

As before, we have the stationary distribution satisfying the right eigenvector equation (i.e., $\boldsymbol{\pi'} = \boldsymbol{P'}\boldsymbol{\pi'}$), where we use $\boldsymbol{\pi'}$ to denote the stationary distribution of the biased random surfer. Note that this methodology adapts and extends our previous work [Geigl et al., 2016a]. However, in this paper we do not bias towards groups of nodes but rather set the probability of traversing a link proportional to structural properties of its target node. Hence, all links of the network are affected by the induced bias as opposed to our previous work, where only links pointing towards selected nodes were affected. In practice this would mean that we highlight each link proportional to a property of the target page (e.g., popularity).

3.4.5. Experimental Setup

Website Coverage. In general, biases allow us to manipulate the link selection process of random surfers and influence the visit probabilities of specific nodes. To investigate the bias effects on the *effective* number of visited pages (i.e., pages with practically relevant visit probability) we calculate three properties of the stationary distribution. First, we analyze the visit probability of the most visited page of each website to see and compare how likely the random surfer can be found on just this single page. In all our datasets we find that the most visited page is always the *home page* (i.e., main/entry page) of the website. Second, we use the complementary cumulative distribution function of the stationary distribution (i.e., $CCDF(\boldsymbol{\pi})$) to determine the number of pages on which

the random surfer can be found with a probability higher than 95%. Third, we analyze the entropy of the stationary distribution H, which measures the uncertainty in the current location of the random surfer. We calculate H as:

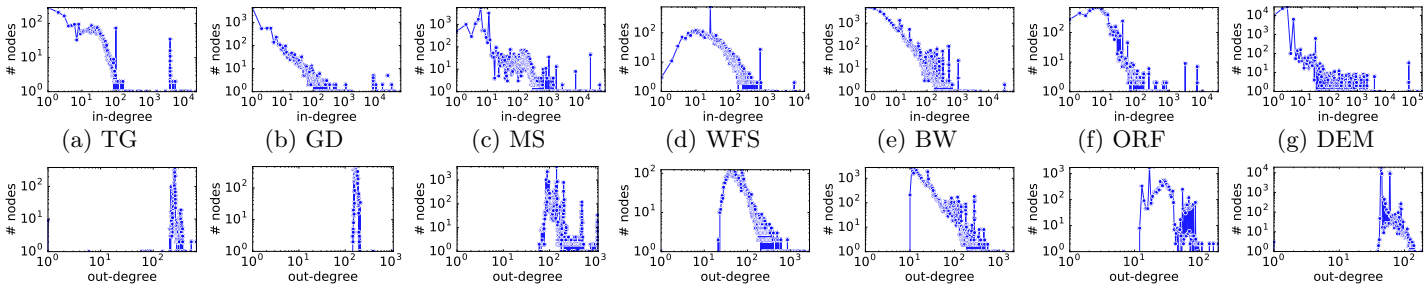
$$H = -\sum_i \pi_i \log_2 \pi_i \,. \tag{3.9}$$

Surfer Guidance. To analyze the dynamics of the link selection process we calculate the entropy rate of the random surfer. Entropy rate is the average entropy of all decisions made by the random surfer in the limit of a large number of steps. Thus, it measures average uncertainty in all the decisions made by a random surfer. We calculate the entropy rate H_{rate} as:

$$H_{rate} = -\sum_{ij} \pi_j P_{ij} \log_2 P_{ij} \,. \tag{3.10}$$

Note that the entropy of each node is weighted with the corresponding value of the stationary distribution. Thus, the uncertainty of the random surfer at a highly visited page has a greater impact on the entropy rate than the one from a less frequently visited page.

Webpage Response. To improve our understanding of changes in the visit probabilities of the random surfer due to different biases, we investigate how each individual page is affected on a microscopic level. We do that by analyzing heat maps which are based on log-scaled scatter plots between stationary distributions of unbiased and biased random surfers.



(a) TG (b) GD (c) MS (d) WFS (e) BW (f) ORF (g) DEM

(h) TG (i) GD (j) MS (k) WFS (l) BW (m) ORF (n) DEM

Figure 3.5.: **In-Degree and Out-Degree Distributions of our Datasets.** The figure depicts the in-degree (*top*) and out-degree (*bottom*) distributions of all datasets. The in-degree distributions are skewed towards few pages with a very large in-degree, which is typical for webgraphs. In contrast, the out-degree distributions are more homogeneously distributed, except for the Wikipedia datasets (*WFS*, *BW*). This is due to the way the websites are designed. In webshops (*TG*, *GD* and *MS*) and online media libraries (*ORF* and *DEM*) most pages are similarly structured and thus contain roughly the same number of outgoing links. On Wikipedia pages widely vary in their length, which is why the out-degree varies more strongly.

3.4.6. Biases

In this section we introduce the investigated biases and the intuitions behind them.

Popularity Bias. With this bias we steer the random surfer towards popular nodes. For example, in webshops operators may want to increase visits (and thus potentially sales) of frequently visited products. In encyclopedias and media libraries, operators may have an interest in further increasing the visibility of popular articles or movies. For popularity we use the *degree* of webpages as a proxy and set $b_i = k_i$, where k_i is the total degree (in and out) of node i.

Unpopularity Bias. To dampen the natural attraction of popular nodes we may want to induce an unpopularity bias. As a webshop operator, this could be used in a strategy to clear out stocks by increasing the visibility of unpopular items. In encyclopedias or media libraries operators may want to ease-up and diversify navigation to specific (mostly unpopular) pages and decrease the visibility of popular pages. For unpopularity bias we use the *inverse degree* and set $b_i = 1/k_i$ for all i.

Eigenvector Centrality. A bias proportional to eigenvector centrality of a node has already been investigated by researchers on *unweighted*, *undirected* networks [Parry, 1964; Demetrius and Manke, 2005; Delvenne and Libert, 2011; Sinatra et al., 2011]. In such networks the eigenvector centrality bias produces the highest possible entropy rate[Sinatra et al., 2011]. Therefore, we include this bias in our experiments as a baseline. Eigenvector centrality is the right eigenvector of the adjacency matrix of a network and satisfies $\boldsymbol{W}\boldsymbol{v} = \kappa_1 \boldsymbol{v}$, where $\boldsymbol{W}$ is the weighted adjacency matrix of the network and κ_1 the largest eigenvalue of $\boldsymbol{W}$. Thus, we introduce the eigenvector centrality bias by setting $b_i = v_i$ for all i.

3.4.7. Datasets

To simulate navigational biases "in the wild", we have crawled webgraphs of seven different websites. In particular, we collected data from three

Table 3.2.: **Network Statistics.** The table displays the basic statistics of our datasets, with n being the number of nodes, m the number of edges, and d the network diameter.

Dataset	Category	n	m	d
ThinkGeek (TG)	webshop	3,884	1,002,226	3
GetDigital (GD)	webshop	8,258	2,101,254	21
Milan-Spiele (MS)	webshop	21,566	3,128,693	70
Wikipedia for Schools (WFS)	encyclopedia	6,796	646,646	4
Bavarian Wikipedia (BW)	encyclopedia	32,734	1,324,839	9
ORF TVThek (ORF)	media library	9,799	301,844	10
Das Erste Mediathek (DEM)	media library	70,063	3,448,513	2274

webshops that deal with "geeky" gadgets or board games (ThinkGeek[8], GetDigital[9] and Milan-Spiele[10]), two online encyclopedias (Wikipedia for Schools[11] and Bavarian Wikipedia[12]), as well as two online media libraries (ORF TVThek[13] and Das Erste Mediathek[14]). In the remainder of the paper we will refer to the datasets using the abbreviations of their names denoted in Table 3.2. The degree distribution of all datasets are depicted in Figure 3.5.

Concerning the crawling process itself, our web crawler recursively extracted and followed all links, starting from the main page of each website. Note that we did not fully render each page individually, resulting in the omission of links generated via (client-rendered) AJAX queries and Flash content. In a post-processing step we have removed self-loops—links from a webpage to itself. Further, we preprocessed and removed links that coincide with several different redundant actions, such as links containing *?sessid=* or *?oCsid=* for session identifiers, *action=review* for displaying the "write a review" box, as well as "add to shopping cart", or "log-in" personalized user account links and parameters. From the cleaned datasets we constructed the corresponding webgraphs.

[8] http://www.thinkgeek.com
[9] http://www.getdigital.eu
[10] http://www.milan-spiele.de
[11] http://schools-wikipedia.org/
[12] https://bar.wikipedia.org
[13] http://tvthek.orf.at/
[14] http://mediathek.daserste.de/

Table 3.3.: **Website Coverage and Surfer Guidance.** This table depicts the results of our experiments for all biases (columns) and all datasets (rows). The highest values for each dataset in each of the four sections (i.e., Home page, 95%, Stationary Entropy, and Entropy Rate) are highlighted in blue, and the lowest are marked in red. All three *Website Coverage* measurements indicate that a popularity bias (pop.) decreases website coverage, whereas the unpopularity bias (unpop.) is able to increase it. Hence, a bias towards popular pages traps the random surfer within a few pages, while the unpopularity bias allows random surfers to effectively visit more pages. The *Surfer Guidance* is represented by the *Entropy Rate*, which is the uncertainty of the random surfer when selecting a link to traverse. All biases are able to increase the certainty of the random surfer.Furthermore, eigenvector centrality biases (e.c.) in directed networks do not produce the highest entropy rate in our datasets, which is caused by the weak correlation between nodes in- and out-degrees.

	Website Coverage									**Surfer Guidance**			
	Home Page			*95%*			*Stationary Entropy H*			*Entropy Rate* H_{rate}			
	unb.	pop.	unpop.	unb.	pop.	unpop.	unb.	pop.	unpop.	unb.	pop.	unpop.	e.c.
TG	0.01%	0.00%	0.01%	1547 (39.83%)	184 (4.74%)	2431 (62.59%)	8.78	7.16	10.66	7.64	6.86	6.41	6.76
GD	2.88%	5.82%	0.29%	177 (2.09%)	68 (0.80%)	1165 (13.74%)	6.86	5.40	9.93	6.38	5.31	5.01	5.35
MS	1.00%	1.17%	0.13%	599 (2.78%)	65 (0.30%)	4082 (18.93%)	7.94	5.99	11.32	6.11	5.61	4.02	5.67
WFS	3.13%	13.04%	0.09%	3229 (47.51%)	22 (0.32%)	4348 (63.98%)	9.65	4.79	12.01	5.61	4.24	4.68	4.16
BW	5.59%	21.62%	0.05%	4563 (13.94%)	23 (0.07%)	14751 (45.06%)	9.28	4.05	13.54	4.98	3.49	2.61	3.51
ORF	12.06%	36.00%	0.13%	1398 (14.27%)	11 (0.11%)	3321 (33.89%)	7.56	3.25	11.04	4.76	2.83	3.03	3.61
DEM	1.41%	1.94%	0.03%	446 (0.64%)	38 (0.05%)	2812 (4.01%)	7.55	5.11	10.75	5.60	4.63	1.94	5.15

For the actual simulations we extracted the largest strongly connected component of the network (i.e., the largest subset of nodes in which every node can be reached from all other nodes) so that the random surfer does not get stuck on pages without outgoing links.

3.4.8. Results & Discussion

Website Coverage

The left part of the Table 3.3 depicts the results for *Website Coverage*. For almost all datasets the popularity biased random surfer achieves (i) the highest probability of being on the home page, (ii) the lowest number of nodes needed to reach an aggregated energy of 95% and (iii) the lowest stationary entropy. These results indicate a low website coverage, meaning that with a high probability we will find the random surfer on just a few nodes of the network. In other words, the random surfer is trapped on just a few pages of the website. On the other hand, we observe the opposite behavior for the unpopularity biased random surfer (cf. Table 3.3). These results follow our intuition. We expect that in a network with an initially skewed stationary distribution, where just a few top nodes possess an aggregated energy of almost 1, adding an additional bias towards these top nodes increases the skewness. The more skewed the distribution becomes, the likelier the random surfer gets trapped within the most popular nodes. On the other hand, a bias towards less popular nodes reduces the skewness of the stationary distribution.

Findings & Implications. The popularity bias decreases the website coverage, whereas the unpopularity bias is able to increase it. To raise the effective website coverage we should counteract the natural skewness of the stationary distribution of a webgraph. We can achieve this by, for example, inducing an unpopularity bias. Such a bias makes particularly sense in the case of online encyclopedias, where users should be able to easily explore the whole content of the website. However, in cases in which website administrators want to reduce costs (e.g., keep just a few movies on expensive, fast accessible storage devices in media libraries such as

ORF or *DEM*), a bias towards popular webpages represents a suitable method.

Surfer Guidance

The second column of Table 3.3 (Surfer Guidance) depicts the entropy rate of all combinations of datasets and biases. We find that the unbiased random surfer consistently exhibits the highest entropy rate across all datasets. This means that the guidance (i.e., certainty in link-selection decisions) of this surfer is low. Note that this is not the case in undirected, unweighted networks, where the eigenvector centrality bias generates the highest (maximum) entropy rates. Random surfers biased by eigenvector centrality exhibit similar entropy rates to the ones biased by popularity across all datasets except for *ORF* and *DEM*. Across all tested biases we achieve the lowest entropy rate for almost all datasets with the unpopularity bias. As steering the random surfer towards unpopular nodes decreases the average number of possible next hops a lower entropy rate is to be expected. However, for *WFS* the eigenvector centrality bias and for *ORF* the degree bias result in the lowest entropy rates.

Both effects—lowest entropy rate for one of the two biases towards popular nodes in *WFS* and *ORF* and the unobserved maximum entropy rate of the eigenvector centrality bias—are caused by specifics of webgraphs topologies. In particular, in our datasets we do not observe a strong correlation between in-degree and out-degree of a node. A possible explanation for this behavior is based on a specific information architecture on the Web and usability considerations. More specifically, websites tend to have a few pages with many incoming links. For example, on many websites all pages contain the website logo on the top, which is linked to the home page. In all our datasets we confirm this assumption by measuring an unweighted in-degree of $n - 1$ for the home page, where n is the number of pages of a website. On the other hand, the majority of other pages have only a few incoming links. Thus, there is a high variability in the number of incoming links. On contrary, the number of outgoing links is much more stable and in a typical cases limited due to usability reasons.

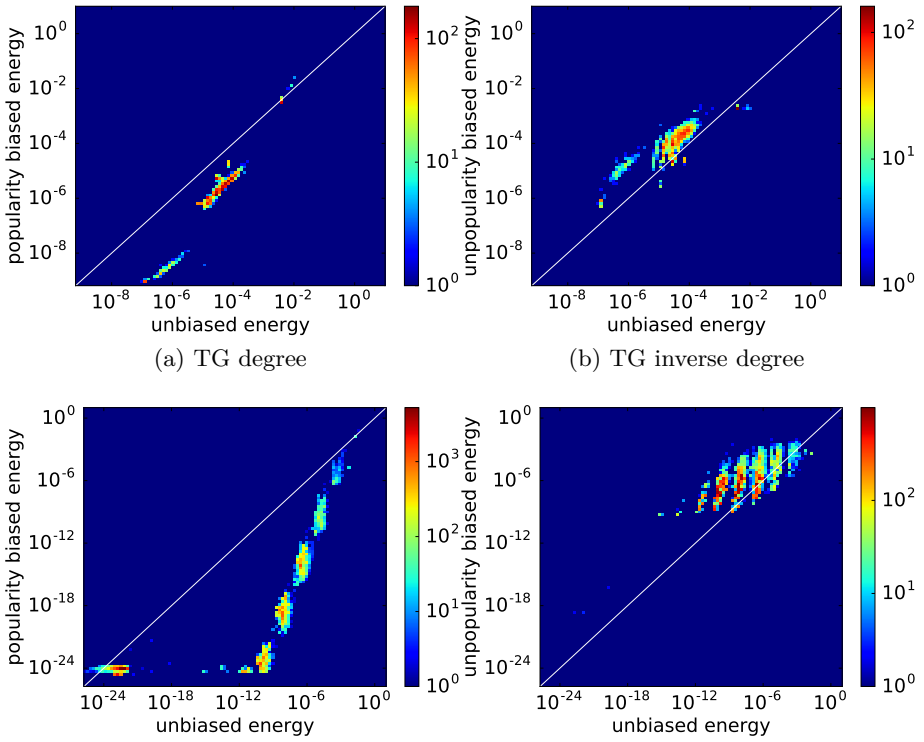
As a consequence of such network topology, the unbiased random surfer often visits nodes with a high in-degree. However, these nodes are often not the ones with the highest out-degree (e.g., the home page of websites often contains only very few outgoing links towards other very popular pages). Consequently, the random surfer has to choose between a few links only. This in turn keeps the uncertainty and the entropy rate low.

Note that in the case of undirected networks the random surfer often visits high-degree nodes, which bear decisions with the highest number of possible outcomes, resulting in highest entropy rates. To find further evidence in favor of our hypothesis we biased the random surfer with node out-degree. This experiment resulted in entropy rates higher than the one of the unbiased random surfer in most datasets.

Findings & Implications. Both popularity and unpopularity bias reduce the entropy rate and at the same time increase certainty for random surfers on directed webgraphs. Consequently, both biases can serve as a way to increase the guidance throughout the website. This finding is in a stark contrast to undirected networks where the popularity biases increase the entropy rate.

Webpage Response

In this experiment we investigate the response of individual pages to a bias. In the case of the popularity bias the majority of pages yields a part of their energy to just a few top pages (see Figure 3.6a and 3.6c). On the other hand, in the case of the unpopularity bias we observe a flow of energy from the top pages towards pages with a low initial energy (see Figure 3.6b and 3.6d). For the popularity bias we observe a slightly left-oriented v-shape in the scatter plots for some datasets (i.e., *MS* and *DEM*). This observation is particularly pronounced for *DEM* (Figure 3.6c). In general, this means that pages with a low initial energy (which are typically more distant to the top pages) are less affected (relatively) by the bias than, for example, pages with an average initial energy (which are closer to the top pages). Note that we can only observe such v-shapes for datasets with a very high (pseudo) diameter (cf. Table 3.2).

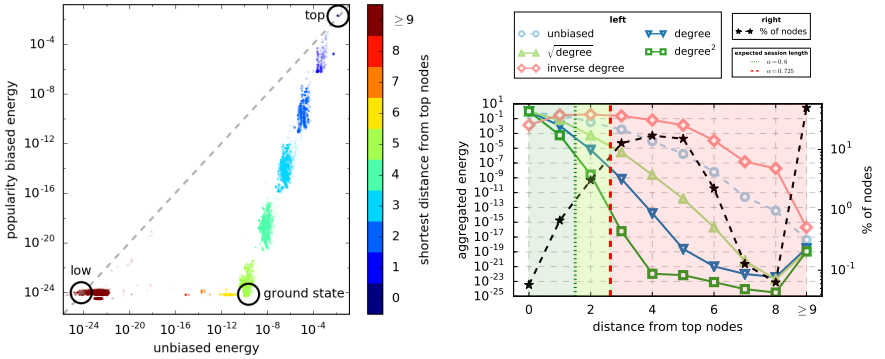


(a) TG degree (b) TG inverse degree

(c) DEM degree (d) DEM inverse degree

Figure 3.6.: **Webpage Response.** The heat maps depict the absolute energy gains and losses of all nodes due to an induced bias. The x-axes correspond to the unbiased energy of a node, whereas the y-axes denote the biased energy. Color refers to the number of nodes observed in that area. The white dashed diagonal marks perfect correlation (i.e., the energy of nodes on this line did not change). For the *TG* dataset (*top*) all biases result in the expected change of the stationary distribution. A popularity bias increases the energy of popular nodes while it decreases the energy of all other nodes. The opposite is true for the unpopularity bias. However, in some datasets, such as *DEM* (*bottom row*), we can see a slightly left-oriented v-shape, where nodes with average initial energies are most affected (relatively) by the bias in the form of decreased energy.

In Figure 3.7a we plot the initial stationary distribution against the one from the popularity biased random surfer. We mark top pages as pages



(a) Degree Bias (b) Distance Energy Level

Figure 3.7.: **DEM Energy Concentration.** These plots describe how energy diffuses when a bias is induced. The *left* plot depicts the stationary distribution of the unbiased random surfer (x-axis) against the one of the popularity biased random surfer (y-axis). We mark pages which increased their energy due to the bias as *top* pages and color each page based on its shortest distance from any of the *top* pages. All pages further away than 6 hops from top pages slide into the ground state of the system—meaning that they will almost never be visited by the random surfer. The minimal increase in energy ($\approx 10^{-20}$) of pages marked as *low* is likely caused by numerical inaccuracies. The *right* plot depicts groups of pages based on their shortest distance from top pages (x-axis) and their aggregated energy (left y-axis). Additionally, we show the fraction of pages (right y-axis; dashed black line with stars) for each distance and the range of expected sessions lengths based on the damping factors α. The colored areas refer to the probability of users reaching nodes of a certain distance, if they start navigating from the home page (i.e., green: very likely, yellow: likely and red: unlikely). We see that, due to the unpopularity bias, a large amount of energy diffuses towards pages being 2 to 3 hops away from top pages. In contrast, all popularity biases (i.e., degree, $\sqrt{\text{degree}}$ and degree^2) concentrate the energy on just a few pages while pushing many other pages into the ground state. This means that small increases in energy of the *top* pages lead to many other pages being pushed into the ground state.

with an increased energy due to the induced popularity bias (see *top* in Figure 3.7a). We then color nodes based on their shortest distance to any of these top pages. The figure shows that these distances have a decisive effect on the biased page energy. A possible explanation for this behavior is that the top pages exhaust energy from all other pages, that is, the energy flows from all other pages towards top pages. The strength of the flow seems to be inversely proportional to the distance from the top nodes—the further away pages are from the top pages, the smaller the flow of energy towards top nodes. After a certain distance (i.e., 6 for *DEM* in Figure 3.7a) some pages reach the lowest possible state of energy and fall into a ground state (see *ground state* in Figure 3.7a). A page in this ground state practically loses all of its energy and thus its visit probability. The pages depicted around the low circle in Figure 3.7b were able to minimally increase their energy ($\approx 10^{-20}$ in *DEM*). We attribute this negligible effect to numerical inaccuracies.

To further analyze the energy flow in a network we group all nodes according to their shortest distance from the top nodes and calculate the energy as a function of this distance. We introduce two new popularity biases. First a bias proportional to square degree and second a bias proportional to square root degree of a node. We assume that the flow of energy towards top nodes must be the fastest in the case of the square degree, followed by the degree bias and then by the square root degree. In Figure 3.7b we see that all popularity biases concentrate the energy on just a few nodes and hence result in most of the other nodes falling into the ground state. We are also able to confirm our energy flow assumption since the flow is strongest for the square degree (the nodes at distance 4 fall into the ground state), followed by the degree bias (the ground state is reached at distance 6). The square root and inverse degree distribute the energy more uniformly over distances (the ground state is reached at distance 8).

To get a better understanding of practical implications of these results we also calculate the expected browsing session length using empirically measured damping factors (i.e., $0.6 \leq \alpha \leq 0.725$ [Gleich et al., 2010]). In particular, we can model session length as a random variable following geometric distribution with the parameter $1 - \alpha$. The expected session

length equals then to $\alpha/(1-\alpha)$. Using empirical damping factors the expected session length lies between 1.5 and 2.64 clicks. Assuming that users start browsing on the home page of a website, the pages that they are expected to visit are within the distance of the expected session length. We mark the range of the expected session length as vertical lines in Figure 3.7b. Only pages that are at distance shorter than the expected session length and have a practically relevant stationary probability can be visited by users. Pages that are further away or are close but are fallen to the ground state will not be visited. Thus, in practice we may be able to increase the visibility of, for example, popular pages but because of a fast energy flow many other pages will practically become invisible to users. Therefore, biasing link selection process includes a trade-off between desirable outcomes and (possibly) unwanted side effects.

Specifically, in Figure 3.7b the upper left area (i.e., aggregated energy higher than 10^{-3} and distance from top nodes smaller than 3) is the most interesting one for a website administrator. Pages in that area have a reasonable probability to be visited while not being too far away from the home page. Website administrators can now utilize our methodology to test different biases and identify the one that best meets their requirements. For example, if the aim is to keep visitors of *DEM* close to the top pages but still enable them to easily explore other pages up to a distance of 2, the squared degree bias would exactly fulfill these requirements (cf. Figure 3.7b).

Please note that in all our calculations the values for the damping factors that we used apply for the general Web and may not hold for a particular website. However, for a given website the operators can determine the damping factors from the actual logfiles.

Findings & Implications. Due to a popularity bias some nodes slide into a ground state in which they are almost never visited by the random surfer. The distance from top nodes determines the final energy of a node. Contrary, the unpopularity bias increases the flow of energy towards nodes with an initially very low energy. This means that an induced popularity bias increases the visibility of already frequently visited nodes and at the same time it shifts many other pages into the ground state. Pages in that

state will hardly be visited. If the aim of a website is to be easily explorable (e.g. Wikipedia datasets *WFS* and *BW*) this should be taken into account. The same applies for webshops, such as *MS*, for which administrators might expect to increase sales of the top products by inducing a popularity bias. However, this will only increase the visits of popular pages—which we find to be mostly overview pages such as *games for* 5 *players* [15]—while putting many actual product pages into the ground state. Nevertheless, taking into account the expected session length of users, it can make sense to concentrate their attention on the top nodes, as they are unlikely to visit nodes further away from the home page.

3.4.9. Conclusions and Future Work

In this paper we presented an approach for measuring the impact of and between biased random surfers and applied it to seven empirical datasets to highlight practical implications for different kinds of networks.

The results gathered from our experiments broaden our understanding of the impact of intrinsic biases for the random surfer on directed webgraphs. Additionally, we found that some combinations of measures and biases (e.g., penalization of popular pages decreases the probability of trapping the random surfer) perform consistently over all datasets. On the other hand, some results highly vary across experiments (e.g., the entropy rate of some biases depend on the structure of the network).

Regarding the *Website Coverage*, we conclude that all used biases highly influence visit probabilities of the random surfer. In particular, we find a consistent pattern based on the type of the bias: Popularity biases tend to trap the random surfer within just a few webpages of the website, whereas biases penalizing popular pages are able to increase website coverage.

The changes in *Surfer Guidance*, due to different biases, are more dependent on the network structure than on the type of the bias itself. However, all biases were able to decrease the entropy rate, which further indicates an increase in guidance.

[15] http://www.milan-spiele.de/nach-anzahl-fuenf-spieler-c-93_98.html

For the *Webpage Response*, in networks with a large diameter we observed a strong side effect. Specifically, the bias puts many pages into a so-called ground state. Pages in that state are barely visited by the random surfer. Thus, website administrators should take these side effects into account.

For future work we plan on analyzing the influence of similarity-based as well as extrinsic biases on random surfers, such as text similarity or categorical mappings between articles (encyclopedias) or products (webshops). Further, we are interested in coloring nodes regarding their type (e.g., product pages, administration pages, types/categories of article pages) and analyzing which type of nodes are favored by different types of biases.

3.4.10. Acknowledgment

This research was in part funded by the FWF Austrian Science Fund research project “Navigability of Decentralized Information Networks” (P 24866-N15).

3.5. Assessing the Navigational Effects of Click Biases and Link Insertion on the Web

With the following article I am answering the third research question. In particular, my co-authors and I compare the differences and commonalities between the emerging effects of click biases and link insertion. To be able to conduct experiments that are capable of answering this question we introduce a method that allows for an equitable comparison between click biases and link insertion. In the experiments conducted for this work we then apply the method to several empirical datasets.

The presented article is based on the assumption that the random surfer is a valid model of human navigation on the Web [Geigl et al., 2015]. In the first step of the experiments we pick uniformly at random a set of target nodes. Subsequently, we try to increase the set's visibility, measured as its sum of visit probabilities over all contained nodes, by either using click biases or link insertion. We incorporate the biases used for the experiments into the model in a similar fashion to the method presented in Geigl et al. [2016b]. With the purpose of modeling link insertion, my co-authors and I introduce a novel approach for this kind of experiments. Specifically, we present a method which inserts new links, with the aim of increasing visits to the set of target nodes, while being still fairly comparable to the method of modeling click biases.

The results gathered in the experiments show, that there seems to be a rule of thumb about when to prefer one manipulation strategy over the other. In particular, if the set of targeted pages has initially had a low cumulative visit probability, link insertion should be preferred over click biases. However, as soon as the initial visit probability of the target set increases, the amplifying effect of click biases starts to work more efficiently. Consequently, in these situations click biases are able to outperform link insertion. Furthermore, we find that for larger initial visit probabilities of the target set click biases lead to more robust effects than those triggered by link insertion. Thus, in terms of reliability, click biases should be preferred over link insertion if, based on the initial situation, both methods are considered to deliver equal results.

3.5.1. Abstract

Websites have an inherent interest in steering user navigation in order to, for example, increase sales of specific products or categories, or to guide users towards specific information. In general, website administrators can use the following two strategies to influence their visitors' navigation behavior. First, they can introduce *click biases* to reinforce specific links on their website by changing their visual appearance, for example, by locating them on the top of the page. Second, they can utilize *link insertion* to generate new paths for users to navigate over. In this paper, we present a novel approach for measuring the potential effects of these two strategies on user navigation. Our results suggest that, depending on the pages for which we want to increase user visits, optimal link modification strategies vary. Moreover, simple topological measures can be used as proxies for assessing the impact of the intended changes on the navigation of users, even before these changes are implemented.

3.5.2. Introduction

Millions of people use the Web on a daily basis to buy products in online shops, perform financial transactions via online banking, or simply browse information systems, media libraries or online encyclopedias, such as IMDb, Netflix or Wikipedia. To find and access relevant information on the Web, people either search, navigate, or combine these two activities. A recent study presented by Gleich et al. [2010] found that 35% of all visits to a website can be attributed to teleports, which are the direct result of clicks on search engine results, navigation through manually typed URLs, or clicks on browser bookmarks. The remaining 65% of the clicks can be attributed to the task of navigating a webpage. In this paper, we direct our attention towards these 65% of actions and tackle the question of what potential effects we can expect if we influence the link selection process of website visitors by simple link modifications. In particular, we are interested in the effects of different link modification strategies on (stochastic) models of web navigation.

Problem. By inserting new links between webpages of a website, we alter the link structure. This has the potential to change user browsing behavior, since new links create new paths for users to explore the website. Alternatively, without changing the link structure of the website, we might be able to influence the link selection process of visitors. Studies have shown that the decisions of users for where to navigate next can be influenced by the layout and the position of the links on a webpage. In particular, due to position bias users are more likely to select links higher up on webpages [Blunch, 1984; Joachims et al., 2005; Murphy et al., 2006; Craswell et al., 2008; Yue et al., 2010; Lerman and Hogg, 2014; Dimitrov et al., 2016]. As a result, inducing click biases, such as repositioning links on a webpage, highlighting the links, or even making them visually more appealing, can affect the users' decision of where to click next on a website, similar to the way that adding new links affects browsing.

In this paper we are particularly interested in investigating and comparing the potential consequences of inserting new links and modifying already existing links on the navigational behavior of users. These newly obtained insights are of a significant practical relevance for website owners, as they can be used, for example, by owners of media libraries to increase visits of specific media files in order to reduce the number of different files that need to be cached on fast storage devices. Another example includes online encyclopedias, where operators may want to guide users towards articles of a specific category over some period of time (e.g., the birthday of an inventor). In some of these cases, link insertion might be more time-consuming than simply changing the layout of the website to increase visibility of specific links and vice versa. Theoretically, we would like to analyze and compare the effects of such link modification endeavors. Practically, new tools are needed to assist website operators in deciding which of the two strategies they should deploy to achieve the desired effects.

Methods. In this paper we study the impact of link modifications on the random surfer, which we apply as a proxy for real user behavior. In the past, a user's decision to click on a link on a webpage was successfully modeled using the random surfer [Brin and Page, 1998; West and Leskovec, 2012a; Helic et al., 2013]. In this model, a user selects one of the links on a

webpage uniformly at random and navigates to the page to which the link points. Apart from the huge success of the Google search engine, whose ranking algorithm is based on the random surfer model, empirical studies have shown that this model provides a very precise approximation of real browsing behavior in many situations and for a variety of applications [Brin and Page, 1998; Geigl et al., 2015]. An important property of a random surfer is its *stationary distribution*, which is the probability distribution of finding a random surfer at a specific webpage in the limit of large number of steps.

In particular, we investigate how the random surfer's stationary distribution of a subset of pages (i.e., *target pages*) of a given website changes as a consequence of (i) modifying already existing links towards them, (ii) introducing new links towards them, or by (iii) combining these two approaches. To that end, we introduce a *click bias*, and a *link insertion* strategy. We model the effects of click biases on the intrinsic attractiveness of a link to the user by increasing the weight of that link. In practice, we may introduce such click biases, for example, by locating the corresponding link on the top of a page. With link insertion, we simply introduce new links between webpages of a website, for example, by linking towards a given target page from the starting page.

We introduce quantitative measures that allow us to address the following research questions:

Navigational Boost. How stable is the stationary distribution with respect to the proposed modification strategies, and what are the limits of stationary distributions that can be achieved for a given set of webpages? Is it (theoretically) possible to achieve a given stationary probability distribution for an arbitrary subset of webpages of a website? What is the connection between simple topological measures of the website network and stationary probability?

Influence Potential. What is the relative gain of the stationary probabilities compared to their unmodified counterparts. This provides us with an answer to the "guidance" potential of a set of webpages, defining to what extent it is possible to increase the relative stationary probabilities as compared to the initial unmodified values.

Combinations. Finally, we are interested how combinations of the two proposed link modification strategies perform in terms of increased stationary probabilities of selected subpages. In particular, we investigate the performance of certain combinations across several different networks and/or selected subpages.

Contributions & Findings. We find that intuitions about how either modification strategy affects navigation are not always correct. Further, our experiments show that the size of a set of targeted subpages is not always a good predictor for the observed effects. Rather, other topological features often better reflect the consequences of a modification. Practically, we provide an open-source framework[16] for website administrators to estimate the effects of link modifications on their website.

3.5.3. Related Work

The random surfer model has received much attention from the research community [Lovász, 1993; Woess, 1994]. While the model is very simple, it became well-established over the last years. It was applied to a variety of problems from graph generators over graph analysis to modeling user navigation. Furthermore, the model has been applied to calculate structural node properties in large networks. HITS [Kleinberg, 1999] and PageRank [Brin and Page, 1998; Page et al., 1999] rank network nodes according to their values in the stationary distribution of the random surfer model. Especially for the later there exists a detailed analysis ranging from the efficiency of its calculation towards its robustness [Langville and Meyer, 2004; Bianchini et al., 2005]. Bianchini et al. [2005] provided an in-depth analysis of how to tweak the cumulative PageRank of a community of websites. They found that splitting up the content of pages onto more highly interlink pages increases the community's cumulative PageRank—since the community is larger it consists of more pages which are able to trap the random surfer for a longer period of time. Moreover, they suggest to avoid dangling webpages (i.e., pages without links to other pages). In this paper we are also interested in the sum of the random surfers visit probabilities in a community, however we do not use (i) teleportation as in the PageRank

[16] https://github.com/floriangeigl/RandomSurfers

model, and (ii) do not modify the network in its size (i.e., number of pages). On the contrary we modify the transition probabilities of certain links and insert new links into the network. Moreover, since all our datasets are strongly connected, we do not face the problem of unwanted high visit probabilities of usually unimportant pages (i.e., dangling nodes) [Bianchini et al., 2005].

A random surfer can be steered towards specific nodes in the network by increasing the probability of traversing links towards those nodes. This can be accomplished by biasing random surfer's link selection strategy so that it is not uniformly random anymore, but biased towards specific nodes. For instance, in the field of information retrieval Richardson and Domingos [2001] successfully applied biased random surfers to increase the quality of search results compared to those achieved using a simple PageRank. At the same time Haveliwala [2003, 2002] biased PageRank towards topics retrieved from a search query to rank the query results. Utilizing this technique the results where more accurate than those produced using a single, generic PageRank. Moreover, Gyöngyi et al. [2004] successfully used trust as bias to detect and filter out spam pages of search results. However, Al-Saffar and Heileman [2007] showed that biased PageRank algorithms generate a considerable overlap in top results with a simple PageRank. Concerning this problem their main suggestion was to use external biases which do not rely onto the underlying link structure of the network. In our paper we randomly decide towards which nodes we bias the random surfer. This allows us to explore the borders of changes in stationary distributions caused by a bias.

In 2013, Helic et al. [2013] compared click trails characteristics of stochastically biased random surfers with those of humans. Their conclusion was, that biased random surfers can serve as valid models of human navigation. Further, Geigl et al. [2015] validated this by showing that the result vector of PageRank and clickdata biased PageRank have a strong correlation in an online encyclopedia. This is especially interesting, since it creates the connection of our simulation to real human navigation on the Web. Additionally, Lerman and Hogg [2014] already showed that it is possible to bias the link selection of users. In particular, they came to the conclusion that users are subject to a *position bias*, making the selection of links

higher up on webpages up to a factor of 3.5 more likely [Blunch, 1984; Joachims et al., 2005; Murphy et al., 2006; Craswell et al., 2008; Yue et al., 2010; Lerman and Hogg, 2014; Dimitrov et al., 2016]. Hence, it is of practical relevance to investigate also the effects of *biases* in the link selection process onto the stationary distribution.

Concerning link insertion there already exists work in literature which makes use of statistical methods to suggest new links in network structures to, for instance, increase the performance of chip architectures [Ogras and Marculescu, 2006]. In particular, the authors use a standard mesh and insert long-range links, converting the network into a small-world network. This reduced packet latency results in a major improvement in throughput. Another field of research where link insertion is of interest are recommender systems for social friendship networks [Li and Chen, 2009; Silva et al., 2010; Moricz et al., 2010; Bian and Holtzman, 2011]. For example, Xie [2010] characterized interests of users in two dimensions (i.e., context and content) and exploited this information to efficiently recommend potential new friends in an online social network. In this paper we focus on the effects of inserted links onto the typical whereabouts of the random surfer.

3.5.4. Methodology

We base our methodology on the calculations of the stationary distribution of a random surfer on the original and manipulated networks. The networks consist of nodes, which represent webpages and directed links between nodes, which represent hyperlinks between webpages. We first calculate the transition matrix and the stationary distribution for the original network as a baseline for comparing the effects of link modifications. Second, we increase the statistical weight of a random surfer visiting a set of predefined nodes (i.e., *target pages* or *target nodes*). We do that either by increasing the link weights towards selected nodes (click bias) or by adding new links pointing towards those nodes (link insertion). Third, we compute the corresponding transition matrix for the modified network. Fourth, we calculate the stationary distribution of the new transition matrices. Finally, we compare the modified stationary distribution with the original

stationary distribution to gain insights into the effects of the different link modifications. Figure 3.8 illustrates these steps on a toy example.

Preliminaries

In what follows we formalize our approach algebraically. We represent a website as a directed network with a weighted adjacency matrix $\boldsymbol{W} \in \mathbb{R}^{n \times n}$, where n is the number of webpages in the website under investigation. We define the element W_{ij} of the weighted adjacency matrix $\boldsymbol{W}$ as the sum of edge weights of all links pointing from node j to node i. For example, $W_{ij} = 1$ if there is a single link from page j to page i with weight 1, and $W_{ij} = 3$ if there are three links pointing from page j to page i each with weight 1.

For our analysis we introduce *target nodes* as the nodes whose stationary probability we want to increase. We use vector $\boldsymbol{t} \in \mathbb{R}^n$ to specify them:

$$t_i = \begin{cases} 1 & \text{if } i \text{ is a } \textit{target node} \\ 0 & \text{otherwise.} \end{cases} \tag{3.11}$$

We further define ϕ as a fraction of target nodes with respect to the total number of nodes n:

$$\phi = \frac{\sum_i t_i}{n} \tag{3.12}$$

Hence, $\phi = 0.1$ means that 10% of nodes from the network are target nodes.

Stationary Distribution

The stationary distribution is a probability distribution over nodes that assigns a probability of finding the random surfer on a given node in the limit of large number of steps. To compute the stationary distribution we first need to construct a diagonal out-degree matrix $\boldsymbol{D}$, with the weighted node out-degrees on its diagonal. Using $\text{diag}(\boldsymbol{v})$ to denote

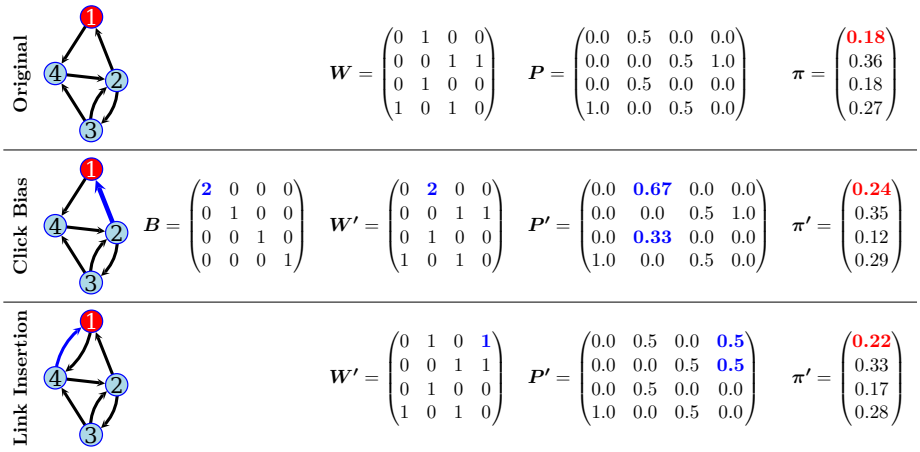


Figure 3.8.: **Modeling Click Bias and Link Insertion - Illustrative Example (Caption Part 1 of 2).** We intend to use different link modification strategies to steer the random surfer towards the red colored node 1 more often. Hence, our target nodes vector consists of only one node: node 1 ($\boldsymbol{t} = \begin{pmatrix} \mathbf{1} & 0 & 0 & 0 \end{pmatrix}^\intercal$). In each row we visualize the corresponding network of the website, where nodes represent webpages and links represent hyperlinks between the webpages. Further, for each of them we show how we calculate its stationary distribution. This involves (from left to right) the weighted adjacency matrix $\boldsymbol{W}$ (unmodified network) or $\boldsymbol{W'}$ (modified networks), the corresponding transition matrix $\boldsymbol{P}$ (unmodified network) or $\boldsymbol{P'}$ (modified networks) and finally the corresponding stationary distribution $\boldsymbol{\pi}$ (unmodified network) or $\boldsymbol{\pi'}$ (modified networks). The blue links in the graphs and the blue matrix elements in bold show the link modifications and their effects on the adjacency and the transition matrix. The red vector elements show the effects of the modifications on the stationary probability (energy) of node 1. **Top row.** Here we depict the original and unmodified network. **Middle row.** We modify the network with a *click bias*. We double the statistical weights of links towards target nodes (bias strength $b = 2$). To calculate the modified adjacency matrix we first construct the diagonal bias matrix $\boldsymbol{B}$ and then compute $\boldsymbol{W'} = \boldsymbol{BW}$. We see an increase in energy of node 1 from 0.18 in the unmodified network to 0.24. (*Caption is continued on the next page*)

Figure 3.8.: **Modeling Click Bias and Link Insertion - Illustrative Example (Caption Part 2 of 2).** *Caption continued from last page.* **Bottom row.** We insert a new link from node 4 to 1 (i.e., blue link in graph and blue element in $\boldsymbol{W'}$) into the original network. Due to the link insertion the energy of node 1 increases from 0.18 in the unmodified network to 0.22 in the modified network. Thus, in this toy example the effects of the click bias are stronger than those of link insertion. Additionally, we see that also elements in the out-component of node 1 (i.e., node 4) profit of an increased energy of node 1 since a significant amount of 1's increased energy flows into node 4.

diagonal matrices with elements of a vector $\boldsymbol{v}$ on their diagonal we define $\boldsymbol{D}$ as:

$$\boldsymbol{D} = \operatorname{diag}\left(\sum_{i=1}^{n} W_{ij}\right). \tag{3.13}$$

Using $\boldsymbol{D}$ matrix we can calculate the transition matrix $\boldsymbol{P}$, which is a left stochastic matrix of $\boldsymbol{W}$ as $\boldsymbol{P} = \boldsymbol{W}\boldsymbol{D}^{-1}$ (in fact this is PageRank matrix without teleportation). The stationary distribution $\boldsymbol{\pi}$ now satisfies the (right) eigenvalue equation for the matrix $\boldsymbol{P}$: $\boldsymbol{\pi} = \boldsymbol{P}\boldsymbol{\pi}$.

Click Bias

To introduce click biases that influence the link selection strategy of the random surfer, we reweigh the links pointing towards target nodes by multiplying their weight by a constant scalar b, which we call bias strength. For example, a bias strength of $b = 2$ doubles the weight of all links towards target nodes. The final probability of the random surfer to traverse a link is then directly proportional to its weight.

Algebraically, we induce biases with a diagonal bias matrix $\boldsymbol{B}$ which we define as $\boldsymbol{B} = \boldsymbol{I} + (b-1) \cdot \operatorname{diag}(\boldsymbol{t})$. The adjacency matrix of a biased network is $\boldsymbol{W'} = \boldsymbol{B}\boldsymbol{W}$. To compute the stationary distribution of the biased network, we first calculate the new transition matrix $\boldsymbol{P'} = \boldsymbol{W'}\boldsymbol{D'}^{-1}$ and then its stationary distribution $\boldsymbol{\pi'}$.

Please note that from the technical perspective, inducing a bias is the same as inserting parallel links towards target nodes—it increases the value of specific elements (i.e., those representing links towards target nodes) in the adjacency matrix. The total weight of newly added parallel links $l(b)$ due to an induced bias b is given by:

$$l(b) = \underbrace{\sum_{ij} W'_{ij}}_{\#\text{ links in } \boldsymbol{W'}} - \underbrace{\sum_{ij} W_{ij}}_{\#\text{ links in } \boldsymbol{W}} \tag{3.14}$$

To allow for a fair comparison between the click bias and the link insertion strategy we insert exactly $l(b)$ new links with weight 1 in the latter case.

Link Insertion

The second link modification strategy consists of inserting new links towards the target nodes from a given set of source nodes. This strategy represents the case where a website administrator inserts links towards target nodes from important subpages of their website. We define the importance of a webpage as its stationary probability in the original network.

To insert a given number $l(b)$ of new links we proceed as follows. We start by sorting nodes by their stationary probability in a descending order. In the next step we insert new links from the top $l(b)/(n \cdot \phi)$ nodes to all target nodes. Here $n \cdot \phi$ is the number of target nodes and we always *ceil* the calculated number of source nodes to ensure that there are enough pairs of nodes. If one of the target nodes is itself designated as a source node we do not insert self-loops—from the practical point of view, it does not make sense to link a webpage to itself. In the rare case where we have connected all possible combinations of source and target nodes but did not reach the required number of links, we simply reiterate the list of the source nodes resulting in parallel links between nodes. Please note that we insert parallel links if a link between a source and a target node has already existed in the original network. However, this happens extremely rarely

because all of our networks are sparse. In fact, in all our experiments the fraction of inserted parallel links was on average less than 1%.

Combinations

Finally, we can combine the two link modification strategies and study the effects of such combinations on the stationary distribution and investigate if an optimal combination of strategies exists, which outperforms the individual approaches. From the practical point of view this means that for optimally steering website users, we combine both, the click bias and link insertion mechanisms.

To create a combined link modification method we first introduce $\alpha \in [0, 1]$, which we call the mixing factor. The mixing factor determines how many of the $l(b)$ links are inserted by the click bias. Then, $1 - \alpha$ defines how many links are inserted by the link insertion strategy:

$$l(b) = \underbrace{\alpha \cdot l(b)}_{\#\text{ biased links}} + \underbrace{(1-\alpha) \cdot l(b)}_{\#\text{ inserted links}} \tag{3.15}$$

With a combined strategy we cannot bias all links towards target nodes—again, we need to select a subset of links towards target nodes. In analogy to the link insertion method we again preferably select links between nodes having higher stationary probability in the unmodified network. Thus, we first compute the probability distribution over the eligible links in the form of matrix $\boldsymbol{L}$, where $\sum_{ij} L_{ij} = 1$. We define matrix $\boldsymbol{L}$ as:

$$\boldsymbol{L} = \text{ diag}(\boldsymbol{\pi}) \cdot \text{ diag}(\boldsymbol{t}) \cdot \boldsymbol{W} \cdot \text{ diag}(\boldsymbol{\pi}). \tag{3.16}$$

The probability of selecting a link is directly proportional to the product of the unmodified stationary probability of its source and target node. Note that due to the multiplicative factor $\text{diag}(\boldsymbol{t}) \cdot \boldsymbol{W}$ only links towards target nodes have a non-zero probability. With $\boldsymbol{L}$ in place we sample $\alpha \cdot l(b)$ links without replacement and multiply their value in $\boldsymbol{W'}$ by b to induce the click bias. To insert the remaining $(1 - \alpha) \cdot l(b)$ links we adopt the link insertion strategy on the matrix $\boldsymbol{W'}$ as described previously.

Measuring the Effects

To measure the effects of link modification strategies we quantify how the stationary probabilities of given target nodes change as a function of the modification. In the remainder of this paper we will refer to a node's stationary probability using the, in the literature established, term ***energy*** [Bianchini et al., 2005]. To that end, we calculate the *energy of target nodes* (π'_t), which is the sum of the modified stationary probabilities of target nodes, as following:

$$\pi'_t = \sum_i \pi'_i \cdot t_i \,, \tag{3.17}$$

where $\boldsymbol{\pi'}$ is the stationary distribution of the modified adjacency matrix.

We further measure the *influence potential*, which is the relative increase in the energy of target nodes due to the modification, as a factor τ:

$$\tau = \frac{\pi'_t}{\pi_t}, \tag{3.18}$$

where π_t is the energy of target nodes of the unmodified network (i.e., $\pi_t = \sum_i \pi_i \cdot t_i$).

3.5.5. Datasets

For our experiments we use three datasets: an online encyclopedia Wikipedia for Schools[17] (*W4S*) and two online media libraries ORF TVthek[18] (*ORF*) and Das Erste Mediathek[19] (*DEM*).

We collected the data by crawling the corresponding websites. Starting from the main page of a website we recursively crawled all subpages by following all outgoing links from a given webpage. Note that we did not follow external links, meaning that we skipped links to pages not belonging

[17] http://http://schools-wikipedia.org/
[18] http://tvthek.orf.at/
[19] http://mediathek.daserste.de/

to a given website. Further, we did not follow links generated via Flash, AJAX or any other client-rendered content.

After collecting the data, we removed self-loops, which are links from a webpage to itself, and special links such as "log-in", "write a review", and all other links that require a session-id. In the next step, we represented each dataset as a directed network—webpages are represented as nodes connected by directed links. For calculating the stationary distribution, we extracted the largest strongly connected component (SCC) of each network, so that in the final network it is possible to navigate from any given node to any other node in the network. These final networks have $4,051$ nodes and $111,795$ links ($W4S$), $9,799$ nodes and $301,844$ links (ORF), and $70,063$ nodes and $3,448,513$ links (DEM).

3.5.6. Experimental Setup

To investigate the effects of manipulating links we first generate sets of target nodes. For this purpose we draw the desired number of nodes uniformly at random from the network without replacement, creating a synthetic set of nodes of a specified size. Note that those sets can consist of unconnected webpages. We conduct all of our experiments with the same initially generated target nodes to reduce the influence of the random node selection process. For making the number of webpages selected as target nodes comparable between datasets we refer to the size of target nodes as ϕ, which is the fraction of target nodes. To generate target nodes we use several values for ϕ which range from 0.01 to 0.2. For each dataset and each ϕ we generate 100 different synthetic sets of nodes (i.e., target nodes).

Limiting (High) Bias Behavior. In our first experiment we are interested in analyzing the impact of an increasing bias strength on the energy of target nodes using either a click bias on already existing links or inserting new links in an informed way. We use bias strengths reaching from $b = 2$ to $b = 200$ to investigate their effects. Note that for the link insertion strategy the number of inserted links is defined by the bias

strength b using Equation 3.14. This ensures a fair comparison between the two methods.

Realistic (Lower) Bias Strengths. In this experiment we investigate practically relevant values for the bias strength b [Lerman and Hogg, 2014; Hogg and Lerman, 2015]. In particular, we iterate over the range 2 to 15 as bias strengths. With this experiment we gain insights into the effects of the proposed modifications, which can be implemented in websites. After the modification of the adjacency matrix we measure the energy of target nodes π'_t. This allows us to investigate the efficiency of both methods for a given bias strength.

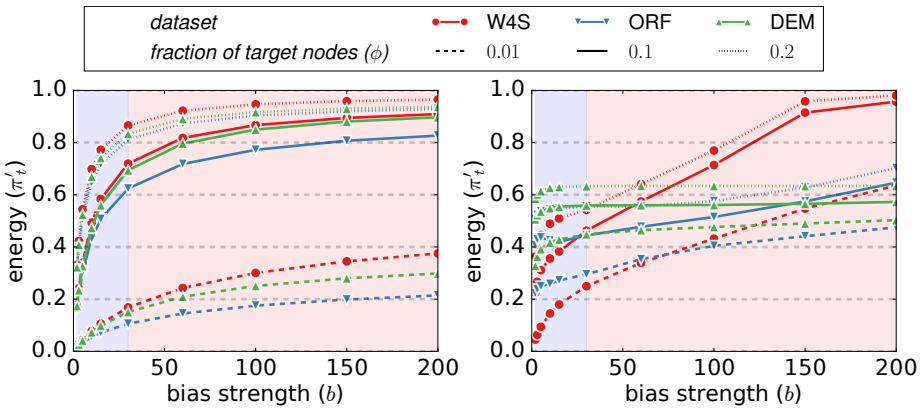
Relative Increase in Stationary Probability. With the previous experiments we analyze changes in the energy of target nodes in absolute terms. For instance, we may learn that for a given set of target nodes we may achieve an energy of $\pi'_t = 0.5$. However, we do not know what the relative increase in their energy is. For example, the set of target nodes may have had $\pi_t = 0.49$ in the unmodified network rendering our efforts futile in *relative* terms. Thus, in this experiment we use τ to measure the influence potential. A higher value for τ means a larger *relative* increase in the energy of target nodes. Again, we compare the results for a given bias strength between our two methods.

Combination of Strategies. Finally, we are interested in investigating if and to what extent the energy of target nodes changes if we combine click biases and link insertion. We vary the mixture factor α from 0 to 1 in steps of 0.1 and measure the energy of target nodes π'_t of the modified networks.

3.5.7. Results & Discussion

Saturation

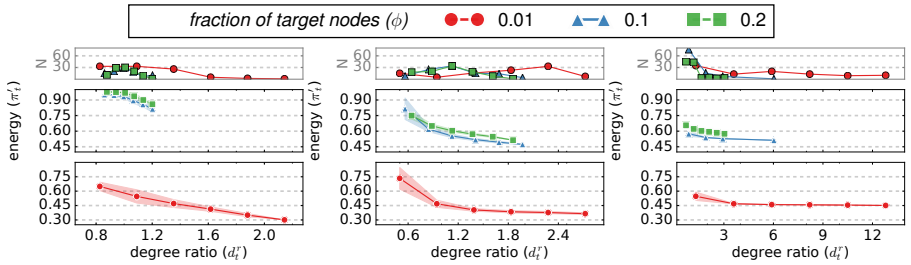
Figure 3.9 depicts the effects of link modifications in our datasets with increasing values of bias strength b and varying fractions of target nodes ϕ (0.01, 0.1 and 0.2).



(a) Click Bias (b) Link Insertion

Figure 3.9.: **Saturation.** The plots depict the connection between bias strength (x-axis) and the increased energy of target nodes due to an induced click bias (left) or link insertion (right). Each marker type and color refers to one dataset. Dashed, solid and dotted line styles refer to fraction of target nodes ϕ 0.01, 0.1 and 0.2 respectively. We can observe that both link modification strategies reach a certain level of saturation—meaning that further increases in bias strength do not result in an increase in energy of target nodes. Therefore, for both strategies we identify two phases: a (i) *navigational boost* phase in which we observe a rapid increase of the stationary probability (blueish region with small values of the bias strength), and a (ii) *saturation* phase (reddish region with larger values of the bias strength).

In the case of click bias we observe the following situation. For small values of b the energy of target nodes π'_t increases very quickly (navigational boost phase, which we analyze in more detail in Section 3.5.7)—this energy saturates for larger values of b (i.e., $b > 35$). This holds for larger ϕ values (0.1 and 0.2), whereas for a smaller ϕ, for example $\phi = 0.01$, the initial growth as well as the saturation are significantly slower and lower respectively. Further, for higher ϕ (0.1 and 0.2) π'_t saturates at an almost identical and very high level (> 0.8)—if the click bias is strong enough we can increase the energy of any fraction of target nodes larger than $\phi > 0.1$.



(a) W4S (b) ORF (c) DEM

Figure 3.10.: **Influence of Target Nodes Degree Ratio onto the Saturation of Their Energy.** The plots depict the due to link insertion achieved energy of target nodes pi'_t as the function of their degree ratio. Each line depicts the results for a given ϕ. For increased readability we group data points into six equally sized bins according to their degree ratio (x-axes). Values on the y-axes represent the averages of the data points falling into the corresponding bin. **Top row.** In the top row we show the distribution of the target nodes degree ratios (in bins; x-axes)—the y-axes denote the number of data points (N) falling into each bin. **Middle row.** Here, we depict the results for medium and large fractions of target nodes. **Bottom row.** For readability we depict small fractions of target nodes separately. Over all datasets and all ϕ we consistently observe a negative correlation between degree ratio and energy π'_t. This means that with an increasing ratio—an increasing out-degree and a decreasing in-degree—the drain of energy increases and this leads to the saturation of the energy of target nodes.

An interesting question in this respect is the height of the energy saturation level. Theoretically, this level is close to 1.0 but as Figure 3.9 shows, in empirical networks this level can not be fully reached. Essentially, due to the directed nature of the network, the target nodes out-component (i.e., the nodes with incoming links from target nodes) will always act as a drain that will take some energy from the target nodes. That amount depends on the size of the out-component as well as its connectivity with other parts of the network—in particular the existence of back-links towards target nodes. This situation is depicted in our toy example Figure 3.8 in the middle row. Node 4, which has an incoming link from node 1, profits

from an induced click bias towards node 1 (cf. original $\pi_4 = 0.27$ and modified $\pi'_4 = 0.29$). Thus, although π'_1 increases with increasing bias strength, node 1 would never reach energy values close to 1.0 because node 4 attracts a certain amount energy to itself.

In the case of the link insertion strategy the results are more diverse (cf. Figure 3.9b). For DEM dataset we observe a quick saturation for all values of ϕ. Differently from the click bias the saturation level is significantly lower for this dataset (i.e., 0.6). For the ORF dataset we do not observe saturation but a monotonous increase in the energy of target nodes for increasing values of ϕ. Finally, for the W4S dataset and larger ϕ (0.1 and 0.2) we can observe saturation at levels higher than 0.9.

As previously, the size of the out-component of the target nodes, combined with the size of their in-component (i.e., the source nodes which point towards target nodes), as well as the ratio of these two quantities provide a possible explanation for this behavior. Basically, we can calculate the average number of newly inserted links as $l(b) = \overline{d} \cdot n \cdot \phi \cdot b$, where $\overline{d}$ is the average degree (i.e., in a directed network average degree $\overline{d}$ corresponds to both, the average in-degree and average out-degree) and n, ϕ, b are as before. Thus, in the networks with a higher average degree we insert more new links. For smaller values of bias strength (blueish region in Figure 3.9b) these new links lead to a navigational boost, resulting in a quick increase in the energy π'_t of target nodes. The navigational boost is higher in networks with a higher average degree—we observe the highest increase in π'_t in DEM with $\overline{d} = 49.22$, the second highest in ORF with $\overline{d} = 30.8$, and the lowest in W4S with $\overline{d} = 27.6$. As mentioned before, in Section 3.5.7 we analyze this navigational boost in more detail. However, for larger values of bias strength (reddish region in Figure 3.9b) the effects of the drain due to the larger size of the out-component become visible—the networks with a higher increase for smaller bias strengths lose their energy now more quickly. Thus, the ordering of the saturation levels for higher bias strengths is reversed to the navigational boost in energy for lower bias strengths, resulting in W4S to now have the highest saturation level, followed by ORF and then by DEM.

To confirm our intuition about the saturation for the link insertion strategy we performed the following analysis. First, we calculated some structural properties for the target nodes. In particular, based on the insights of Ding et al. [2002, 2004], we define the *in-degree* of target nodes as the sum of the weights of links pointing towards target nodes $d_t^- = \sum_{ij} (\text{ diag}(\boldsymbol{t}) \cdot \boldsymbol{W})_{ij}$. The *out-degree* of target nodes is the sum of the weights of outgoing links of target nodes $d_t^+ = \sum_{ij} (\boldsymbol{W} \cdot \text{ diag}(\boldsymbol{t}))_{ij}$. Finally, the *degree ratio* of target nodes is a ratio between the previous two measurements (i.e., $d_t^r = d_t^+ / d_t^-$). Although, it has been shown that properties, such as the simple count of in-links of a node, are bad approximations for PageRank on a large scale [Pandurangan et al., 2002], they proved to be a good indicator for the random surfer behavior on our datasets.

In our experiments, DEM has on average by one order of magnitude higher both target node in-degree and out-degree than the other two datasets. This explains a quick increase of π_t' for smaller bias strengths. However, degree ratio is typically larger in DEM target nodes than in ORF or W4S target nodes and this explains a higher drain of energy and a lower saturation level in the DEM dataset (cf. Figure 3.10).

Finding 1: For larger fractions ϕ of target nodes their energy π_t' achieved through a click bias quickly saturates across all datasets at very high levels (> 0.8). Boost and saturation of the energy is significantly slower for smaller fractions ϕ. The saturation level is determined by the out-degree of the target nodes and reciprocity of outgoing links from the target nodes. For link insertion saturation existence, speed, and levels vary between datasets and ϕ values. The average degree of the original networks as well as the ratio between out-degree and in-degree of target nodes significantly influences those effects.

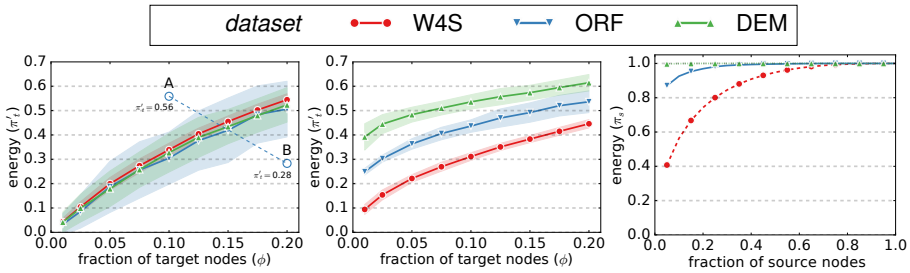
Implications. In case of medium ($\phi = 0.1$) and large ($\phi = 0.2$) fractions of target nodes we reach high saturation levels with both link modification methods even with small bias strengths. For example, if we would like to increase visibility of a large category in, for example Wikipedia, we can

achieve this by either slightly increasing the font size of the links towards the articles of that category or by simple creating some new links towards those articles. Click bias reaches very high visibility levels consistently across several different datasets, whereas link insertion is dependent on the network structure—in datasets with a smaller average number of links we can achieve larger changes. This follows our intuition—in a network with a smaller number of links, each new link affects the network more significantly. However, to match the effects of the click bias we need to insert a very large amount of new links. On the other hand, in case of small ($\phi = 0.01$) fractions of target nodes, we can achieve larger changes by using link insertion—we are able to reach higher saturation levels more consistently and more quickly regardless of the dataset. Again, we can explain this intuitively—small fractions of target nodes have, on average, only few links pointing towards them. Hence, inserting a new link from a top webpage achieves larger changes than highlighting an existing (and probably negligible) link.

Navigational Boost

The blueish region from Figure 3.9b corresponds to smaller and more realistic bias strengths. In practice, increasing the visibility of a link (e.g., by repositioning or highlighting) by more than a factor of 15, meaning that it would receive 15 times more clicks than before, seems quite unrealistic. In particular, users position bias is estimated to be lower than 3.5 [Lerman and Hogg, 2014; Hogg and Lerman, 2015]. Hence, we focus on bias strengths ranging from 2 to 15 (the blueish region in Figure 3.9b) where we can observe a phase of quick increase in the energy of target nodes. We call this phase *navigational boost* phase. The results for all bias strengths from 2 to 15 are quite similar and therefore we report only the results for bias strength $b = 5$.

For click bias we observe a robust performance across datasets, see Figure 3.11. The energy of target nodes π'_t increases almost linearly with the fraction ϕ of target nodes. However, at higher ϕ (i.e., $0.15 \leq \phi \leq 0.2$) the linear trend tends to flatten. This is due to a transition to the stationary phase (cf. Section 3.5.7). Further, we observe a rather high variance of π'_t



(a) Click Bias (b) Link Insertion (c) Lorenz Curve

Figure 3.11.: **Navigational Boost.** Left center figures depict the energy of target nodes after modifying the network through either inducing a click bias or link insertion respectively. The x-axes correspond to the fraction ϕ of target nodes, whereas on the y-axes we denote the energy of target nodes π'_t. Each line represents the average of 100 samples for each ϕ of one dataset. The areas filled with the same color denote the corresponding standard deviations. **Left.** Inducing a click bias is robust across datasets. However, the variability within values of ϕ is rather high. The high variance is caused by the presence or absence of one or more nodes with high original energy in the target nodes. Thus, in cases where such nodes are present in target nodes (depicted as point A in the plot) even smaller fractions of target nodes are able to outperform larger fractions of target nodes without such a top node (depicted as point B in the plot). **Center.** On the contrary, the performance of link insertion varies over datasets but is stable across various ϕ values, which is signified by the low standard deviation suggests over ϕ. **Right.** We plot the Lorenz curve of the datasets' original stationary distributions. We can observe that for different datasets these distributions are differently skewed. In particular, in DEM the energy of just a few nodes is close to 1, whereas in ORF and W4S we need far more nodes to reach the same level (i.e., 0.4 and 0.7 respectively). This explains why we can achieve the highest effect with link insertion in DEM, followed by ORF and W4S. Thus, the performance of link insertion depends on the initial stationary distribution of the network, whereas click biases are robust across datasets. Moreover, for smaller fractions of target nodes link insertion constantly outperforms click biases.

over ϕ and different sets of target nodes. For example, we measure the following average standard deviations over ϕ: W4S = 0.023, ORF = 0.103 and DEM = 0.068. This high variance can be attributed to situations in which smaller fractions of target nodes are often able to outperform larger ones. We depict one such extreme situation of two outlier samples marked as A and B in Figure 3.11. Target nodes depicted with A with $\phi = 0.1$ reach an energy that is almost twice as high as those of the target nodes depicted with B with $\phi = 0.2$.

One potential explanation for these observations is that if the energy of target nodes of the unmodified network is already quite high, that is, the target nodes include one or more nodes with a substantial energy, then the click bias acts as an *amplifier* further magnifying the energy of target nodes. On the other hand, target nodes with a small unmodified energy receive indeed the amplifying effect but are never able to reach the same (high) levels of the modified energy. Therefore, it is possible for smaller fractions of target nodes with one or more nodes with high starting energy to outperform larger fractions of target without such nodes. This can be further attributed to the target nodes structural properties, such as out-degree, in-degree and degree ratio, which we introduced in the previous sections. Basically, starting energy positively correlates with in-degree of target nodes, and therefore we can expect that the click bias is able to amplify target nodes with a higher in-degree more than the target nodes with a lower in-degree. In particular, to confirm this finding we conducted a similar correlation experiment as depicted in Figure 3.10, but used a combination of the target nodes in-degree and energy achieved due to a click bias. However, due to limitations in space, we do not report the experimental details here.

Finding 2: The fraction ϕ of target nodes does not have a decisive effect on navigational boost. Often, smaller ϕ exhibit larger effect sizes. Click bias acts as an *amplifier* that only magnifies what is already present in the target nodes.

In the case of link insertion, navigational boost appears to be highly dataset dependent (see Figure 3.11b). However, the variance of each dataset individually is very low with average standard deviations of 0.017 for W4S, 0.029 for ORF and 0.034 for DEM. Across all datasets we can observe a quick increase in the energy of target nodes with an increasing fraction of target nodes, which then experience a transition towards a stable saturation phase.

To explain the difference in performance between different datasets we have plotted the Lorenz curves of the stationary distributions of our datasets (see Figure 3.11c). We see that for W4S, a very small fraction of top nodes (0.01) only possesses 0.4 of energy. Diversely, for ORF and DEM the same fraction of top nodes already possesses energy higher than 0.85. As the out-component of a specific set of nodes acts as a drain for the energy of source nodes, connecting source nodes with high energy to target nodes leads to a flow of energy from those source nodes towards target nodes. Thus, the initial energy of source nodes plays a crucial role in this process. Through link insertion from top source nodes towards target nodes we attach the target nodes as drains to such top nodes. Consequently, target nodes receive a huge amount of energy and experience a large navigational boost (i.e., ORF and DEM). In other words, we can say that link insertion induces *diffusion* of the energy of top nodes towards target nodes. Given the average degree of the network and the fraction of source nodes (which increases with the fraction ϕ of target nodes), we can use the Lorenz curves to approximately predict the point where the performance across datasets becomes similar. For example, the Lorenz curves of DEM and ORF meet around a fraction of 0.4 of source nodes and we can expect that the performance of those two datasets will become similar for all fractions of source nodes larger than 0.4. In the case of W4S, we need a larger fraction of source nodes (0.7) to reach a similar behavior (cf. Figure 3.11c).

Comparing link insertion with click bias we find that the former outperforms the latter for smaller fractions ϕ of target nodes. For example, in the DEM dataset, link insertion reaches four times higher energy values for the target nodes with $\phi = 0.01$. However, for higher values of ϕ the click bias exhibits a similar performance as link insertion. Further, in the

case of the W4S dataset, the click bias even outperforms link insertion (see W4S in Figure 3.11a and Figure 3.11b) at $\phi = 0.2$).

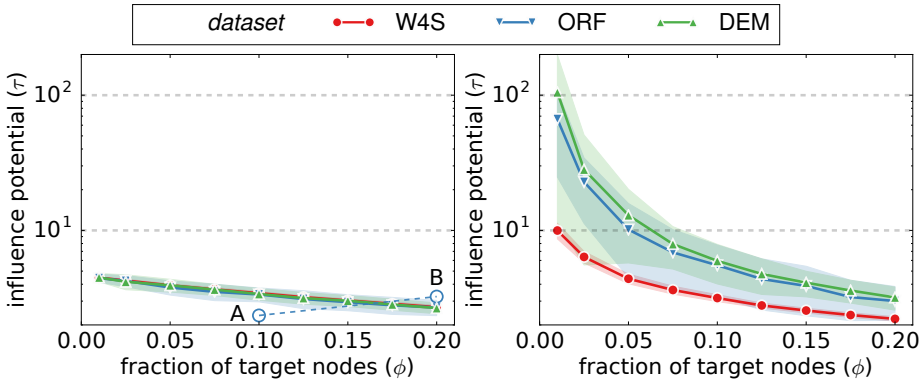
Finding 3: The performance of link insertion varies across the datasets and depends on the skewness of the initial stationary distribution in a dataset. Inserting links from other important webpages towards a given set of webpages results in a higher navigational boost than with the click bias. This is due to the induced *diffusion* of the energy from top nodes towards target nodes.

Implications. If it is possible to insert new links on a website (especially if the fraction of target nodes is small) we should prefer the link insertion over the click bias. However, creation and insertion of such links may be problematic in practice. For example, on Wikipedia it may be difficult and semantically unjustified to insert new links to completely unrelated articles since this may have opposite and contrasting effects on the navigational behavior of users, such as confusion and dissatisfaction. In those cases we may rather choose to increase the transition probability of an already existing link by, for example, highlighting that link (i.e., using CSS[20]) or repositioning it to the webpage's top area. In some other scenarios (i.e., birthdays of famous inventors) implementing a banner which contains links towards a given set of webpages may be an easy way to insert thousands of new links instantly. In those cases, such user interface modifications may prove to have higher lasting effects on the stationary probability than, for example, highlighting links.

Influence Potential

Figure 3.12 depicts the effects of link modifications strategies on the relative increase of the energy of target nodes (i.e., influence potential). Again, we concentrated in this experiment on realistic settings for the bias strength from the interval $[2, 15]$. Since we got comparable results

[20] cascading stylesheets



(a) Click Bias (b) Link Insertion

Figure 3.12.: **Influence Potential.** The figure depicts the relative increase in the energy of target nodes τ with a fixed bias strength of $b = 5$ over different ϕ and datasets. **Left.** Inducing a *click bias* performs robustly and similarly over all datasets (curves for different datasets overlap each other in the plot). Influence potential correlates negatively with the fraction of target nodes, that is, the relative increase in energy is higher for small fractions of target nodes than for large fractions. **Right.** With *link insertion*, we find a significant variance in performance across our datasets. This confirms our findings from the previous section—the skewness of the original stationary distribution determines the effectiveness of the link insertion strategy in a dataset. Similarly to the click bias, the influence potential decays with an increasing fraction of target nodes.

over that complete interval we present only the results for bias strength $b = 5$.

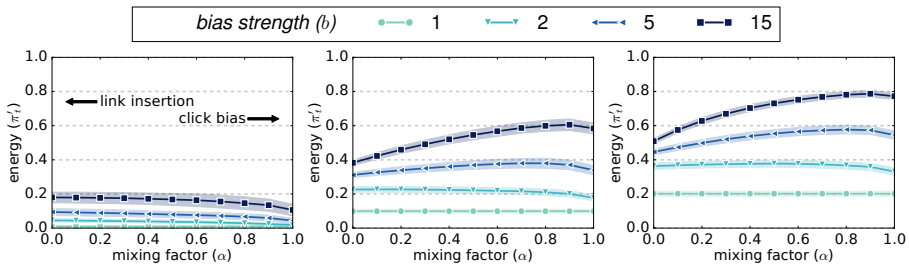
The performance of the click bias is robust across datasets and different ϕ with a low variance in both dimensions (cf. Figure 3.12a). We observe a negative correlation between influence potential and fraction ϕ of target nodes, meaning the smaller fractions of target nodes profit more from an induced click bias than larger fractions. Our calculations of the influence potential confirm once more the results from the previous section, in which smaller fractions with top energy nodes are able to outperform larger fractions of target nodes without top nodes. We once more depict two such

examples from Figure 3.11a. Target nodes depicted by A with $\phi = 0.1$ reach an energy that is almost twice as high as those depicted by B with $\phi = 0.2$. However, nodes A start with a larger initial energy and nodes B with a smaller one. Therefore, in relative terms nodes B have a higher influence potential than nodes A (cf. Figure 3.12a).

Performance of link insertion is again strongly dependent of the dataset. However, similarly to the click bias we observe over all datasets that smaller fractions of target nodes profit significantly more from the link insertion than the larger ones. For example, in DEM dataset for $\phi = 0.01$ we measure an average influence potential of more than 100, whereas for $\phi = 0.2$ influence potential is less than 4 (cf. Figure 3.12b). A similar decay, although not as pronounced as in DEM can be seen in the other two datasets. Similarly to the navigational boost this high influence potential of smaller fractions of target nodes in the case of link insertion can be explained through the skewness of the initial stationary distributions (cf. Figure 3.11c).

As previously, we investigated more closely the relation between influence potential of small fractions of target nodes and their structural properties such as in-degree, out-degree and degree ratio. Target nodes with a high degree ratio (i.e., a small in-degree, a large out-degree or both) have the largest influence potential. Intuitively, such target nodes start with a very small initial energy and therefore can achieve a significant relative increase. On contrary, in absolute terms such target nodes keep a rather small energy even after the modification, whereas target nodes with a large initial energy (a low degree ratio) are experiencing a significant navigational boost in absolute terms but possess relatively low influence potential.

Finding 4: The influence potential of small fractions of target nodes is very high regardless of the link modification strategy. For click bias the influence potential is limited by the bias strength, whereas for link insertion we do not observe such a limit and influence potential can become as high as 100. With increasing fraction of target nodes the influence potential decays drastically.



(a) $\phi = 0.01$ (b) $\phi = 0.1$ (c) $\phi = 0.2$

Figure 3.13.: **Combinations of Link Modification Strategies.** The plots depict average results of 100 sets of target nodes of W4S for three ϕ: 0.01 small on the left, 0.1 medium in the middle, and 0.2 large on the right. On the x-axes we denote α, which defines the combination of the two link modification strategies, whereas on the y-axes we denote the energy of target nodes. We see that for smaller values of ϕ $\alpha = 0$ (100% link insertion) outperforms all the others over all used bias strengths. However, for medium and large ϕ values with higher bias strengths this sweet spot shifts towards higher combinations ($\alpha = 0.7$). In the other two datasets we can observe similar results.

Implications. As previously, if possible we should prefer link insertion over click bias in cases where we are interested in utilizing the influence potential of the target nodes. Our findings suggest that in practice there is a trade-off that we need to make between optimizing for influence potential and for navigational boost. For the former, we need to aim at target nodes with a high degree ratio and for the latter at target nodes with a low degree ratio.

Combinations

In the previous experiments we found that in some situations link insertion should be preferred over click bias (e.g., small fraction ϕ of target nodes), whereas sometimes the opposite represents an optimal approach (e.g., large ϕ). For that reason we want now to shed more light onto combinations of both strategies, that is, we are interested in the navigational effects of simultaneously applying click bias and link insertion to varying extent.

Figure 3.13 depicts the results of this experiment. We find consistent best performing mixtures over all datasets. In particular, we observe that for small fractions ϕ of target nodes, exclusive link insertion outperforms any other combination (see Figure 3.13a). For medium sized target nodes (i.e., $\phi = 0.1$) we observe a shift of best performing combinations towards $\alpha = 0.9$ for higher bias strengths (i.e., $b = 5$ and $b = 15$). This combination consist of 90% click bias and 10% link insertion. For combinations of large fractions of target nodes (i.e., $\phi = 0.2$) and small bias strengths ($b = 2$) the best performing combination is around $\alpha = 0.5$ (50% click bias and 50% link insertion) and further shifts towards $\alpha = 0.9$ (90% click bias and 10% link insertion) with an increased bias strength.

These results confirm our insights from the previous experiments. Thus, click biases act as an amplifier and only work well if target nodes initially possess valuable incoming links. This is highly likely for larger and medium-sized fractions of target nodes, and very unlikely for the case of smaller fractions of target nodes. On the other hand, link insertion diffuses a large portion of the energy of top nodes towards target nodes. Hence, it works especially well for combinations of small fractions of target nodes and datasets with a highly skewed stationary distribution.

Finding 5: For small fractions of target nodes with initially low energy, pure link insertion should be preferred over any other combination. However, with increasing bias strength and larger fraction of target nodes, combinations consisting of 90% click bias and 10% link insertion performs best.

Implications. Smaller sets of webpages (i.e., small ϕ) should focus on introducing new links to achieve the highest browsing guidance. The bigger the set of webpages and the used bias strength becomes, the more this preference shifts towards a combination of 0.9, meaning that 90% of the modifications should be invested in increasing the transition probability of already existing links towards target nodes (e.g., highlighting in the user interface). The remaining 10% should be used to insert new links towards target nodes.

Stationary Versus Transient User Behavior

The random surfer which navigates forever (stationary behavior) may look like a rather unrealistic behavior of users. More realistically, a single user visits a website clicks a couple of times on various links and leaves the website again (transient behavior). However, our calculations of the stationary distribution show that, at least on the networks that we have investigated in this paper these two behaviors are quite similar to each other.

The stationary distribution is calculated with the power-iteration introduced by method Golub and Van Loan [2012]. Thus, we initialize a probability vector representing an initial probability to find a random surfer on each particular node in the network. We initialize this vector (i) with a uniform distribution and (ii) by setting the visit probability of the home page to 1. The former initialization accounts for the assumption that initially each page is equally likely to be visited by users, whereas the latter models users entering the website over the home page. Afterwards, we iterate by recalculating the probabilities for the next click of the random surfer. Thus, one iteration step of the power-iteration method can be interpreted as a step or a click performed by the random surfer moving from the current node to one of its neighbors. Hence, the number of iteration steps that are needed until there are no significant changes in the node probabilities, that is, the convergence rate of the power-iteration method, can be interpreted as the number of clicks needed to model the stationary user behavior. In other words the random surfer does not need to navigate forever—it only needs to navigate through the network until the point where the next click does not change the observed stationary distribution.

In all our datasets, all networks that we generated and modified for these datasets, all combinations of fractions of target nodes ϕ and the bias strength b our calculations converge within 8 iterations regardless of the initialization. Thus, the stationary user behavior is in fact a behavior of users who navigate 8 pages in a website at most. We believe that these 8 clicks are within realistic boundaries for user behavior in the cases in which users decide to explore and browse a website. However,

since many users leave a website immediately upon arrival or within only a single or a small number of clicks this still represents a limitation in our work. This limitation can be easily remedied by introducing a small teleportation probability of jumping to an arbitrary page without following the underlying network structure (i.e., calculating PageRank vector instead of the stationary distribution). We have already experimented with the calculations of PageRank and our first results are quite similar to results that we have presented in this paper. However, we plan to address this question in more details in our future work.

3.5.8. Conclusions

In this paper we have analyzed the effects of two link modification strategies used to influence the typical whereabouts of the random surfer. We investigated how an induced click bias towards a set of webpages changes the stationary distribution (i.e., energy) of those pages. Additionally, we compared those effects with the consequences of altering the network structure by inserting new links. We find that both strategies have a high potential to modify the stationary distribution and that for certain situations there exist constantly high performing link modification strategy. In particular, click biases work well on sets of webpages containing already highly visible webpages, whereas link insertion should be preferred for sets of webpages consisting of pages with low visibility. Further, we showed that a simple structural property of target nodes, namely degree ratio, provides a valuable basis for the estimation of the effects of both link modification strategies.

Assuming that the random surfer is a realistic model of user behavior on the Web—which previous studies seem to confirm [Helic et al., 2013; Geigl et al., 2015]—website operators can use our approach and open-source framework to determine the best strategy for their settings without having to implement and test all the different strategies. Such strategies include but are not limited to altering link positions (bias [Blunch, 1984; Joachims et al., 2005; Murphy et al., 2006; Craswell et al., 2008; Yue et al., 2010; Lerman and Hogg, 2014; Dimitrov et al., 2016]) or creating new links using a recommender systems [Herlocker et al., 2004] (link insertion).

An important practical issue that we have not addressed in this paper is usability. Usability considerations limit the number of new links that we can insert or how we can reposition links. In future, we plan to account for usability by extending our model and investigating limitations induced by various usability restrictions. Also, including the existing user link selection bias derived from user clickstreams into the model would further improve the practical relevance of our method.

3.5.9. Acknowledgments

This research was in part funded by the FWF Austrian Science Fund research project “Navigability of Decentralized Information Networks” (P 24866).

4. Conclusions

The World Wide Web has seen an enormous growth over the last decades. Starting with the first version of this technology created by Tim Berners-Lee in 1989, the number of websites and users has been steadily increasing. Today, due to the sheer number of websites available on the Web, its constantly changing structure, and the lack of a centralized index, the number of websites on the Web can only be estimated. At the time of writing, the website WorldWideWebSize.com[1] estimates the Web to encompass 4.5 billion pages. The popularity of the Web has led it to be an important business factor, and commercial websites constantly compete to attract as many visitors as possible. Moreover, companies strive to steer users towards certain webpages on their own website to, for example, increase sales of a specific product in a webshop, or just to facilitate navigation and assist users in accomplishing their objectives.

Previous studies have shown that humans exhibit certain biases while browsing the Web, such as the well-known position bias [Blunch, 1984; Joachims et al., 2005; Murphy et al., 2006; Craswell et al., 2008; Yue et al., 2010; Lerman and Hogg, 2014; Dimitrov et al., 2016; Lamprecht et al., 2016]. However, the potential effects of such biases had not been investigated until recently. Furthermore, it had been unclear which of the two basic manipulation strategies, namely click biases and link insertion, should be preferred over the other regarding their efficiency to steer users on a website.

In this thesis, I have investigated, if human navigation can be approximated by the random surfer model and how click biases and link insertion affect this model. To that end, I have shown that, from a macroscopic point of view, the random surfer is able to imitate human navigation on the Web

[1] http://www.worldwidewebsize.com/

(i.e., stationary distribution). Extending the model to incorporate click biases has allowed me to investigate their effects on the typical whereabouts of users on several empirical datasets. The extension of the model consists of a theoretical solid way to incorporate click biases and link insertion into the random surfer model. Additionally, to broaden our understanding about how to steer visitors of a website more efficiently, I have presented a method that allows to fairly compare click biases and link insertion.

The remainder of this chapter is structured as follows. In Section 4.1 I give an overview of the results and contributions of this thesis. Subsequently, in Section 4.2 implications of this work are summarized before discussing the limitations in Section 4.3. Finally, in Section 4.4 I provide ideas and potential new avenues for future work.

4.1. Results and Contributions

In the following section I answer the research questions stated in Section 1.5.

4.1.1. Can we model human navigation using random surfers?

The analysis of human navigation on the Web has been the focus of a great deal of research attention, and has been studied in many details. The analyzes range from investigations on a microscopic level (e.g., whether or not a user's link selection process is influenced by previously visited webpages [Singer et al., 2014b]) to experiments conducted on a macroscopic level (e.g., distribution of path lengths [Gleich et al., 2010; Helic et al., 2013; Lamprecht et al., 2015b]). However, website administrators are often solely interested in the number of visits each of their websites' pages receives. Put differently, they want to know the typical whereabouts of their visitors. Yet, it is often the case that no empirical data (e.g., number of page views) has been collected to answer these kinds of questions. Examples of such situations include, but are not limited to, newly published websites for which no data exists.

To tackle this question, I have compared the stationary distribution of the random surfer to empirical data [Geigl et al., 2015]. In particular, I have investigated the webpage visits of humans navigating through an online encyclopedia. By conducting various experiments, I have found that the stationary distribution of random walks strongly correlates with the distribution derived from empirical user data. However, this only holds true when the landing pages included in the empirical dataset are not being considered. This means that the random surfer is not capable of mimicking humans who utilize search engines and frequently reach specific subpages on a website directly. To the advantage of the presented model, previous research has shown that, despite the increased usage of search engines, the larger fraction of clicks on the Web still originates from users following static links [Gleich et al., 2010]. Consequently, the findings of the conducted experiments are of utmost relevance and show that human navigation can be meaningfully modeled using random surfers.

4.1.2. How can we model navigational biases of humans?

Based on the answer to the first research questions we now know that the random surfer produces human-like distributions of page views. However, from a practical perspective, it is not only interesting how we can synthesize human behavior, but rather if and how we can steer this process. To answer this question, I have analyzed random surfers biased towards two properties observed in empirical data, namely homophily and popularity [West and Leskovec, 2012a].

In a first step, I have presented a novel approach that allows modeling arbitrary biases towards specific pages while not altering the underlying link structure. More specifically, the method increases the transition probabilities towards pages based on a predefined property of the target page (e.g., popularity). As a second step, I have applied this approach to several websites obtained by a web crawler to shed light on the efficiency of click biases and their effects.

The results of these experiments indicate that click biases can drastically alter the distribution of visits over pages of a website, and are thus an

effective means to steer users. Furthermore, I have found that click biases potentially give rise to unintended side effects, such as drastically reducing the visit probabilities of the majority of webpages to a state in which users barely view them [Geigl et al., 2016b]. This suggests that we need to carefully evaluate the introduction of these biases with reference to both the target objectives and the accompanying ramifications.

Moreover, I have found that, contrary to undirected graphs [Sinatra et al., 2011], on directed webgraphs all investigated biases increase the certainty in the random surfer's decisions. I believe that these results are of utmost importance for website administrators wishing to actively manipulate the link selection process of users navigating their website. The circumstance that we already know how to implement such biases on websites (e.g., positioning of links [Blunch, 1984; Joachims et al., 2005; Murphy et al., 2006; Craswell et al., 2008; Yue et al., 2010; Lerman and Hogg, 2014; Dimitrov et al., 2016; Lamprecht et al., 2016]) further underlines the practical relevance of these results.

4.1.3. How do navigational biases compare to structural modifications of networks?

In the second research question I have shown that click biases have a high potential to alter the distribution of visits over pages of a website. Apart from click biases, alternative ways to increase the visibility of webpages exist. One such method is the insertion of new links leading to these pages. This action enables users to reach the targeted webpage, that is the page of which one wants to increase the visit probability, from a larger subset of other pages. Based on these considerations, I have investigated which of the two manipulation strategies—click biases or link insertion—bears the higher potential in steering users towards a predefined subset of pages of a website.

As a first step to tackle this question, I have presented an approach that allows for a fair and intuitive comparison of the two manipulation strategies under investigation. To analyze the effects triggered by each method, the

approach measures the relative and absolute increase in visit probabilities of the targeted subset of webpages.

I have executed several experiments on a range of real-world datasets to explore the differences and commonalities between the two manipulation strategies across datasets. The findings show that the visibility of targeted pages with an initially low visit probability can be increased more efficiently when utilizing link insertion. However, as the initial visit probability of targeted pages increases, click bias start to outperform the method of link insertion. Contrary to link insertion, the effects of click biases showed a rather robust performance. Overall this indicates that the decision about which of the two manipulation strategies to apply is strongly dependent on the initial state of the targeted subset of webpages. To allow website administrators and researchers to account for this, I have open-sourced the tool[2] used to produce the presented results. Website administrators can use this tool to encode and set up their initial situation. Subsequently, the tool simulates both manipulation strategies to provide actionable insights about which strategy should be preferred.

4.2. Implications and Potential Applications

To aid website administrators in structuring and creating websites tailored to their needs, a better understanding of how humans navigate the Web is essential. The effects arising from the manipulation of this process are relevant for website owners as well as scientists who study human behavior on the Web. With this thesis, I have provided a first stepping stone towards this larger goal. I believe that the results of this thesis provide important and actionable insights, and that the obtained knowledge provides a valuable basis for further research in this area. The remainder of this section discusses direct implications of the presented results and possible applications of the methods developed in this thesis.

[2] https://github.com/floriangeigl/RandomSurfers

4.2.1. Random Surfers as Model of Human Navigation on the Web

In the literature, a great amount of work has been dedicated to finding a method that models the decisions of humans navigating the Web as accurately as possible. However, in practice it is often sufficient to analyzes the typical whereabouts of users. Showing that the random surfer is a sufficient model for this task equips researchers with a computationally cheap and easily extensible approach for further research in this area. From a practical point of view, the approach allows website administrators to investigate effects emerging due to structural changes of their website without the need to execute tests that potentially disperse valuable visitors. For example, A/B testing a feature could confront some users with a new structure that is not to their liking and potentially drive them away.

4.2.2. A Method for the Simulation and Comparison of Click Biases and Link Insertion

Extending the well-studied random surfer model to simulate click biases [Geigl et al., 2016b] and structural changes [Geigl et al., 2016a] of a website enables researchers to study various effects arising due to this manipulation. The model of human navigation is simple, computationally cheap and yet suffices for many use cases. I firmly believe that the simplicity of the presented model and the theoretical solid way how to encode different scenarios can assist researchers who study human behavior on the Web in their work. Furthermore, to the best of my knowledge, the model is the first of its kind that allows for a direct comparison between click bias and link insertion, all while doing so in a fair manner.

4.2.3. Side Effects of Click Biases

Implementing interface changes on a website to exploit human biases, such as the well-known position bias, has already been shown to be an efficient method to manipulate a user's decision of which link to follow next [Lerman and Hogg, 2014]. In this thesis, I have shown that such

click biases affect not only the targeted pages but also the pages these pages link to. In some cases, this effect can spread throughout the entire website—leading to an unintended and drastic change in the distribution of visits over pages of a website. To increase the awareness of website owners for such pitfalls, this thesis has presented an approach that allows to simulate click biases in an offline setting.

4.2.4. Click Biases Versus Link Insertion

The application of our method to several empirical datasets has shown that there exists a rule of thumb stating when to prefer one method over the other. In particular, I would recommend website administrators to prefer link insertion over click biases if the pages for which they want to increase visits are barely visited by users. However, if the targeted pages are already frequently visited, they should exploit the amplifying effect of click biases to outperform the method of inserting new links while also achieving more robust results.

4.3. Limitations

This chapter discusses the limitations entailed by the analysis conducted in this thesis.

- **Generality of Empirical Findings.** Commercial and privacy concerns did not allow me to get access to logfiles from all of the websites I have investigated in this thesis. For this reason, many of the empirical results are based on datasets that I have crawled independently. Some of these automatically crawled websites included content rendered on the client side (e.g., flash & JavaScript), which I was not able to incorporate into the crawling process. Consequently, the crawled link structures might not cover all links a user is able to actually click on when visiting the website in a browser which supports rendering of such content. However, based on a substantial amount of time spent on manually inspecting the websites used in

the experiments, I am confident that any missing links are few and far between, and therefore very likely negligible.

- **Restrictions Based on the Choice of Datasets.** The first part of this thesis has validated the random surfer as a model of human navigation on the Web. The results were obtained using data from an online encyclopedia. Consequently, it is not clear whether or not this holds true for other websites, or even the entire Web. Yet, the experiments have been conducted on a large set of user click trails which, in my belief, makes them still relevant and sufficient for further research in this area.
- **Methodological Restrictions.** The methods presented by this thesis allow to examine the emerging effects of click biases and link insertion on the typical whereabouts of humans browsing a website. However, in the conducted experiments, I did not consider the inherent biases arising from the layout of a website. Thus, the initial state assumes that each outgoing link of a webpage is selected with equal probability. According to the literature in this field, the positioning of links on a webpage drastically influences a user's decision about which link to follow. Hence, this inaccuracy should be addressed in future experiments. Based on the current state of the presented model, this can be easily incorporated if one has access to empirical click trails of users and/or the exact positions of all links.
- **Availability of Datasets.** As a consequence of crawling many of the datasets, I am not permitted to make them publicly available for research purpose. This problem stems from the fact that Austrian law permits to crawl such datasets to conduct studies, but prohibits sharing them with others. Nevertheless, whenever I have used such datasets I additionally provided a detailed explanation on how I crawled them. Consequently, everyone can implement a similar crawler to gather the same datasets and validate the presented results.

4.4. Future Work

To conclude this thesis, I want to discuss new avenues for potential future work in this section.

4.4.1. Random Surfer as a Model of Human Navigation

To further validate the assumption that the random surfer is an appropriate, or at least sufficient, model of human navigation on the Web, further studies on empirical datasets should be conducted. Specifically, experiments using empirical data from other types of websites (e.g., webshops, movie platforms and blogs) are an interesting new direction for future work. This would also improve the understanding of the model's validity.

4.4.2. Calculation of Biases

At the time of writing, the presented method of inducing various biases into the random surfer model is missing a way to fit it to empirical data. Such a method would permit to measure to what extent humans are biased while navigating the Web. Moreover, the resulting bias of these experiments would directly represent the modeling error of the random surfer model as compared to real humans navigating the Web. A good starting point in this direction would be the article presented by Kumar et al. [2015].

4.4.3. New Types of Biases

The research presented by this thesis has modeled biases using properties derived from webgraphs, such as the degree of a node (i.e., number of in- and outgoing links of a webpage). However, these characteristics are only proxies of information exploited by users during navigation. An interesting path for future research to get a better understanding of human biases would be to model extrinsic biases, such as the semantic similarity between pages. Additionally, this would provide us with insights about whether

or not proxies derived from the network structure are able to reflect the corresponding extrinsic biases correctly.

4.4.4. Microscopic Analysis

The experiments conducted throughout this thesis have mostly covered the macroscopic view of typical whereabouts of users. However, it would be of practical relevance to examine the results on a more fine-grained level. An example of this would be to categorize pages and analyze which categories profit from click biases most. An interesting question in this context would be whether or not a general pattern across various empirical datasets exists.

4.5. Closing Words

I hope that with this thesis I encourage and facilitate future research investigating human navigation on the Web. Furthermore, I am confident that website administrators and researchers can benefit from the open-source tool developed for this thesis. Finally, it is my own long-term vision that this thesis will lay out the basis for a deeper understanding of the effects of navigational human biases on the Web.

List of Figures

List of Tables

Bibliography

Adamic, L. A., Lukose, R. M., Puniyani, A. R., and Huberman, B. A. (2001). Search in power-law networks. *Phys. Rev. E*, 64:046135.

Al-Saffar, S. and Heileman, G. (2007). Experimental bounds on the usefulness of personalized and topic-sensitive pagerank. In *Web Intelligence, IEEE/WIC/ACM International Conference on*, pages 671–675.

Alstott, J., Bullmore, E., and Plenz, D. (2014). powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9(1):e85777.

Ambite, J. L., Lerman, K., Fierro, L., Geigl, F., Gordon, J., and Burns, G. (2017). Bd2k erudite: The educational resource discovery index for data science. *4th WWW Workshop on Big Scholarly Data*.

Bates, M. J. (1989). The design of browsing and berrypicking techniques for the online search interface. *Online review*, 13(5):407–424.

Berners-Lee, T., Cailliau, R., Groff, J.-F., and Pollermann, B. (1992). World-wide web: The information universe. *Internet Research*, 20(4):461–471.

Berners-Lee, T., Fischetti, M., and Foreword By-Dertouzos, M. L. (2000). *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. HarperInformation.

Bian, L. and Holtzman, H. (2011). Online friend recommendation through personality matching and collaborative filtering. *Proc. of UBICOMM*, pages 230–235.

Bianchini, M., Gori, M., and Scarselli, F. (2005). Inside pagerank. *ACM Trans. Internet Technol.*, 5(1):92–128.

Blum, A., Chan, T.-H. H., and Rwebangira, M. R. (2006). A random-surfer web-graph model. *ANALCO*, 6:238–246.

Blunch, N. J. (1984). Position bias in multiple-choice questions. *Journal of Marketing Research*, pages 216–220.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. In *Proc. of the 7th international conference on World Wide Web*, WWW, pages 107–117, Amsterdam, The Netherlands, The Netherlands. Elsevier Science Publishers B. V.

Brockmann, D., Hufnagel, L., and Geisel, T. (2006). The scaling laws of human travel. *Nature*, 439(7075):462–465.

Buscher, G., Cutrell, E., and Morris, M. R. (2009). What do you see when you're surfing?: Using eye tracking to predict salient regions of web pages. In *Proc. of the SIGCHI conference on human factors in computing systems*, pages 21–30. ACM.

Chi, E. H., Pirolli, P., Chen, K., and Pitkow, J. (2001). Using information scent to model user information needs and actions and the web. In *Proc. of the SIGCHI conference on Human factors in computing systems*, CHI '01, pages 490–497, New York, NY, USA. ACM.

Chierichetti, F., Kumar, R., Raghavan, P., and Sarlos, T. (2012). Are web users really markovian? In *Proc. of the 21st international conference on World Wide Web*, pages 609–618. ACM.

Craswell, N., Zoeter, O., Taylor, M., and Ramsey, B. (2008). An experimental comparison of click position-bias models. In *WSDM '08: Proc. of the international conference on Web search and web data mining*, page 87–94, New York, NY, USA. ACM.

Delvenne, J.-C. and Libert, A.-S. (2011). Centrality measures and thermodynamic formalism for complex networks. *Phys. Rev. E*, 83:046117.

Demetrius, L. and Manke, T. (2005). Robustness and network evolution—an entropic principle. *Physica A: Statistical Mechanics and its Applications*, 346(3–4):682 – 696.

Dimitrov, D., Singer, P., Lemmerich, F., and Strohmaier, M. (2016). Visual positions of links and clicks on wikipedia. In *Proc. of the 25th International Conference on World Wide Web*, WWW '16 Companion, New York, NY, USA. ACM.

Ding, C., He, X., Husbands, P., Zha, H., and Simon, H. D. (2002). Pagerank, hits and a unified framework for link analysis. In *Proc. of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 353–354. ACM.

Ding, C. H., Zha, H., He, X., Husbands, P., and Simon, H. D. (2004). Link analysis: Hubs and authorities on the world wide web. *SIAM review*, 46(2):256–268.

Dodds, P. S., Muhamad, R., and Watts, D. J. (2003). An experimental study of search in global social networks. *science*, 301(5634):827–829.

Fronczak, A. and Fronczak, P. (2009). Biased random walks in complex networks: The role of local navigation rules. *Physical Review E*, 80(1):016107.

Fu, W.-T. and Pirolli, P. (2007). Snif-act: A cognitive model of user navigation on the world wide web. *Hum.-Comput. Interact.*, 22(4):355–412.

Gastwirth, J. L. (1971). A general definition of the lorenz curve. *Econometrica*, 39(6):pp. 1037–1039.

Geigl, F. and Helic, D. (2014). The role of homophily and popularity in informed decentralized search. *Dynamic Networks and Knowledge Discovery*, 1229:49.

Geigl, F., Lamprecht, D., Hofmann-Wellenhof, R., Walk, S., Strohmaier, M., and Helic, D. (2015). Random surfers on a web encyclopedia. In *Proc. of the 15th International Conference on Knowledge Technologies and Data-driven Business*, i-KNOW '15, pages 5:1–5:8, New York, NY, USA. ACM.

Geigl, F., Lerman, K., Walk, S., Strohmaier, M., and Helic, D. (2016a). Assessing the navigational effects of click biases and link insertion on

the web. In *Proc. of the 27th ACM Conference on Hypertext and Social Media*, HT '16, pages 37–47, New York, NY, USA. ACM.

Geigl, F., Moik, C., Hinteregger, S., and Goller, M. (2017). Using machine learning and RFID localization for advanced logistic applications. In *2017 IEEE International Conference on RFID (RFID) (IEEE RFID 2017)*, Phoenix, USA.

Geigl, F., Walk, S., Strohmaier, M., and Helic, D. (2016b). Steering the random surfer on directed webgraphs. In *Proc. of the International Conference on Web Intelligence*, pages 280–287. IEEE/WIC/ACM.

Ghosh, R. and Lerman, K. (2012). Rethinking centrality: The role of dynamical processes in social network analysis. *CoRR*, abs/1209.4616.

Gleich, D. F., Constantine, P. G., Flaxman, A. D., and Gunawardana, A. (2010). Tracking the random surfer. In *Proc. of the 19th international conference on World wide web - WWW '10*, page 381.

Goldhirsch, I. and Gefen, Y. (1987). Biased random walk on networks. *Phys. Rev. A*, 35:1317–1327.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU Press.

Granka, L. A., Joachims, T., and Gay, G. (2004). Eye-tracking analysis of user behavior in www search. In *Proc. of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '04, pages 478–479, New York, NY, USA. ACM.

Gulli, A. and Signorini, A. (2005). The indexable web is more than 11.5 billion pages. In *Special interest tracks and posters of the 14th international conference on World Wide Web*, pages 902–903. ACM.

Gyöngyi, Z., Garcia-Molina, H., and Pedersen, J. (2004). Combating web spam with trustrank. In *Proc. of the Thirtieth international conference on Very large data bases-Volume 30*, pages 576–587. VLDB Endowment.

Hasani-Mavriqi, I., Geigl, F., Pujari, S. C., Lex, E., and Helic, D. (2015). The influence of social status on consensus building in collaboration

networks. In *Proc. of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015*, pages 162–169. ACM.

Hasani-Mavriqi, I., Geigl, F., Pujari, S. C., Lex, E., and Helic, D. (2016). The influence of social status and network structure on consensus building in collaboration networks. *Social Network Analysis and Mining*, 6(1):80.

Haveliwala, T. and Kamvar, S. (2003). The second eigenvalue of the google matrix. *Stanford University Technical Report*.

Haveliwala, T. H. (2002). Topic-sensitive pagerank. In *Proc. of the 11th international conference on World Wide Web*, pages 517–526. ACM.

Haveliwala, T. H. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, IEEE Transactions on*, 15(4):784–796.

Helic, D. (2012). Analyzing user click paths in a wikipedia navigation game. In *MIPRO, 2012 Proc. of the 35th International Convention*, pages 374–379. IEEE.

Helic, D. and Geigl, F. (2015). Importance of network nodes for navigation with fractional knowledge. *Proc. of the 38th International Convention on Information and Communication Technology, Electronics and Microelectronics*.

Helic, D., Strohmaier, M., Granitzer, M., and Scherer, R. (2013). Models of human navigation in information networks based on decentralized search. In *Proc. of the 24th ACM Conference on Hypertext and Social Media*, HT '13, pages 89–98, New York, NY, USA. ACM.

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

Hill, N. and Häder, D.-P. (1997). A biased random walk model for the trajectories of swimming micro-organisms. *Journal of Theoretical Biology*, 186(4):503–526.

Hinteregger, S., Kulmer, J., Goller, M., Galler, F., Arthaber, H., and Witrisal, K. (2017). UHF-RFID backscatter channel analysis for accurate wideband ranging. In *2017 IEEE International Conference on RFID (RFID) (IEEE RFID 2017)*, Phoenix, USA.

Hogg, T. and Lerman, K. (2015). Disentagling the effects of social signals. *Human Computation Journal*, 2(2):189–208.

Hwang, S., Lee, D. S., and Kahng, B. (2012). Effective trapping of random walkers in complex networks. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, 85(4).

Joachims, T., Granka, L., Pan, B., Hembrooke, H., and Gay, G. (2005). Accurately interpreting clickthrough data as implicit feedback. In *Proc. of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161. Acm.

Juvina, I., Karbor, P., Pauw, B., et al. (2005). Toward modeling contextual information in web navigation. In *Proc. of the Cognitive Science Society*, volume 27.

Kamvar, S. D., Haveliwala, T. H., Manning, C. D., and Golub, G. H. (2003). Extrapolation methods for accelerating pagerank computations. In *Proc. of the 12th international conference on World Wide Web*, pages 261–270. ACM.

Kan, G. (2001). Harnessing the benefits of a disruptive technology. *O'Reilley & Associates*.

Kitajima, M., Blackmon, M. H., and Polson, P. G. (2000). A comprehension-based model of web navigation and its application to web usability analysis. In *People and computers XIV—Usability or else!*, pages 357–373. Springer.

Kitajima, M., Polson, P. G., and Blackmon, M. H. (2007). Colides and snif-act: Complementary models for searching and sensemaking on the web. In *Human Computer Interaction Consortium (HCIC) 2007 Winter Workshop*.

Kleinberg, J. (2000a). The small-world phenomenon: An algorithm perspective. In *Proc. of the thirty-second annual ACM symposium on Theory of computing*, STOC '00, pages 163–170, New York, NY, USA. ACM.

Kleinberg, J. (2001). Small-world phenomena and the dynamics of information. In *Advances in Neural Information Processing Systems (NIPS) 14*, pages 431–438, Cambridge, MA, USA. MIT Press.

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

Kleinberg, J. M. (2000b). Navigation in a small world. *Nature*, 406(6798):845.

Kumar, R., Tomkins, A., Vassilvitskii, S., and Vee, E. (2015). Inverting a steady-state. In *Proc. of the Eighth ACM International Conference on Web Search and Data Mining*, pages 359–368. ACM.

Lamprecht, D., Geigl, F., Karas, T., Walk, S., Helic, D., and Strohmaier, M. (2015a). Improving recommender system navigability through diversification: A case study of imdb. In *Proc. of the 15th International Conference on Knowledge Technologies and Data-driven Business*, page 21. ACM.

Lamprecht, D., Lerman, K., Helic, D., and Strohmaier, M. (2016). How the structure of wikipedia articles influences user navigation. *New Review of Hypermedia and Multimedia*, pages 1–22.

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

Langville, A. N. and Meyer, C. D. (2004). Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380.

Lerman, K. and Hogg, T. (2014). Leveraging position bias to improve peer recommendation. *PLoS ONE*, 9(6):e98914.

Leskovec, J. and Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network. In *Proc. of the 17th international conference on World Wide Web*, pages 915–924. ACM.

Li, N. and Chen, G. (2009). Multi-layered friendship modeling for location-based mobile social networks. In *MobiQuitous, 2009. 6th Annual International*, pages 1–10.

Lovász, L. (1993). Random walks on graphs: A survey. *Combinatorics, Paul erdos is eighty*, 2(1):1–46.

Meiss, M., Duncan, J., Gonçalves, B., Ramasco, J. J., and Menczer, F. (2009). What's in a session: Tracking individual behavior on the web. In *Proc. of the 20th ACM conference on Hypertext and hypermedia*, pages 173–182. ACM.

Moricz, M., Dosbayev, Y., and Berlyant, M. (2010). Pymk: Friend recommendation at myspace. In *Proc. of the 2010 ACM SIGMOD International Conference on Management of Data*, SIGMOD '10, pages 999–1002, New York, NY, USA. ACM.

Muchnik, L., Itzhack, R., Solomon, S., and Louzoun, Y. (2007). Self-emergence of knowledge trees: Extraction of the wikipedia hierarchies. *Phys. Rev. E*, 76:016106.

Murphy, J., Hofacker, C., and Mizerski, R. (2006). Primacy and recency effects on clicking behavior. *Journal of Computer-Mediated Communication*, 11(2):522–535.

Newman, M. (2010). *Networks: An Introduction*. Oxford University Press, Inc., New York, NY, USA.

O'Day, V. L. and Jeffries, R. (1993). Orienteering in an information landscape: How information seekers get from here to there. In *Proc. of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 438–445. ACM.

Ogras, U. Y. and Marculescu, R. (2006). " it's a small world after all": Noc performance optimization via long-range link insertion. *Very Large Scale Integration (VLSI) Systems, IEEE Transactions on*, 14(7):693–706.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Pandurangan, G., Raghavan, P., and Upfal, E. (2002). Using pagerank to characterize web structure. In *Computing and Combinatorics*, pages 330–339. Springer.

Parry, W. (1964). Intrinsic markov chains. *Transactions of the American Mathematical Society*, 112(1):pp. 55–66.

Pirolli, P. (1997). Computational models of information scent-following in a very large browsable text collection. In *Proc. of the SIGCHI conference on Human factors in computing systems*, CHI '97, pages 3–10, New York, NY, USA. ACM.

Pirolli, P. and Card, S. (1999). Information foraging. *Psychological Review*, 106(4):643–675.

Pons, P. and Latapy, M. (2005). Computing communities in large networks using random walks. In *International Symposium on Computer and Information Sciences*, pages 284–293. Springer.

Qiu, F. and Cho, J. (2006). Automatic identification of user interest for personalized search. In *Proc. of the 15th international conference on World Wide Web*, pages 727–736. ACM.

Richardson, M. and Domingos, P. (2001). The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *NIPS*, pages 1441–1448.

Rosvall, M. and Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *Proc. of the National Academy of Sciences*, 105(4):1118–1123.

Silva, N., Tsang, I.-R., Cavalcanti, G., and Tsang, I.-J. (2010). A graph-based friend recommendation system using genetic algorithm. In *Evolutionary Computation (CEC), 2010 IEEE Congress on*, pages 1–7.

Simsek, O. z. and Jensen, D. (2005). Decentralized search in networks using homophily and degree disparity. In *Proc. of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, pages 304–310, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Sinatra, R., Gómez-Gardeñes, J., Lambiotte, R., Nicosia, V., and Latora, V. (2011). Maximal-entropy random walks in complex networks with limited information. *Physical Review E*, 83(3):030103.

Singer, P., Helic, D., Hotho, A., and Strohmaier, M. (2014a). Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web. In *Proc. of the 24th International Conference on World Wide Web*.

Singer, P., Helic, D., Taraghi, B., and Strohmaier, M. (2014b). Detecting memory and structure in human navigation patterns using markov chain models of varying order. *PloS one*, 9(7):e102070.

Singer, P., Niebler, T., Strohmaier, M., and Hotho, A. (2013). 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.

Sudarshan Iyengar, S., Veni Madhavan, C., Zweig, K. A., and Natarajan, A. (2012). Understanding human navigation using network analysis. *Topics in cognitive science*, 4(1):121–134.

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

Walk, S., Helic, D., Geigl, F., and Strohmaier, M. (2016). Activity dynamics in collaboration networks. *ACM Transactions on the Web (TWEB)*, 10(2):11.

Walk, S., Singer, P., Noboa, L. E., Tudorache, T., Musen, M. A., and Strohmaier, M. (2015). Understanding how users edit ontologies: Comparing hypotheses about four real-world projects. In *International Semantic Web Conference*, pages 551–568. Springer.

Walk, S., Singer, P., Strohmaier, M., Tudorache, T., Musen, M. A., and Noy, N. F. (2014). Discovering beaten paths in collaborative ontology-

engineering projects using markov chains. *Journal of biomedical informatics*, 51:254–271.

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

West, R. and Leskovec, J. (2012a). Automatic versus human navigation in information networks. In *ICWSM*.

West, R. and Leskovec, J. (2012b). Human wayfinding in information networks. In *Proc. of the 21st International Conference on World Wide Web*, WWW '12, pages 619–628, New York, NY, USA. ACM.

White, R. W. and Huang, J. (2010). Assessing the scenic route: Measuring the value of search trails in web logs. In *Proc. of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 587–594, New York, NY, USA. ACM.

Woess, W. (1994). Random walks on infinite graphs and groups-a survey on selected topics. *Bulletin of the London Mathematical Society*, 26(1):1–60.

Xie, X. (2010). Potential friend recommendation in online social network. In *Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int'l Conference on Int'l Conference on Cyber, Physical and Social Computing (CPSCom)*, pages 831–835.

Yue, Y., Patel, R., and Roehrig, H. (2010). Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proc. of the 19th international conference on World wide web*, pages 1011–1018. ACM.

Zlatić, V., Gabrielli, A., and Caldarelli, G. (2010). Topologically biased random walk and community finding in networks. *Physical Review E*, 82(6):066109.

Appendices

A. Complete List of Own Publications

A.1. Journal Articles

- **Journal 1:** [Walk et al., 2016] Walk, S., Helic, D., Geigl, F. and Strohmaier M. (2016). Activity Dynamics in Collaboration Networks. *ACM Transactions on the Web*

- **Journal 2:** [Hasani-Mavriqi et al., 2016] Hasani-Mavriqi, I., Geigl, F., Pujari, S., Lex, E., and Helic, D. (2016). The Influence of Social Status and Network Structure on Consensus Building in Collaboration Networks. *Social Network Analysis and Mining*

A.2. Conference Proceedings

- **Article 1:** [Geigl et al., 2016a] Geigl, F., Lerman, K., Walk, S., Strohmaier, M. and Helic, D. (2016). Assessing the Navigational Effects of Click Biases and Link Insertion on the Web. *27th Conference on Hypertext and Social Media*

- **Article 2:** [Geigl et al., 2016b] Geigl, F., Walk, S., Strohmaier, M. and Helic, D. (2016). Steering the Random Surfer on Directed Webgraphs *International Conference on Web Intelligence*

- **Article 3:** [Geigl et al., 2015] Geigl, F., Lamprecht, D., Hofmann-Wellenhof, R., Walk, S., Strohmaier, M. and Helic, D. (2015). Random Surfers on a Web Encyclopedia. *15th International Conference on Knowledge Technologies and Data-driven Business*

- **Article 4:** [Hasani-Mavriqi et al., 2015] Hasani-Mavriqi, I., Geigl, F., Pujari, S., Lex, E., and Helic, D. (2015). The Influence of Social Status on Consensus Building in Collaboration Networks. *International Conference on Advances in Social Networks Analysis and Mining*

- **Article 5:** [Lamprecht et al., 2015a] Lamprecht, D., Geigl, F., Karas, T., Walk, S., Helic, D., and Strohmaier M. (2015). Improving Recommender System Navigability Through Diversification: A Case Study of IMDb. *15th International Conference on Knowledge Technologies and Data-driven Business*

- **Article 6:** [Helic and Geigl, 2015] Helic, D. and Geigl, F. (2015). Importance of Network Nodes for Navigation with Fractional Knowledge. *38th International Convention on Information and Communication Technology, Electronics and Microelectronics*

A.3. Workshop Articles

- **Workshop 1:** [Ambite et al., 2017] Ambite, J., Lerman, K., Fierro, L., Geigl, F., Gordon, J. and Burns, G. (2017). BD2K ERuDIte: The Educational Resource Discovery Index for Data Science *4th WWW Workshop on Big Scholarly Data*

- **Workshop 2:** [Geigl and Helic, 2014] Geigl, F. and Helic, D. (2014). The Role of Homophily. *2nd International Workshop on Dynamic Networks and Knowledge Discovery*

A.4. Poster

- **Poster 1:** [Geigl et al., 2017] Geigl, F., Moik, C., Hinteregger, S. and Goller, M. (2017). Using Machine Learning and RFID Localization for Advanced Logistic Applications. *11th Annual IEEE International Conference on RFID*