Aspects of Information Access: Modeling Navigation on Wikipedia

Dimitar Dimitrov

Institute for Web Science and Technologies
University of Koblenz–Landau
&
GESIS – Leibniz Institute for the Social Sciences
dimitar.dimitrov@gesis.org

December 2018

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

Doktor der Naturwissenschaften (Dr. rer. nat.)

genehmigte Dissertation.

PhD thesis at the University of Koblenz-Landau

Datum der wissenschaftlichen Aussprache:	17.12.2018
Vorsitz des Promotionsausschusses	Prof. Dr. Thomas Burkhardt
Berichterstatter:	JProf. Dr. Claudia Wagner
Berichterstatter:	Prof. Dr. Markus Strohmaier
Berichterstatter:	Prof. Dr. Denis Helic

Abstract

Navigation is a natural way to explore and discover content in a digital environment. Hence, providers of online information systems such as Wikipedia—a free online encyclopedia—are interested in providing navigational support to their users. To this end, an essential task approached in this thesis is the analysis and modeling of navigational user behavior in information networks with the goal of paving the way for the improvement and maintenance of web-based systems. Using large-scale log data from Wikipedia, this thesis first studies information access by contrasting search and navigation as the two main information access paradigms on the Web. Second, this thesis validates and builds upon existing navigational hypotheses to introduce an adaptation of the well-known PageRank algorithm. This adaptation is an improvement of the standard PageRank random surfer navigation model that results in a more "reasonable surfer" by accounting for the visual position of links, the information network regions they lead to, and the textual similarity between the link source and target articles. Finally, using agent-based simulations, this thesis compares user models that have a different knowledge of the network topology in order to investigate the amount and type of network topological information needed for efficient navigation. An evaluation of agents' success on four different networks reveals that in order to navigate efficiently, users require only a small amount of high-quality knowledge of the network topology. Aside from the direct benefits to content ranking provided by the "reasonable surfer" version of PageRank, the empirical insights presented in this thesis may also have an impact on system design decisions and Wikipedia editor guidelines, *i.e.*, for link placement and webpage layout.

Zusammenfassung

Navigation ist ein natürlicher Weg zur Erforschung und Erkundung von Inhalten in einer digitalen Umgebung. Aus diesem Grund sind Betreiber von Online-Informationssystemen, wie digitale Enzyklopädien (bspw. Wikipedia), daran interessiert, ihren Nutzern eine Navigationsunterstützung zu bieten. Eine wesentliche Aufgabe, die in dieser Arbeit behandelt wird, ist die Analyse und Modellierung des navigatorischen Nutzerverhaltens in Informationsnetzwerken mit dem Ziel, den Weg für die Verbesserung und Wartung von webbasierten Systemen zu ebnen. Unter Verwendung von umfangreichen Log-Daten aus Wikipedia wird in dieser Arbeit als erstes der Informationszugriff untersucht, indem Suche und Navigation als die zwei Hauptinformationszugriffsparadigmen auf der Web gegenübergestellt werden. Als nächstes werden bestehende Navigationshypothesen validiert und für eine Anpassung des bekannten PageRank-Algorithmus erweitert. Die vorgestellte Anpassung ist eine Verbesserung des Standard-PageRank-Random-Surfer Navigationsmodells hin zu einem "vernünftigeren Surfer", der die visuelle Position der Links, die Informationsnetzwerkregionen, zu denen sie führen, und die textuelle Ähnlichkeit zwischen den Linkquellen und den Zielartikeln berücksichtigt. Diese Arbeit beschäftigt sich schließlich mit agentenbasierten Simulationen und vergleicht Benutzermodelle mit unterschiedlichen Kenntnissen der Netzwerktopologie, um die Menge und Art der topologischen Netzwerkinformationen zu ermitteln, die für eine effiziente Navigation benötigt werden. Die Bewertung des Erfolgs der Agenten in vier verschiedenen Netzwerken zeigt, dass Benutzer nur eine geringe Menge an qualitativ hochwertigen Kenntnissen über die Netzwerktopologie benötigen, um effizient navigieren zu können. Neben den direkten Vorteilen für das Ranking von Inhalten, die durch die eingeführte "vernünftige" Surfer-Version von PageRank bereitgestellt werden, können auch die in dieser Arbeit vorgestellten empirischen Erkenntnisse Auswirkungen auf Systemdesignentscheidungen oder bestehende Richtlinien für Wikipediaeditoren, bspw. für die Linkplatzierung und das Webseiten-Layout haben.

I dedicate this work to my beloved grandparents
Dimitar and Maria Dimitrovi, and Iliya and Jordanka Ilievi.

Acknowledgments

I can barely express my gratitude to Markus Strohmaier, who has guided me throughout this dissertation. In all of our meetings, he posed the tough and challenging questions that helped me to develop and advance the research ideas for this dissertation, while providing enough room for me as a PhD student to decide in which direction to proceed. I would like to thank him for his initial curiosity about my very raw and preliminary thoughts that established the foundation of the follow-up collaboration.

I am also particularly grateful to Philipp Singer and Florian Lemmerich, who collaborated with me on almost all papers building the backbone of this dissertation. Working with Phillip and Florian on a daily basis has always been a pleasure. Although handling these two enthusiastic PostDocs as a PhD student can be extremely challenging, I consider myself truly lucky and privileged to have access to their expertise. Further, I would like to thank Denis Helic for giving me the opportunity to work with him as a visiting researcher at the Technical University of Graz. This visit resulted in my very first paper and allowed me to meet Daniel Lamprecht. Together with Robert West from EPFL, Daniel helped me to further strengthen the topic of this dissertation, for which I am very grateful. I would also like to say thank you to Fabian Flöck for always putting on the reviewer's hat and for his (unfortunately doomed) attempts at improving my writing style. I am especially obliged to Steffen Staab and the team of the "massage the diss" session at the off-campus meeting, after which I focused my work around one word, namely "navigation". I also owe a special thank you to Claudia Wagner for the numerous scientific discussions, administrative support associated with this work, and entertaining pre- and post-conference trips together with Philipp. I would also like to show my gratitude to several colleagues and friends at GESIS I have met and collaborated with. I am especially grateful to my office buddy Erdal "Komschi" Baran, who has assumed responsibility for the work on the da|ra project. Without his support I would not have had the time to work freely on my dissertation. I would like to thank my office neighbors and friends Johann "Wanja" Schaible, Zeljko Carevic, and Matthäus Zloch, who made me laugh at my own research and

implementation mistakes and also endured my frustration during the dark publishing moments of this dissertation. I would also like to thank all of my colleagues in the WTS and CSS departments at GESIS. A very special thanks to Daniel Hienert, Katarina Boland, and Lisette Espin Noboa.

I also wish to thank my clique of buddies in Bulgaria who recharged my batteries every time I returned home for holidays. Special thanks to Martin Rachev, Dima Bogdanova, Eleonora Kabachieva, Petar Taskov, Nikolay Stoyanov, and Maria Petkova, with whom I have been in constant contact. It's been a blast.

Finally, I would like to thank my parents Georgi and Elena, and my sister Maria for their constant and unconditional support. Obtaining a PhD degree is extremely hard work that occupies so much time that could have certainly been better spent with them.

Contents

Chapter 1

Introduction

1.1 Motivation

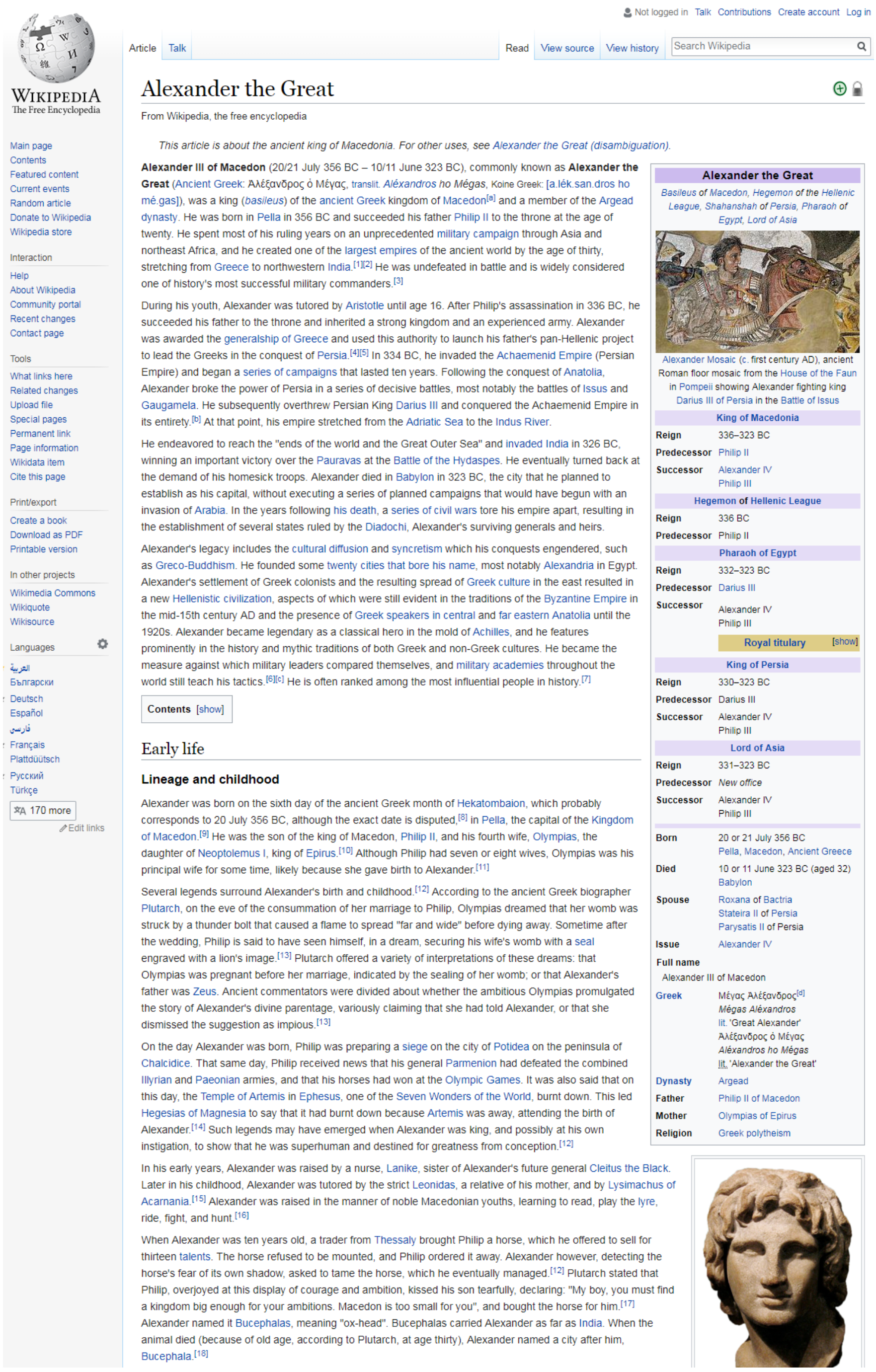
With the invention of the printing press around the year 1440, Johannes Gensfleisch zur Laden zum Gutenberg laid down the foundation of the printing revolution in Europe. This new technology made it easier to spread knowledge and ideas, and thus paved the way for significant historical developments such as the Renaissance and Reformation. Gutenberg's invention is a shining example of how a new technology can change society by enhancing the availability and accessibility of information. Now more than ever, society relies on the efficient creation, distribution of, and access to information. Significant advances in information technology over the past decades have led to the development of Hypertext [Nel65], allowing us to not only easily produce information, but also connect individual information pieces over the Internet [Mar97] to the World Wide Web [BLFFBD00]. The Web represents an information network in which documents are nodes and the hyperlinks connecting them are edges. In the early days of the Web, information was mainly accessed by following links between documents, or, in other words, by *navigating* through the information network. Since its inception, however, the Web has been rapidly growing, which has led to the development of modern search engines [BP12, PBMW99]. Using Web crawlers (programs that automatically follow links), search engines have created an index of the Web that can be queried, or *searched*, by users in order to access information. Both information access paradigms, *i.e.*, navigating and searching, offer a drastically different way to consume information compared to the strictly linear reading, which dominates offline information consumption. For example, search is preferred by those users who are looking for the location of very specific information. However, the returned results come at the high cognitive cost: users must find the best keyword to describe the information needed [FLGD87]. Moreover, search results might not be fully satisfying, and the user context is lost [Law00]. This is precisely where navigation has its biggest advantages: Apart from context preservation, users are not required to explicitly formulate their information needs, but only to select the appropriate path through the information space. This is especially useful when the information needed cannot be captured by keywords, the required keywords are not part of the user's vocabulary, or when users change their information needs during the information seeking process [Hea09, TAAK04]. Navigation is of significant importance in cases for which a single page does not sufficiently satisfy an information need, which can then only be fulfilled by a specific path through the network. If users do not have a specific target or goal in mind but are rather interested in exploring the information space, navigation is also referred to as *browsing*.

Different characteristics of web-based systems, *i.e.*, diversity and the amount of content and users, and often the collaborative nature of content production make it difficult to ensure efficient information access and optimal user support. With more than 5 million articles connected by roughly 400 million links, Wikipedia—a free online encyclopedia—is a comprehensive source of information on the Web. The social significance of Wikipedia has been recognized by users from all over the world, some of which donate on a regular basis[1]. Moreover, Wikipedia receives a significant amount of Web traffic and is one of the most visited websites[2]. Combined, these characteristics make Wikipedia a central object of study: This thesis analyses and models information access in information networks with a focus on navigation. Navigational models have numerous applications and can help to improve and maintain web-based systems; for example, they can be used to rank content, predict the next click, or provide recommendations. Another important application is identifying shortcomings in webpage design and layout. Aside from the direct benefit of better supporting Web users, modeling navigation also has advantages with respect to the organization and structuring of large document collections into an information network that can be efficiently explored.

The rest of this chapter is structured as follows: I first provide an overview of the information access specifics in the Wikipedia context (*cf.* Section 1.2). Subsequently, in Section 1.3, I present the problem statement of this thesis, its objectives, and the followed approach. In Section 1.4, the problem statement is decomposed into three research questions. Section 1.5 gives a overview of the main publications used in this thesis, while contributions and findings are summarized in Section 1.6. Finally, the thesis structure is outlined in Section 1.7.

[1] `https://wikimediafoundation.org/wiki/Fundraising_reports` Accessed: 2018-04-17

[2] `https://www.alexa.com/siteinfo/wikipedia.org` Accessed: 2018-01-12



Alexander the Great

From Wikipedia, the free encyclopedia

This article is about the ancient king of Macedonia. For other uses, see Alexander the Great (disambiguation).

Alexander III of Macedon (20/21 July 356 BC – 10/11 June 323 BC), commonly known as **Alexander the Great** (Ancient Greek: Ἀλέξανδρος ὁ Μέγας, translit. *Aléxandros ho Mégas*, Koine Greek: [a.lék.san.dros ho mé.gas]), was a king (*basileus*) of the ancient Greek kingdom of Macedon[a] and a member of the Argead dynasty. He was born in Pella in 356 BC and succeeded his father Philip II to the throne at the age of twenty. He spent most of his ruling years on an unprecedented military campaign through Asia and northeast Africa, and he created one of the largest empires of the ancient world by the age of thirty, stretching from Greece to northwestern India.[1][2] He was undefeated in battle and is widely considered one of history's most successful military commanders.[3]

During his youth, Alexander was tutored by Aristotle until age 16. After Philip's assassination in 336 BC, he succeeded his father to the throne and inherited a strong kingdom and an experienced army. Alexander was awarded the generalship of Greece and used this authority to launch his father's pan-Hellenic project to lead the Greeks in the conquest of Persia.[4][5] In 334 BC, he invaded the Achaemenid Empire (Persian Empire) and began a series of campaigns that lasted ten years. Following the conquest of Anatolia, Alexander broke the power of Persia in a series of decisive battles, most notably the battles of Issus and Gaugamela. He subsequently overthrew Persian King Darius III and conquered the Achaemenid Empire in its entirety.[b] At that point, his empire stretched from the Adriatic Sea to the Indus River.

He endeavored to reach the "ends of the world and the Great Outer Sea" and invaded India in 326 BC, winning an important victory over the Pauravas at the Battle of the Hydaspes. He eventually turned back at the demand of his homesick troops. Alexander died in Babylon in 323 BC, the city that he planned to establish as his capital, without executing a series of planned campaigns that would have begun with an invasion of Arabia. In the years following his death, a series of civil wars tore his empire apart, resulting in the establishment of several states ruled by the Diadochi, Alexander's surviving generals and heirs.

Alexander's legacy includes the cultural diffusion and syncretism which his conquests engendered, such as Greco-Buddhism. He founded some twenty cities that bore his name, most notably Alexandria in Egypt. Alexander's settlement of Greek colonists and the resulting spread of Greek culture in the east resulted in a new Hellenistic civilization, aspects of which were still evident in the traditions of the Byzantine Empire in the mid-15th century AD and the presence of Greek speakers in central and far eastern Anatolia until the 1920s. Alexander became legendary as a classical hero in the mold of Achilles, and he features prominently in the history and mythic traditions of both Greek and non-Greek cultures. He became the measure against which military leaders compared themselves, and military academies throughout the world still teach his tactics.[6][c] He is often ranked among the most influential people in history.[7]

Contents [show]

Early life

Lineage and childhood

Alexander was born on the sixth day of the ancient Greek month of Hekatombaion, which probably corresponds to 20 July 356 BC, although the exact date is disputed,[8] in Pella, the capital of the Kingdom of Macedon.[9] He was the son of the king of Macedon, Philip II, and his fourth wife, Olympias, the daughter of Neoptolemus I, king of Epirus.[10] Although Philip had seven or eight wives, Olympias was his principal wife for some time, likely because she gave birth to Alexander.[11]

Several legends surround Alexander's birth and childhood.[12] According to the ancient Greek biographer Plutarch, on the eve of the consummation of her marriage to Philip, Olympias dreamed that her womb was struck by a thunder bolt that caused a flame to spread "far and wide" before dying away. Sometime after the wedding, Philip is said to have seen himself, in a dream, securing his wife's womb with a seal engraved with a lion's image.[13] Plutarch offered a variety of interpretations of these dreams: that Olympias was pregnant before her marriage, indicated by the sealing of her womb; or that Alexander's father was Zeus. Ancient commentators were divided about whether the ambitious Olympias promulgated the story of Alexander's divine parentage, variously claiming that she had told Alexander, or that she dismissed the suggestion as impious.[13]

On the day Alexander was born, Philip was preparing a siege on the city of Potidea on the peninsula of Chalcidice. That same day, Philip received news that his general Parmenion had defeated the combined Illyrian and Paeonian armies, and that his horses had won at the Olympic Games. It was also said that on this day, the Temple of Artemis in Ephesus, one of the Seven Wonders of the World, burnt down. This led Hegesias of Magnesia to say that it had burnt down because Artemis was away, attending the birth of Alexander.[14] Such legends may have emerged when Alexander was king, and possibly at his own instigation, to show that he was superhuman and destined for greatness from conception.[12]

In his early years, Alexander was raised by a nurse, Lanike, sister of Alexander's future general Cleitus the Black. Later in his childhood, Alexander was tutored by the strict Leonidas, a relative of his mother, and by Lysimachus of Acarnania.[15] Alexander was raised in the manner of noble Macedonian youths, learning to read, play the lyre, ride, fight, and hunt.[16]

When Alexander was ten years old, a trader from Thessaly brought Philip a horse, which he offered to sell for thirteen talents. The horse refused to be mounted, and Philip ordered it away. Alexander however, detecting the horse's fear of its own shadow, asked to tame the horse, which he eventually managed.[12] Plutarch stated that Philip, overjoyed at this display of courage and ambition, kissed his son tearfully, declaring: "My boy, you must find a kingdom big enough for your ambitions. Macedon is too small for you", and bought the horse for him.[17] Alexander named it Bucephalas, meaning "ox-head". Bucephalas carried Alexander as far as India. When the animal died (because of old age, according to Plutarch, at age thirty), Alexander named a city after him, Bucephala.[18]

Alexander the Great	
Basileus of Macedon, Hegemon of the Hellenic League, Shahanshah of Persia, Pharaoh of Egypt, Lord of Asia	
Alexander Mosaic (c. first century AD), ancient Roman floor mosaic from the House of the Faun in Pompeii showing Alexander fighting king Darius III of Persia in the Battle of Issus	
King of Macedonia	
Reign	336–323 BC
Predecessor	Philip II
Successor	Alexander IV Philip III
Hegemon of Hellenic League	
Reign	336 BC
Predecessor	Philip II
Pharaoh of Egypt	
Reign	332–323 BC
Predecessor	Darius III
Successor	Alexander IV Philip III
	Royal titulary [show]
King of Persia	
Reign	330–323 BC
Predecessor	Darius III
Successor	Alexander IV Philip III
Lord of Asia	
Reign	331–323 BC
Predecessor	*New office*
Successor	Alexander IV Philip III
Born	20 or 21 July 356 BC Pella, Macedon, Ancient Greece
Died	10 or 11 June 323 BC (aged 32) Babylon
Spouse	Roxana of Bactria Stateira II of Persia Parysatis II of Persia
Issue	Alexander IV
Full name	
Alexander III of Macedon	
Greek	Μέγας Ἀλέξανδρος[d] *Mégas Aléxandros* lit. 'Great Alexander' Ἀλέξανδρος ὁ Μέγας *Aléxandros ho Mégas* lit. 'Alexander the Great'
Dynasty	Argead
Father	Philip II of Macedon
Mother	Olympias of Epirus
Religion	Greek polytheism

Figure 1.1: Example. The article of Alexander the Great shows how rich Wikipedia articles can be with respect to the type of content (*e.g.*, text, pictures, animations, etc.) and available opportunities for further navigation, *i.e.*, number of outgoing links.

1.2 Accessing Information on Wikipedia

Wikipedia content is organized in articles connected by hyperlinks within an information network. These articles can be extensive, *e.g.*, an article can cover the life of a notable person by providing a detailed textual description, pictures, and links to other relevant articles in this context (*cf.* Figure 1.1). This content richness suggests different user reading preferences and information access strategies that shape Wikipedia traffic. As on the Web, search and navigation represent the major information access strategies on Wikipedia. In order for navigation to take place, users must first land on a Wikipedia page. This can be achieved by (i) directly typing the URL of a Wikipedia article into the address bar of a browser, (ii) typing a related query into a search engine, or (iii) following a link from a webpage outside of Wikipedia (*e.g.*, from a social media or blog post; *cf.* Figure 1.2). As search engines are responsible for directing a large share of external (Web) traffic to Wikipedia articles that is then transformed into system-internal traffic, they directly influence navigation within the Wikipedia network. For example, users might reach pages through search, but then leave them immediately if they were able to satisfy their information need. However, they may also leave if the search engine did not properly account for the semantics of the keywords entered and the provided search results were consequently not adequate. Alternatively, after search, a user may continue navigating. In this scenario, there are two further possibilities: (i) a search has provided a good starting point, however, the user is not completely satisfied and starts navigating; (ii) the search has been satisfying, however, the user is captivated, *e.g.*, by the topic, and starts exploring the information network further. The latter example illustrates typical Wikipedia usage and the interplay between search and navigation. The amount of digital traces produced as a result of this interplay is especially useful for developing navigational models, as it reflects user information access behavior, *i.e.*, on which pages users land after search and which links users select to proceed with navigation.

In this thesis, I look at information access on Wikipedia as repeated search and navigation activities that both complement and influence each other. I also argue that among other factors, *e.g.*, external events [RFF+10] and user motivation [SLW+17b], modeling navigational user behavior in information networks such as Wikipedia depends on several aspects. The *local* and *global* structure of the network, *i.e.*, the local choices presented to users to proceed navigating and their global network connectivity (*cf.* Section 1.2.1), *what* is potentially interesting to users, *i.e.*, the content offered (*cf.* Section 1.2.2), and *how* it is presented, *i.e.*, the general visual appearance of the article and the visibility of a promising links (*cf.* Section 1.2.3), are all relevant components in this context.

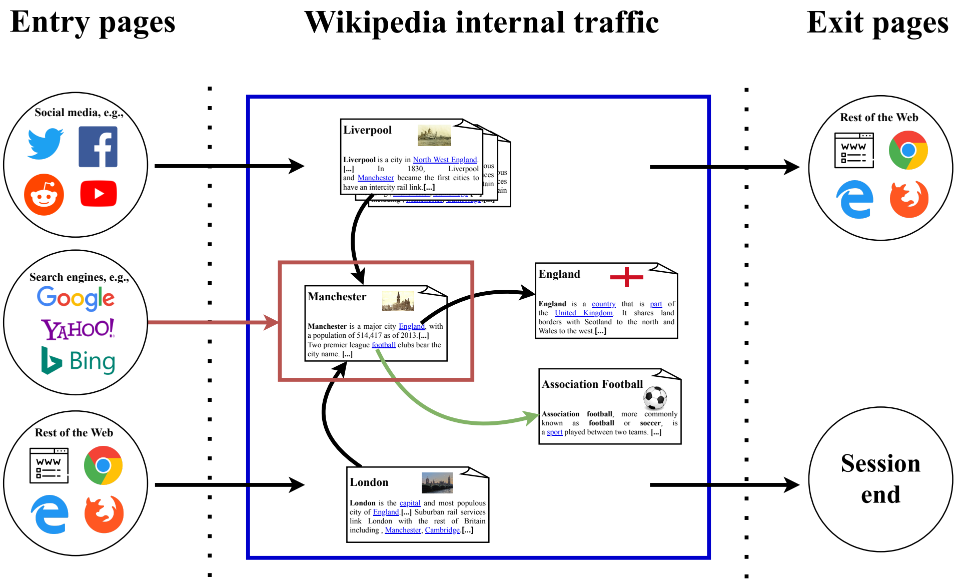

Entry pages
Wikipedia internal traffic
Exit pages
Social media, e.g.,
Search engines, e.g.,
Google
YAHOO!
Bing
Rest of the Web
WWW
Liverpool
Liverpool is a city in North West England. [...] In 1830, Liverpool and Manchester became the first cities to have an intercity rail link.[...]
Manchester
Manchester is a major city England, with a population of 514,417 as of 2013.[...] Two premier league football clubs bear the city name. [...]
England
England is a country that is part of the United Kingdom. It shares land borders with Scotland to the north and Wales to the west.[...]
Association Football
Association football, more commonly known as football or soccer, is a sport played between two teams. [...]
London
London is the capital and most populous city of England.[...] Suburban rail services link London with the rest of Britain including , Manchester, Cambridge.[...]
Rest of the Web
WWW
Session end

Figure 1.2: Information access on Wikipedia. The figure shows typical Wikipedia usage. While different webpages link to Wikipedia and act as entry points, external traffic is injected into the network mainly through search engines. Internal Wikipedia traffic is shaped by user sessions capturing navigational behavior that either end on Wikipedia articles or continue on webpages outside Wikipedia. This thesis focuses on three aspects of information access on Wikipedia: (i) at article level (red): interplay between search and navigation, (ii) at link level (green): competition between links, and (iii) at network level (blue): topological network properties ensuring high navigability (*cf.* Section 1.4).

1.2.1 Network Structure

While navigating, users must make decisions based on the available options in order to satisfy their information needs. In the context of Wikipedia, these options represent the edges of the information network. In an early work, Kamps and Koolen highlighted characteristics of Wikipedia's link structure [KK09], which inevitably plays a crucial role when modeling navigation (*cf.* Section 2.2.4). For example, well-connected network nodes with multiple incoming and outgoing edges act as *hubs* for Wikipedia's navigational traffic. Such nodes are often popular articles describing general concepts. Furthermore, in a network, information is presented in the context of the neighbor nodes that often connect themselves, forming a *cluster*. This makes edges leading out of network clusters especially useful for navigation, as they offer new avenues. The network topology can also be described in terms of the *k-core* measure [BZ02], which decomposes a network into core and periphery regions. Among others, these topological network measures capture different aspects useful for designing navigational hypotheses. Influential contributions to the connection between the network topology and its navigability were made with the introduction of the *small world* and *bow tie* network models (presented in more detail in Section 2.2). These models reveal the relationships between local and global network characteristics important for classifying networks with respect to their intrinsic ability to guide navigation.

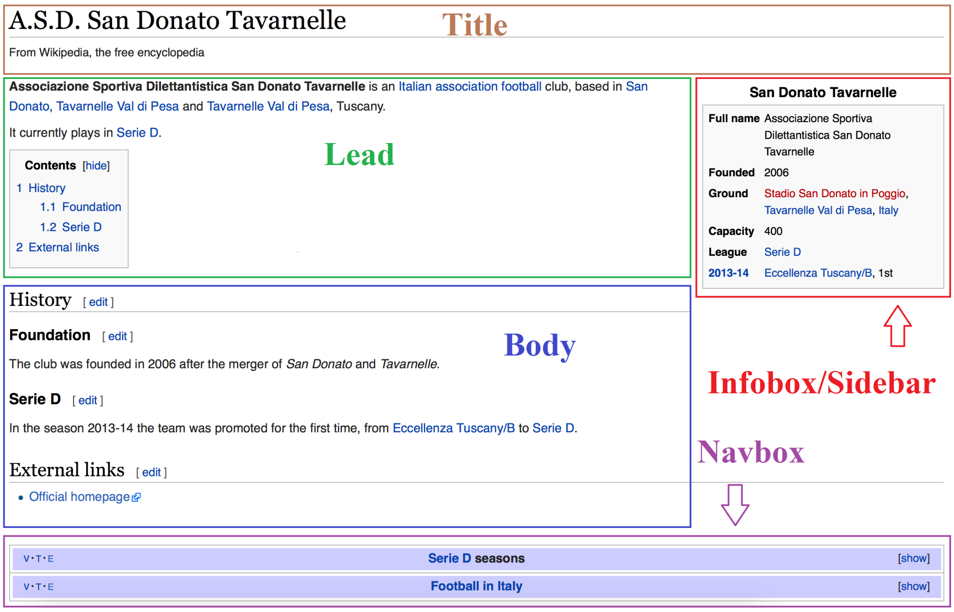
1.2.2 Content and Topics

Wikipedia is one of the first Web platforms on which users can collaborate and contribute content themselves. Collaborative content production has played an important role in Wikipedia's success, and has made it a valuable source of information covering numerous fields and topics. Wikipedia content is produced as a community effort following the *five pillar principles*[3] that define Wikipedia as a free online encyclopedia that anyone can read, edit, and distribute. The Wikipedia community has also established guidelines of cooperation and collaboration between editors, and, most importantly, defined its own core content policies: (i) natural point of view, (ii) verifiability, and (iii) no original research. The quality and type of content available on Wikipedia is governed to a large degree by these policies.

1.2.3 Content Presentation

The Wikipedia community has also established guidelines governing the appearance of articles and how content is presented to readers. Figure 1.3 shows the typical structure of a Wikipedia article, which consists of a *title*,

[3] `https://en.wikipedia.org/wiki/Wikipedia:Five_pillars` Accessed: 2018-01-12



A.S.D. San Donato Tavarnelle

From Wikipedia, the free encyclopedia

Associazione Sportiva Dilettantistica San Donato Tavarnelle is an Italian association football club, based in San Donato, Tavarnelle Val di Pesa and Tavarnelle Val di Pesa, Tuscany.

It currently plays in Serie D.

Contents [hide]

1 History
  1.1 Foundation
  1.2 Serie D
2 External links

San Donato Tavarnelle	
Full name	Associazione Sportiva Dilettantistica San Donato Tavarnelle
Founded	2006
Ground	Stadio San Donato in Poggio, Tavarnelle Val di Pesa, Italy
Capacity	400
League	Serie D
2013-14	Eccellenza Tuscany/B, 1st

History [edit]

Foundation [edit]

The club was founded in 2006 after the merger of *San Donato* and *Tavarnelle*.

Serie D [edit]

In the season 2013-14 the team was promoted for the first time, from Eccellenza Tuscany/B to Serie D.

External links [edit]

- Official homepage

v·t·e **Serie D seasons** [show]

v·t·e **Football in Italy** [show]

Figure 1.3: Visual structure of a typical Wikipedia page. This figure shows the main elements of a typical page: Title (yellow), Lead (green), Body (blue), Infobox/Sidebar (red), Navbox (purple).

a *lead* section, a *body*, *infoboxes*, and *navboxes*. The lead section is displayed directly before the table of contents and the first heading of the article containing the article title. It provides introductory information, establishes the subject of the article, and summarizes the most important facts. The body, or the remaining article sections, gives more detailed information about the subject at hand. Infoboxes provide an overview of the most important information for each article in tabular form, and are usually located in the upper right corner. Navboxes are tables containing Wikipedia internal links to related articles carefully selected by editors to put an article into context and ease further navigation.

1.2.4 Challenges and Opportunities

As the two main information access strategies, search and navigation have been extensively studied both in the context of the Web and specifically for Wikipedia. For example, Spoerri studied popular Wikipedia content, *i.e.*, articles as well as topics, and compared it to Web search results [Spo07]. Similarly, Weller analyzed Australian search traffic to Wikipedia pages [WJ11]. More recently, McMahon *et al.* studied the interdependence of Wikipedia and other information technology ecosystems, *i.e.*, Google's search engine and knowledge graph [MJH17]. The authors showed that Wikipedia content improved search engine results. Moreover, while Google forwards large amount of Web traffic to Wikipedia, it is also responsible for a significant traffic reduction to Wikipedia. With respect to navigation on Wikipedia, two influential papers based on analyses of click data from navigational games (artificial navigation scenarios) have been published. West and Leskovec showed that users experience a trade-off between article popularity in terms of its degree and textual similarity to the target article [WL12]. Helic *et al.* analyzed targeted navigation on Wikipedia and proposed a user model capturing the change in user intuition to be on the right path [HSGS13].

Several promising research directions remain uncharted. For example, navigation has been previously studied without considering the influence of search, which is responsible for a significant proportion of the Web traffic flowing to Wikipedia. Furthermore, the manifold navigational hypotheses and models presented in related literature are based on navigational click data from artificial scenarios[4] [WPP09a]. As one of the main goals of Wikipedia is to offer access to high quality content, the Wikimedia Foundation have begun to publish aggregated clickstream data [WT]. This data provides researchers with the opportunity to study user information access behavior in a large information network for the first time. Understanding where people enter and exit Wikipedia can also assist in developing hypotheses explaining users' navigational behavior, which is a challenging task. Using clickstream data, I have the opportunity to validate and build upon existing navigational hypotheses in this work. Through this, one could significantly improve node rankings, *e.g.*, produced by state-of-the-art algorithms such as PageRank. The clickstream data allows researchers to contrast both information access forms, and also provides a characterization of access patterns with respect to different topics, network regions, and edit activity. Gaining empirical insights into the usage of search and navigation strategies could reveal opportunities to make information consumption more efficient. These potential insights can be leveraged in order to address passive readers and convince them to become editors.

1.3 Problem Statement, Objectives, and Approach

This section introduces the main problem statement of this thesis. Furthermore, I present the objectives and general approach to tackling the challenges of the main problem.

Problem statement. The constantly changing and increasing amount of content, diverse and large number of users, as well as collaborative content production pose numerous challenges across different dimensions of a web-based system such as Wikipedia. For example, maintaining and improving

[4] `https://research.thewikigame.com` Accessed: 2018-06-12

the hyperlink structure of the network or offering an adequate content presentation fitting the preferred information access form are just two problems spanning the article, link, and network level of information systems. Improving information access in a web-based system has always been challenging for computer scientists. Although a significant amount of work on information access through both search and navigation on the Web exists, we still do not fully understand how exactly search and navigation interplay at article level, nor what makes a link successful at link level, nor which network structural properties ensure high navigability at the network level. Finding answers to these questions can have a significant impact on the design of information systems, *i.e.*, on the way data is structured, presented, ranked, and cached.

Objectives. The main objective of this thesis is to (i) make progress towards understanding user interactions through search and navigation on information networks and (ii) build models to improve user experience by supporting design decisions and making suggestions how to better produce and present information on Wikipedia. Of special interest is also how existing algorithms can be improved in order to reflect access patterns of a digital environment represented as an information network. Furthermore, this thesis aims to understand how the structure of an information network affects the way information is accessed in relation to the way information is presented for consumption or visualized.

General approach. To reach these objectives, this thesis follows a data driven approach. Using log data, this work starts by characterizing the interplay of search and navigation and their access patterns on different topics. Heatmaps are used to visualize the screen position and collective usage of hyperlinks representing the edges of information networks. The obtained insights are then used to design different navigational hypotheses capturing users' navigational behavior in information networks, which are then tested and compared. Subsequently, the results are applied to improve existing ranking algorithms. Finally, agent-based simulations are applied to study the general navigability of information networks and their intrinsic ability to guide navigation.

1.4 Research Questions

In this section, I present the three research questions derived from the general problem this thesis addresses, as well as the opportunities of the newly-published, large-scale user behavior data. This thesis aims to (i) characterize the functional roles (entry, exit, relay points) Wikipedia articles play in information network traffic, (ii) model the success of a link in an information network, and (iii) use agent-based simulations to assess the navigability of information networks. In the following, I present a detailed overview of each research question. First, I contrast search and navigation processes in order to characterize both access forms by focusing on the traffic each article receives. I then endeavor to identify which links are most successful and what their distinguishing characteristics are. After modeling link success for links with different target articles, I model the competition between links with the same target article. Finally, I simulate navigation in order to quantify and qualify the amount and type of network structural information needed by users to navigate efficiently.

RQ1: How do user reading preferences shape the external and internal traffic on Wikipedia?

Problem. Content diversity in web-based information systems such as Wikipedia requires users to access content by searching, navigating, or a combination of both. Although previous work has thoroughly studied information access through search [Wal11, MJH17, Spo07] and navigation [DSLS17, GY17, LMBL$^+$14, LDHS16, LLHS17], there is a lack of understanding of how search and navigation interplay in a digital environment such as Wikipedia. Studying user access patters and the interplay between search and navigation can highlight cases in which web-based systems fail to support users. To this end, I am interested in contrasting search and navigation, *i.e.*, how do user reading preferences shape the external and internal traffic on Wikipedia? Given an article, I want to know how its status as a search entry point is related to (not) relaying navigation traffic to Wikipedia, and vice versa. Beyond these general system characteristics, I also aim to identify which specific article properties influence their role in the search-vs-navigation ecosystem. I therefore ask: Which article features (*i.e.*, topic, network, content and edit features) are indicative of specific information access behavior?

Approach. Using large-scale, openly available log data from the English version of Wikipedia, this thesis contrasts search and navigation as the main ways to explore Wikipedia's content. Initially, the problem is approached by looking at the interplay between search and navigation on the total amount of article views, *i.e.*, for the top of the views distribution. Further, I develop two features capturing individual article traffic behavior: (i) searchshare—the amount of views an article received by search—, and (ii) resistance—the ability of an article to channel traffic into and through Wikipedia. Depending on both searchshare and resistance, articles are assigned to four groups describing user access behavior on articles: (i) *search-exit* articles with high searchshare that are often accessed from external search engines, but fail to animate users to navigate further (high resistance); (ii) *navigation-exit* articles that mainly receive traffic from internal navigation, but cannot channel traffic to other pages; (iii) *navigation-relay* articles that are mainly accessed from within Wikipedia, but pass on traffic internally; and (iv) *search-relay* articles that are often searched for while simultaneously contributing to further navigation. This thesis analyzes general traffic behavior for Wikipedia and highlights the major differences between the usage of articles from different topics. Subsequently, it characterizes article traffic behavior with respect to article features, *i.e.*, network properties, page content and edit history properties. In addition to a descriptive analysis, a gradient boosting model is fit to determine the impact of these article features on the preferred user access behavior.

Findings and contributions. The main contributions are the following: (i) Regarding general (collective) access behavior on Wikipedia, this thesis provides empirical evidence for the dominance of search over navigation for the most-viewed articles in the number of articles accessed and views received. For the tail of the views distribution, however, navigation appears to become more and more important. (ii) This thesis also links article properties such as position in the Wikipedia network, number of article revisions, and topic to preferred access behavior, *i.e.*, search or navigation. Finally, (iii) this thesis quantifies the strength of the relationship between article properties and preferred access behavior. The conducted analysis suggests that: (i) search and navigation are used to access and explore different articles, and both types of information access are crucial for Wikipedia, (ii) the data indicates that exit points of navigation sessions are located at the periphery of the network, whereas entry points are located at the core, and (iii) edit activity is strongly related to the ability of an article to relay traffic, and thus to the preferred access behavior. The presented results might have applications in, *e.g.*, improving and maintaining the visual appearance and hyperlink structure of articles, or identifying articles exhibiting changes in access behavior pattern, for example due to vandalism or other online misbehavior. This analysis is an initial step toward a better understanding of how search and navigation interplay to shape the traffic on platforms such as Wikipedia, and on websites in general.

RQ2: What makes a link successful on Wikipedia?

Problem. Accessing information in web-based systems relies on adequate content presentation, efficient link structures, and content ranking algorithms such as the well-known PageRank algorithm. Each of these three dimensions can benefit from understanding which link features affect user click behavior. For example, by casting all links as equal, PageRank assumes random navigation and disregards potential user preferences towards specific links. To this end, I am interested in understanding the relationship between link properties and link popularity as measured by large-scale transitional click data.

Approach. This thesis utilizes openly available large-scale datasets pertaining to the English version of Wikipedia, *i.e.*, the network of internal links within Wikipedia articles and navigational data capturing user transitions between articles. Motivated by previous work, I study three types of features that are computed from the link network and full article texts: (i) *structural features* (*e.g.*, users might prefer to navigate to central nodes in the information network) [WL12], (ii) *semantic features* (*e.g.*, users might prefer to navigate between semantically similar articles) [SNSH13], and (iii) *visual features* (*e.g.*, users might prefer links displayed at the top of the screen) [DSLS16]. In other words: Given a Wikipedia article and its outgoing links, I am interested in understanding which link features (*e.g.*, visual position) can better predict actual link transitions (popularity). Methodologically, this analysis begins with gaining descriptive insights into the success of links. Next, the corresponding effects are statistically modeled using mixed-effects hurdle models [PB06, HPN11], which allow to account for the heterogeneity in the data. Subsequently, the obtained insights are integrated into existing stochastic models of human navigation. To that end, I utilize first-order Markov chain models—likely the most popular models for this task—in two separate analyses. First, I craft hypotheses regarding human navigation based on previous insights, and subsequently integrate these into a Bayesian inference process as priors; *cf.* [SHHS15]. Bayesian model selection is then used to identify whether a given hypothesis can improve the Markov chain model fit as compared to an uninformed model. Second, I aim to further utilize these hypotheses in order to advance the classic, well-known PageRank algorithm in a weighted variation, relaxing the basic assumption of an uninformed random surfer.

In order to explore empirical insights into visual link and click positions, this thesis adopts heatmaps, which are calculated by dividing the screen into equally sized bins, and then counting the number of times (i) a link exists and (ii) a link is clicked within the respective bin. Links that are not visible due to CSS styling are ignored. If multiple links with the same target exist on the same page, the actual click count is divided by the number of link occurrences on the page to account for link ambiguity. To gain further understanding of user click preferences when multiple links with the same target article compete for attention, this thesis resorts to click data containing click counts for links with unresolved article redirects. Using a random forest approach, a link disambiguation model is then trained and evaluated in terms of mean absolute error (MAE) and earth mover's distance (EMD).

Findings and contributions. The main contributions are threefold: (i) This thesis provides empirical evidence that Wikipedia users have a preference for choosing links leading to the periphery of the underlying topological link network, preferring links leading to semantically similar articles, and that links positioned at the top and left-side of the screen have a higher likelihood of being clicked on. (ii) It integrates this evidence with first-order Markov chain models to demonstrate an improvement of respective model fits by incorporating the assumed behavior of human navigation into the inference process. (iii) Finally, this thesis establishes the utility of these findings by adapting the well-known PageRank algorithm to better reflect human navigation behavior. This enhancement leads to a significant improvement over an uninformed baseline algorithm that assumes random navigation by evaluating obtained ranking against actual article page views on Wikipedia. The methodological framework provided in this work can also be applied to other transitional click data.

RQ3: What kind of and how much network structural information is needed for efficient navigation?

Problem. Knowing which network nodes are important, and how to identify them, is crucial for navigation. To this end, this thesis studies the challenge of properly exposing the structure of the information space through an interface. This problem has two dimensions: (i) Which are the important structural properties that contribute to properly exposing the structure of the information space, and (ii) how much knowledge of these properties is required in order to navigate efficiently? Such knowledge could reduce the amount of and alter the nature of information needed for improving users' understanding of the information space, resulting in improved navigational efficiency. Consequently, I derive and discuss two sub-research questions: (i) What kind of and (ii) how much structural information is needed for efficient navigation? Firstly, I am specifically interested in identifying important structural properties of the information space that should be exposed through an interface in order to properly guide user navigation. Related work has suggested that the degree—as a proxy of a node's popularity—is a suitable navigational feature in networks with a power law degree distribution [ALPH01]. However, little is known about the suitability of the clustering coefficient as a navigational feature. The clustering coefficient may also be feasible as a navigational feature due to its importance for the emergence of the small world property of a network [Kle00b, WS98]. With this work, I investigate whether nodes with specific topological properties, *i.e.*, low clustering coefficients or high node degree, have an impact on navigation. Secondly, I am interested in determining the amount of structural information needed for navigation and whether this amount depends on the quality of the structural information.

Approach. This thesis approaches the above-mentioned research questions by analyzing the structural properties of four different networks. In order to gain an initial understanding of the topology of these networks, they are characterized using their shortest path distance, degree, and clustering coefficient distributions. In order to study what kind of and how much information is needed to efficiently navigate an information network, this thesis introduces a *partially informed* version of the *decentralized search* algorithm. Decentralized search is a message-passing algorithm that has been demonstrated to be useful for modeling navigation in information networks in a natural way [HSGS13]. The modified algorithm models a user limited in his/her exposure to the structure of the information space, and thus has only a weak or limited understanding of its topology. I study two strategies for selecting important nodes with regard to their popularity and clustering coefficient. With simulations, this thesis compares partially informed decentralized search, random search, and fully informed decentralized search algorithms. In this setting, random search corresponds to a user clicking at random without intuition, whereas fully informed search models a user with a complete overview of the network. The thesis also compares the two strategies for node selection in order to assess the importance of a user's exposure to the underlying structure of the information.

Findings and contributions. This work empirically demonstrates the sensitivity of decentralized search as a navigational model on the type of structural information utilized for guiding navigation within a network. The navigational performance of decentralized search appears to depend on the amount of high quality structural information provided. The main finding of the presented analyses is that a surprisingly small amount of structural information is needed for efficient navigation, which can be further reduced by intelligent node selection based on clustering coefficient and node degree.

1.5 Main Publications

The core chapters of this thesis are based on results from the following publications:

- Article 1 [DLFS18]: Dimitar Dimitrov, Florian Lemmerich, Fabian Flöck and Markus Strohmaier. Query for Architecture, Click through Military: Comparing the Roles of Search and Navigation on Wikipedia. In *Proceedings of the 10th ACM Conference on Web Science*, pages 371–380. ACM, 2018.

- Article 2 [DSLS17]: Dimitar Dimitrov*, Philipp Singer*, Florian Lemmerich and Markus Strohmaier. What Makes a Link Successful on Wikipedia? In *Proceedings of the 26th International Conference on World Wide Web*, pages 917–926. International World Wide Web Conferences Steering Committee, 2017.

- Article 3 [DSLS16]: Dimitar Dimitrov, Philipp Singer, Florian Lemmerich and Markus Strohmaier. Visual Positions of Links and Clicks on Wikipedia. In *Proceedings of the 25th International Conference*

Companion on World Wide Web, pages 27–28. International World Wide Web Conferences Steering Committee, 2016.

- Article 4 [DLWS17]: Dimitar Dimitrov, Daniel Lamprecht, Robert West and Markus Strohmaier. Link Disambiguation in Wikipedia: From Transition Counts Between Articles to Click Probabilities of Individual Links. (under submission)

- Article 5 [DSHS15]: Dimitar Dimitrov, Philipp Singer, Denis Helic and Markus Strohmaier. The Role of Structural Information for Designing Navigational User Interfaces. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 59–68. ACM, 2015.

In addition to these core publications, the following publications contributed to formulating the basic ideas of this thesis:

- Article 6 [Dim17]: Dimitar Dimitrov. Modeling Navigation in Information Networks. In *Doctoral Consortium at the 10th ACM International Conference on Web Search and Data Mining*, pages 845–845. ACM, 2017.

- Article 7 [DHS18]: Dimitar Dimitrov, Denis Helic and Markus Strohmaier. Tag-based Navigation and Visualization. In: Peter Brusilovsky, Daqing He (eds) *Social Information Access: Systems and Technologies*, LNCS, vol 10100, pages 181–212. Springer International Publishing, 2018.

An asterisk (*) indicates equal contributions to a paper. As the first author of the paper "What Makes a Link Successful on Wikipedia?", published at the International Conference on World Wide Web [DSLS17], I designed the experiments together with my co-authors and implemented them myself, apart from the regression analysis and the calculation of the *text similarity* feature, which were conducted by Philipp Singer. I also wrote the first draft of this paper, excluding the text passages concerning the regression analysis (also contributed by Philipp Singer). For all other publications listed here, I designed the presented the experiments together with my co-authors and implemented them myself. I also wrote the first draft of each publication, which was then iteratively improved by my co-authors. By using the plural form "we" in the remaining chapters of this work, I intend to honor my co-authors as well as the readers of this manuscript.

1.6 Contributions and Implications

The main contributions of this thesis can be summarized as follows:

- First, this thesis presents an empirical comparison of information access behavior on Wikipedia articles with respect to search and navigation using aggregated article log data. It shows that the articles accessed through search differ from the articles accessed through navigation across numerous dimensions, *i.e.*, topics, position in the network, and edit activity. Furthermore, it quantifies the strength of the relationship between article properties and preferred access behavior.

- Second, this thesis presents empirical evidence for user preferences towards specific links in an article. It also studies the success of individual links by systematically validating and building upon existing navigational hypotheses using aggregated link log data. Results suggest that users tend to select links located at the top of a page, links leading towards the periphery of the network, and links leading to semantically similar articles. These findings are then incorporated into a PageRank-based model mimicking a "reasonable" Wikipedia surfer.

- Third, this thesis studies the role of network structural information needed for efficient navigation. It introduces a partially informed version of the decentralized search algorithm to model users with various levels of knowledge of the network topology. By simulating navigation on various networks (including information networks), the amount and type of required network structural information is estimated. Results suggest that with an intelligent node selection based on clustering coefficients and node degree, the amount of knowledge of the network topology needed for efficient navigation can be notably reduced.

This thesis analyses and models information access on Wikipedia with a focus on navigation. The presented results and findings may serve different purposes. For example, they can be utilized by the Wikipedia community to improve existing content production processes and re-think the way information is presented to users. Additionally, these results have implications for algorithms structuring and exposing the information space through navigational interfaces, as well as for ranking algorithms, *i.e.*, PageRank. My hope is that the software framework[5] developed during the course of this thesis, which aligns the clickstream dataset by the Wikimedia Foundation with the corresponding Wikipedia edition, can contribute to further research on user behavior in the Wikipedia context.

1.7 Structure of this Thesis

The remainder of this thesis is organized as follows: I begin with an overview of related work and foundations in Chapter 2. In Section 2.1, I first focus

[5] `https://github.com/trovdimi/wikilinks` Accessed: 2018-06-12

on relevant theories of human navigation on the Web. Next, in Section 2.2, I discuss the theoretical navigability of an information space structured as a network. Subsequently, I focus on established frameworks for modeling navigation such as Markov chains and decentralized search (*cf.* Section 2.3), as well as state-of-the-art hierarchical clustering algorithms (*cf.* Section 2.4). This chapter concludes with a summary of important empirical studies (*cf.* Section 2.5) and applications of navigational models and click data (*cf.* Section 2.6).

Table 1.1 provides a structural representation of the relationship between the presented research questions covered by the individual chapters of this work. Moreover, this table provides a short summary of the methods and data used to approach the issues as described. More specifically, Chapter 3 studies user access behavior by contrasting search and navigation at the article level, while Chapter 4 and Chapter 5 focus on the competition between links within articles. Finally, Chapter 6 concentrates on the network itself, and uses agent-based simulations to estimate the amount and type of network structural knowledge needed for efficient navigation.

The thesis concludes with Chapter 7, in which I summarize the main results and contributions (*cf.* Section 7.1). Furthermore, I discuss important implications (*cf.* Section 7.2) as well as limitations of this work (*cf.* Section 7.3). Finally, I provide a short overview of several promising future directions (*cf.* Section 7.4).

Table 1.1: Structural overview of the content of this thesis. This table gives a structural overview of this thesis by highlighting the relationships between the individual research questions, which are described in terms of their main network entity of interest, type of data, and methods used. Additionally, the pursued goals are summarized.

Chapter	**RQ**	**Focus**	**Goals**	**Data**	**Methods**
Chapter 3 based on [DLFS18]	RQ1	articles, nodes	Characterize the interplay between search and navigation and the access patterns they produce for Wikipedia articles	internal and external traffic towards articles, aggregated empirical article access data	Topic Modeling, Gradient Boosting
Chapters 4 and 5 based on [DSLS17] and [DLWS17]	RQ2	links, edges	validate and build upon existing navigational hypotheses, tune PageRank	internal navigation traffic, aggregated empirical link transitions data	Markov chains, HypTrails, Mixed Effects Models, Random Forest
Chapter 6 based on [DSHS15]	RQ3	network	Gain insights on the type and amount of network structural knowledge required for efficient navigation	simulated traffic, synthetic navigational sessions	Agent-based simulations, Decentralized Search, Small world analysis

Chapter 2

Related Work

The following chapter puts this thesis and its contributions into the broad context of research conducted to date. Literature closely related to the individual research questions of this thesis is discussed separately in the corresponding chapters.

2.1 Information Access on the Web

In this section, I provide an overview of the most prominent information seeking theories and how they have been applied to both the design of navigational interfaces and the improvement of information system access. These theories cover both search and navigation as the two main paradigms for accessing information in a digital environment. The work conducted in this thesis complements this line of research with an empirical characterization of the interplay between search and navigation by way of illustrating a comprehensive picture of the general access behavior concerning Wikipedia articles (*cf.* Chapter 3). Further, it shows how different article properties (*e.g.*, topic, network position, editor activity) influence the preferred access, which in turn have a strong influence on both external (search) and internal (navigation) article traffic.

2.1.1 Information Seeking Theories

Information seeking is a complex cognitive process driven by activities rooted deeply in general human behavior. Several standard models attempt to describe this process from different perspectives. The most basic models treat information seeking as cycle of several user activities [SBC97, SE98], such as Marchionini and White who introduced one of the most fine-grained models describing human information seeking [MW07]. Their model includes the following steps: (i) information need recognition, (ii) accepting the challenge of satisfying the information need, (iii) formulating the problem, (iv) expressing the information need by either formulating a query or following a link, (v) examining the results, (vi) potential reformulation of the problem and the expression of the information need, (vii) application of the results. In step four, the authors consider a possible change in the information need as defined during the search triggered by new insights. This dynamic change of information need, however, is even better represented by the *berrypicking* model of information seeking introduced by Bates [Bat89]. According to her model, the process of information seeking is similar to the process of picking berries: as with berries, pieces of information that satisfy curtain information need not occur in bunches, but are typically scattered about the information space. Furthermore, only one piece of information (berry) can be picked at a time, and each pick could potentially result in a change of the information need.

Berrypicking therefore describes a dynamic form of information seeking. There are a substantial amount of observational studies and surveys that support this model of information seeking in the existing literature [HdH84, Sto82, Sto84]. Probably the most prominent study is by O'Day and Jeffries who reported common operations that users perform while seeking information [OJ93]. The authors focused on three different settings: (i) monitoring a familiar topic over time, (ii) following a predefined information seeking plan, and (iii) general exploration of a topic without a predefined target. One important result from this study is that users perform a sequence of dependant and evolving searches that tend to be initially rather broad, but gradually become more focused over time. Moreover, they identified triggers that cause a change in users' information seeking strategies.

Pirolli and Card proposed the information foraging theory to describe human information seeking in the digital environment [PC99]. In subsequent work, their original theory has been adapted to model user navigation on the Web [PF03, FP07] and to an elementary social information foraging model (SIF) [Pir09]. The information foraging theory is inspired by the optimal foraging theory, which describes animal behavior during the food search [PPC77]. According to the former, similar to food, information has a *scent* that can be followed by users while searching for information [Pir97, CPCP01]. Another central idea is that information seeking involves costs. Therefore, while users search for information, they are constantly examining the available options to proceed based on their information "scent". Users then select the most rewarding option in terms of potential information gain and corresponding acquisition costs.

2.1.2 Applications and Systems

Chi *et al.* utilized the notion of information scent for developing user flow visualizations in web-based information systems [CPP00]. In their work, the authors introduced the Web User Flow by Information Scent (WUFIS) algorithm to predict user behavior in the presence of imperfect navigational cues and a given information need. To analyze the information scent of links, WUFIS uses information retrieval techniques and spreading activation. In order to infer user information need from their actions, Chi *et al.* also developed the Inferring User Need by Information Scent (IUNIS) algorithm. As the WUFIS algorithm can be used to simulate user interactions with a website, it has been applied to infer the navigability of a website [CRS+03].

Several user interfaces and web-based systems have been introduced in order to ease navigation and browsing. For example, some user interfaces attempt to achieve better integration between search and navigation [ERG+89, FR98]. Other systems alter the underlying static topology of a website by dynamically introducing hyperlinks to better reflect the current user's needs [Gol97, PE00, YJGMD96]. Another way to improve the navigability of a website is to annotate links in order to provide better browsing cues, and thus strengthen their information scent. Furnas pointed out the benefit of highlighting links according to a user's needs as specified in the query [Fur97]. This approach proved to be promising, and several systems, including Web-Based Intermediaries (WBI) [BMK97, CM99] and ScentTrails [OC03], have since implemented it. The work on WBI annotates links based on their latency and suggests that the highlighting of links has a significant impact on navigational efficiency. ScentTrails introduced an algorithm to determine the strength of the highlighting of links with respect to their relevance to a given user query. The highlighted links provide effective browsing cues and suggest short(er) paths to the relevant documents. When compared to searching and browsing separately, ScentTrails has been shown to reduce the time necessary to find one particular piece of information.

In the Web-Watcher project by Joachims *et al.*, users specify their information need in the form of keywords [JFM+97]. While navigating, links accumulate keywords describing user information needs each time a user traverses them. Unlike ScentTrails, in the Web-Watcher Project links are highlighted using reinforced learning techniques and the previously accumulated keywords.

To measure the navigational potential of a website, Levene and Wheeldon developed a theoretical framework for Web navigation based on the concept of *potential gain* [LW04]. Similar to the concept of information scent, the potential gain framework allowed the authors to measure the usefulness of a webpage as a starting point independent of the user's information need. As with ScentTrails, they also developed an algorithm called BestTrails that detects useful, short paths through the Web given a specific query [WL03].

2.2 Navigability of Networks

In the context of this thesis, the navigability of a network refers to the extent to which the individual network nodes are reachable by traversing edges. In the following section, I discuss the small world experiment and small world network models. The small world experiment is a prominent example of navigation in a social network. The presented network models capture the topological properties of small world networks, which are known to be especially navigable. I also discuss how these models can be used to measure network navigability. Finally, I provide an overview of the bow tie model, which sheds light on the network characteristics of the Web and is fundamental for tasks that rely on navigation, such as web crawling. The section concludes with an overview of existing literature comparing the network properties of the Web with Wikipedia.

2.2.1 Small World Experiment

In 1967, Stanley Milgram conducted the so-called "small world" experiment to study the structure of American society [Mil67]. Milgram had been interested in discovering the probability of two random people being connected; in other words, the length of the chain of acquaintances that connects two random individuals. To that end, he asked a random group of Nebraska residents to forward a letter to a specific person in Boston. The participants were only allowed to use first name acquaintances and a short description of the target person to forward the letter. Surprisingly, this experiment produced a very unexpected result: the average length of the chain connecting two random individuals was only six. As a consequence, Milgram concluded that the American society represents a "small world" social network in which any individual can be reached in six steps.

This experiment inspired Jon Kleinberg to propose the *decentralized search* algorithm, which exemplifies how navigation in a network can be modeled by passing along a message [Kle00a]. In decentralized search, the message holder passes the message to one of its immediate network neighbors until the recipient of the message is found. As finding a path to a node is, algorithmically, a very simple task if the structure of the network is known, Kleinberg constrained the algorithm to use only local structural information as in the small world experiment. Decentralized search is a widely adopted method for modeling navigation in information networks, and discussed more extensively later on in this work (*cf.* Section 2.3.2). As observed by Kleinberg, the small world experiment also reveals that individuals are able two find short network paths with limited information, implying that they possess some latent knowledge of the topology of their social network that allows them to navigate it efficiently.

2.2.2 Small World Network Models

In the following, I provide an overview of two different types of models that generate small world networks, broadly divided into grid-based and hierarchy-based models. Both model types are intended to capture societal structures while generating efficiently navigable networks via decentralized search. As significant parallels between the structural organization of societies and information exist, a deeper understanding of these models is required in order to assess the navigability of information networks.

Grid-based models. In a pioneer work motivated by the small world experiment, Watts and Strogatz defined the class of "small world" networks based on the characteristic path length and clustering coefficient [WS98]. They showed that not only social networks, but many different network types possess small world properties. The authors proposed a model for generating small world networks that rewires each edge of a regular lattice at random with a probability p. Small world networks are easy to navigate, as every node in the network can be reached within a few steps. In order to navigate efficiently, one needs proper intuition of which of the candidate nodes is the closest to the target node (similar to the participants in the small world experiment).

Although the Watts and Strogatz model was the first to capture the properties of small world networks, it also has some limitations regarding navigation modeled as decentralized search. The procedure of random rewiring the authors proposed did not respect the local structure of the lattice, resulting in a lack of navigational clues for a decentralized search algorithm attempting to minimize the lattice distance to the target node in a greedy manner (the message is always forwarded to the node with the shortest lattice distance to the target node). To tackle this problem, Kleinberg proposed a new model for generating easy to navigate small world networks [Kle00b, Kle00a]. Kleinberg's model adds long range connections to a two dimensional lattice. The probability of a long range connection between two nodes in a network $p = r^{\alpha}$ is a function of their lattice distance r and the parameter α called clustering exponent. This model generates small world networks in which the correlation between the local structure and the long range connections can be used to guide navigation. This correlation can be assumed to capture the latent knowledge of the participants of the small world experiment. Kleinberg showed that for a specific value of the clustering exponent $\alpha = 2$, a lattice distance greedy decentralized search can achieve a delivery time (number of steps needed to find the target) bounded by a polynomial in $logN$, where N is the number of nodes in the network.

Hierarchical models. The lattice models of Watts and Strogatz and Kleinberg represent an initial step towards modeling small worlds that account for the *spatial* organization of individuals. People, however, break down society *hierarchically* into layers. The top layer is the most general and captures the entire world, whereas each of the underlying layers cognitively divide the society into groups with specific properties. This observation has been independently utilized by Watts *et al.* [WDN02] and Kleinberg [Kle02] to propose hierarchical models for generating small world networks:

The hierarchical model suggested by Kleinberg constructs a network from a complete b-ary tree with n leaves [Kle02]. The *tree distance* between two leaves v and w $h(v, w)$ is defined as the height of their lowest common ancestor in the tree. For each leaf v, the model creates k links by choosing the endpoint w independently with a probability proportional to $b^{-\alpha h(v,w)}$. Repetitions of w are allowed. The parameter α determines how the tree (hierarchy) and the generated network are related. The value of α encodes the probability of linking to more "nearby" nodes. For $\alpha = 0$, the target nodes are selected uniformly at random, whereas for larger α values, edges are created between nodes that are "closer" to each other. For $\alpha = 1$, the model produces a network in which a tree distance greedy decentralized search can achieve a delivery time bounded by $\mathrm{O}\big(log(n)\big)$.

Watts *et al.* also implemented the idea that individuals cluster society in a hierarchical way in order to propose a small world network model [WDN02]. Theoretically, the clustering can be terminated at the level at which each individual forms his/her own unique group. However, to better reflect reality, Watts *et al.* assumed that this process terminates at a level at which individuals form a group of a cognitively manageable size g. As in the model proposed by Kleinberg, the hierarchy used to create the actual network is a cognitive construct characterized by its depth l and constant branching ratio b. The similarity x_{ij} between two individuals i and j is defined as the height of the lowest common ancestor of the groups to which they belong in the hierarchy. If i and j belong to the same group, their similarity is set to one. In this model, the probability of a friendship between two individuals decreases with decreasing similarity of the groups to which they belong. The model implements this relationship by randomly selecting a node i and a link distance x with a probability $p(x) = ce^{-\alpha x}$. While c is a normalizing constant, the parameter α represents the homophily of the generated social network; in other words, the tendency of individuals to associate with similar individuals. The second node j is selected uniformly among all nodes at distance x from i. This process is repeated until each node in the network has an average number of neighbors z. As in the grid model by Kleinberg, the parameter α controls the network's clustering coefficient by producing either a network consisting of fully isolated cliques ($e^{-\alpha} << 1$), or a uniform random network ($e^{-\alpha} = b$). The model proposed by Watts *et al.* can also generate small world networks using multiple hierarchies. This is a notable extension of Kleinberg's hierarchical model that accounts for the fact that individuals cluster society across different dimensions, *e.g.*, geography (covered by the lattice models presented above), occupation, or income. Watts *et al.* assumed that all hierarchies H in their model are independent. Thus, if a node is close to another node in one hierarchy, this does not necessarily mean that the two nodes are close to each other in another hierarchy. The identity of a node is then represented as an H-dimensional vector with elements indicating the position of the node in each hierarchy. The position of each node in a given dimension (hierarchy) is assigned in advance at random. In the case of multiple hierarchies, the links between nodes are formed as described above by randomly choosing the dimension for each link.

Assessing network navigability. Along with the proposed model for small world networks, Watts and Strogatz also propose one approach for assessing whether a network constitutes a small world [WS98]. The authors classify networks based on their characteristic path length L and clustering coefficients C. The characteristic path length is defined as the number of edges in the shortest path between two nodes, averaged over all node pairs in the network. The clustering coefficient is defined as the fraction of the number of existing links between the neighbors of a node and all possible links between them. Both metrics have natural interpretations in a social network: L is the average number of friendships in the shortest path connecting two persons, whereas the clustering coefficient represents to what extent an individual's friends are also friends, or to what extent an individual's ego network is a fully connected graph. Watts and Strogatz consider a network to be a small world network if it possesses almost the same characteristic path length as a random network, and its clustering coefficient is much larger than the clustering coefficient of a random network.

A very similar approach for classifying networks regarding their navigability was proposed by Boguña *et al.*, in which the exponent of the power law degree distribution of a network γ and its the clustering coefficient C are used [BKC09]. They considered a network to be greedy navigable if it possesses a low γ and high clustering coefficient, whereas networks with low clustering coefficient and high γ are shown to be greedy non-navigable.

Networks can be also classified with respect to their navigability according to the hierarchy model defended by Watts *et al.* [WDN02]. Using two of the model's parameters, one can create the $H - \alpha$ space of navigable networks, where as we have mentioned the parameter α measures the tendency of individuals to connect to similar individuals and H represents the number of different social dimensions. Theoretically, almost all searchable networks have $\alpha > 0$ and lie in between the networks of fully isolated cliques and random networks. According to the model, searchable networks are mainly homophilous with a modest number of long range connections to different cliques in the network.

2.2.3 Bow Tie Model

The small world models presented in the previous section are especially valuable for understanding the navigability of a network. In order to even better understand a network's structural properties and how they influence navigability, I give an overview of the work conducted by Broder *et al.* in this section [BKM$^+$00]. In their work performing several experiments on a large-scale Web crawls, the authors have been able to validate previous results regarding the exponent of power law in- and out-degree distribution of the Web [BA99, AH00]. Furthermore, they introduced a model that describes the graph structure of the Web as a bow tie. Similar to a real bow tie, the Web graph can be decomposed into several parts. The first part of the bow tie model is the strongly connected component (SCC), which Broder *et al.* refer to as the "heart of the Web". In the SCC, all pages are mutually reachable using directed links. The IN component accommodates only nodes which can reach the SCC, but cannot be reached from it. The opposite is valid for the OUT component, which consists of nodes that are reachable from SCC, but have no back connections to it. The nodes in the TENDRILS are nodes with neither connections leading to the SCC nor back connections from it. If a TENDRIL connects the IN with the OUT component, it is called TUBE, as it bypasses the SCC. The last component is the DISCONNECTED component, consisting of nodes which exist on their own and have no connections to the previous components. The empirical analysis regarding the sizes of each component on the Web crawls revealed that SCC, IN and OUT components, and TENDRILS are almost the same size. The bow tie model develops an understating of the network structure that is beneficial to the design Web crawlers [CGM00], navigation algorithms [PBMW99, Kle99], and predictions for the future evolution of the Web itself [KRRT99].

2.2.4 Wikipedia Network Properties

The network structure of Wikipedia is similar to the structure of the Web and resembles it properties [ZBŠD06, CSC$^+$06]. Moreover, the power law exponents of the degree distributions are almost identical to the exponents reported by Broder *et al.* [BKM$^+$00], while the power law fit for the degree distributions fits the data significantly better than other distributions [ZBŠD06]. Such degree distributions are normally produced by preferential attachment processes [BA99, New01]. Using Wikipedia history, Capocci *et al.* identified that preferential attachment mechanisms are driving the growth of the Wikipedia network [CSC$^+$06], which has been shown to exhibit properties crucial to small world navigation [Hel12]. With respect to the bow tie structure, there are several studies reporting different sizes of the bow tie components. For example, Capocci *et al.* reported that the SCC of Wikipedia contains about 67% of its nodes, while Zlatic *et al.* reported a much larger SCC (about 86%). This difference might be explained by the introduction of article categories to Wikipedia in the time between the two studies. More recent studies showed that Wikipedia is more densely interlinked than, for example, the .GOV collection used at the TREC Web tracks [KK09].

2.3 Modeling Frameworks

In this section, I present the two main frameworks for modeling navigation used in this thesis: Markov chains and decentralized search.

2.3.1 Markov Chain Models

Markov chains are named after the Russian mathematician A. A. Markov (1856-1922) who first proposed the theory of *stochastic*, or random, processes. Markov chains have a long tradition and variety of applications in different branches of science, *e.g.*, biology, chemistry, physics, and finance. A Markov chain model is defined by a set of states and a set of transitions with associated probabilities. For each state, a distribution over the next possible states is defined by the transitions originating from it. A Markov chain is a special case of a stochastic process that satisfies the Markov property, meaning the next state of the Markov process depends on the current state, but is conditionally independent of the previous states of the stochastic process. The Markov property is also referred to as the *memoryless* property of a stochastic process. Furthermore, Markov chains describe discrete time stochastic processes. To model navigation (*cf.* Chapter 4), only a discrete state space $S = s_1, s_2, ..., s_m$ with $m = |S|$ is considered. For a sequence of random variables $X_1, X_2, ..., X_t$, a Markov chain is then defined as follows:

$$\Pr(X_{t+1} = s_j | X_1 = s_{i_1}, X_2 = s_{i_2}, \ldots, X_t = s_{i_t}) = \\ \Pr(X_{t+1} = s_j | X_t = s_{i_t})) = p_{i,j}$$

A Markov chain model is often defined as a stochastic transition matrix P with elements $p_{i,j}$. Each matrix element describes the transition probability between two states s_i to s_j . The stochastic nature of the transition matrix implies that the probabilities of each row i sum to one.

Markov chains can also be of higher-order, meaning that the current state also depends on the states before the previous. Markov chains have important adaptations such as Hidden Markov Models (HMM) and Monte Carlo Markov Chains (MCMC) as well as many applications such as speech recognition and natural language processing. In contrast to simple Markov chains in which states can be observed, HMM are useful when states cannot be directly observed; in other words; are *hidden.* MCMC methods are often used for sampling from a probability distribution by constructing a Markov chain with a equilibrium probability distribution as the desired probability distribution. MCMC methods gained popularity in recent years due to their advantages for statistical inference, *i.e.*, Bayesian inference. Another adaptation of Markov chains are the Markov Random Fields (MRF). They are often used in the fields of computer vision and graphics. A MRF is a graphical model of a joint probability distribution and it consists of an undirected graph in which the nodes represent random variables.

Markov chains model navigation on the Web by assigning transition probabilities between webpages. The first-order Markov chain model has a long tradition and is currently recognized as a state-of-the-art model for human navigation on the Web [HPPL98, LBL01, PC99, LW04, SHHS15]. The most prominent example of a Markov chain-based navigational model is the PageRank algorithm, which implements a random surfer [PBMW99]. First-order Markov chains have also been used for link prediction and path analysis [Sar00], as well as for mining navigational patterns [May12]. Markov chain models have also been applied to predicting Web users' next click [Dav04, DK04, ZAN99]. Although higher-order Markov chain models are not commonly used to model navigation, they have been previously applied in that context. For example, Ching *et al.* used third-order Markov chains to improve the pre-fetching of webpages in Web servers. This work extends the existing literature on the reduction of server load and network traffic using Markov chains to model user navigation [AZN99, Bes96]. Markov chain-based models for navigation have also been applied to infer semantic relatedness between concepts [SNSH13, WPP09a, YRM$^+$09, DNLH16].

2.3.2 Decentralized search

Decentralized search is an algorithm designed by Jon Kleinberg to explain the ability of humans to efficiently search for other people in large social networks [Kle00a, Kle00b]. To model navigation, the *decentralized search* algorithm passes messages between network nodes. With this algorithm, the message holder forwards a message to one of his/her immediate neighbor nodes until the intended message recipient (the target node) is found. In order to select the next step in the navigation process, the decentralized search algorithm resorts to *background knowledge* to rank the current node's neighbor nodes (also called candidate nodes) and then forward the message to one of them. This algorithm was inspired by the small world experiment presented earlier, in which background knowledge consists of a short description of the target person and participants' knowledge of their social network.

Adamic and Adar demonstrated that the quality of background knowledge plays an important role for modeling navigation [AA05]. They simulated navigation modeled as decentralized search in an organizational e-mail network and an online social network of university students. By using different hierarchies as background knowledge, they showed that efficient navigation dependents on the type of background knowledge available.

When modeling navigation with decentralized search, hierarchical background knowledge is often employed [HST$^+$11, TSHS12, HKG$^+$12, HS11].

At each navigational step, the message is passed to the neighbor node j with the shortest hierarchical distance $d(j, t)$ to the target node t. However, this distance function may be arbitrary as long as it provides a ranking for a set of candidate nodes with respect to the target node. For Markov chains, this function is given by the transition probabilities between states. In the case of hierarchical background knowledge, these probabilities are defined using the distance distributions between the nodes of the hierarchy. In Section 2.4, I provide an extensive overview of state-of-the-art algorithms for hierarchical clustering.

As shown by Adamic *et al.*, decentralized search may also exploit the degree distribution of a network in order to guide navigation [ALPH01]. In the same work, the authors showed that in power law degree distributed networks, random walks are very efficient as they naturally select high degree nodes. As random walks can be seen as a version of decentralized search without any background knowledge, Adamic *et al.* suggested a version of decentralized search that intentionally selects high degree nodes. This version scales sublinearly with the number of nodes in the network if the degree distribution of the network follows a power law. In networks with other degree distributions, *i.e.*, Poisson, the algorithm is not as efficient. Compared to Kleinberg's version, the algorithm presented by Adamic *et al.* does not depend on the target node and its position in the network. Simsek and Jensen presented another version building on the insights of Kleinberg in which navigation is guided by homophily (the tendency of like to associate with like) and on the observation of Adamic *et al.* regarding the degree distribution, or popularity, of nodes [SJ+05].

Ke and Mostafa experimented with decentralized search guided by degree, similarity, and a combination of both, similar to Simske and Jensen to study the performance of each decentralized search version under different information network sizes and clustering conditions [KM10, KM09]. Background knowledge can be indogenous and exogenous knowledge of the network in which navigation takes place. In the case of hierarchical background knowledge extracted from the network, Trattner *et al.* showed that decentralized search guided by indogenous knowledge results in navigational paths more similar to human paths than paths created by decentralized search guided by exogenous knowledge [TSHS12]. Following the same line of research, Lamprecht *et al.* showed that ontologies are suitable as background knowledge and that they also produce paths very similar to human paths [LSH+15].

Banerjee and Basu presented a significant extension of the decentralized search algorithm by introducing a social query model [BB08]. In their work, nodes have different levels of expertise, correctness, response rates, and routing policies (similar to the action selections presented next). The authors argue that an optimal message passing policy exists for all nodes in the network, and their algorithm computes it in linear time.

Numerous of the decentralized search versions presented above operate using network-endogenous background knowledge. In such cases, the message is often passed along in a greedy manner, *i.e.*, to the neighbor with the shortest hierarchy distance to the target node. This action selection tries to utilize the topology of the network in the most efficient way in order to deliver the massage. However, as shown by Helic *et al.*, this greedy action selection is not appropriate in information networks such as Wikipedia, as it does not reflect the uncertainly of the user to be on the right path [HSGS13]. To this end, the authors developed and evaluated the following action selections from which the decaying e-greedy action selection performed best:

e-greedy rule: The e-greedy action selection approach chooses the candidate node j with the shortest distance to the target node t with a probability of $1-e$. With a probability of e, another candidate node is chosen uniformly at random.

Softmax rule: The softmax rule [Bri90, DOD+06] chooses a candidate node with the shortest distance to the target node with a probability of $p(j) \propto e^{cf(j)}$. Here, $f(j)$ represents the fitness function calculated from the distances $d(j,t)$, with c as the user's confidence in his/her navigational intuition. For high values of c, the softmax rule selects the candidate node with the shortest distance to the target nodes, thus reduces to greedy selection. For small c values, the softmax rule is tuned to select other candidate nodes based on $f(j)$, thus to model a user with low navigational confidence.

Inverse distance rule: The inverse distance rule [MTC+12] is very similar to the softmax rule as it selects the candidate node with a probability of $p(j) \propto f(j)^{-c}$. The parameter c expresses confidence. The main difference to the softmax rule is the different probability distribution.

Decaying e-greedy rule: The decaying e-greedy rule [HSGS13] is based on the idea that humans do not possess sufficient navigational intuition at the beginning of the navigation process, but that their intuition continuously improves while navigating. This rule is based on a decay function that adapts e at every step of navigation. Different decay functions are possible, but generally $e(t) = e_0\lambda^{-t}$ is used. Here, e_0 is the initial value of e and λ is a decaying factor at step t.

In Chapter 6, I propose a *partially informed* version of the decentralized search algorithm. Compared to the decentralized versions as presented thus far, this version operates with very limited network structural knowledge. Partially informed decentralized search is most similar to the decentralized search algorithm suggested by Simsek and Jensen due to its reduction property. While Simsek and Jensen's version reduces to degree greedy navigation in the case of no homophily, the partially informed decentralized search algorithm reduces to a random walk when no background information is available.

2.4 Hierarchical Clustering

One convenient way to represent knowledge is by organizing it as a hierarchy. The nodes of the hierarchy can be seen as, for example, Wikipedia concepts, and the hierarchical distance between two nodes can be interpreted as a probability describing the similarity of the two concepts. In this way, a hierarchy defines a probability distribution over the nodes of a information network that can be utilized as background knowledge for a Markov chain or for decentralized search models. Therefore, I would like to give an overview of several state-of-the-art algorithms for extracting hierarchies. The following three hierarchical clustering algorithms are covered in a more detailed fashion: Hierarchical K-Means [DFG01a], Affinity Propagation [PLG10a] and Generality in Tag Similarity Graph [HGM06a]. Among the other algorithms shortly presented in this section, these three are commonly used, simple to implement, and exist in different variations. Helic *et al.* conducted initial studies on the theoretical suitability of hierarchies created by these algorithms to guide navigation [HST$^+$11]. The authors compared the structural properties of hierarchies extracted by the different algorithms by comparing their distance distributions to two synthetic distance distributions mimicking the two extreme cases of Watts' small world model—random (short and long range) and homophily (the distance distribution of isolated cliques). Their results showed that algorithms that produce hierarchies with distance distributions with many short range links mixed with a few long range links are optimal. As they point out, their results are in-line with Watts' model. The work conducted in this thesis has implications for the design of algorithms that extract hierarchical background knowledge (*cf.* Chapter 6) and is a natural continuation of this line of research. For example, I present two strategies for identifying network nodes of special navigational importance. Redesigning the algorithms from this section in such a way that these nodes are placed in the top levels of the hierarchies would result in a background knowledge of higher quality. Thus, users would be able to explore and discover content in a more efficient way.

2.4.1 Hierarchical K-Means

K-Means is likely the most prominent clustering algorithm [Llo82, For65] and exists in many variations. For example, Zhong introduced a spherical online version of K-Means [Zho05], while Dhillon *et al.* adapted the algorithm to work with textual data by replacing Euclidian distance with cosine similarity [DFG01a]. A combination of these two K-Means versions creates a hierarchy in a top-down manner. To begin, the algorithm splits the input data into ten clusters. Clusters with more than ten samples are then processed iteratively in the same manner, whereas clusters with less than ten samples are considered leaf clusters. One particular case has been introduced to handle clusters with eleven samples, which initially would have been split into ten clusters. This case supports a more free partitioning, as it allows for the division of clusters with eleven samples not into ten, but into three clusters, with each node in the hierarchy represented by the nearest object to the centroid. This object is then removed from the actual objects contained in a cluster if the cluster is further partitioned.

2.4.2 Affinity Propagation

Affinity propagation was originally proposed by Frey and Dueck [FD07]. The input of the algorithm is defined as a set of similarities between data samples provided in a matrix. The diagonal of this matrix contains the self-similarity values representing the suitability of the data sample to serve as a cluster center, also known as "preferences". The specification of a certain number of desired clusters is not necessary, however, there is a correlation between the preference values and the number of clusters (lower preference values imply a low number of clusters, and vice versa). Affinity propagation characterizes each data sample according to its "responsibility" and "availability" values. Responsibility expresses the sample's ability to serve as an exemplar for other samples, whereas availability indicates the suitability of other data samples as exemplars for a specific data sample. Affinity propagation exchanges messages between data samples and iteratively updates the responsibility and availability values of each sample with the parameter λ as the update factor.

There are different ways to create a hierarchy using Affinity propagation. For example, Plangprasopchok *et al.* adapted Affinity propagation to create a taxonomy by adding structural constraints to the algorithm's global objective function [PLG10a]. Another approach to induce a hierarchy is to use the original Affinity propagation version recursively in a bottom-up manner. The algorithm then begins with a matrix containing the top ten cosine similarities between the objects in any given dataset. The minimum of those similarities acts as the preference for all data samples. The clusters are then produced by selecting examples with associated data samples. Depending on an adjustable parameter specifying the ratio between the desired number of clusters and the data samples, the results are either returned, or another iteration begins. If the number of selected clusters in the previous run was too high, the preference values are lowered. Otherwise, they are increased. The sum of the connected data samples normalized to unit length represents the centroid of the cluster, while cosine similarities between the centroids of the clusters serve as the input matrix for the next iteration. This process is repeated until the top level is reached. As the output of the algorithm should take the form of a hierarchy, each node in the hierarchy needs to represent a unique dataset object. To this end, the nearest object to the centroid is selected as the object representing the node. Additionally, the selected object is removed from the actual objects contained in the leaf cluster and it cannot be used in lower hierarchy levels. The update factor λ can be dynamically adjusted in each iteration. The algorithm terminates when a specified number of iterations is reached, or if the clusters are stable for at least ten iterations.

2.4.3 Generality in Similarity Graph

Introduced by Heymann and Garcia-Molina, this algorithm receives a *similarity graph* as input [HGM06a]. The similarity graph is an unweighted graph in which each object is represented by a node, and two nodes have an edge between them if the similarity between their respective objects is above a specified threshold. The algorithm begins by setting the most general node (central node in the similarity graph) as the root of the hierarchy. All other nodes are added to the hierarchy in descending order of their centrality in the similarity graph. For each candidate node, the similarity between it and all currently present nodes in the hierarchy is calculated. The candidate node is then added as a child of the most similar node in the hierarchy if their similarity is above a given threshold. Otherwise, the candidate node is added as a child of the root. As pointed out by Heymann and Garcia-Molina, more control over the hierarchy properties is attainable by dynamically adjusting the similarity threshold. Further, the authors mention the possibility of using different similarity measures as well as different centrality measures. Typical versions of this algorithm are degree centrality as centrality measure, and co-occurrence as similarity measure (DegCen/Cooc), and closeness centrality and cosine similarity (CloCen/Cos).

2.4.4 Other Algorithms

Muchnik *et al.* discussed an algorithm for condensing a hierarchy based on metrics for estimating the hierarchy level of single nodes in a network [MISL07]. Clauset *et al.* presented a general approach for extracting hierarchies from network data, demonstrating that the existence of a hierarchy can simultaneously explain and quantitatively reproduce several commonly observed topological properties of networks, *e.g.*, right-skewed degree distributions, high clustering coefficients, and short path lengths [CMN08]. Lancichinetti *et al.* proposed an approach for discovering hierarchies based on overlapping network community structures [LFK09]. The authors introduced a fitness function for estimating the quality of cover, and implemented it to find the most appropriate community for each network node. Helic *et al.* adapted the generality in similarity graph algorithm by Heymann and Garcia-Molina to control the breadth of the top levels of the created hierarchy, as they identified this as a navigability reducing factor [HS11]. Benz *et al.* also changed the generality in similarity graph algorithm to create more balanced hierarchies by reducing the number of orphaned nodes [BHSS10]. Schmitz introduced a subsumption-based algorithm for inducing hierarchies [Sch06]. This algorithm uses co-occurrence statistics and builds a graph of possible parent-child relationships. For each node, the best path to the root is calculated under consideration of reinforced possible parents. These paths are then compiled into a tree structure.

2.5 Empirical Studies and Navigational Hypotheses

The algorithms from the previous section are useful for modeling navigation; however, the extracted background knowledge is endogenous to the information network and does not consider individual users operating on the network. Initial studies of human navigation behavior conducted by Catledge and Pitkow captured client side user events of NCSA's XMosaic system [CP95]. By studying event frequencies, they classified the system's users with respect to their content access strategies into three predefined groups [CW88]: (i) serendipitous browsers, who avoid the repetition of long click sequences, (ii) general purpose browsers, who perform as expected, and (iii) searchers, who, on average, produce long navigational trails. Huberman *et al.* defined the "law of surfing", which states that a user will request an additional page if its utility is equal to the utility of the current page plus a normally distributed error. With this simple stochastic approach, they were able to obtain good fits for the data collected by Catledge and Pitkow as well as for AOL Web users. By running simulations with spreading activation algorithms, they confirmed previous findings stating that the probability distribution of the number of requests for a specific webpage follows Zipf's law. The "law of surfing" model also suggests that the probability of surfing follows power law. The model presented by Huberman *et al.*, however, ignores the topology of the Web, an issue addressed by Levene *et al.* [LBL01]. In this work, user navigation on the Web graph is modeled with Markov chains. Interestingly, again, the power law distribution modeled the probability of surfing best. Huberman's "law of surfing" also ignores also webpage re-visitation. This issue has been studied by Adar *et al.*, who combined log and survey data in order to map re-visitation patterns to user intent [ATD08]. In a subsequent work, Adar *et al.* studied the relationship between the type, amount, and frequency of content change and the re-visitation patterns of webpages [ATD09]. Kumar and Tomkins characterized online browsing behavior using toolbar log and search log data [KT10]. They also defined a taxonomy for webpage types, *i.e.*, content, communication, and search. Their analysis showed the half of the requests on the Web are content, one-third are communication, and the rest are search requests. By using this taxonomy, they were able to study the re-visitation and burstness of different page types in order to further review navigational behavior on the Web.

2.5.1 Simple Navigational Hypotheses

The previously presented empirical studies have led to the formulation of several simple navigational hypotheses, which I present in this section. Some of these hypotheses have laid the foundation for powerful navigation models and have been widely discussed in previous literature. Others have been discussed in the context of social networks or information retrieval, but have yet to be evaluated and compared with each other in a coherent way based on large-scale click data. This is one of the principle contributions of this thesis (*cf.* Chapter 4).

Random surfer. The “random surfer” navigational model is based on the most simplistic navigational hypothesis, stating that while navigating the Web, users select links completely at random. This hypothesis treats all links equally and makes no assumptions about the underlying network topology, the users’ information need, or the presentation of the content. It builds the foundation for the very powerful node ranking algorithm PageRank by only allowing the random surfer to teleport to a random page if he/she gets bored [PBMW99]. Several improvements of PageRank have been proposed, which attempt to better reflect browsing behavior and, thus, better rank webpages. An extensive survey of existing literature discussing PageRank has been published by Langville *et al.* [LM04]. Due to its importance, PageRank has not only been studied from a theoretical perspective [BGS05, HK03], but also from a more technical angle with the goal of accelerating computation [Hav99, KHMG03, KCN06, McS05]. For modeling navigation, PageRank has been modified in various ways. For example, several topic-based extensions of PageRank have been suggested [Din11, Hav02, Hav03]. Two studies have analyzed PageRank’s damping factor, which accounts for the toleration rate of the random surfer [BSV05]. Becchetti *et al.* showed that PageRank’s distribution follows power law only for specific damping factor values [BC06]. There have also been several adaptations of PageRank to better reflect user-specific webpage ratings [JW03, HKJ03].

Semantic similarity. As mentioned, connections between nodes in social networks emerge based on homophily, which reflects the tendency of people to become friends with others to whom they are similar with respect to curtain social dimensions. In information networks such as Wikipedia, connections between articles representing concepts emerge if they are from the same topic or textually similar. This notion of semantic similarity is related to the notion of information scent, as it provides users with intuition as to which direction to continue navigating. CoLiDeS (Comprehension-based Linked Model of deliberate Search) is a prominent example of a cognitive model based on semantic similarity [KBP00].

Popularity. User navigational choices can also be guided by structural knowledge of the network. For example, if a node has a high degree, it is considered very popular within the relevant network. Popular nodes represent traffic hubs, as their high degree offers different options for proceeding with navigation. Additionally, they are highly reachable due to the high number of incoming connections. In information networks, popular nodes offer a lot of and very diverse content, and often represent very general concepts. The article about Alexander the Great is a good example of a hub node on Wikipedia due to its high in- and out-degree.

Visual appearance and layout of webpages. Apart from the content itself and its underlying the network structure, the presentation of said content may also guide users' navigational behavior. For example, human have been found to be influenced by the presentation of information depending on their level of expertise [MJ02]. In the context of information retrieval, several studies show that users focus their attention on information presented at the top the search results page [CZTR08]. Eye-tracking studies have even revealed an F-shaped information consumption pattern that indicates that users focus their attention on the top, left areas of webpages [Nie06]. Studies on tag-cloud usage in social tagging systems also highlight the importance of the presentation of individual tags in the cloud. More specifically, tag color, size, and ordering is found to influence tag selection [BGN08].

2.5.2 Complex Navigational Behavior

In this section, I give an overview of existing literature dealing with more complex navigational user behavior in information networks such as Wikipedia. These works cover additional aspects of information seeking theories faced by users during navigation, concepts that are not completely or directly addressed by the simple navigational hypotheses presented in the previous section, *i.e.*, trade-offs and uncertainty. The work conducted in this thesis naturally extends this line of research, as it intelligently combines those simple hypotheses to create more complex ones. These complex hypotheses can then build the foundation for a "reasonable" surfer model, which, as it is able to better reflect the collective navigational behavior on Wikipedia, can better rank Wikipedia nodes (*cf.* Chapter 4).

Trade-off between semantic similarity and degree. Human click behavior has been studied using click data from online navigational games such as Wikispeedia [WPP09a]. In such games, players are asked to find a path between Wikipedia pages by clicking as few links as possible. After a comprehensive analysis of the produced click paths, West and Leskovec concluded that humans make progress very easily if they are either far or close to the target [WL12]. Nodes with very high degree—hubs—are found to be crucial at the beginning of the navigation process, and to have a negative correlation with the produced path length. Hubs are especially useful as they offer links leading to many different network regions topologically further away from the starting point. Furthermore, human click behavior appears to be predictable at the beginning of the navigation process (when users are far from the target) and towards the end (when they are close to the target). The conceptual distance to the target is found to steadily decrease over the course of navigation, independently of the way the distance is produced, *i.e.*, the cosine between the TF-IDF vectors of two articles, or category tree distance. West and Leskovec summarize these observations by stating that during navigation, users experience a trade-off between semantic similarity and degree.

Human behavior in the presence and absence of intuition. Helic *et al.* studied human navigation in information networks using click data from TheWikiGame[1] [HSGS13]. The authors compared navigation in social networks (*i.e.*, Milgram experiment) to that in information networks, and identified several key differences. For instance, in information networks, navigation is centralized , *i.e.* an agent operates on the network, whereas in social networks, agents are part of the network and decide to whom a message will be forwarded in a decentralized fashion (independently from each other). In information networks, users have limited knowledge about the local neighborhood of a node. In social networks, on the other hand, nodes (the navigating agents in this setting) have richer knowledge about their neighbors. To this end, users' intuition guiding navigational behavior is also different regarding type and strength. In social networks, navigation is guided by social intuition, whereas in information networks, among other, topical intuition can be a guiding force. Furthermore, Helic *et al.* observed an association between the degree of a node during navigation and the current number of hops (steps) which revealed a relationship between users' exploration need and their position in the network. This observation suggests that users act greedyily if they have good intuition about their next step, whereas they seem to operate at random if they are not aware of their position in the network and/or the position of the needed piece of information. This behavior can be rooted in the structure of the network itself, as it exhibits two phases: The first phase, the exploration phase, is where users gain impressions of the environment and their position in it. The second phase, the exploitation phase, is where the real goal-seeking is conducted. Both phases are characteristic for greedy navigation, and are also referred to as zoom-out and zoom-in phases [BKC09]. Greedy navigational behavior in the presence of intuition confirms the observations by West and Leskovec with respect to the steady decrease of of the conceptual distance to the target and the predictability of clicks very close and far away from the target. Random behavior in the absence of intuition, as pointed out by Helic *et al.*, is reflected by Scaria *et al.*, who studied unsuccessful navigational game

[1] `https://research.thewikigame.com` Accessed: 2018-06-12

missions [SPWL14]. They also reported that backtracking plays an essential role in both the success and failure of a navigational mission.

2.6 Applications of Navigational Models and Click Data on Wikipedia

Navigational models have a number of applications, of which ranking content is perhaps the most important. However, these models have been also used to compute measures of semantic relatedness between concepts [SHTS14], to give recommendations [WASB16, WS04], to enhance webpage design [CRS$^+$03], as well as to improve advertising [Moe03]. Another notable application that I would like to discuss in more detail is the improvement of the link structure of information networks. As one of the largest online collaboration platforms and presumably the most dynamically-changing, Wikipedia faces sizable challenges in maintaining an up-to-date, efficiently-nevigable network structure. To this end, Wikipedia depends on automated solutions to identify broken, incorrect, or invalid links, as well as to suggest new ones. Moreover, for websites that cannot rely on a loyal community of volunteers, the development of automated approaches is of even higher value than for Wikipedia. Maintaining the link structure of a network is a task related to the problem of link prediction, which has a long tradition in social as well as complex networks. In social networks, link prediction approaches concentrate on the network structure and resort, *i.e.*, to similarity between nodes based on common neighbors, or Jaccard's coefficient [LNK07]. Furthermore, several approaches rely on a combination of network features and different machine learning techniques [LHK10, AHCSZ06, ZLZ09]. Clauset *et al.* presented an approach for inferring hierarchical structures from networks, and used these hierarchies to predict missing links [CMN08] (*cf.* Section 2.4). In information networks, several content-based approaches using text clustering, keyword extraction, and matrix factorization are prevalent [MWDR12, MW08, MC07, AdR05, WPP09b, WPP10]. These approaches cover the missing link scenario when a source article is specified and the task is to predict the target article of a link. Navigational models, *i.e.*, supervised random walkers, have also been used to recommend links [BL11]. The main advantage of these models over the previous approaches is that they naturally combine the present network structure with external network node attributes. West *et al.* presented an approach based on navigational click data from Wikispeedia and TheWikiGame that addresses another scenario, aiming to identify source articles that potentially mention a given target article and would benefit from linking to it directly [WPL15]. Their approach consists of three steps: (i) collecting navigational paths with a given target, (ii) generating pairs of start and target nodes, and (iii) filtering existing links and pairs for which the source has no established textual relationship to the target article. The authors performed automatic evaluation using ground truth based on navigational click data complemented with evaluation by humans.

Paranjape *et al.* introduced an approach for improving the link structure of web-based information systems by automatically identifying useful links using server logs [PWZL16]. Compared to the previous approaches, Paranjape *et al.* introduces an algorithm for link prediction under budget constraints, *i.e.*, the maximum allowable number of links that can be added to a webpage. This approach produces a ranking of useful links, reducing the required manual effort for human editors. The authors propose two navigational models (single and multi-tab browsing models), complemented by three objective functions to determine the quality of the suggested links. They demonstrated the generality of their approach by evaluating two distinct Web platforms: Wikipedia and Simtk. Their approach is language-independent, as it relies on strong empirical signals to build navigational models that are able to assess not only the usefulness of existing links, but also the potential usefulness of currently nonexistent links.

The internal navigability of eight different Wikipedia language editions has been examined by Lamprecht *et al.*, who applied navigational views reflecting typical Wikipedia usage, *i.e.*, users traversing links located at the top of the screen more frequently (*cf.* Chapter 4) [LDHS16]. Their results show that visual restrictions strongly limit the navigability of the underlying information networks.

Despite existing approaches for growing small Wikipedia language editions, *e.g.*, via recommendations based on content mismatches [WWZL16], significant differences in the kind and amount of available content across different Wikipedia language editions are still present. From navigational point of view, this issue is addressed by Moreira and Moreira, who focused on the creation of links in order to improve the cross-language transition on Wikipedia [OKU^{+}08]. They developed a method for identifying missing cross-language links using existing links between categories, a newly developed Cross-language Link Transitivity feature, and further textual article features.

The navigational models and insights obtained through this thesis are also applicable in a variety of contexts. For example, I propose and evaluate a weighted version of PageRank that represents the browsing behavior of Wikipedia users (*cf.* Chapter 4). In this version, the edges of the information network are weighted according to collective user click preferences. The insights from Chapter 6 have implications for the design of navigational user interfaces, *i.e.*, which and how many information network nodes must be exposed through user interfaces in order to ensure efficient content exploration.

Chapter 3

Query for Architecture, Click through Military: Comparing the Roles of Search and Navigation on Wikipedia

3.1 Introduction

Before the age of the World Wide Web [BLFFBD00], information was predominantly consumed in a linear way, *e.g.*, starting at the first page of a book and following the laid out narrative until the end. With the introduction of hypertext [Nel65] in digital environments, the way people consume information changed dramatically [LECZ$^+$12, CD07, LCH$^+$05, Man08]. While on early websites, users still predominantly visited a main page through a fixed address and were sometimes even bounded by a more directory-like navigation structure, the rise of search engines and tighter interlinking of websites have corroded the linear consumption paradigm even further. Today, users access a single website through a multitude of webpages as entry points and can usually choose from numerous paths through the available linked content at any time. In such a setting, understanding at which (kind of) pages users typically begin and end their journey on a given website, vs. which pages relay traffic internally from and to these points, provides several useful insights. On one hand, it has high practical importance since it provides the first and last contact opportunity; pages could be shaped to leverage their function as an entry point (*e.g.*, by prioritizing improvements of navigational guidance for these pages to retain visitors), or as an exit point (*e.g.*, by surveying visitors for their user experience before leaving, or by providing increased incentives to continue navigation). On the other hand, knowledge about entry, relay and exit points is also closely tied to the relation of the major information seeking strategies, *i.e.*, search and navigation: the first page visited in a session on a website is frequently reached via search engine results, after a query formulation, while navigation has been often used when the exact information need cannot be easily expressed in words [Fur97, FLGD87]. Understanding under which circumstances search or navigation dominate the users' information seeking behavior can help in developing an agenda for improving the web content in order to optimize visitor rates and retention.

Scope and research questions. Information consumption on the Web has been of special interest to researchers since the Web's earliest days [KT10, KT09]. While both search [Wal11, MJH17, Spo07] and navigation [DSLS17, GY17, LMBL^{+}14, LDHS16, LLHS17] have been investigated thoroughly in related work, they were mostly looked at separately. Consequently, so far little is known about which parts and content types of a specific website (inter)act in which structural roles, begetting different information access patterns.

In this work, we analyze how these patterns manifest on the online encyclopedia Wikipedia. With more than 5 million articles, Wikipedia is one of the primary information sources for many Web users and through its openly available pageview data provides an essential use case for studying information seeking behavior, as made apparent by numerous studies [DSLS16, LLHS17, PWZL16]. Yet, there is a lack of understanding how search and navigation as the two major information access forms *in combination* shape the traffic of large-scale hypertext environments, such as the world's largest online encyclopedia. To this end, we are interested in answering the following research questions: (i) How do search and navigation interplay to shape the article traffic on Wikipedia? Given an article, we want to know how its acting as a search entry point is related to (not) relaying navigation traffic into Wikipedia, and vice versa. This also addresses the issue of how search and navigation contribute to the article's popularity. Beyond these characteristics of the system in general, we also examine which specific properties of articles influence their roles in the search-vs-navigation ecosystem. We hence ask: (ii) Which article features (*i.e.*, topic, network, content and edit features) are indicative of specific information access behavior?

Materials, approach and methods. Building our analysis on large-scale, openly available log data for the English edition of Wikipedia, we propose two metrics capturing individual traffic behavior on articles, *i.e.*, (i) searchshare—the amount of views an article received by search—, and (ii) resistance—the ability of an article to channel traffic into and through Wikipedia (*cf.* Section 3.2).

We use searchshare and resistance to first explore the relation between search and navigation and their effect on the popularity of articles independent of their content (*cf.* Section 3.3). Depending on these two measures, we assign articles to four groups describing the role they assume for attracting and retaining visitors. Subsequently, we characterize the influence of several article attributes, including the general topical domain, edit activity and content structure on the preferred information access form (*cf.* Section 3.4). Finally, we fit a gradient boosting model to determine the impact of these article features on the preferred user access behavior (*cf.* Section 3.5).

Contributions and findings. Our contributions are the following: (i) Regarding the general (collective) access behavior on Wikipedia, we provide empirical evidence that for the most viewed articles search dominates navigation in the number of articles accessed and received views. For the tail of the view distributions, navigation appears to become more and more important. (ii) We link article properties, *i.e.*, position in the Wikipedia network, number of article revisions, and topic to preferred access behavior, *i.e.*, search or navigation. Finally, (iii) we quantify the strength of the relationship between article properties and preferred access behavior.

Our analysis suggests that (i) while search and navigation are used to access and explore different articles, both types of information access are crucial for Wikipedia, and (ii) that exit points of navigation sessions are located at the periphery of the link network, whereas entry points are located at the core. (iii) Edit activity is strongly related with the ability of an article to relay traffic, and thus with the preferred access behavior.

Our results may have a variety of applications, *e.g.*, improving and maintaining the visual appearance and hyperlink structure of articles, identifying articles exhibiting changes in access behavior patterns due to vandalism or other online misbehavior. We consider our analysis as an initial step to better understand how search and navigation interplay to shape the user access behavior on platforms like Wikipedia and on websites in general.

3.2 Transition Data and Definitions

Below, we give an overview of the used dataset capturing the traffic on Wikipedia articles and define *searchshare* and *resistance* as our main metrics for describing the individual article traffic behavior.

3.2.1 Transition Data

For studying the access behavior on Wikipedia articles, we use the clickstream dataset published by the Wikimedia Foundation [WT]. The used dataset contains the transition counts between webpages and Wikipedia articles in form of (*referrer, resource*) pairs extracted from the server logs for

August, 2016, and is limited to pairs that occur at least 10 times. The referrer pages are either external (*e.g.*, search engines, social media), internal (other Wikipedia pages), or missing (*e.g.*, if the article is accessed directly using the browser address bar). The navigation targets are purely internal pages.[1] Since we are interested in contrasting Wikipedia article access from search engines and navigation (see also our discussion in Section 3.7), we focus our analyses only on those articles in the clickstream dataset that have received views through search or internal navigation, setting aside remaining view sources (mostly "no referrer"). Accordingly, we define *total views* of an article as the sum of all page accesses by either search or navigation.

The resulting dataset consists of 2,830,709 articles accessed through search 2,805,238,298 times and 14,405,839 transitions originating from 1,370, 456 articles and accounting for 1,251,341,103 views of 2,149,104 target articles. In total, the dataset consists of 3,104,702 articles viewed 4,056,579,401 times, with a ratio of 69% stemming from search and 31% from internal navigation—in-line with previous reports on the clickstream data [DSLS17, LDHS16].

3.2.2 Definitions

To achieve a fundamental understanding of the parts that search and navigation each play for the distribution of views in Wikipedia, we take a look at the functional roles articles can assume for the overall traffic flow in respect to their *searchshare* and *resistance.*

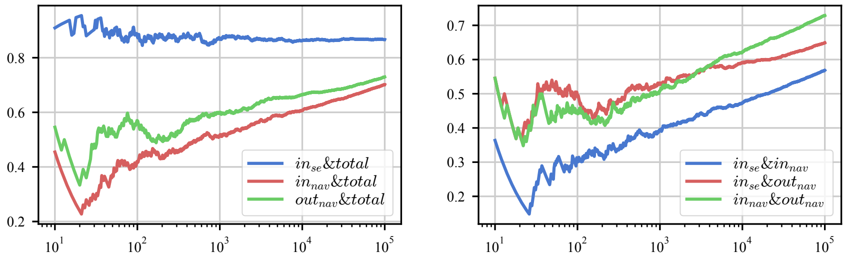
Searchshare. A high *searchshare* value indicates that search is the predominant paradigm of accessing an article, and thus that the article acts as an *entry point* for a site visit. In contrast, articles with a low value receive most of their views from users visiting them by means of navigation. The *searchshare* metric is defined as

$$searchshare(a) = \frac{in_{se}(a)}{in_{se}(a) + in_{nav}(a)} \tag{3.1}$$

where $in_{se}(a)$ is the number of pageviews an article a received directly from search engine referrers, and $in_{nav}(a)$ is the number of views from navigation as recorded in the Wikipedia clickstream.

Resistance. A low *resistance* value signals that an article forwards most of its received traffic to other articles within Wikipedia, hence does not block the flow of incoming traffic onward. A high value in turn indicates that an article acts as an *exit point.* Thus, it rarely relays users to other Wikipedia articles. These articles are traffic sinks in the Wikipedia information

[1] Leaving a Wikipedia page is treated as the end of the visit in the logs, whether by clicking on an external link or closing the page.



(a) Search and navigation vs. total

(b) Search vs. navigation

Figure 3.1: Ranking overlap. Four rankings are shown, according to the total number of pageviews (*total*), the number of pageviews coming from search (in_{se}) as well as in- (in_{nav}) and out-navigation (out_{nav}). The y-axis indicates overlaps between pairs of rankings, considering the top-k articles of each ranking as marked on the x-axis (log-scaled, top articles on the left). As a result, the overall ranking of total pageviews shows a very high overlap with the incoming search ranking. The top pages by search and navigation differ substantially. Notably, being a distribution point of traffic (high out_{nav}, *cf.* (b)) is correlated most to receiving search, but only for top out_{nav} articles, with lower ranks being supplied with traffic predominantly through in_{nav}.

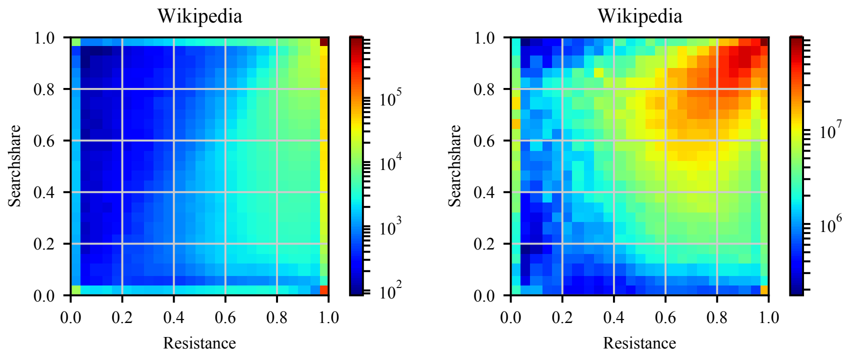
network. We define the resistance metric as

$$resistance(a) = 1 - \frac{out_{nav}(a)}{in_{se}(a) + in_{nav}(a)} \tag{3.2}$$

where $out_{nav}(a)$ is the number of pageviews that had article a as a referrer. Additionally, we restrict the values to be in the interval [0,1]. This is necessary since a small number of articles generates more out-going traffic than they receive pageviews, *e.g.*, due to a user opening several links in a new tab each.

3.3 General Access Behavior

In this section, we investigate how exogenous and endogenous traffic contribute to article popularity on Wikipedia, and we study the distribution of traffic features. We provide a first overview of the general access behavior on Wikipedia regarding search and navigation, aided by a division of articles into four groups with respect to searchshare and resistance; in Section 3.4, we will subsequently take a deeper look at dissimilarities between different types of articles.



(a) Number of articles

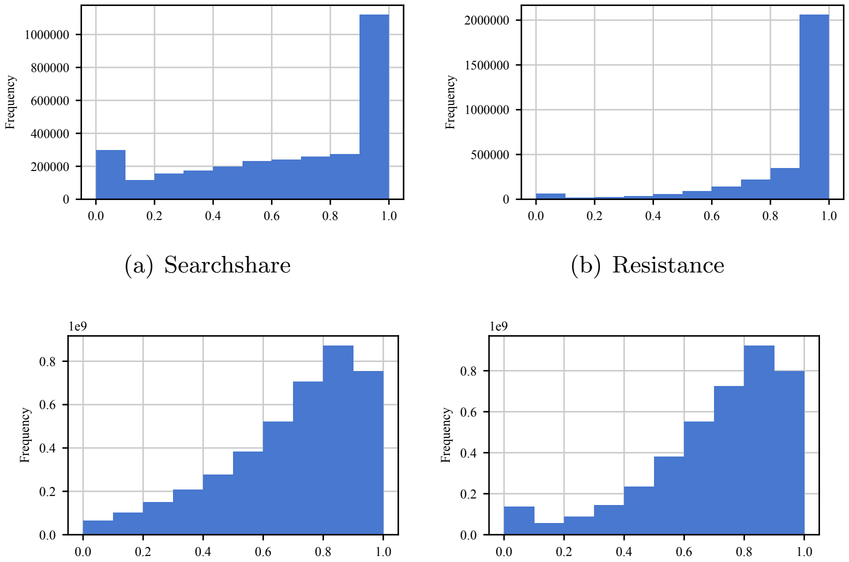
(b) Total views

Figure 3.2: Articles and article views by access behavior. For a given searchshare (y-axis) and resistance (x-axis), the figure shows (a) the number of articles and (b) the sum of their views in each heatmap square bin. Warm colors denote high values, using a logarithmic scale. We observe that search dominates navigation in terms of number of accessed articles (note the single top data bin in (a)) and that a substantial amount of articles exhibits high resistance values. When focusing on views, we see a more spread-out pattern, evidencing that a relatively small amount of articles attracts a substantial amount of search views and channels them onward to other articles (upper left side of (b)), corresponding to the *search-relay* group (*cf.* Table 3.1).

Search and navigation in relation to total views. As can be expected from related research on Wikipedia and similar online platforms, the distribution of pageviews over articles is long-tailed with a heavy skew towards the head (80% views generated by the top-visited 5.2% of all articles). To better investigate the relationship between search and (incoming and outgoing) navigation on the articles popularity, we calculate the cumulative overlap (intersection) of the descendingly ranked articles at each rank k, divided by k; this is an adaptation of the Rank Biased Overlap[2] measure.

Figure 3.1(a) shows that the top k articles ordered by search traffic (top-k-search) are highly overlapping with the top articles by total views (top-k-total) at any k, underscoring the general importance of search as a driver of incoming views. In-navigation, in contrast, is not a deciding factor to belong to the top most visited pages, but sees an extreme increase in the influence on overall views for articles up to top-k-total around 8000, at

[2] Rank Biased Overlap [WMZ10] is a common metric for similarity between rankings using cumulative set overlap in cases where the two lists do not necessarily share the same elements (as is the case here). Top-weighting as can be specified for RBO is neither suited nor necessary for the distinction of different k that we aim for here.



(a) Searchshare

(b) Resistance

(c) Searchshare-weighted

(d) Resistance-weighted

Figure 3.3: Traffic feature distributions. Figures (a) and (b) show an unweighted histogram of searchshare and resistance, while (c) and (d) respectively weight articles by their pageview counts. Most articles have a very high value for searchshare and resistance. However, extreme values close to 1.0 in (a), (b) stem mostly from rarely visited pages.

which point the increase continues, but levels off. Apparently, while search is the overall main driver for traffic, in-navigation rapidly becomes a more central source of traffic beyond the extremely popular articles. Turning to navigation passed on *from* articles to other articles, we can glean from Figure 3.1(a) that (i) while the very top of viewed articles contribute little in relation to their accumulated views to the internal traffic flow of Wikipedia (low overlap for $out_{nav}\&total$), we (ii) see a rapid and constant drop in the amount of traffic “dying” at a given page with increasing top-k-total.

Further, while it is not surprising that the outgoing traffic accumulates generally in-line with the overall received views, up until around top-k-total 1,500,000 it is generated at a rate *surpassing* the relative increase of total views, with the highest ranks of top-k-total contributing comparably little to it, just as to in-navigation. These observations are in-line with Figure 3.1(b), where we see that a higher rank in receiving navigation - rather than from search - is more strongly correlated with distributing views to other articles for the largest portion of pages, after top-k-total 3000; up until that point, the largest share of channeled traffic stems from search views. As bottom line, we see a pattern that points to a small number of pages at the extreme top of the pageview counts that are mostly searched, but *in relation to their* *popularity* rather isolated in terms of navigation; with in- and out-navigation similarly gaining notably in correlation with overall views for lower top-k-total ranks.

Traffic feature distributions. Figure 3.3 depicts the system-wide distribution of searchshare and resistance. Pages are generally much more searched than navigated to (searchshare median = 0.74, mean = 0.66) as seen in Figure 3.3(a). It is also apparent from Figure 3.3(b) that most articles do not tend to forward much of their received traffic internally, with the median for resistance for all articles lying at 1.0 and the mean at 0.88. This general tendency prevails when these scores are weighted by their received views (Figures 3.3(c) and 3.3(d)), but a notably less skewed distribution emerges, implying that—even when accounting for regression-to-the-mean effects—a majority of views is acquired via search and that a majority of views hits rather high-resistance targets.

Relation between searchshare and resistance. We observe a light positive correlation (pearson = 0.26, spearman = 0.33) indicating that the more likely an article is used to start a session, the more likely it is also to be the last article accessed in a session. Figure 3.2 depicts this association for all articles in our dataset.

To explore this relation further, we assign each article to one of four groups, determined by the *mean* of both searchshare and resistance as the thresholds.[3] We label each group according to its traffic behavior, *i.e.*, (i) *search-relay* articles that are often searched while simultaneously contributing to further navigation (above-mean searchshare, below-mean resistance); (ii) *search-exit* articles with above-average searchshare that are often accessed from search but do not lead to users navigating further (above-mean searchshare, above-mean resistance); (iii) *navigation-exit* articles that receive their traffic mostly from navigation but cannot channel traffic to other pages (below-mean searchshare, above-mean resistance); (iv) *navigation-relay* articles that are mainly accessed from within Wikipedia and able to pass traffic on internally (below-mean searchshare, below-mean resistance). Table 3.1 reports the share of articles and views pertaining to each group.

[3] A delimitation by median yields groups with the sole resistance value 1.0 and was therefore not used. Cut-offs at 0.5 would have created extremely unbalanced groups.

Table 3.1: Article group sizes and views. For each group, the table shows the percentage of articles and their received views. The majority of the articles are less visited and act as exit points of user session, whereas only popular articles are able to further relay traffic.

	search-exit	search-relay	nav.-relay	nav.-exit	total
articles	43%	9%	21%	27%	100%
views	17%	37%	39%	7%	100%

We observe that a small group of highly visited articles is able to inject considerable amounts of traffic (search-relay) into Wikipedia while about a fifth of the articles' role is mainly to channel traffic internally (nav.-relay). On the other hand, exit points receive less views while covering a much bigger portion of Wikipedia articles. Overall, these observations are in-line with Figure 3.3.

Summary. Our analysis shows that search dominates navigation with respect to the number of articles accessed and visit frequency. However, the less viewed an article is, the more significant navigation becomes as an information access form. Further, only popular articles are able to relay traffic while the majority of the articles acts as exit points for user search and navigation sessions.

3.4 Characterizing Access Behavior

In the previous section, we analyzed the general Wikipedia information access behavior, setting aside individual page attributes. However, Wikipedia articles have different properties that may influence the way they are retrieved (*cf.* Section 3.4.1). To this end, we analyze the general Wikipedia access behavior dependent on the article network (*cf.* Section 3.4.2), and content and edit properties (*cf.* Section 3.4.3). Subsequently, we highlight differences between general access behavior on Wikipedia and on Wikipedia topics dominated by search and navigation, respectively (*cf.* Section 3.4.4).

3.4.1 Wikipedia Article Data and Features

To study the influence of the content on the preferred access behavior, we focus on a snapshot of all Wikipedia articles contained in the main namespace of the English language version from August, 2016[4]. We obtained the articles using the Wikipedia API[5]. The collected article data represent the HTML version of each article on which the transitions data used to study the Wikipedia traffic has been generated (*cf.* Section 3.2.1). By parsing and rendering the HTML version of the articles, we are able to extract article features capturing aspects related to the content of the articles. The dataset contains roughly 5 million articles connected by 391 million links.

For these Wikipedia articles, we determine a wide variety of features describing their characteristics. We categorize these features into three different groups, *i.e.*, (i) network properties, (ii) content and edit properties and (iii) article topics. The network features consist of *in-, out-* and *total degree* of the article in the Wikipedia link network as well as the k-core value for this network as a typical centrality measure. Regarding the content and

[4] `https://archive.org/details/enwiki-20160801`

[5] `https://www.mediawiki.org/wiki/API:Main_page`

edit properties, we calculated for each article the *number of sections*, the *number of figures* and the *number of lists* contained in the article. These features capture visual appearance of the article, whereas the *number of revisions and editors* represent content production process. We also consider the article *age* measured in years to account for differences between mature and young articles. To account for the amount of information provided in an article, we calculate its *size in kilobytes*. The features capturing the content production process are extracted from the TokTrack dataset [FEA17] and consider the period between article creation and the end of August 2016. As the Wikipedia article categories are often too specific[6], we fit a Latent Dirichlet Allocation (LDA) [BNJ03] model on article texts using Gensim [ŘS10] bag of words article vectors with removed stop words. To allow for manual interpretation of the topics, we fit a model for 20 topics. Subsequently, we asked five independent researchers to provide topic labels based on the top words and Wikipedia articles for each topic and summarized their labels. Section 3.4.4 describes the extracted topics. The following analyses are based on a random sample of 50000 articles.

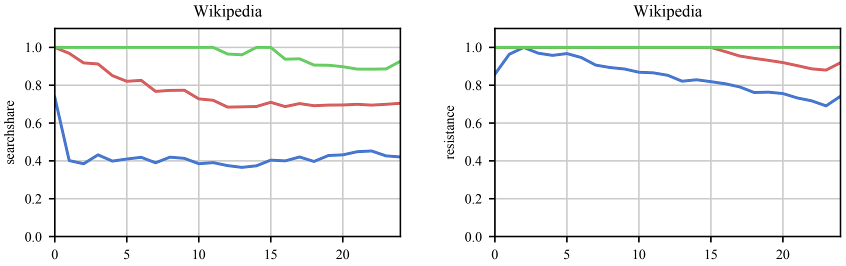
3.4.2 Network Features

To understand the role of the network features, we compute the median of the features for each of the four article groups *search-exit*, *search-relay*, *nav.-relay*, and *nav.-exit* (*cf.* Section 3.3). The results are shown in Table 3.2. We can observe that articles with below-average searchshare and resistance (article group *nav.-relay*) have higher median values across all

[6] I.e., very specific categories of articles are not linked to the relevant super-category; in other cases, two conflicting categories are linked or fitting categories are missing completely.

Table 3.2: Network features. For each network feature, the table shows the *median* feature values of the articles in the respective group. The article network properties influence the preferred access behavior. Nav.-relay articles act as intersections for the traffic as they occupy central network positions and provide lots of in- and outgoing links. Search-relay articles are similarly well-connected, which is important for injecting traffic into Wikipedia. Exit points (search-exit and nav.-exit articles) lack connectivity and are unable to channel external and internal traffic, respectively.

M	search-exit	search-relay	nav.-relay	nav.-exit	overall
in-degree	14	38	54	18	22
out-degree	33	56	71	35	41
degree	51	105	131	57	69
k-core	44	76	95	49	57



(a) K-core vs. searchshare (b) K-core vs. resistance

Figure 3.4: Network position. The figure shows the first (blue), second (red), and third (green) quartile of searchshare (a) and resistance (b) as function of the article position in the network indicated by its k-core. K-core values are divided into 25 bins. The access behavior on articles is influenced by their position in the network. The more central an article, the lower its searchshare and resistance—*i.e.*, the more traffic it relays through the network.

network features, *i.e.*, they are located more in the center of the network and consistently have more incoming and outgoing links. Although search-relay articles are not as well connected as nav.-relay, their relatively central position in the network and high number of outgoing connections is important in order to inject traffic into Wikipedia. By contrast, articles that are often used as exit points (*search-exit* and *nav.-exit*) are located more in the periphery of the network (low k-core value), are less often linked to, and contain less out-links themselves, which eventually results in higher resistance values, signifying the termination of user sessions.

For further analysis, we sort the articles according to their k-core value and discretize them into 25 equally-sized bins. For each bin, we compute the quartiles for searchshare and resistance, as seen in Figure 3.4. Looking at the median (center red line), we find that for articles with increasing k-core values the searchshare indeed decreases (*cf.* Figure 3.4(a)). However, this effect stops at around 50% of the dataset, *i.e.*, for half of the articles, which are located in high k-core network layers, the searchshare is mostly independent from the exact centrality. Regarding the resistance, there exists a substantial amount of nodes with a resistance of 1.0 for all k-core values, *cf.* the green line indicating the upper quartile. However, for the more central nodes, an increased number of pages have a significantly lower resistance (*cf.* Figure 3.4(b)).

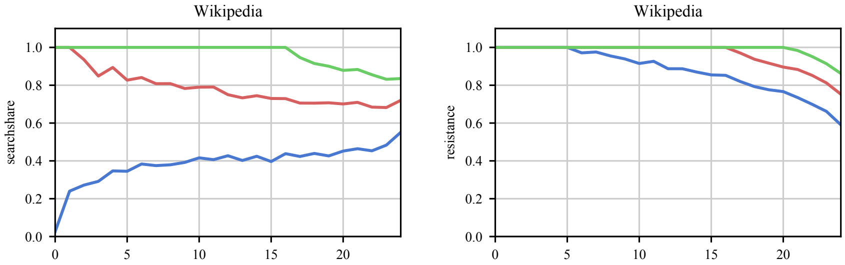
3.4.3 Content and Edit Features

Next, we characterize the article groups in terms of the article content and edit history which account for the content presentation and content production process. Table 3.3 reports the median values of these features in the four article groups. We can observe that the content features (number of tables, number of sections, size of the article) are modestly increased for relay articles, *i.e.*, articles that contain more content tend to be less often exit points of navigation sessions. By contrast, the revision history plays a more important role: we can see that articles in the *search-relay* group have (as a median) more than twice the number of editors and revisions compared to exit articles, and tend also to be somewhat older. Articles in the group *nav.-relay* show similar, but slightly lower values with the same tendency. The median feature values for both "exit" article groups are very similar and show slightly lower editor and revision numbers. Overall, content and edit features provide strong indicators for articles relaying traffic (as opposed to being exit points), but only weak indicators for being accessed by search or by navigation.

We will have a more detailed look at an exemplary edit feature, *i.e.*, the number of revisions. Analogously to above (network features), we assign the articles to one of 25 bins according to their revision count, compute the distribution of searchshare and resistance for each bin, and plot the quartiles. The results are shown in Figure 3.5. We can see that the median searchshare continuously decreases with increasing number of revisions. The effect is in particular significant for very low number of revisions (*cf.* Figure 3.5(a)). Additionally, the spread of the distribution—measured by the interquartile range (IQR)—also substantially decreases the more revisions an article has. This can likely be explained by *regression to the mean* since articles with less revisions receive overall less views, making more extreme searchshare values more likely. With regard to the resistance, we can observe that specifically high number of revisions correlate with a lower resistance scores (*cf.* Figure 3.5(b)). The number of editors, and the age of an article is highly

Table 3.3: Content and edit features. For each content and edit feature, the table shows the median feature values of the articles in the respective group. The content production process influences the access behavior as search- and nav.-exit points have low edit activity, and offer less content. On the other hand, relay articles are more frequently edited, and congruently, are generally more extensive.

M	search-exit	search-relay	nav.-relay	nav.-exit	overall
editors	21	52	46	21	25
revisions	38	97	86	37	46
sections	6	7	7	4	6
tables	3	3	4	3	4
age	9	11	10	8	9
size	41	50	54	41	44



(a) Revisions vs. searchshare

(b) Revisions vs. resistance

Figure 3.5: Edit activity. The figure shows the first (blue), second (red), and third (green) quartile of searchshare (a) and resistance (b) as function of the article editors' activity indicated by the number of revisions. Revisions values are divided into 25 bins. Except for the most edited articles, high edit activity has a negative effect on the resistance, which on the other hand has a positive effect on navigation indicated by the lower searchshare.

correlated with the number of revisions and reveal a very similar behavior with respect to searchshare and resistance.

3.4.4 Topic Features

Search-related popularity, navigability as well as other characteristics related to traffic might be highly dependent on the topical domain of an article. We hence investigate the access behavior across Wikipedia's numerous article themes, represented by the 20 topics we have extracted. Table 3.4 provides descriptive statistics of these topics. With 32% "TV and Movies" is the topic with the most views while consisting of a mere 7.5% of all articles on Wikipedia. "Technology, Stubs" and "Architecture" show an opposing dynamic, providing a large amount of articles, but relatively few views.[7] Overall, the amounts of articles and view counts are not strongly correlated. Consistent with previous research, we also observe that the popular articles are in general longer, relative old, and revised more often by more editors [Spo07].

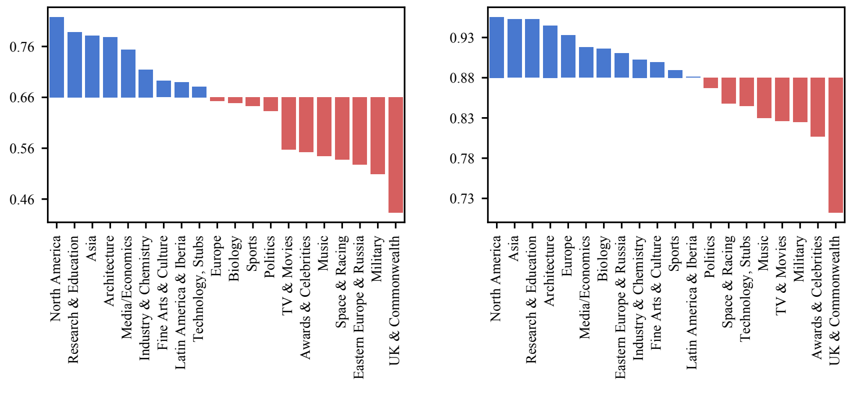
A look at the distribution of searchshare and resistance in the overview provided by Figure 3.6 reveals the different access behaviors for Wikipedia topics. To examine these pronounced differences further, we set out to highlight the dissimilarities of the overall searchshare vs. resistance distribution for total views—as shown in Figure 3.2(b)—with the same distribution for the individual topics. To do so, we create heatmaps pinpointing the *relative*

[7] "Technology, Stubs" is a compound of general Wikipedia:Stub articles and often short technology articles that were not sufficiently distinguishable by LDA. We exclude it from discussion here due to its ambivalent nature.

differences of each topic to the baseline of the overall distribution. This is achieved by performing a bin-wise division of a topic's normalized view count for a given searchshare-resistance bin with the respective normalized bin for the general Wikipedia traffic behavior. The resulting heatmaps are shown in Figure 3.7 for selected topics. They draw a clear picture of the over- and under-representation of certain article types (in terms of views) in each topic against the whole-system baseline. “Architecture” in Figure 3.7 stands as one representative for a group of topics (“Biology”, “Industry & Chemistry”, Research & Education, “Space & Racing”) that all exhibit a very similar distribution with their article views occurring at high searchshare and high resistance, *i.e.*, these topics are mostly searched and not used for further navigation. In stark contrast, views for “Military” topics occur to the largest part in comparably low-resistance articles, that are mostly navigated to (views for “UK & Commonwealth” are distributed almost analogously). “Sports” reveals a similar inclination for *nav-relay* types of articles attracting views, yet sports articles also frequently get accessed by search and abandoned immediately (closely related patterns: “Fine Arts

Table 3.4: Topic statistics. The table shows the percentage of articles and views for each topic. Additionally, it reports the median age in years, number of editors and revisions, and the size of the articles in kB. Not surprisingly, popular articles are generally longer in terms of text, edited more and by higher number of editors, and relatively old.

topic	% articles	% views	*M* age	*M* editors	*M* revisions	*M* size
Technology, Stubs	19.3	7	7	20	36	38
Architecture	12.4	5	8	31	61	56
Sports	12.0	8	7	37	86	68
Politics	8.1	8	8	47	103	60
TV&Movies	7.5	32	8	96	197	55
Fine Arts&Culture	7.2	6	8	49	100	49
Biology	7.0	6	7	29	57	46
Music	6.9	9	8	64	136	53
Research&Education	4.8	2	7	40	87	47
Media/Economics	3.3	4	8	54	115	49
Military	3.1	4	8	47	105	65
Industry&Chemistry	3.0	6	9	66	126	55
North America	1.2	0	9	23	39	52
Space&Racing	1.2	2	8	52	114	73
Europe	0.9	0.0	7	19	36	52
Asia	0.7	0.0	5	8	14	60
Latin America&Iberia	0.5	0.0	7	20	33	58
UK&Commonwealth	0.5	0.0	7	19	40	43
Eastern Europe&Russia	0.4	0.0	7	15	24	49
Awards&Celebrities	0.0	0.0	6	22	42	41



(a) Searchshare

(b) Resistance

Figure 3.6: Access behavior for all topics. Topics are ordered from highest (left) to lowest (right) for searchshare (a) and resistance (b). Values over (blue) and below (red) the respective mean value are colored respectively. There are pronounced differences in the dominant access behavior on different Wikipedia topics.

& Culture" and "Politics"). Lastly, "Music" and "TV & Movies" exhibit remarkably idiosyncratic distribution patterns, not mirrored by another topic. "Music" attracts many views in a *search-relay* fashion, but on the other hand also explicitly acts as a "dead end" for internal navigation.

As "maximally different" topics in respect to these traffic patterns and with overall high view counts, we select "Architecture" for search-heavy topics, and "Military" for navigation-heavy topics to conduct a deeper analysis regarding article network, content and edit properties. While "Architecture" includes articles covering popular buildings, landmarks and municipalities, "Military" consists of articles covering significant historic events often associated with violence such as wars and notable battles, along with many articles dedicated to military units, personnel and equipment (*cf.* [SLW$^+$17a]). For the general access behavior concerning the network, content and edit features, we again assign the articles to one of 25 bins according to their k-core and revision counts, compute the distribution of searchshare and resistance for each bin, and plot the quartiles (*cf.* Figure 3.8). For "Architecture", searchshare (a) initially decreases for increasing k-core but sees an uptick for very central nodes, and a very similar behavior can be observed for resistance (b). "Military" is characterized by generally lower levels of both metrics, yet shares the trend of decreasing resistance with increasing k-core (e), meaning that for both topics, the more central articles in the network are able to channel visitors into Wikipedia, with the top-most central nodes excluded from this trend. Being edited more implies decreasing resistance for both topics ((d), (h)), although this trend reveals itself only for much

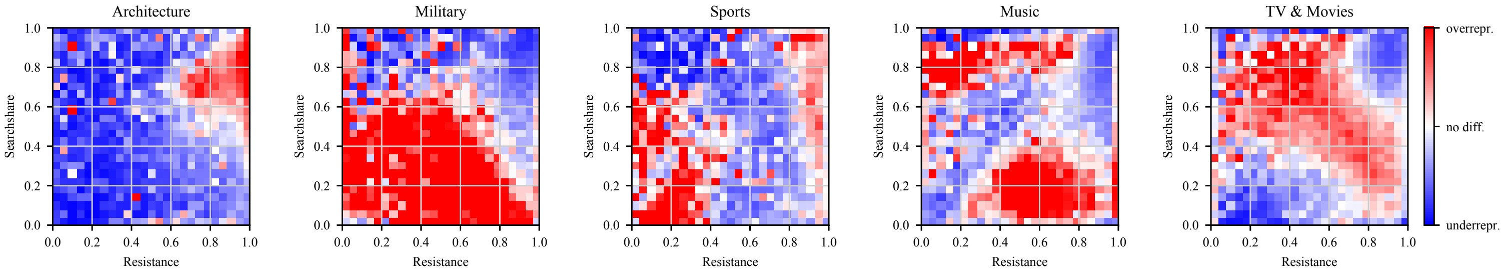
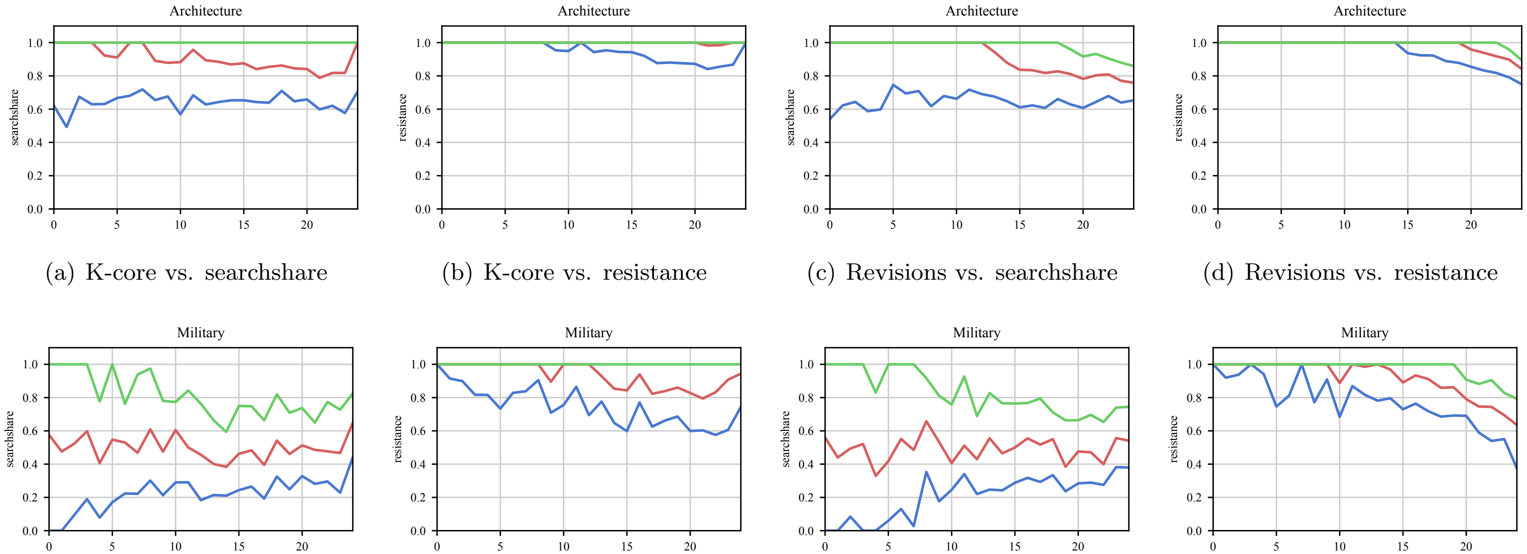


Figure 3.7: Relative difference of individual topics to the overall view distribution of searchshare vs. resistance (*cf.* Figure 3.2(b)). White denotes no relative difference, blue denotes underrepresentation (down to 0), while red denotes overrepresentation (max. over all topics at 2). The figure highlights the differences between search-heavy and navigation-heavy topics compared to the all-articles baseline. “Architecture”, exhibiting above-mean searchshare and resistance (*cf.* Figure 3.6) stands representative for six similarly distributed topics and mainly attracts search hits that it cannot pass on. “Military” shows an almost inverted pattern, mostly receiving as well as producing internal navigation. The bi-focal distribution of “Sports” can be found in “Politics” and “Fine Arts & Culture” as well, while patterns for “Music” and “TV & Movies” are more unique.



(a) K-core vs. searchshare

(b) K-core vs. resistance

(c) Revisions vs. searchshare

(d) Revisions vs. resistance

(e) K-core vs. searchshare

(f) K-core vs. resistance

(g) Revisions vs. searchshare

(h) Revisions vs. resistance

Figure 3.8: Relation of traffic features with network and content features. For a topic dominated by search (“Architecture”) and one dominated by navigation (“Military”), the figure shows the first (blue), second (red), and third (green) quartile of the article searchshare and resistance as function of its position in the network indicated by its k-core and editors’ activity indicated by the number of revisions. Articles are divided into 25 bins by k-core and revision values. Apart from base-level differences of searchshare and resistance, the topics exhibit comparable trends, with the exception of searchshare not being influenced as much by network position or edit activity features for “Military” articles.

higher revision counts for "Architecture", most likely to its generally higher resistance. Edit counts have no clearly distinguishable influence on "Military" articles' searchshare, for "Architecture" it, however, implies lower searchshare.

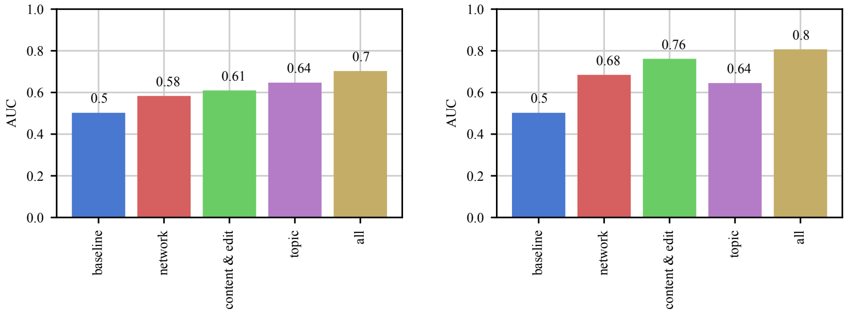
Summary. The results presented in this section suggest that the content heavily influences the access behavior on Wikipedia. Particularly, topical domains are accessed differently, *i.e.*, users prefer to access articles about architecture and landmarks mainly through search, whereas more historical articles about military actions are navigated. Moreover, mature articles with high revision numbers and article located in the core of the network are more likely to channel traffic through Wikipedia, whereas articles located at the network periphery act as exit points.

3.5 Modeling Access Behavior

Our previous analysis characterized the user access behavior on Wikipedia articles with respect to their traffic from search and navigation dependent on the article features. However, this analysis does not reveal the impact of the feature groups on the access behavior. To this end, we set out to model the access behavior on articles in order to measure the influence of each feature group. The higher the predicative performance of a feature group, the higher the influence of the group is on the role articles play with respect to the traffic (entry-exit and relay articles), and thus on the preferred information access form (search and navigation).

Modeling searchshare. We ask, given a Wikipedia article, if it is possible to classify it as dominated by search, *i.e.*, $searchshare > 0.66$ or dominated by navigation, *i.e.*, $searchshare \leq 0.66$. The threshold used for the separation is the searchshare mean (*cf.* Section 3.3). In our experiments, we consider four different sets of article features: (i) network features, *i.e.*, in-, out-degree, k-core, (ii) content & edit features, *i.e.*, number of revisions and editors, article size, number of tables, pictures, lists, and (iii) topic. For predicting the preferred access form, we fit a model using gradient boosting and evaluate the model's performance with ROC AUC. The model is trained using 10-fold cross validation at a balanced dataset. On this dataset random guessing results in 50% accuracy, which is also used as baseline. Figure 3.9(a) shows the individual performance for each feature group, as well as the performance for the combination of all features. We observe that modeling searchshare is difficult even with all features (AUC = 0.70). As expected, the network features are the least indicative (AUC = 0.58). Further, the topic feature predicts searchshare best (AUC = 0.64), which suggests strong user preferences for specific information access form, *i.e.*, search or navigation for different topics.

Modeling resistance. To model the resistance of Wikipedia articles, *i.e.*,



(a) Searchshare (b) Resistance

Figure 3.9: Results. The figure shows the model performance (ROC AUC) for (a) searchshare and (b) resistance. Predicting searchshare is more challenging than predicting resistance. The article topic determines the preferred access behavior (search or navigation). However, position in the network, content maturity and presentation of the article are indicative of resistance, and thus if an article will be an entry-exit point or a relay point for the traffic.

they ability to relay traffic, we treat an article as a relay point if $resistance \leq 0.88$ and an exit point if $resistance > 0.88$. Again, the separation of the articles is based on the resistance mean (*cf.* Section 3.3). We consider the same feature groups as for modeling searchshare and use random guessing as our baseline. For classifying the articles, we again utilize 10-fold cross validation to train a gradient boosting model on a balanced dataset. The performance is measured in terms of ROC AUC. Figure 3.9(b) shows the individual classification performance for that task for each feature group. The content and edit features are the most important (AUC = 0.76). This makes the case for an influence of the way content is presented to the user on lowering or increasing the resistance of a page. Unlike for searchshare, the network features are indicative of the resistance of an article (AUC = 0.68). This suggests that the network position of an article influences the extent to which it channels traffic. The topic plays only a small role, which again highlights the importance of the quality of the content presentation and production process.

Summary. In general, modeling article resistance is easier than modeling searchshare as suggested by the higher ROC AUC values. Modeling searchshare is challenging due to the influence of external events (*e.g.*, the transition data exhibits high view numbers on articles about the Summer Olympics 2016), and the content diversity. However, investing in diverse content from different topics seems to be the best way for Wikipedia to attract people as the article's topic is the most indicative feature regarding searchshare. On the other hand, the content presentation, and the article's position in the network are decisive for its ability to relay traffic.

3.6 Related Work

Since the inception of the Web, researchers have been studying the user content consumption behavior. Initially, content has been accessed by traversing hyperlinks on the Web [KT10]. This navigational user behavior on the Web and on Wikipedia is often modeled using well-established methods such as Markov chains [CKRS12, SHHS15, SHTS14, PBMW99, PP99] and decentralized search models [DSHS15, HSGS13]. Numerous navigational hypotheses on Wikipedia have also been presented based on, *e.g.*, click traces stemming from navigational games and on click data from server logs. For example, West and Leskovec observed a trade-off between similarity and popularity to the target article in the user sessions of Wikispeedia [WL12]. Lamprecht *et al.* studied the general navigability of several Wikipedia language editions and showed how the Wikipedia article structure influences the user click behavior [LDHS16, LLHS17]. By constructing a navigational phase space from transition data, Gilderslave and Yasseri studied internal navigation on Wikipedia and identified articles with extreme, atypical, and mimetic behavior [GY17]. Web content can be also discovered by formulating and executing a search query. Kumar and Tomkins performed an initial characterization of the user search behavior [KT09], while Weber and Jaimes studied the search engine usage with respect to the users demographics, topics, and session length [WJ11]. Earlier Wikipedia reading behavior studies focused on explaining bursts, dynamics of topic popularity and search query analysis to Wikipedia [tTVLK12, RFF$^+$10, Spo07, Wal11, LMBL$^+$14]. A more recent study by Singer *et al.* investigated the Wikipedia readers motivations [SLW$^+$17b]. By complementing a reader survey with server log data, they discovered specific behavior patterns for different motivations, *i.e.*, bored readers tend to produce long article sequences spanning different topics. McMahon *et al.* focused on the interdependence between search engines, *i.e.*, Google and Wikipedia [MJH17]. They showed that Google is responsible for generating high traffic to Wikipedia articles, although, in some cases traffic is reduced due to the direct inclusion of Wikipedia content in search results. Compared to our work, McMahon *et al.* concentrate on the peer production site and not on the content consumption. While there is a long line of research with respect to search and—more so—navigation, they have rarely been studied together which is the focus of this chapter.

3.7 Discussion

As a general observation, our results shed light on the different roles of articles with respect to traffic entering and leaving Wikipedia. On one hand, an overwhelming amount of pages attracts mostly direct search traffic and only little internal navigation, thanks to Wikipedia's strong symbiotic relationship with web search engines. Yet, notably, most of that traffic goes to articles that act mainly as exit points, *i.e.*, users to not continue visiting Wikipedia directly afterwards. This is congruent with, but not necessarily because of, a pure "look up" nature of search. Only a very small share of searched articles is responsible for relaying disproportionally large amounts of traffic into the rest of Wikipedia. We see that these articles are well-connected, more edited and more extensive than their exit counterparts, although we cannot yet conclude whether this is because of a "worn path" paradigm, wherein links and content are built because of the natural thematic positioning and suitability of an article to act as an entry point *and* as a bridge to more content, or because the a-priori structure of these article facilitates the observed navigational patterns. A longitudinal study, which we plan for future work, could obtain more detailed insights on this co-evolution of structural features and navigation. Furthermore, our data shows that articles, which are able to forward traffic, sit mostly at the very (k-)core of the link network. However, this is not necessarily the case for being a receiver of navigation traffic, with searchshare values stabilizing already at lower k-cores—and with inlinks not being more highly correlated with k-core than outlinks. This hints to the fact that—to some extent—users enter Wikipedia by search on more central articles, and then navigate outwards to articles in the periphery of the information network (cf. Chapter 4).

Regarding articles with different topical alignments, we see certain evidence that the thematic domain of a user's information pursuit seems connected with the "mode" of how this information is attained. While the highly aggregate data used in this work does not allow for direct inferences as to the type of information retrieval in the continuum between a targeted and well-defined lookup and a completely serendipitous discovery process, we can nonetheless discern distinct patterns between article topics. Although "Architecutre" articles are not more devoid of in- or out-navigation opportunities than "Military" ones, they show far higher amounts of search views and exit points, while the latter one is navigated at a constantly high level, regardless of their connectedness. A possible explanation of the navigation-heavy behavior on "Military" articles is that people like follow paths through events in order to understand historical developments.

For our analysis, we utilize publicly available clickstream data about Wikipedia. However, due to privacy restrictions, the data contains only (referrer, resource) pairs that occurred at least ten times during the data collection period. This could lead to a skewed view on the access behavior when contrasting search and navigation. For example, if an article is navigated in total much more than ten times over different links, but each individual link is transitioned less than ten times, all of these transitions will not be included in the data. In this case, the searchshare for this article might get substantially overestimated. Since this might occur specifically for articles with overall few page views, it may be a potential explanation for some findings, *e.g.*, that article in the periphery of the link network show a stronger prevalence of search.

3.8 Conclusion and Future Work

In this work, we studied the prevalence of user access preferences across articles on Wikipedia. For that purpose, we introduced *searchshare* and *resistance* as two key features to characterize article traffic. While we can identify search as the more dominant access paradigm compared to navigation on Wikipedia overall, we observe heterogeneous behavior at different types of articles. That is, depending on the article topic and other article properties, the share of navigation and search strongly varies, as well the amount of traffic an article relays to other Wikipedia pages. For example, articles on topics such as “Military” exhibit above average access by navigation, while topics such as “Architecture” show a strong prevalence of search. Furthermore, edit activity on a an article and its position in the network is strongly correlated with its ability to relay traffic on Wikipedia. Thus, we find overall that both, search and navigation play a crucial role for information seeking on Wikipedia.

In the future, we plan to extend our studies over time intervals and to other language editions in order to further explore cultural differences in the identified access patterns.

Chapter 4

What Makes a Link Successful on Wikipedia?

4.1 Introduction

Even though links are omnipresent on the Web, only a minority of them get regularly clicked by humans. In the previous chapter, we studied how people access information by contrasting search and navigation on Wikipedia. Our results suggest that there are notable user preferences regarding the information access form, which potentially leads to user preferences towards specific links as well. For example, on Wikipedia only around 4% of all existing links are clicked by visitors more frequently than 10 times within a month (*cf.* Section 4.3). This phenomenon demonstrates the importance of a deeper understanding of human navigation behavior in order to provide efficient and effective navigational support for users. While a variety of navigational regularities, patterns and strategies have been identified in previous work [WL12, SHHS15, PWZL16, LLHS16], our research community still lacks a systematic understanding of *factors that make a link successful in information networks.* Towards that end, we turn our attention to Wikipedia, one of the most popular information sources on today's Web.

Problem and objectives. In particular, we are interested in understanding *the relationship between link properties and link popularity* as measured by *large-scale transitional click data.* Understanding what link features affect user click behavior can have implications for how we place links on Wikipedia, but can also have consequences for fundamental Web algorithms such as Google's PageRank that assume random navigation casting all links as equal and do not account for potential preferences towards specific links.

Materials, approach and methods. We utilize openly available large-scale datasets about the English Wikipedia: the network of internal links within Wikipedia articles and navigational data capturing user transitions between articles (*cf.* Section 4.2). Motivated by previous work, we study three types of features that we compute from the link network and full texts of the articles: (i) *structural features* (*e.g.*, humans might prefer to navigate to central nodes in the information network) [WL12], (ii) *semantic features* (*e.g.*, humans might prefer to navigate between semantically similar articles) [SNSH13], and (iii) *visual features* (*e.g.*, humans might prefer links on the top of the screen) [DSLS16]. For example, in Figure 4.1, given the Wikipedia article “Manchester” and the outgoing links to the articles “England” and “Association Football”, we are interested in understanding which link features (*e.g.*, the visual position of the link) can better predict the actual transitions (popularity) of a link.

Methodologically, we start by gaining visual and descriptive insights into the success of links before we statistically model corresponding effects using mixed-effects hurdle models [PB06, HPN11] allowing us to account for the heterogeneity in the data (*e.g.*, links in the article “USA” might in general be more frequently used compared to links in the article “Manchester”). Subsequently, we integrate obtained insights into existing stochastic models of human navigation. To that end, we utilize the first-order Markov chain model—probably the most popular model for this task—in two separate analyses. First, we craft hypotheses about human navigation based on our insights and integrate these into a Bayesian inference process as priors; *cf.* [SHHS15]. Based on Bayesian model selection, this allows us to identify whether given hypotheses can improve the Markov chain model fit compared to an uninformed model; additionally, we can obtain a ranking of hypotheses allowing us to confirm our statistical results within a coherent research approach. Second, we aim at further utilizing these hypotheses to advance the classic well-known PageRank algorithm in a weighted variation relaxing the basic assumption of an uninformed random surfer; we evaluate the ranking of articles against actual page views.

Contributions and findings. The main contributions of this chapter are threefold: (i) We provide empirical evidence that Wikipedia users have a preference of choosing links leading to the periphery of underlying topological link network, that they prefer links leading to semantically similar articles, and that links positioned at the top and left-side of an article have a higher likelihood of being used. (ii) We integrate this evidence with first-order Markov chain models demonstrating an improvement of respective model fits by incorporating the assumed behavior of human navigation into the inference process (Section 4.5.1). (iii) Finally, we demonstrate the utility of these findings by adapting the well-known PageRank algorithm to better reflect human navigation behavior. Our enhancements lead to a significant improvement over the uninformed baseline algorithm that assumes random navigation by evaluating obtained ranking against actual article page views on Wikipedia (Section 4.5.2). The methodological framework provided in this work can also easily be applied to other transitional click data apart

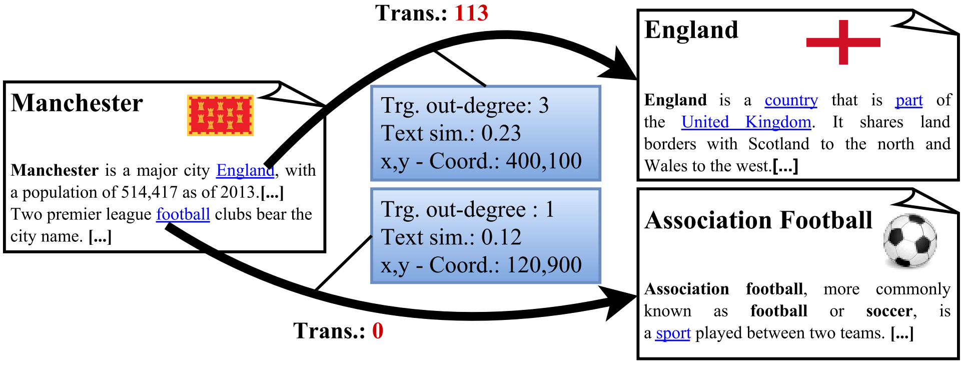

Trans.: 113
England
England is a country that is part of the United Kingdom. It shares land borders with Scotland to the north and Wales to the west.[...]
Manchester
Manchester is a major city England, with a population of 514,417 as of 2013.[...] Two premier league football clubs bear the city name. [...]
Trg. out-degree: 3
Text sim.: 0.23
x,y - Coord.: 400,100
Trg. out-degree : 1
Text sim.: 0.12
x,y - Coord.: 120,900
Association Football
Association football, more commonly known as football or soccer, is a sport played between two teams. [...]
Trans.: 0

Figure 4.1: Example. Wikipedia pages are connected by links for which we can compute a variety of features that are categorized in *network features* (*e.g.*, target article's out-degree), *semantic features* (*e.g.*, text similarity between source and target articles), and *visual features* (*e.g.*, position of the link on the screen). This work aims at understanding what makes a link successful—*i.e.*, which link properties best explain observed numbers of user transitions (shown on edges).

from Wikipedia.

4.2 Description of Data

This section describes the utilized datasets and extracted features. We make an implementation of the data extraction process as well as a sample of the data publicly available online[1].

4.2.1 Wikipedia link and transition data

Wikipedia data. In this work, we focus on all articles contained in the main namespace of the English Wikipedia as extracted from the public XML dump from March, 2015[2]. To obtain authentically rendered pages, we retrieved the corresponding static HTML pages by using Wikimedia's public API [3]. In contrast to using readily available link dumps, this allowed also for considering links that are indirectly included in a page, *e.g.*, by templates. A tiny part ($< 0.01\%$) of articles could not be retrieved and had to be excluded from the analysis. With this data, we created the Wikipedia link network D_{wiki} using articles as nodes and unique links as directed edges; D_{wiki} contains ~4.8 million articles connected by ~340 million distinct links.

[1] `https://github.com/trovdimi/wikilinks`

[2] `https://archive.org/details/enwiki-20150304`

[3] `https://www.mediawiki.org/wiki/API:Main_page`

Transition data. For measuring actual usage of links, we utilize openly available transition data from Wikipedia from February 2015 [WT]. It contains aggregated page requests extracted from the server log for the English desktop version of Wikipedia in the form of *(referrer, resource)* pairs, *i.e.*, transitions, and their respective transition counts. The data has already been pre-processed to filter bots and web crawlers and transitions occurring less than 10 times, see [WT] for details. Since in this chapter we are only interested in studying *internal* navigation on Wikipedia, we excluded article requests from outside of Wikipedia (*e.g.*, users arriving directly from search engines) and focus only on transitions corresponding to an internal link in our network D_{wiki}. These amount for about 1 billion or 31% of all page requests and cover ~13.6 million distinct links between ~1.4 million source and ~2.1 million target articles. With this information we can weight edges in D_{wiki} by their transition counts to obtain an edge-weighted network D_{trans_w}. Additionally, we denote the sub-network of D_{wiki} that contains only edges (links) that we observe in the transition data—*i.e.*, those links having a weight larger than 0 in D_{trans_w}—as D_{trans}.

4.2.2 Link features

For studying link success, we focus on three types of link features: network features, semantic similarity features, and visual features.

Network features. Here, we capture the centrality of the source (*src*) and target articles (*trg*) of a specific link in the Wikipedia link network D_{wiki} considering the *in-degree*, the *out-degree*, the overall *degree*, the *pagerank* [PBMW99] and the *k-core* [BH03] measure.

Semantic similarity features. Here, we detect the semantic similarity between the source and target article of a specific link. First, we computed a *text similarity* score by utilizing a sub-linear scaled tf-idf weighted vector space model capturing word tokens of Wikipedia articles [SHHS15]. For dimensionality reduction, we used sparse random projection retaining the pairwise distance between concepts with some small error. The text similarity between two pages was then determined by the cosine similarity of corresponding vectors. We also computed a *topic similarity* score with the cosine similarity between Wikipedia categories assigned to source and target.

Visual features. The third group of features is based on the placement of links within the source article, *cf.* also [DSLS16]. For that purpose, we rendered the HTML of each articles on a screen for a resolution of 1920 × 1080, a common resolution for desktop users. This enables deriving the exact screen position of links in terms of their *x/y-coordinates*. Based on visual position, CSS classes, and HTML tags, we additionally assigned each link to one of the following visual regions:

(1) *lead*: all links in the first section of the article excluding infobox, (2) *body*: all other links in the main text, (3) *left-body*: links in the body that are displayed in the 10% most left part of the screen, (4) *right-body*: links in the remaining right part of the body, (5) *infobox*: all links in the infobox, (6) *navbox*: all links in a table in the last section of the article placed by editors for facilitating navigation. As links can occur multiple times on an article, we derived the visual position of a link based on its first occurrence on the screen. This goes hand-in-hand with our empirical results elaborated in Section 4.4.2. A more detailed discussion about this facet of our data is provided in Section 4.7.

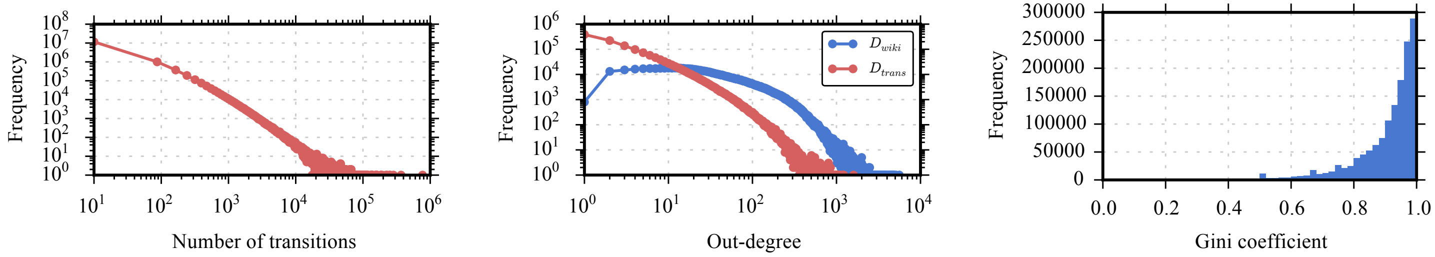
4.3 Focus of Attention

We start our analysis by investigating the focus of attention with respect to link usage in Wikipedia, *i.e.*, we study how strongly user transitions are concentrated on the few most visited links.

Distribution of transition counts. As a first key statistic, we observe that of ~340 million links in D_{wiki} only about 13.6 million links (or about 4% of all links) are regularly used, *i.e.*, are visited at least 10 times within a month. However, even when considering only these 4% regularly used links, we can observe that users focus heavily on a comparatively small set of links: about 50% of all transitions stem from only about $600,000$ links. A frequency plot of the distribution, see Figure 4.2(a), reveals a long-tailed distribution, where most of the links receive very low amounts of transitions.

Out-degree distribution. A similar effect can be observed at the level of individual articles. We compare the out-degree distributions of nodes in D_{wiki} and D_{trans}, *i.e.*, we compare the number of available links on articles in D_{wiki} with the number of regularly used links in D_{trans}. For a fair comparison, we consider only the out-degrees of nodes that occur in both networks. The results in Figure 4.2(b) reveal clear differences. The out-degree distribution of D_{wiki} reaches a maximum frequency for articles having eleven out-links indicating a decent amount of links to choose from. However, the out-degree distribution of D_{trans} clearly reveals that for most articles, users only visit a few articles regularly with the maximum frequency of only a single link. Formal comparison of both distributions with respect to (truncated) power-law, exponential, and log-normal distributions by their goodness of fit reveals that for D_{trans} the out-degree distribution resembles a truncated power-law most closely, while for D_{wiki} a log-normal distribution is preferred.

Gini coefficient of transition counts. The previous analyses have suggested a discrepancy between the link network D_{wiki} and actually transitioned links in D_{trans} and D_{trans_w}. The *Gini coefficient* is a measure of statistical dispersion allowing us to quantify the inequality of link popularity



(a) Transitions counts

(b) Out-degrees

(c) Gini coefficients

Figure 4.2: Focus of attention. This figure illustrates the focus of users' attention towards specific links. In (a), we see the frequency distribution of link transition counts. The long tail distribution shows that majority of links have low popularity. Figure (b) compares the out-degree distributions of the Wikipedia link network D_{wiki} (blue) and the transition network D_{trans} (red) for nodes contained in both networks. We can observe that only a few links are visited regularly on articles (peak at out-degree of one in D_{trans}) while more links are available (peak at out-degree of eleven in D_{wiki}); Figure (c) shows the histogram of Gini coefficients calculated on article vectors of link transition counts revealing clear inequality of link popularity within articles.

within articles. We calculate the Gini coefficient for each article based on a vector containing the transition counts for all links (including zeros) in this article from D_{trans_w}. A Gini coefficient of 0 indicates complete equality, while a result of 1 would mean that only a single link gets the whole attention. We visualize the histogram of all coefficients for each article in Figure 4.2(c); results indicate that there is a clear inequality of link popularity within articles, *i.e.*, attention focuses on just a few links.

Summary. In this section, we have presented empirical evidence that on Wikipedia user attention focuses on a small set of links while navigating. This hints towards a clear *preference towards a few links* and demonstrates the importance of finding indicators of link success, on which we focus in the remainder of this chapter.

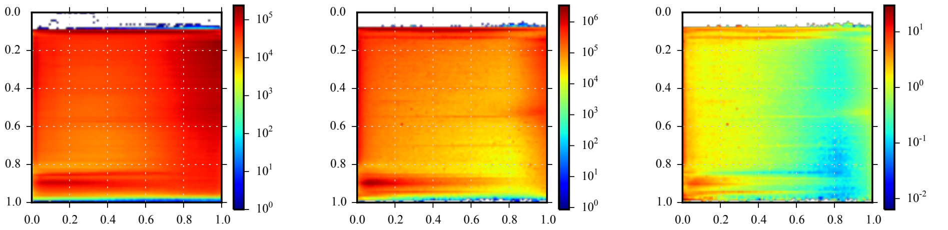
4.4 What makes a link successful?

Motivated by the discrepancy between link presence and link usage, we now study link success. That is, as illustrated in Figure 4.1, given a Wikipedia article, we aim at identifying effects of link features (network features, semantic features, or visual features) on counts of transitions to the respective target article. This allows us to answer questions such as: "Are links leading to peripheral articles more popular and thus, more successful?" We propose a methodology in Section 4.4.1 and discuss results in Section 4.4.2.

4.4.1 Methodology

For studying effects of link features on the success of links, we employ a visual analysis, descriptive statistics and *hurdle mixed-effects regression models*; we outline these approaches next.

Visual analysis. For empirical insights into visual link and click positions, we adopt heatmaps that are calculated by dividing the screen into 100×100 equally sized bins and then counting the number of times (i) a link exists and (ii) a link is clicked in the respective bin. In order to normalize screen width and height, we divide the x coordinate of a link by the screen width (1920) and the y coordinate of a link by the length of the respective page. We ignore links from the HTML that are not visible due to CSS styling. When multiple links with the same target exist on a page, we divide the actual click count by the number of link occurrences on the page as we do not know which of the links people actually clicked. This mimics the random surfer model as it assigns each link an equal amount of attention. For (iii) studying which regions produce more or less clicks per link, we re-use the heatmaps for clicks and links and perform an element-wise division of the corresponding bin counts. The counts in all bins of the three heatmaps are logarithmically scaled for the color mapping (*cf.* Figure 4.3).



(a) Links heatmap

(b) Clicks heatmap

(c) Clicks/Links heatmap

Figure 4.3: Heatmaps. (a) shows links positions, (b) clicks positions, and (c) clicks/links. The links heatmap (a) indicates high link density in the lead, the infobox/sidebar, and the navbox regions. The clicks heatmap (b) shows the regions with high click frequency—the lead, the infobox/sidebar, navbox and left body. The clicks/links heatmap (c) highlights the preference of users clicking on the left side of the screen, exceeding expectations implied by the presence of links. All heatmaps are logarithmically scaled.

Figure 4.3(a) shows the position of links on the screen. When comparing the results to the visual structure of a typical Wikipedia page as shown in Figure 1.3, we can identify four main concentrations of links on the screen: (i) the lead section on the top, (ii) the section on the right hand side containing infoboxes and sidebars, and (iii) the bottom area mostly containing navboxes (specifically, the left part). In comparison, the main body shows lower density with exception of (iv) the outer left part which might be explained by the multiple presence of lists with many links in Wikipedia articles. Figure 4.3(b) shows the location of clicks on the screen. People seem to prefer to click on those links that are located (i) at the top of the screen in the lead section, (ii) on the right sidebar (focus on infoboxes), (iii) in the bottom navboxes (focus on left side), as well as (iv) the left side of the body of a page. Not surprisingly, these patterns are overall similar to those observed in Figure 4.3(a) since users can only click on links that exist. Yet, some differences can be detected such as a general disfavor of regions located at the right hand side of the screen. To further investigate these differences, Figure 4.3(c) displays the number of clicks per links in a region. Here, hot colors indicate regions with high numbers of clicks per link, cool colors the opposite. The results confirm our intuition that people prefer to click on links in the left side of the screen, and this preference exceeds what one would expect from the link count in that area. By contrast, the links in the sidebar on the right hand side (infoboxes) are less often followed as the pure number of links there would suggest. Please note that the heatmaps are created using the complete dataset.

Descriptive analysis. We start our descriptive analysis with an overview on the effects of our features on transition counts. For an intuitive picture, we investigate boolean expressions on our features and determine for each feature the number of links, the number of transitions, and the average number of transitions per link. We focus on easy-to-interpret expressions: for network features, we compare the centrality of the target node with the centrality of the source node. For semantic similarity features, we compare the similarity between source and target article of the link with the median of the respective values for all links within the same source article. Finally, for the visual features, we divide the screen position in three areas for each dimension, that is the left, the middle, and the right third of the screen, and links in the top-half, the bottom-half and links that are only visible after vertical scrolling (assuming a resolution of 1920 × 1080) respectively. Furthermore, we use the position of the link in a specific part of the article such as the lead or the infobox.

Mixed-effects hurdle model. Additionally, we aim at statistically modeling the effects of all link features on link success given a certain start node. Our setting (*cf.* Figure 4.1) implies that individual links represent data points with the transition counts being the outcome to predict, and the respective link features the effects at interest. Yet, links in the dataset are not independent of each other, but rather nested within Wikipedia articles. Heterogeneous effects might be present as *e.g.*, apart from generally different amounts of total transitions, the article of "Manchester" might pose different navigational behavior than the article of "USA". Pooling all links into a single dataset for ordinary regression can lead to biased results; consequently we resort to *mixed-effects models* (also known as multilevel models or hierarchical models), allowing us to account for the nested structure—specifically the resulting random variations—between links (level 1) and Wikipedia pages (level 2). We define our mixed-effects model following [HMvdS10]; for a broader introduction we point to [PB06]:

$$Y_{i,j} = \gamma_{0,0} + \gamma_{1,0}X_{1,i,j} + u_{0,j} + u_{1,j}X_{1,i,j} + e_{i,j} \tag{4.1}$$

Here, $Y_{i,j}$ refers to the transition count of a link i (level 1) on a Wikipedia page j (level 2); $\gamma_{0,0}$ refers to the overall intercept and $\gamma_{1,0}$ to the overall regression coefficient for the level 1 predictor $X_{1,i,j}$, *e.g.*, the degree of the target node. These are the fixed effects. The remainder refers to the random effects: $u_{0,j}$ is the error component of the intercept allowing the group intercepts to differ, $u_{1,j}$ refers to the error component of the slope allowing the group slopes to differ, and $e_{i,j}$ is the classic random error of predictions.

To study the effects of network, semantic similarity and visual features, we model each feature separately. As shown in Section 4.3, only 4% of all links have been used at least 10 times, resulting in a large amount of zeros in the data outcomes. To account for that, we employ *hurdle models* [HPN11] in a two-step approach. (i) First, we model the binary data in D_{trans};

i.e., the link values are set to 1 iff at least 10 transitions for these links have been observed, and to 0 otherwise. We then model this data with a mixed effect binomial logistic regression and make inference on the fit. (ii) In a second step, we only take the data from D_{trans_w} capturing those links that have been clicked and their respective transition counts. As our outcome variable is generated by count data, basic assumptions for *linear* mixed-effects models are violated. Consequently, we resort to zero-truncated negative binomial mixed-effect models that do not allow the outcome to be zero (as all outcome counts are at least 10). The choice of model also enables modeling the presence of overdispersion in the data. For comparison, we also fitted Poisson mixed-effects models with an observation-level random effect that allows for overdispersion [Har14] but could not find core differences in inference. Thus, we overall build two models for each feature: one to model *if* a link has been used at least 10 times in a month, and a second to model the transition *counts* for these regularly used links.

We fit our models using the "Template Model Builder" framework [KNB+15] that evaluates and maximizes the Laplace approximation of the marginal likelihood utilizing automatic differentiation of the joint likelihood; we use the R package "glmmTMB"[4]. For improving convergence, we scale non-binary features by subtracting the mean and dividing by their standard deviation. For judging significance of individual effects, we follow an incremental model comparison approach [BMB+14]. That is, we compare the model including the effect of interest with a restricted model that neglects the effect using a *likelihood ratio chi-square test.* Results could be confirmed by model comparison via AIC and BIC.

4.4.2 Results

For computational reasons, we applied the methodology to a random sample of 10,000 articles containing at least one outgoing link in the transition data D_{trans} considering all links and their features. The sample contains 1,028,704 links with 6,686,581 transitions. We made sure that the sample captures similar power-law degree distributions as the overall dataset (*cf.* Section 4.3); multiple sampling iterations did not change the results. For reproducibility reasons, we make the full sample available[5]. We present the descriptive results in Table 4.1 and the mixed-effects model results in Table 4.2. For the statistical models, we only report fixed effects at interest but make the full results available in additional material[6]. We discuss results for our three major feature groups next.

Network features. The results in Table 4.1 highlight that there appears to be a preference of users to choose links leading to target Wikipedia arti-

[4] `https://github.com/glmmTMB/glmmTMB`

[5] `https://github.com/trovdimi/wikilinks/raw/master/sample.csv.gz`

[6] `https://github.com/trovdimi/wikilinks/tree/master/notebooks`

cles that have a smaller in-degree, out-degree and degree compared to the source article. This is specifically imminent when controlling for the presence of links as the last column indicates. Similar behavior can be seen for the k-core and pagerank of nodes. Our mixed-effects model results shown

Table 4.1: Descriptive results. This table provides a descriptive overview on link success; it shows for each boolean feature the number of links, the overall number of transitions these links accumulated, and the average transition count per link. Results show that links leading to the periphery of the network, to semantically similar articles, and links positioned in the lead and left area of an article accumulate relatively more transitions than opposite links.

Feature	# Links	# Transitions	Mean Trans. per Link
trg_degree < src_degree	462,147	3,944,747	8.54
trg_degree ≥ src_degree	566,557	2,741,834	4.84
trg_in_degree < src_in_degree	412,653	3,831,100	9.28
trg_in_degree ≥ src_in_degree	616,051	2,855,481	4.64
trg_out_degree < src_out_degree	545,980	3,912,604	7.17
trg_out_degree ≥ src_out_degree	482,724	2,773,977	5.75
trg_kcore < src_kcore	283,668	3,477,828	12.26
trg_kcore ≥ src_kcore	745,036	3,208,753	4.31
trg_page_rank < src_page_rank	414,124	3,833,596	9.26
trg_page_rank ≥ src_page_rank	614,580	2,852,985	4.64
text_sim > median of article	511,871	4,335,445	8.47
text_sim ≤ median of article	516,833	2,351,136	4.55
topic_sim > median of article	392,955	3,246,766	8.26
topic_sim ≤ median of article	635,749	3,439,815	5.41
position = lead	93,546	2,489,342	26.61
position = body	373,407	3,442,017	9.22
position = left-body	64,320	934,766	14.53
position = right-body	309,087	2,507,251	8.11
position = navbox	502,079	99,498	0.2
position = infobox	59,672	655,724	10.99
screen_x_coord: left third	211,549	2,358,899	11.15
screen_x_coord: middle third	291,260	1,422,183	4.88
screen_x_coord: right third	525,847	2,905,499	5.53
screen_y_coord: top half	221,980	3,991,912	17.98
screen_y_coord: bottom half	174,500	998,287	5.72
screen_y_coord: scroll needed	632,176	1,696,382	2.68
Overall	1,028,704	6,686,581	6.5

Table 4.2: Mixed-effects hurdle modeling results. This table presents the results for the mixed-effects hurdle models based on a basic model specification of: transition$_{i,j} = \gamma_{0,0} + \gamma_{1,0}$feature$_{1,i,j} + u_{0,j} + u_{1,j}$ feature$_{1,i,j} + e_{i,j}$. For each feature (first column) and a given transformation (second column), results for both parts of the hurdle model are reported. The binomial models the binary outcome whether a link has been used at all, and the zero-truncated negative-binomial models the transition counts of regularly used links. For each model, we report the fixed effect coefficient of the fitted model, the likelihood ratio test statistic compared to a reduced nested model without the fixed effect, and the significance according to χ^2-test. Overall, the results confirm the descriptive results of Table 4.1 suggesting a preference of links leading to the periphery of the network, links targeting semantically similar articles, and links being positioned at the beginning and left-side of articles.

		Binomial Model			Zero-truncated NB Model		
Feature	Transformation	Fixed effect	LRT	Pr(>Chi)	Fixed effect	LRT	Pr(>Chi)
trg_degree	scale	-9.0704	9441	< 1.0 e-300 ***	-0.1353	174	7.98 e-40 ***
trg_in_degree	scale	-9.2925	9456	< 1.0 e-300 ***	-0.1341	185	4.2 e-42 ***
trg_out_degree	scale	-0.6425	4052	< 1.0 e-300 ***	-0.0375	32	1.6 e-08 ***
trg_kcore	scale	-9.4873	9470	< 1.0 e-300 ***	-0.1296	180	4.5 e-41 ***
trg_pagerank	scale	-8.8103	9384	< 1.0 e-300 ***	-0.1457	196	1.7 e-44 ***
text_sim	scale	0.2944	617	3.4 e-136 ***	0.238	1179	1.9 e-258 ***
topic_sim	scale	0.2052	482	8.6 e-107 ***	0.1212	401	3.7 e-89 ***
position = lead	none	1.9682	6146	< 1.0 e-300 ***	0.2858	498	2.7 e-110 ***
position = body	none	0.2351	93	5.9 e-22 ***	-0.4581	1478	< 1.0 e-300 ***
position = left-body	none	1.106	739	9.4 e-163 ***	-0.1671	123	1.7 e-28 ***
position = right-body	none	-0.3631	231	4.4 e-52 ***	-0.4389	1236	1.0 e-270 ***
position = infobox	none	-0.1955	25	7.2 e-07 ***	0.1365	37	1.4 e-09 ***
position = navbox	none	-6.3237	9226	< 1.0 e-300 ***	-0.9963	688	1.5 e-151 ***
screen_x_coord	scale	-0.2974	1081	4.4 e-237 ***	0.0225	16	6.0 e-05 ***
screen_y_coord	scale	-4.258	10588	< 1.0 e-300 ***	-0.4004	419	3.987 e-93 ***

in Table 4.2 statistically confirm this observation. The coefficients for both parts of the hurdle model (binomial and zero-truncated negative binomial models) show negative signs for all network features at hand meaning that the higher those features are (*i.e.*, the more central the target nodes are) the less likely it is that those links are clicked. Interestingly, the effect sizes and model fits (LRT) are quite similar for all features except the out-degree which can be explained by the fact that in general the out-degree is not a good indicator regarding the centrality of a node in the network as it just depends on the number of links present on a given page. Overall, the results suggest a preference for users to navigate to peripheral nodes in the network in contrast to navigating towards the core of the network. A potential explanation of this behavior could be that pages located in the core of the network are more general, whereas pages located in the periphery of the network are more specific and thus possibly more interesting to the user on average. Moreover, more general Wikipedia pages are likely more often returned as results to search engine queries, making them the entry point to Wikipedia (*cf.* Chapter 3). Then, with the next click within Wikipedia, the user narrows down her information need. Such a behavior—looking up specific facts or information—would be in-line with a common use of encyclopedias in general.

Semantic similarity features. The results in Table 4.1 and Table 4.2 indicate that links connecting two semantically related Wikipedia articles—both for text and topic similarity—are more likely to be transitioned compared to those links that indicate less semantic similarity. This is imminent from the higher ratio of transitions to high similarity target articles, and specifically, by the positive mixed-effects model coefficients for the hurdle models at hand. Overall, the results suggest a preference towards navigating semantically similar articles supporting previous research [WL12, SNSH13, SHHS15].

Visual features. The descriptive results in Table 4.1 indicate user preference towards choosing links on the left side of the screen and specifically links that are at the beginning of a Wikipedia article. The model results in Table 4.2 confirm these observations. The positional lead feature has strongly positive coefficients in both the binomial and zero-truncated negative binomial model indicating that links in the lead are more frequently used. Additionally, we can see that the right-body feature (links in the right 90% body region) indicates negative coefficients while the left-body feature (links in the left 10% body region) has a positive coefficient for the binomial model. The models for the infobox are contradicting each other, but overall show much less model improvements over the baseline model compared to our other visual feature models (lower LRT). Furthermore, we can clearly see that the navbox feature has a strong negative coefficient suggesting that links positioned at the bottom of the page in the navigation boxes are only seldomly visited. The coordinate features summarize the more fine-grained results in a broader level indicating that there is a preference towards choosing links that are at the beginning and also on the left side of the screen. The results from the descriptive analysis and statistical modeling are by and large in-line with the visual analysis and confirm previous findings on the F-shaped way of information consumption [DSLS16, Nie06, BCM09].

Summary. Our results suggest three indicators having an effect on the success of links on Wikipedia. First, humans appear to have a *preference towards choosing links leading to the periphery* of the underlying topological link network. Second, *links connecting semantically related articles tend to be clicked more frequently* than those connecting semantically unrelated articles. Third, the visual position of a link appears to have an impact of the link's success. Links that are located on the top of the screen (*i.e.*, lead of an article) and on the left of the screen are preferred by humans navigating Wikipedia suggesting a *visual top-left preference.*

4.5 Integration in Markov Models

While the results presented in Section 4.4 provide evidence of factors effecting the success of links on Wikipedia, it is an open question how we can integrate these findings into existing models of human navigation. To that end, we investigate the influence of link features on models of human navigation in this section. We focus on first-order Markov chains as the most wide-spread models for this task [BL00, BP98, SHTS14]. First-order Markov chains are stochastic systems that assume a memoryless generative process, *i.e.*, the probability of the next state (Wikipedia article) depends only on the current state. We utilize these models in two separate analyses. First, in Section 4.5.1, we craft hypotheses about human navigation based on our insights of Section 4.4.2 and incorporate these into a Bayesian inference process. This allows us to (a) identify whether given hypotheses can better explain observed transition behavior compared to a uniform structural baseline hypothesis and thus, potentially improve the Markov chain model fit, and (b) the model comparison provides an additional ranking of given hypotheses helping in confirming our previous results within a coherent research approach. Second, in Section 4.5.2, we aim at utilizing these insights to improve the classic PageRank algorithm in a weighted variation and validate its precision against the observed visit counts of Wikipedia articles.

4.5.1 Bayesian integration

Next, we explore the impact of integrating our findings in Markov chain navigation models; for that task we resort to a recently proposed Bayesian approach called HypTrails [SHHS15].

Methodology. We aim to integrate hypotheses about navigational behavior, which we form based on our results obtained in Section 4.4.2, as prior knowledge into the inference process for Markov chains. HypTrails [SHHS15] provides a coherent framework for this task. First, one can express a hypothesis, *i.e.*, beliefs in transition probabilities of a Markov chain, as a matrix $M = (m_{ij})$, where a higher value expresses higher belief in a transition between articles i and j. Then, these hypotheses matrices are incorporated as Dirichlet priors into the Bayesian model inference procedure. Due to the sensitivity of a Bayesian model to the priors on its parameters, more plausible hypotheses—*i.e.*, hypotheses that are in-line with the observed data—induce higher marginal likelihood (evidence) values used for model comparison. Marginal likelihoods can be compared along different values of a parameter κ expressing the strength of belief in a hypothesis. Higher values reflect a higher belief in given hypothesis; *i.e.*, a stronger probability mass concentration of the Dirichlet prior. Formally, model comparison is done by calculating a *Bayes factor* [KR95], which is the ratio of the marginal likelihoods for two hypotheses H_1 and H_2; if positive, the first hypothesis is judged as more plausible—strength of Bayes factors can be checked in an interpretation table of Kass and Raftery [KR95].

Comparing the marginal likelihood of a specific hypothesis (*e.g.*, that humans prefer to navigate to the periphery of the network) with a structural baseline hypothesis (*i.e.*, random preference of out-links) allows us to detect whether integrating given hypothesis into the inference process enhances the general model fit. Additionally, we can compare hypotheses with each other to obtain a ranking of hypotheses. For more methodological details, we point the reader to [SHHS15]. We describe how we construct hypotheses for our data next.

Hypotheses. We formalize and compare several hypotheses based on our main findings of Section 4.4 providing evidence of feature effects on link success. As mentioned, hypotheses are expressed as matrices $M = (m_{ij})$ indicating belief in transition probabilities. For finding good formalizations of these hypotheses, we experimented with different scaling variations such as logarithmic, square-root, and exponential feature scaling, but only report on the best formalization for each of the three feature groups due to limited space.

As a baseline, we use the *structural hypothesis* expressing the belief that someone navigating in the Wikipedia network chooses any out-link of an article randomly. Here, we set $m_{ij}^{structure} = 1$ if there is a link from article i to article j and $m_{ij}^{structure} = 0$ otherwise. If other hypotheses do not reveal higher marginal likelihoods than this baseline hypothesis (*i.e.*, positive Bayes factors), they indicate no improvements to the model fit—*i.e.*, they do not capture the actual human navigation behavior well.

For the network features, we focus on the *k-core* having shown promising results in Section 4.4. Since we want to formulate a hypothesis that expresses higher beliefs in transitions to articles with lower k-cores (*i.e.*, the periphery of the network), we set the entries of the hypothesis matrix to $m_{ij}^{kcore} = \frac{1}{\sqrt{trg\_kcore}}$. Regarding semantic similarity, we use the *text_sim* feature as it is, *i.e.*, $m_{ij}^{text\_sim} = text\_sim$. This reflects a belief that transitions to similar articles (w.r.t. to the text similarity) are more likely. For visual features, we concentrate on the *position* feature. Our visual hypothesis states that links that are in the lead, in the most left part of the body, or in the infobox are more likely to be visited than the others (as the results of Section 4.4 suggest). Thus, we express this hypothesis as $m_{ij}^{visual} = 1$ in these cases, and $m_{ij}^{visual} = 0$ otherwise. For these hypotheses, we add the structural baseline hypothesis for smoothing in order to guarantee a minimum belief in all transitions to linked articles.

In addition to these individual hypotheses, we also include combinations of them in our comparison. In this direction, we form the hypotheses *kcore+text_sim*, *kcore+visual*, *text_sim+visual*, and *kcore+text_sim+visual* by adding the respective matrices element-wise to each other—here, no smoothing is necessary.

Results. We report the resulting Bayes factors for the comparison between expressed hypotheses and the structural baseline hypothesis (for varying hypotheses weighting factors) in Figure 4.4; for a detailed description of this figure, please resort to the figure caption. The primary observation is that all investigated hypotheses improve the fit of the Markov chain model in comparison to the structural baseline hypothesis as imminent from the Bayes factors for all values of κ. Complementary hypotheses (*e.g.*, navigational preference towards the core of the network) result in negative Bayes factors falling below the baseline (not shown here). The relative ranking of hypotheses also provides interesting insights. Comparing the single hypotheses (*text_sim*, *kcore* and *visual*) reveals that the *visual* hypothesis appears to be the most relevant, followed by the *kcore* and then the *text_sim* hypothesis. Additionally, by combining these hypotheses, we can further improve the model. Overall, a combination of all three hypotheses has the highest evidence for higher values of κ. Yet, the combination of belief of peripheral navigation and preference towards choosing links in the visual top and left position already provides a similarly large improvement.

Summary. By and large, the results of this section have successfully demonstrated that we can improve models of human navigation on Wikipedia—*i.e.*, the Markov chain model—by integrating hypotheses (beliefs) about human navigational behavior into the inference process of the models. In this case, we have utilized Bayesian inference to incorporate expressed hypotheses based on our insights of Section 4.4 as priors. All hypotheses improve the model fit. Additionally, a ranking of hypotheses reveals the relative

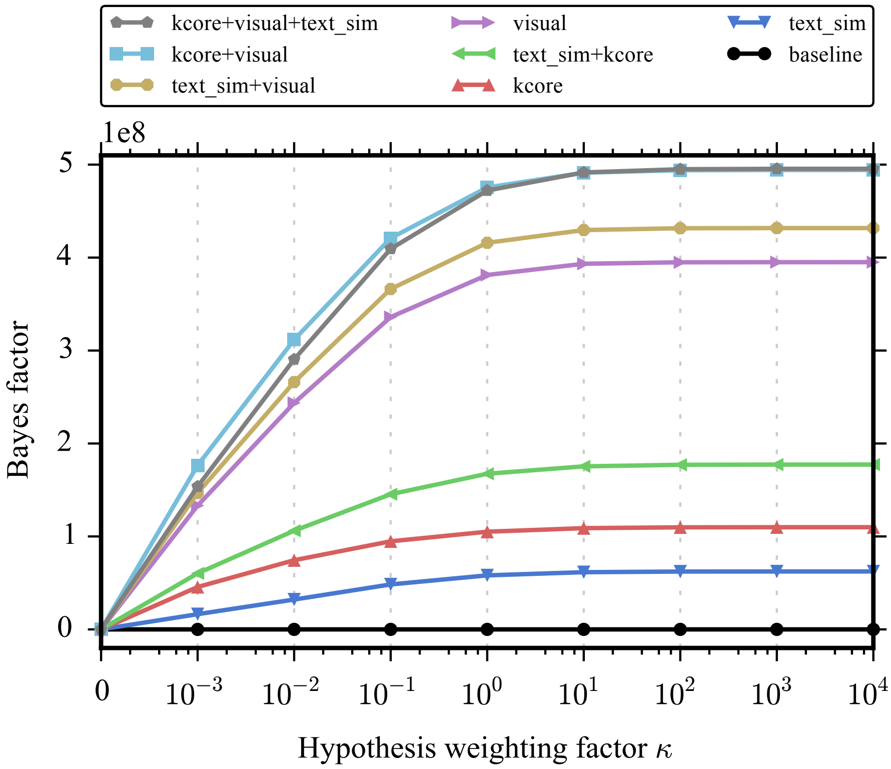


Figure 4.4: Bayesian integration results. This figure reports the results for our experiments on integrating navigational hypotheses as Bayesian priors in Markov chain models. The x-axis depicts different hypothesis weighting factors κ—higher values reflect stronger belief in a given hypothesis. The y-axis denotes the Bayes factor for a given hypothesis (different lines); the Bayes factor compares the fit of the respective hypothesis model to the fit of the baseline, *i.e.*, the structural hypothesis. Higher Bayes factors reflect higher plausibility of given hypothesis; all Bayes factors in this plot are strongly decisive. All hypotheses improve the fit of the Markov chain model when integrated as prior assumptions. Hypotheses labels are in order of their ranking. A combination of visual and network features leads to the strongest improvement.

plausibility and combinations show further improvements. Next, we want to further make use of these insights to improve the well-known PageRank algorithm utilizing the Markovian framework.

4.5.2 PageRank integration

In this section, we investigate if our insights can be used for improving the well-known PageRank algorithm for Wikipedia.

Weighted PageRank. The PageRank algorithm [PBMW99] is one of the most popular Web algorithms. It computes a centrality measure—pagerank—for each node in the network, such that a node receives a high value if it has many incoming links from other important nodes. This can be interpreted as the probability of a random surfer in the network to land on respective page—*i.e.*, the stationary distribution of a Markov chain model. While for the classic PageRank algorithm, weights in the network are propagated equally through all links, we propagate weights according to the eight hypotheses that we formalized in Section 4.5.1 based on our main findings. That is, we model the landing probabilities of a random surfer that chooses the next node in the network proportional to the entries in the respective hypothesis matrix $M = (m_{ij})$. This weighted version of PageRank implements a *reasonable surfer model* [DAB10], *i.e.*, a model where a random surfer does not choose links uniformly, but is influenced by link features such as the link position. Formally, we compute the weighted pagerank of a node as:

$$PR(j) = \frac{1-\alpha}{N} + \alpha \sum_{i \in \Gamma^-(j)} \frac{PR(i) \cdot m_{ij}}{Z_i}$$

where N is the number of nodes in the network, $\Gamma^-(j)$ is the set of nodes linking to the node with index j, α is the damping factor and $Z_i = \sum_{j' \in \Gamma^+(i)} m_{ij'}$ is a normalization factor for each matrix row i.

For evaluation, we determine the Spearman rank correlation of respective weighted PageRank with the sum of incoming transitions for a page from D_{trans_w}, *i.e.*, the number of views of an article stemming from internal navigation. As a baseline, we use the classic unweighted PageRank algorithm assuming random navigation.

Results. Table 4.3 shows the resulting Spearman correlation coefficient for the different evaluated hypotheses and the unweighted PageRank as a baseline for different damping factors α. We observe that most, but not all hypotheses achieve an improvement of the correlation between the aggregated transitions per target article and the pagerank values. In particular, the *kcore* and *visual* hypotheses lead to improvements. The best result can be obtained by combining these two hypotheses: the *kcore+visual* hypothesis achieves the best correlation score, which is about 0.1 better than the baseline. In contrast, and unexpected considering the previous results, the *text_sim* hypothesis is not able to increase the correlation, and also does not lead to improvements when combined with other hypotheses. This might be caused by the fact that PageRank assumes an infinite random walk while people navigate in-line with this hypothesis maybe only in certain phases of their Wikipedia visit, *e.g.*, at the end of their visit, *cf.* [WL12] and Section 4.7. Please note that the correlation increases with the increase of the damping factor. This is a natural behavior, since we base our analysis on Wikipedia internal transitions and higher damping factors reduce the chances of teleportation.

Apart from that, results are mostly in-line with the ranking that we obtained from the Bayesian integration of the hypothesis, *cf.* Section 4.5.1. All correlation coefficients reported in this chapter are strongly significant according a t-test with a null hypothesis that two datasets are uncorrelated. Additionally, we formally test whether the improvements of the weighted PageRank algorithms in terms of their correlation coefficients with view statistics are significant compared to the unweighted algorithm. To that end, we employ Steiger's one-tailed hypothesis test for assessing the difference between two paired correlations [Ste80]; the results reveal clear significance for all correlations when compared to the baseline.

Summary. The results presented in this section suggest that a PageRank model with an intelligent edge weighting reflecting the users' transition behavior explains the page views stemming from internal navigation better than the standard random surfer model. The reasonable surfer implemented through the edge weighting that performs best tend to select links that are leading *out of the network core to articles in the periphery and located in the top of the screen, on the left side of the screen and in the infoboxes.*

Table 4.3: Weighted PageRank results. Comparison of the hypotheses specified in Section 4.5.1 with respect to the Spearman correlation between the transitions (aggregated per target article) and the pagerank values of the corresponding articles in D_{wiki} for different damping factors. The correlation coefficient for the unweighted PageRank (marked italic) is used as a baseline. We see that most, but not all hypotheses achieve an improvement (marked bold); the best correlation with an improvement of ~ 0.1 across the three dumping factors is achieved by the hypothesis kcore+visual.

Damping factor α/Hypothesis M	0.80	0.85	0.90
baseline	*0.421*	*0.428*	*0.436*
kcore	**0.434**	**0.440**	**0.447**
visual	**0.507**	**0.516**	**0.526**
text_sim	0.400	0.407	0.415
text_sim+kcore	0.407	0.412	0.417
text_sim+visual	**0.489**	**0.500**	**0.513**
kcore+visual	**0.530**	**0.538**	**0.545**
kcore+visual+text_sim	**0.494**	**0.505**	**0.517**

4.6 Related Work

Since the inception of the Web, our research community has been interested in studying human navigational click data on the Web—*e.g.*, see [HPPL98]. In this line of research, a variety of models has been proposed including the well-known Markov chain model utilized in this work [SHTS14, PBMW99, PP99, SHHS15], or models such as decentralized search [HSGS13, DSHS15] motivated by small-world navigation [Kle00b].

Insights have been utilized to infer missing links [WPL15], to predict break-ups of the navigation process [SPWL14], for recommendations [WDY08], or to improve the link structure of a website [PWZL16]. For the latter, Paranjape *et al.* [PWZL16] highlighted the importance of improving hyperlink structures based on their usage due to the large amount of unused links. Consequently, they proposed an algorithm for suggesting useful links and estimate their success based on clickthrough rates.

Based on this wide range of studies aiming at understanding human Web navigation, a series of navigational regularities, patterns and strategies have been suggested. For example, West and Leskovec [WPP09a] found trade-offs between similarity and degree in navigational behavior suggesting different phases in user sessions, namely an exploration (orientation) and an exploitation (goal-seeking) phase [HSGS13]. Subsequent research has suggested that humans prefer to navigate between semantically similar websites [WL12, SNSH13, SHHS15], have preferences for choosing links at the beginning of pages [LLHS16, PWZL16, DSLS16], and that navigational patterns exhibit regularities with respect to underlying network characteristics [WL12, PWZL16, SHHS15, LLHS16].

While, as elaborated, a large amount of studies on human navigational behavior exist and a variety of navigational hypotheses have been proposed, there still has been a lack of a systematic understanding of what makes a link successful in information networks. To that end, we have extended previous work covered in this section by providing a study on effects of link properties on actual user transition behavior on the complete English Wikipedia utilizing large-scale data. Instead of providing a global view on clickthrough rates, we have also focused our attention on the clear setting of given a current Wikipedia article, which features best predict the popularity of links on that article. The features of interest have been motivated by previous work and our results broadly confirm previous hypotheses. Also, our analyses allow for more fine-grained insights such as navigational preference towards peripheral nodes or visual preferences on detailed screen coordinates. We have also demonstrated the importance of better understanding human navigational behavior and link popularity by successfully enriching existing Markov chain models and the PageRank algorithm with supplementary behavioral hypotheses.

Our work is also broadly connected to research studying click data in other contexts, *e.g.*, characterizing user behavior and sybil detection in online social networks [BRCA09, WKW$^+$13], improving search engine ranking functions [Joa02, XZC$^+$04, Joa03, ABD06, JHW07], and marketing and next purchase prediction [CHN03, BS09, BS03, MLSL04]. However, these do not cover the specifics of large-scale information networks such as Wikipedia.

4.7 Discussion

In this section, we discuss limitations and future work.

No session information. In this chapter, we have focused on aggregated click counts for link transitions. While this was perfectly suited for our task at hand, we cannot differentiate between potentially varying phases within navigation sessions as *e.g.*, zoom-in and zoom-out phases postulated in [WL12]. Also, we have only looked at *internal navigation* while observed behavior might differ when people navigate between different platforms (*e.g.*, Google to Wikipedia).

Multiple link occurrence. Although the Wikipedia editor guidelines discourage placing multiple occurrences of links to the same target article within one source article, this is sometimes ignored by editors. Additionally, links can be repeated in the infoboxes or overview tables for improved usability. Unfortunately, the data at hand does not disambiguate which link instance has been clicked if a link occurs multiple times. While this has no implications for the network and semantic features as they are independent of the link position, it influences the analysis of the visual features. Moreover, the utilized click data only provides an approximation of clicks. Future work could also contrast these studies by analyzing more fine-grained data such as eye-tracking or mouse movement studies. For our visual analysis, we have addressed the multiple links issue by assigning each link an equal amount of attention. However, more refined approaches are warranted in future work (*cf.* Chapter 5). We chose a specific screen resolution in this work for deriving and studying the visual link positions. Future work should extend this research to various resolutions for more detailed insights. Our empirical insights give an aggregated view without distinguishing between individual pages. In future work, we plan a more detailed study aiming at the impact of link positions on an individual page level. For the Bayesian and PageRank integration (*cf.* Section 4.5), we have decided to use the visual feature of the first occurrence of a link on the screen based on its x- and y-coordinates. This choice is justified by additional experiments on a restricted dataset filtering out links with multiple occurrences. As for the main data, a strong preference for links in the top and the left hand side of the page can be found, *cf.* also Section 4.4.2. The results for the restricted dataset confirm our main findings, but also reveal a bias: the popularity of links at the top of the page is less pronounced (but still substantial) in this variation. For further analysis, we included the number of times a link occurs on an article as a nuisance parameter to our regression model, but this did not change the main inference. Thus, we are confident that our main findings are robust with respect to that issue.

Different screen resolutions. In our analysis, we chose the specific screen resolution of 1920 × 1080 for deriving visual features. While this adheres to our data only containing click information from desktop users only, different users may have different visual perception of the same article. Since this may affect presented results, future work could set out to tackle this limitation by, *e.g.*, repeating the visual analysis for different screen resolutions. Another issue of special interest would be to complement the analysis using transitions data capturing the mobile user behavior on Wikipedia.

Definition of success. We defined success of a link as the click popularity derived from user transitions. However, other forms of link success can be studied. For example, instead of explaining transition preference of outgoing links, one could modify the research question of this study and look at incoming links. Then, the question of interest would be: Given a set of incoming links to a given article, which ones are more popular than others and how can we explain them? Also, in this work, we have not considered the narrow textual context of a link which is tightly related to the concept of information scent [Pir07]. Further research can concentrate on this aspect and study the relationship between diminishing information scent [Hea09] caused, *i.e.*, by editors and the success of a link. Apart from that, the success of a link could also be defined in light of how much a link contributes to the different types of navigation. One could analyze navigability of the networks D_{trans} and D_{trans_w} on the edge and node level, *i.e.*, by studying the flow in and out of a node or by removing nodes and edges. Link success could be also examined in terms of its ability to enable and encourage further navigation and exposing the user to a more richer content or to the strongly connected component of the network.

Markov chain hypotheses formulation. The results of Section 4.5 depend on how we code the hypotheses as priors. In this chapter, we handcrafted hypotheses that are in-line by our data-driven insights of Section 4.4. We acknowledge that more fine-grained engineering and learning of good priors and weightings might further improve results. Additionally, even though in different context and with different methodology, one might argue that some results of Section 4.5 could have been expected given our earlier findings. Yet, the main goal here was to demonstrate the utility of incorporating assumptions about transition behavior into existing navigational models (*i.e.*, Markov chain and PageRank) that predominantly, often have assumed uniform behavior. Additionally, obtained results helped us to further confirm our empirical insights of this chapter. Nonetheless, we hope that in the future more studies on additional datasets, *e.g.*, on other language editions of Wikipedia, are conducted in order to enrich our understanding of human navigation and link success.

Explaining and modeling heterogeneity. While we have accounted for basic heterogeneity by utilizing mixed-effects models and Markov chain models that both model transitions given a current start article, more fine-grained analyses in that direction are warranted in future. For example, there might be specific user groups, positions within sessions, or specific times that exhibit deviating navigation behavior. In this direction, a detailed analysis of the random effects in the mixed-effects models might lead to better understanding of heterogeneous effects. Additionally, pattern mining approaches could be employed to detect interpretable subgroups of Markovian transition behavior [LBS+16]. Obtained insights could then eventually be used to further advance existing models by integrating mixtures of navigational hypotheses into the modeling process [BLS+16].

4.8 Conclusions

In this chapter, we studied what makes a link successful on Wikipedia. To that end, (i) we investigated link features and their effects on link popularity utilizing visual and descriptive analyses supplemented by mixed-effects hurdle models. Results suggest that Wikipedia users prefer to navigate to articles that are in the periphery of the Wikipedia link network, to articles that are semantically similar, and to articles that are linked at the top or at the most left-hand side of the source article. (ii) Based on these findings, we integrated hypotheses about human navigational behavior into the well-known Markov chain model utilizing Bayesian inference. By doing so, we could improve respective model fits compared to a uniform baseline model. (iii) For further demonstrating the utility of our findings, we adapted the well-known PageRank algorithm that assumes random navigation by accounting for observed navigational preferences in a weighted variation. An evaluation of resulting pagerank values obtained by link weighting against actual view statistics revealed significant improvements over the unweighted baseline algorithm.

Our work is relevant for researchers interested in studying human navigational behavior and link structures on the Web. Obtained insights can be utilized for providing qualitative link structures and for supplementing existing algorithms utilizing these structures to better adhere to observed transition behavior. In addition to future work directions discussed in Section 4.7, we also aim at extending our studies to different language versions of Wikipedia and newer time frames for identifying similarities and differences. Also, apart from Wikipedia, the provided methodological framework can be applied to other platform data in a straight-forward manner.

Chapter 5

Link Disambiguation in Wikipedia: From Transitions Counts Between Articles to Click Probabilities of Individual Links

5.1 Introduction

Wikipedia is one of the richest sources of information on the Web. To ease and coordinate the collective production and curation of content, the Wikipedia community has established detailed guidelines in a manual of style. Specifically, a link between Wikipedia articles should only be placed if the target article helps to understand the source article while placing multiple hyperlinks to the same target is discouraged[1]. However, these guidelines are not always followed, and there exists a range of special cases allowing multiple links, such as for occurrences in infoboxes, overview tables, or for the first occurrence of a term after the lead section of the article. As a result, there are about 122 million multiple links in the English Wikipedia, comprising 50% of all link targets (*cf.* Section 5.3). These links receive 75% of the clicks as measured in the Wikipedia click stream dataset [WT], thus they are of high relevance to the encyclopedia, *i.e.*, for studying the human click behavior as well as for designing methods and models for navigation on Wikipedia. In the previous chapter, we have made assumptions regarding the click distribution among multiple links connecting two articles. For example, we assigned the whole attention to the first occurring link instance on the screen based on y-coordinates [DSLS17], or divided the attention

[1] `https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style/Linking`

equally between the link instances [DSLS16]. However, given the different properties that links exhibit (*e.g.*, their visual regions), these assumptions fail to hold in general. For pairs of links comprising two link instances with the same target article in Wikipedia, Table 5.1 shows the distribution of link pairs with respect to the visual region of both links. Given that the visual region in which a link appears, *i.e.*, lead, body, infobox, navbox affects click probabilities [DSLS16] and the distinct position bias of the click distribution among links [PWZL16], the table also emphasizes the need for a general, unified model that can *disambiguate* between multiple links and accurately predict the click probability of each link respecting its properties, *e.g.*, visual position and appearance on the screen.

Problem and objectives. In this work, we study the problem of *link disambiguation* in Wikipedia. More precisely, for a set of links connecting two articles, we are interested in understanding the user click behavior when multiple links with the same target article compete for attention in terms of user clicks. Developing a model that can disambiguate between links—*i.e.*, predict their click probabilities—can help to create precise click heatmaps from transition data for individual Wikipedia articles similar to [DSLS16, Wes15, AL08].

Materials, approach and methods. Our work makes use of openly available article datasets of the English Wikipedia and the Wikipedia clickstream (*cf.* Section 5.3). We develop and analyze a large set of link features that capture not only local link properties but also global properties that contribute to the probability of a link being selected. To account for heterogeneity between the articles, we also include article features in our analysis. To be able to fairly compare links, we use redirects to study multiple links to the same page for which the exact location and click count are known. We then train a random forest model for link disambiguation and evaluate it in terms of mean absolute error (MAE) and earth mover's distance (EMD). Finally, for the purpose of creating heatmaps from transition data, we show how our model can be tuned based on the distance between the links, as measured on screen and in the running text.

Table 5.1: Percentage of link pairs with links appearing in the visual regions indicated by the rows and columns for a random sample of 100k link pairs.

	lead	body	infobox	navbox
lead	0.8%	5.9%	0.9%	4%
body		32.9%	3.5%	16.6%
infobox			0.8%	5.8%
navbox				28.8%

5.2 Related Work

Modeling human click behavior has a long tradition. For example, Huberman *et al.* studied the usage patters of web pages and showed their strong regularities [HPPL98]. The random surfer model assumes that people click at random and is fundamental for ranking nodes in a network [PBMW99]. Markov chains and decentralized search models have been widely adopted for simulating agent based navigation or for navigational hypothesis testing on information networks such as Wikipedia [HSGS13, SHHS15, SHTS14]. Different hypotheses about the driving factors behind the human click behavior have been proposed, *i.e.*, that humans follow semantically similar links [WL12, SHTS14], click on links located at the top of a web page [DSLS17, DSLS16], or click on links by constantly evaluating the cost and value of information with respect to the current information need [PP99]. Studying human click behavior and navigation on Wikipedia using transition data has a variety of application, *i.e.*, inferring missing links between articles [WPL15], improving the hyperlink structure [PWZL16], and extracting semantic relatedness between articles [SHTS14]. While we have highlighted just a small fraction of related work, in the majority of the studies, click behavior is modeled as navigation on a network by mapping links to edges while ignoring multiple link occurrences. To that end, these models account for competition between links targeting different articles, thus possibly more or less interesting content to the user. In our work, we focus on modeling click behavior when multiple links with the same target compete for a user's attention. In this sense, the model we propose controls for the inherent importance of a link independent of the content of its target article.

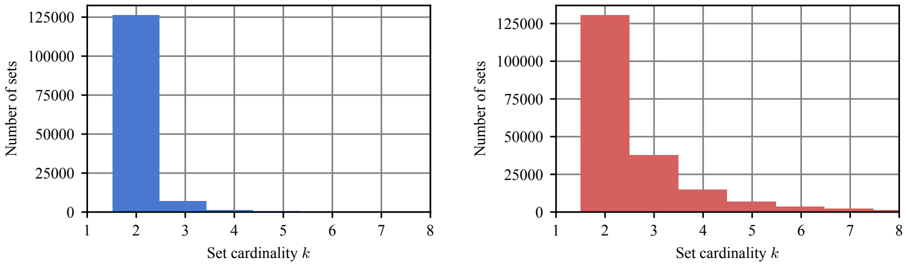
5.3 Datasets

In this section, we describe the data at hand which consists of the transition data (including redirects transitions) and Wikipedia link data containing link instances features.

Wikipedia data. For this work, we focus on a snapshot of the English Wikipedia from August, 2016[2], for which we obtained the HTML versions of the articles using the Wikipedia API[3]. By parsing and rendering the HTML version of the articles, we are able to extract article features and link features capturing aspects related to the content of the articles and the visual appearance of the links and the articles themselves. The dataset contains roughly 5 million articles and 513 million links, 122 million of which occur multiple times.

[2] `https://archive.org/details/enwiki-20160801`

[3] `https://www.mediawiki.org/wiki/API:Main_page`



(a) Redirect link instances

(b) Link instances

Figure 5.1: Multiple link instances for (a) the redirects dataset and (b) for a random sample of 200k links in Wikipedia. A link disambiguation model for $2 \leq k \leq 4$ would be able to cover the vast majority of multiple links.

Transition data including redirects. Transition data between articles in Wikipedia is published in aggregated form by the Wikimedia Foundation. This data contains the transition counts between articles in Wikipedia in form of (*referrer, resource*) pairs extracted from the server logs. Bots and web crawlers are filtered out, and the data contains only entries with more than ten transitions between articles. Unlike other versions of the Wikipedia clickstream, the dataset from August 2016 that we use in this work does not resolve redirects [WT]. Instead, it captures the exact click counts for multiple link instances linking to the same target article if one of the link instances points to a redirect page. For example, the article of *United_States* that is linked twice from the article of *Barack_Obama* is referred 20 times from a direct link and 12 times from a link leading to the article of *USA* which redirects to *United_States*. This allowed us to extract a dataset containing sets of link instances (*i.e.*, a direct link and additional links via redirects) and their exact click counts. To compile the dataset, we used all links appearing in $k \geq 2$ non-resolved redirect links, each of which appeared exactly once in the source article. This left us with 286,607 links emerging from 87,503 articles. We will refer to this dataset as the redirects dataset. For the redirects dataset (a) and for a random sample of about 200k multiple link instances from Wikipedia (b), Figure 5.1 shows the distributions of the link sets across different link set cardinalities k. With the data at hand, we can study link disambiguation only for maximum four link instances (*cf.* Figure 5.1(a)). However, this would cover the majority of multiple link instances in Wikipedia (*cf.* Figure 5.1(b)).

5.4 Modeling Link Disambiguation

As shown in Section 4.4.2 and from related literature we know that the structure of Wikipedia articles influences the click behavior of users [LLHS17]. To this end, to predict the click probability for a set of multiple links, we start by proposing a set of link and article features capturing different aspects of the appearance of links and articles. Furthermore, we perform statistical analysis on selected features to gain a first impression of their nature. In this section, we also give a detailed description of the model fitting including parameter optimization, data splitting, baselines, and the performance measures used in our experimental setup.

5.4.1 Link and Article Features

In this work, we introduce a set of link and article features for modeling link disambiguation (*cf.* Table 5.2). The link features we propose can be divided into *local* and *global* link features. The local features express the relation between a link and its immediate surroundings, whereas the global link features capture the global context in which the link is embedded. The article features capture information regarding the layout and content of the article.

Statistics on selected features. For the most-represented case of multiple links with the same target in our redirects dataset ($k = 2$), we provide some statistical insights on the features *visual_region* and *in_table*. We report the number of links and clicks, and the median number of clicks calculated on the training data (*cf.* Section 5.4.2 for exact data splits). In general, links located in the text are clicked more frequently than links occurring in a table (*cf.* Table 5.3). However, the type of the table and its position on the screen matter (screen top for infobox and screen bottom for navbox). Links presented in infoboxes are clicked almost twice as often as links presented in navboxes (*cf.* Table 5.4). These initial observations show the nature of the features and their possible importance to our model as they capture essential aspects of the click behavior of the Wikipedia users.

5.4.2 Experimental Setup

In this section, we describe the experimental setup for the task at hand, *i.e.*, modeling competition between Wikipedia links with the same target article. More specifically, we provide an overview of baselines, performance measures and the model fitting including parameter optimization and data preprocessing.

Baselines. To approach the problem of multiple link instances, previous work has generally resorted to two simple models, which represent our baselines. The *first link occurrence* model assigns all the attention to the first link instance either based on the position in text or based on the position on the screen. This is the simplest possible model. However, it also implies

Table 5.2: Link and article features description.

	feature	description
local	in_visual_element	A binary feature capturing if a link occurred in a visual element (*i.e.*, table, list, figure caption, or hatnote, a short note placed at the top of an article or section).
	position_in_vis_element	A numerical feature capturing the link position in a visual element, *e.g.*, in a table.
	position_in_section	The position of a link in the text of the article section.
	position_in_section_only	The position of a link in the text of the article section excluding all links in visual elements.
	#_of_links_in_section	The number of links in section.
	#_of_links_in_vis_element	The number of links in a visual element.
	sem_similarity	The semantic similarity between the section containing a given link and the link's target article; calculated using the link-based method proposed by Witten and Milne [WM08].
	avr_clicks_in_section	The overall user click activity around a link.
	anchor_text_length	The length of the anchor text of a link.
global	x/y_coordinates	The exact screen position of a link.
	visual_region	The type of visual region in which a link appears. The regions are defined as follows [DSLS16]: (1) *lead*: all links in the first section of the article excluding the infobox, (2) *body*: all other links in the main text, (3) *infobox*: all links in the infobox, (4) *navbox*: all links in a table in the last section of the article placed for facilitating navigation.
	position_in_text	The global position in the text of the article.
	position_in_text_only	The global position in the text of the article excluding all links in visual elements.
	section_number	This features captures in which section the link appears.
	visual_element_number	This features captures in which visual element number the link appears in, *e.g.*, in table three out of four tables in the article.
article	page_length	The page length of the article in pixels.
	#_of_sections	The length of the article in terms of the number of sections.
	#_of_links	Total number of links.
	#_of_links_in_text	Total number of links in the text only, links in visual elements are excluded.
	#_of_visual_elements	For each visual element type (*i.e.*, table, list, figure caption, or hatnote), this feature counts the times it appears in an article.

a very skewed distribution, which makes this model not applicable in the general case. The *equal attention* model distributes the click probabilities equally among all link instances. However, this may also not always be a good fit as suggested by the statistics reported in the previous section, *i.e.*, with respect to the visibility of a link.

Performance measures. To assess the extent to which the proposed model is able to predict the click probability for each link instance, we employ two measures. By the mean absolute error (MAE), we measure the difference between the predicted and the empirical distribution for each set of links. On the other hand, the earth mover's distance (EMD) measures the distance between two probability distributions over a region D [RTG00] and for the task at hand it captures how much probability mass has to be moved between the links in order to match the empirical probability values for each link. For the region D, we utilize the positions of the links on the screen (*y-coordinate*) and in the text of the article (*position_in_text*). We evaluate our models in terms of the average MAE and EMD over all data points.

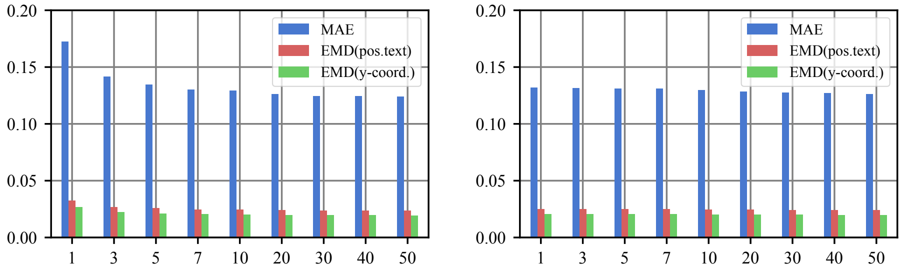
Learning algorithm and model fitting. Our task of interest is to learn the probability distribution for a set of links with the same target, which are in direct competition for user attention. Although, there are several possible learning algorithms that can be applied to this setting, we decided to use random forest, as it is well-equipped to model non-linear effects between a set of features while efficiently handling large high-dimensional datasets. Since most ambiguous link instances in our data occur with multiplicity between $2 \leq k \leq 4$, we focus on link sets with this cardinality. For each k, we fit a separate random forest model using the R package *randomForestSRC*[4]. For each set of links, we build a feature vector containing the features of all links in the set. The corresponding response vector captures the probability distribution of the links in the set. For growing regression forests, RandomForestSRC uses the composite normalized mean squared error as splitting rule and it is one of the few available software packages that is able to grow a forest for such a scenario with multivariate responses.

As mentioned, a possible application of our model is the generation of precise click heatmaps from transition data. In this case, making smaller errors for links located far from each other is a desirable property. We therefore also trained a version of the model tuned by weighting the training

[4] https://github.com/kogalur/randomForestSRC

Table 5.3: Statistics for *in_table*. In general, links in tables are clicked less than links in running text.

in_table	# Links	# Clicks	Median clicks
✓	33,173	2,646,074	27
✗	90,447	6,614,178	32



(a) Trees

(b) Samples

Figure 5.2: Parameter optimization. We performed a grid search for (a) the number of trees in the forest and (b) number of samples in the terminal nodes.

data using the distance between the links—the further away two links are, the higher the probability of a sample being selected by the bootstrapping mechanism of the random forest algorithm. Possible distances between the links are given by the difference of the feature values for *position_in_text* and *y-coordinate*.

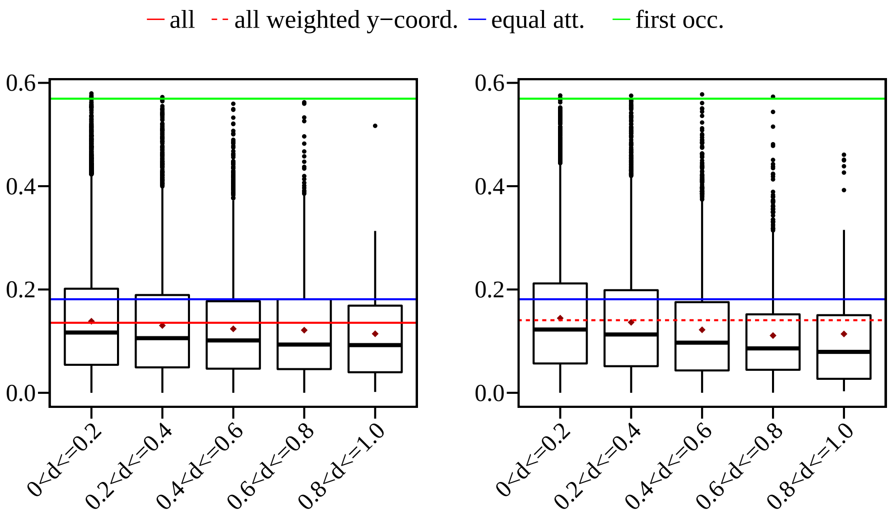
Preprocessing and data split. As RandomForestSRC cannot deal with missing values, we imputed the median for all features containing missing values for links. We also experimented with mean imputation and normalization of the numerical features, however, the achieved results were very similar.

For each k, we randomly split the redirects dataset into sets for training (50%), validation (20%) and testing (30%). We trained our models on the training dataset and conducted parameter optimization, model tuning and feature engineering on the validation data. After the optimization, we evaluated our models on the test data.

Parameter optimization. We performed a grid search to optimize the performance over the random forest parameters: *number of trees* and *number of samples in terminal nodes*. Figure 5.2 shows the performance of the model with respect to MAE and EMD for the features *position_in_text* and

Table 5.4: Statistics for *visual_region*. The majority of the links are located in the body. The lead and the infobox have the highest median. The type of the table (*e.g.*, infobox or navbox) plays a role for a link to be followed.

visual_region	# Links	# Clicks	Median clicks
lead	25,415	2,374,901	42
body	75,816	5,459,661	30
infobox	10,780	1,017,772	34
navbox	11,609	407,918	18



(a) MAE vs. y-coord.

(b) MAE vs. weighted y-coord.

Figure 5.3: MAE by distance d between the links. Red dots indicate the mean. MAE decreases with growing distances between the links as measured by the difference in *y-coordinate*. Weighting with the distance between the links reduces the error for larger distances. Similar results are achieved for the feature *position_in_text* (not shown here).

y-coordinate. We see that increasing the number of trees in the forest increases the performance of the model with respect to all three metrics (*cf.* Figure 5.2(a)). Likewise, we optimized for the number of samples in the terminal nodes finding that this parameter has very small to no effect on the performance (*cf.* Figure 5.2(b)). In our experiments, we used the default bootstrap protocol of the RandomForestSRC package, which bootstraps the data by sampling with replacement at the root node before growing the tree. Alternative protocols such as node level sampling did not change the performance. To keep the model performance at reasonable levels and model fitting time low, we performed our experiments with seven trees in the forest and five samples in the terminal nodes (number recommended for regression in RandomForestSRC).

5.5 Results

Next, we report the achieved results for different link set cardinalities k.

Pairs of link instances (k = 2). The redirects dataset contains 126,143 link pairs for k = 2. We built a separate model for each feature group, namely the article, local and global link features. Table 5.5 shows the results with respect to MAE and EMD. Each feature group is able to outperform the first occurrence and the equal attention model. The best-performing feature group is the global link features. This group alone is able to achieve a performance almost as good as the combination of all features. In terms of EMD, the model makes smaller errors for *y-coordinate* than for *position_in_text*. Figure 5.3 shows the MAE as function of the distance between the links. In general, the error decreases as the distances grow (*cf.* Figure 5.3(a)). After weighting with the distances between the links, we are able to tune the model and reduce the resulting errors for larger distances even more (*cf.* Figure 5.3(b)). However, this comes at the cost of slightly increasing the overall error (*cf.* Table 5.5). Studying the contribution of the distance features for approximating mouse-tracking or eye-tracking heatmaps is a possible future work. As mentioned, we are interested in a general, unified model for link disambiguation. To this end, we evaluated our model for pairs of visual regions. To account for the different numbers of link pairs in regions, we calculated 95% confidence intervals for the MAE using bootstrapping (*cf.* Table 5.6). Overall, the model makes the smallest errors when one of the links is located in a navbox. Links in navboxes are rarely clicked, at the bottom of the page, and in a table, which makes them easy to disambiguate. On the contrary, the largest MAE of about 0.15 occurs when both links are located in the lead. This is likely due to the fact that links in the lead are located near to each other and in running text. When both links are in the body, which is the largest fraction of cases in our data, the MAE is about 0.14. In general, if one of the links is in a visual element, *i.e.*, infobox or navbox, the model makes smaller errors.

Multiple link occurrences ($k > 2$). Table 5.7 shows the MAE with 95% confidence intervals for $k = 3$ and $k = 4$. For these cases, we have much less data in our redirects dataset (6,812 and 1,180 samples for $k = 3$ and $k = 4$, respectively, *cf.* Figure 5.1(b)). Overall, we observe that with increasing k—representing an increase in the visibility of the target page—the equal attention model fits the data better (*cf.* Table 5.7). The same is valid for the first occurrence model. As for $k = 2$, however, this model performs much worse than the equal attention model. For a random forest model with all features, we are able to beat the baselines for $k = 3$. For four link instances, however, our model outperforms only the first occurrence model. Tuning the models with respect to the distance between the links is also not as straightforward as for $k = 2$ and requires to define a way to measure a distance between more than two links. We therefore trained separate binary models ($k = 2$) for each possible combination between the links. To obtain multivariate responses, we combined the binary models using the PKPD method [WLW04]. With this approach, we achieved similar performance as with direct modeling while tuning the model by weighting samples as for $k = 2$.

Table 5.5: Feature group results for ($k = 2$). Each feature group outperforms both baselines (marked in italic). The global link features group performs best compared to the article and local link features. Combining the three feature groups performs best (marked in bold). Weighting with the distance between the links slightly reduces the performance.

	MAE	EMD (pos. text)	EMD (y-coord.)
first occ. —	*0.5693*	*0.1243*	*0.1048*
equal att. —	*0.1811*	*0.0421*	*0.0354*
article	0.1792	0.0382	0.0319
local link	0.1667	0.0328	0.0272
global link	0.1416	0.0271	0.0225
all —	**0.1355**	**0.0253**	**0.0210**
all weighted pos. text	0.1378	0.0255	-
all weighted y-coord. - -	0.1405	-	0.0214

Table 5.6: MAE and 95% CIs for the model based on all features for $k = 2$, separated by visual regions. Both baselines fall outside the CIs for all combinations of regions.

	lead	body	infobox	navbox
lead	0.1515 (0.1468, 0.1562)	0.1407 (0.1385, 0.1431)	0.1394 (0.1348, 0.1440)	0.0993 (0.0950, 0.1035)
body		0.1404 (0.1389, 0.1420)	0.1493 (0.1451, 0.1535)	0.1156 (0.1125, 0.1186)
infobox			0.1371 (0.1248, 0.1490)	0.0835 (0.0778, 0.0890)
navbox				0.0727 (0.0638, 0.0812)

Table 5.7: MAE and 95% CIs for all features models. Beating the equal attention baseline is not possible for $k = 4$.

	MAE ($k = 3$)	MAE ($k = 4$)
first occ.	*0.4898*	*0.4003*
equal att.	*0.1431*	*0.1112*
all	**0.1252** **(0.1220, 0.1284)**	0.1132 (0.1051, 0.1173)

5.6 Discussion

In this work, we model competition between Wikipedia links with the same target article, which are ambiguously represented in the transition data extracted from server logs. Although the exact click locations could also be determined by installing adequate logging routines for Wikipedia, redesigning the logging systems would be cumbersome and expensive for a large information system such as Wikipedia, rendering it virtually infeasible. Moreover, the standard logging procedures of common web servers such as Apache do not include this possibility. The model presented by this work is therefore an especially useful addition for researchers and practitioners who wish (i) to have an easy way of assessing the distribution of clicks to areas of a webpage, or (ii) to inspect historic log data for which no data on click positions exists.

Understanding the exact location of clicks to Wikipedia, one of the most-visited websites worldwide, can help to build tools for the visualization of human click behavior from transition data. These tools could benefit Wikipedia editors by helping them to precisely identify screen regions or text passages with shortcomings, or of special interest to readers. Furthermore, our model can be applied for automatically estimating the optimal text spans between links in an article.

5.7 Conclusion and Future Work

In this chapter, we have proposed a machine learning approach for link disambiguation in Wikipedia. The proposed method resorts to global and local link features, and to article features for modeling click probabilities for a set of multiple links with the same target in a Wikipedia article. We have trained a random forest model that outperforms existing models, *i.e.*, the *first link occurrence* and *equal attention* models with respect to MAE and EMD. We have also shown how our model can be tuned for the purpose of visualizing user click behavior by means of heatmaps. In future work, our model can be applied to create heatmaps to approximate mouse-tracking or potentially eye-tracking heatmaps for Wikipedia articles using only transition data.

Chapter 6

The Role of Structural Information for Designing Navigational User Interfaces

6.1 Introduction

With the increasing amount of information made available to people on the Web every day, it has become increasingly difficult to build information systems that can be navigated in an efficient way. Information systems that deliver strong intuition about the choices made available to their users through the interfaces are efficient at guiding the user to the needed piece of information. Thus, they are considered good at supporting activities such as *navigation* or *browsing*. In order to improve navigability, new interfaces—*e.g.*, tag clouds, breadcrumbs, subcategories—have been introduced. In Figure 6.1, we see an example of a tag cloud. Besides other aspects of tag cloud design [SKK$^+$08], tag clouds—as well as all other kinds of user interfaces—are only useful for augmenting navigation to the extent to which they are able to expose the underlying structure of the information space [Hea09]. Yet, little is known about what kind of and how many topological clues should be integrated in navigational user interfaces.

Problem. Consequently, in this chapter, we want to study the problem of properly exposing the topological structure of the information space through an interface. This problem has two dimensions: (i) Which are the important structural properties that contribute to properly exposing the hidden structure of the information space and (ii) how much should we know about them in order to navigate efficiently? Knowing which nodes in a network are important and how to identify them is crucial for navigation. Such knowledge could reduce the amount and nature of information needed for improving the users' understanding about the information space resulting in better navigational efficiency. Subsequently, we next derive and discuss

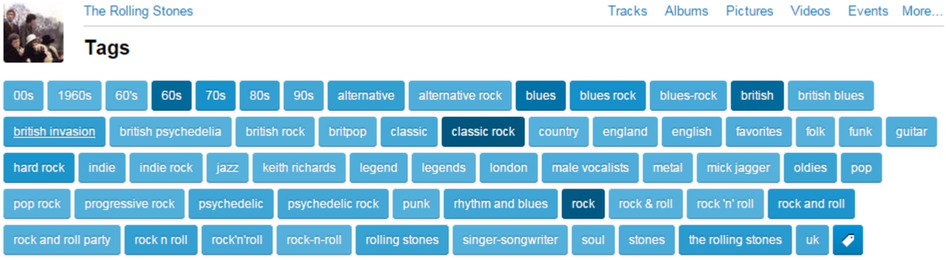


Figure 6.1: A tag cloud enabling navigation from The Rolling Stones page on last.fm. Exemplary user interface used for navigation in many online information systems. The tag clouds among other web interfaces are useful to the extent that they expose the underlying structure of the information space. Identifying the most important tags *from a navigational perspective* is crucial for providing efficient support.

the two main research questions that we want to tackle in this article.

Research questions. (i) What kind of and (ii) how much structural information is needed for efficient navigation? Regarding the first research question, we are specifically interested in deriving important structural properties of the information space that should be exposed through an interface in order to properly guide users' navigation. Related work [ALPH01] has suggested that the degree—as a proxy of a node's popularity—is a very good navigational feature in networks with a power law degree distribution. Yet, little is known about the effect of the clustering coefficient as a navigational feature on the efficiency of navigation. The clustering coefficient may be feasible as navigational feature due to its importance for the emergence of the small world property of a network. Small world networks are known to be particularly navigable [Kle00b, WS98]. In this chapter, we investigate whether nodes with a specific clustering coefficient have an impact on navigation and we study how the clustering coefficient can be used to identify them. Furthermore, regarding the second research question, we are interested in determining the amount of structural information needed for navigation and if this depends on the quality of the structural information.

Approach and methods. We approach the research questions by analyzing the structural properties of four different networks. Initially, we take a look at their shortest path distance, degree and clustering coefficient distributions, and classify them by their expected navigability according to [BKC09]. To model navigation, we use the message-passing algorithm *decentralized search* which is inspired by the small world experiment by Stanley Milgram [Mil67]. Several versions of the algorithm can be found in literature [Kle06, Kle00b, Kle00a, WDN02, ALPH01, AA05]. Decentralized search has already been demonstrated to be useful for modeling navigation in information networks [HSGS13]. For studying which and how much information is needed for efficiently navigating a network, we utilize an adaption of the algorithm which we call *partially informed* decentralized search. The partially informed decentralized search models a user who is limited in her exposure to the structure of the information space and thus, has just a weak or limited understanding of the topology of the information space. We study two strategies for selecting important nodes with regard to their popularity and clustering coefficient. With both strategies, the algorithm navigates by popularity. With simulations, we compare the partially informed decentralized search with the random search and the fully informed decentralized search. In our setting, random search corresponds to a user who is clicking at random and has no intuition. We also make a comparison between the two strategies for node selection to test the importance of the exposure of the user to the underlying structure of the information.

Findings and contributions. The most prominent finding is the surprisingly small amount of structural information needed for efficient navigation and the supportive properties of the clustering coefficient for identifying nodes important for navigation. By and large, our findings suggest that only a limited amount of high quality structural information needs to be exposed through the navigational user interface. Additionally, we empirically demonstrate the sensitivity of decentralized search as a navigational model on the kind of structural information utilized. The navigational performance of decentralized search appears to depend on the amount of high quality structural information provided.

Structure. The rest of this chapter is organized as follows. After discussing related work in Section 6.2, we present an adaptation of decentralized search and two strategies for selecting nodes with high structural importance used in the experimental setup in Section 6.3. In Section 6.4, we give detailed overview of the used datasets. In Section 6.5, we present our results and formulate our findings. Next, Section 6.6 discusses the findings and their implications for the design of navigational user interfaces. Finally, we conclude the chapter and provide some directions for future work in Section 6.7.

6.2 Related Work

The decentralized search algorithm is inspired by research conducted in the 1970s by Stanley Milgram who studied the structure of the American society and conducted the famous *small world experiment* [Mil67]. For this experiment, Milgram asked randomly selected people from Nebraska to forward a packet to a stock broker in Boston. If participants did not know the target personally, they were asked to forward the packet to personal contact that they thought might know the target better. These persons then should repeat this process. Even though there were quite some restrictions, the experiment showed that the average chain length of letter trails that reached the target was around six.

Motivated by this small world experiment, researchers [Kle06, Kle00b, Kle00a, WDN02, ALPH01, AA05] have developed the so-called *decentralized search algorithm* that tries to find a path between a *start node* and a *target node* in a network by passing a message from a node to one of its immediate neighbors also called *candidate nodes*. What information is available and how it is used for selecting one of the candidate nodes is decisive for the success of the search. For a detailed description of the decentralized search algorithm, please refer to Section 6.3.1. Next, we delve into related work and discuss navigation using homophily (Section 6.2.1), navigation using popularity (Section 6.2.2), models for user navigation (Section 6.2.3) and the role of clustering for navigation (Section 6.2.4).

6.2.1 Navigation Using Homophily

There are different models based on node similarity or homophily for generating small world networks in which decentralized search is very effective. The two main models are *grid-based* and *hierarchy-based*. The first *grid-based model* was proposed by Watts and Strogatz in [WS98]. This model places nodes on a two-dimensional grid in a way that nodes with high similarity have small grid distance. In order to assure the emergence of the small world property, the model puts long links between the nodes that are similar, but still locally far away on the grid. This model was improved by Kleinberg in [Kle00a, Kle00b] where he concentrated on the length of the long links. He showed that efficient search is only possible for certain values of the clustering exponent of the model which is responsible for placing the long link connections between the nodes.

The *hierarchical model* was proposed independently by Kleinberg [Kle02] and by Watts *et al.* [WDN02]; these models are also generative. In hierarchical models, similar nodes are placed near to each other in a hierarchy. The probability of two nodes being connected in the hierarchical model not only decreases with their hierarchical distance but also it decreases exponentially. Another generative model was proposed by Boguña *et al.* in [BKC09] where they assumed that nodes form a *hidden metric space*. The topology of the metric space determines the distance between the nodes in the metric space and models the probability of a link between them in the generated network. The model also possesses a parameter that is responsible for the clustering in the network. This clustering parameter, like the clustering exponent in Kleinberg's grid model, is also responsible for expressing the homophily of the nodes in the network. The main limitation of these models is the global information about the node's position on the grid or in the hierarchy.

6.2.2 Navigation Using Popularity

Since estimating similarity between nodes is not easy, Adamic *et al.* [ALPH01] concentrated on the degree of nodes. They proposed an algorithm for efficient search in power law networks which makes use of the power law degree distribution to support the node selection. The algorithm keeps track of a node's identity and uses information about the node's degree and the node's neighbors' degree. The biggest difference to the models elaborated in Section 6.2.1 is the absence of global information about the target node and its position in the network. Adamic *et al.* showed that degree-based navigation works fairly well in power law degree distributed networks in comparison to Poisson degree distributed networks. Additionally, in power law degree distributed networks, random walks tend to select high degree nodes and achieve good results in those kinds of networks.

6.2.3 Model for User Navigation

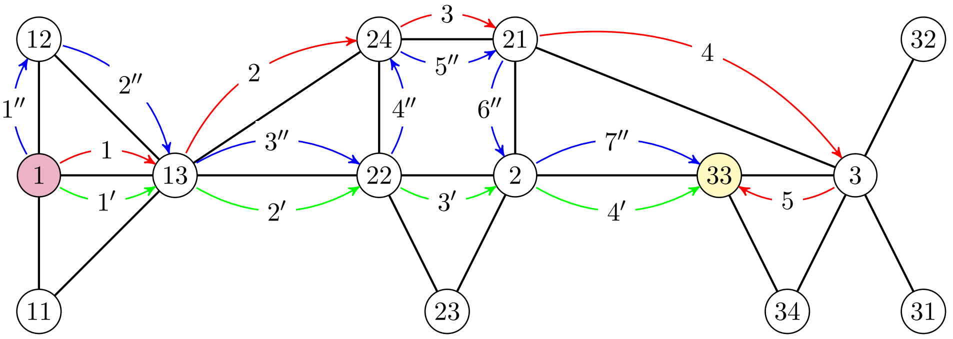
Decentralized search has a long tradition as a model for user navigation in different types of networks. In [HSGS13], Helic *et al.* showed that decentralized search can be used to model user navigation in information networks. The differences and the similarities between the click traces produced by decentralized search with hierarchical background knowledge and actual user navigation were studied by Trattner *et al.* [TSHS12]. Research on the navigational efficiency of different types (broad and narrow) of hierarchical background knowledge conducted by the authors showed that both types are useful. However, broader hierarchies performed better under the limitations introduced by the user interface [HKG$^+$12].

6.2.4 The Role of Clustering

In [WS98], Watts and Strogatz used the characteristic path length and the clustering coefficient to define the class of navigable networks. The characteristic path length is the averaged shortest path length over all nodes in the network. The clustering coefficient can be interpreted as the probability of a link to exist between two randomly picked neighbors of a node [New10]. In a network $G = (V, E)$, where V is a set of nodes and E is a set of edges, $E \subseteq V \times V$. Let $N(u)$ be the neighborhood of the node u and d_u the degree of the node u. The local clustering coefficient $C(u)$ is then defined as the fraction of pairs of neighbors of the node u that are themselves neighbors:

$$C(u) = \frac{|e_{vw} \in E : v, w \in N(u)|}{d_u(d_u - 1)/2}. \tag{6.1}$$

An alternative definition of the class of navigable networks was given by Boguña *et al.* [BKC09] who showed how the navigability of a network depends on its degree distribution and its clustering coefficient. In the models



Popularity/Node	1	11	12	13	2	21	22	23	24	3	31	32	33	34
Degree (fully informed, $I = V$)	3	2	2	5	4	3	4	2	3	5	1	1	3	2
Degree (partially informed, $I \subset V$)	-	-	-	5	-	-	-	-	3	5	-	-	-	-

Figure 6.2: Different versions of decentralized search. *Green:* The arrows show the path produced by a *fully informed decentralized search*. *Red:* The arrows show the path produced by a *partially informed decentralized search*. *Blue:* The arrows represent the path produced by a *uninformed random walker*. The table shows the information provided to the algorithm for selecting the next step. The first row of the bottom table contains the popularity scores of all nodes $I = V$ provided to the fully informed decentralized search and the second row contains only a small portion of all popularity scores $I \subset V$. Fully informed and partially informed decentralized search apply greedy neighbor selection. The partially informed search selects a random node when no information is available. Although finding the shortest path between the nodes 1 (purple node) and 33 (yellow node) is possible with both versions of decentralized search, in general this is not the case because the algorithm can take an informed decision only on the local level.

described in Section 6.2.1 and Section 6.2.2, the clustering exponent plays an important role for the emergence of the small world networks and it is crucial for navigation.

In [KM09, KM10], the authors studied the impact of the clustering exponent on the navigability of a network, *i.e.*, they showed for different network sizes how the change of the clustering exponent affects the effectivity and the efficiency of four different decentralized search versions. In the next Section 6.3, we will present an adaptation of decentralized search—partially informed decentralized search—and we will use the degree distribution and clustering coefficient of the networks to identify the nodes for which the partially informed decentralized search will be able to make an informed decision.

6.3 Methodology

Decentralized search is an established model for navigation. Our goal is to estimate the amount and type of structural information that allows efficient navigation. To this end, we extend the decentralized search algorithm in a way that allows us to simulate navigation with limited amount and different kinds of structural information. By doing so, we can tackle the research questions posed in Section 6.1. Next, we describe (partially informed) decentralized search in Section 6.3.1 before we discuss strategies for node selection in Section 6.3.2 and conduct our experiments in Section 6.3.3.

6.3.1 Decentralized Search

In Figure 6.2, we see an example of both a *fully informed* as well as a *partially informed decentralized search* in a network. The goal is to find the path between node 1 (purple) and node 33 (yellow). The fully informed version of decentralized search uses the degree information as shown in the first row of the table presented in Figure 6.2 and navigates greedy by degree. Let I be an *informed set* of nodes for which the algorithm can take an informed decision regarding the degree of the candidate nodes. In the case of fully informed search $I = V$, this means that the algorithm possesses the degree information about all nodes in the network. This allows it to rank all candidate nodes by their degree and to select the node with the highest degree. The green arrows show how navigation proceeds for this version of the algorithm. The red arrows show a path produced by the partially informed version of decentralized search. In this version, we only have a fraction of the popularity information as shown in the second row of the table in Figure 6.2 and the informed set I is a proper subset of V. The partially informed decentralized search ranks the nodes by their degree and selects the node with the highest one only if the set of candidate nodes C contains nodes whose popularity value is available in $I \cap C \neq \emptyset$; otherwise, it picks one node at random. In both versions of the algorithm, we avoid already visited nodes and we terminate the search if the target node is in the set of candidate nodes. For completeness, Figure 6.2 also highlights an example path of an *uninformed random walker* (blue arrows) that simply picks adjacent nodes at random for navigating.

With the partially informed version of decentralized search, we can estimate the amount of information really needed for navigation in a network. By varying the fraction of the nodes where the popularity is available, we can derive the sensitivity of the algorithm to the amount of popularity information. Thus, this allows us to study our research questions at interest regarding what kind of and how much structural information is necessary for efficient navigation.

Using the methodological concepts explained, we conduct our experiments in Section 6.3.3. We focus on using the degree of the candidate nodes to model the popularity of nodes. Degree corresponds to the number of links attached to the node [New10] and it is a local metric:

$$d_u = \sum_{v \in V} a_{uv} \tag{6.2}$$

Thus, when we speak about fully and partially informed decentralized search, we speak about fully and partially informed on a local level. If the algorithm was informed on the global level—in other words, if we possessed the adjacency matrix A of the network G—we would be able to calculate the shortest path, which is highly unlikely for real user navigation in large information networks on the web.

6.3.2 Strategies for Node Selection

In the following, we define two strategies for selecting structurally important nodes: the popularity strategy and the clustering strategy. The nodes selected by these two strategies are elements of the informed set of nodes for which the partially informed decentralized search is going to possess the information about their popularity (*i.e.*, degree) in the network. With these strategies, we can study how the kind of structural information affects navigation.

Popularity Strategy. We sort the nodes by popularity in descending order and take just the top k% of the sorted list. For these nodes, the algorithm will make an informed decision regarding the popularity of the nodes. The idea behind the popularity strategy for node selection is the same as the idea to navigate by popularity, namely highly popular nodes are very well connected. Selecting a highly popular node increases the probability of finding the target node under the nodes' neighbors.

Clustering Strategy. We sort the nodes by clustering coefficient in ascending order and take just the top k% of the sorted list. For these nodes, we again provide the popularity value of the nodes to the algorithm. Consider that with this strategy the algorithm also navigates greedy by degree.

The rationale behind the clustering strategy for node selection is that nodes with low clustering reduce the probability of a link to exist between two random neighbors of a node. This means that selecting a node with low clustering will provide nodes where the neighbors are not connected. The absence of a link between two neighbors of a node can be interpreted in the way that the neighbors are just too different. This would imply that selecting nodes with low clustering would provide nodes whose similarity between the neighbors is very small and this would allow navigation between clusters in the network. On the other hand, low clustering means that in this network region there is a *structural hole* as defined by Burt in [Bur09]. The absence of connections between the nodes in these regions of the network will give even a higher importance to the existing connections resulting in a higher importance of the nodes in these regions.

6.3.3 Evaluating Navigational Efficiency

As emphasized, we conduct experiments with two distinct strategies for selecting the node members of the informed set having also different informed set sizes. With the popularity and clustering strategy (*cf.* Section 6.3.2), we examine how the exposure of the structure of the information space through the interface affects the efficiency of navigation. Furthermore, with the size of the informed set, we investigate how much structural information is needed for efficient navigation.

We conduct experiments on four different networks (*cf.* Section 6.4): (i) Wikipedia for schools (topological link network), (ii) Facebook (ego network), Twitter (ego network) and (iv) DBLP (co-authorship network). The four datasets can be seen as representatives of popular networks in information system on the web. For each network, we generate thousand navigational missions containing of one *start node* and one *target node* chosen randomly with at least one path between them. The goal for the algorithm is to reach the target nodes. We break up the search after 20 iterations on the small networks (Wikipedia for schools and Facebook) and 50 iterations on the big networks (Twitter and DBLP). We conduct experiments with degree as local popularity metric.

Note that for a set of size 0% of all nodes in the network, we navigate without any structural information. With this setting, the partially informed decentralized search reduces to a uninformed random walker (*cf.* Figure 6.2) which can serve as a baseline for our experiments as it corresponds to a third (random) strategy for selecting important nodes. For a set size of 100%, we navigate with all available information. This means that the partially informed search upgrades to a fully informed decentralized search.

As we are interested in examining the impact of the amount and kind of structural information provided to the partially informed decentralized search algorithm, we also need to evaluate the efficiency of the algorithm. To that end, we focus on two metrics: the *success rate* and the *stretch.* Success rate and stretch respectively measure the effectivity and efficiency of the search. We calculate the success rate as:

$$s = \frac{|W|}{|P|} \tag{6.3}$$

It is the fraction of the set of successful missions W and the set of all missions in the simulation P. The success rate measures the percentage of cases in which the algorithm was able to find the target node. Thus, the success rate measures the effectivity of the algorithm. To measure the efficiency of the algorithm, we consider the stretch defined as:

$$\tau = \frac{1}{|W|} \sum_{s,t \in W} \frac{h(s,t)}{l(s,t)}. \tag{6.4}$$

Technically, the stretch is calculated by dividing the length of the path produced by the algorithms $h(s,t)$ with the length of the shortest path $l(s,t)$ between the start and the target nodes and then averaging over all nodes.

6.4 Dataset Description

In this section, we give a thorough description of the studied datasets and their structural properties. We analyze four different networks (*cf.* Table 6.1) taken from the Stanford Large Network Dataset Collection[1]. The *Wikipedia for schools* network represents the topological hyperlink network derived from Wikipedia articles for teaching purposes referred to as Wikipedia for schools (Wikifs). The *Facebook* and *Twitter* datasets are ego-networks. Finally, the *DBLP* dataset represents a co-authorship network.

Navigability of networks. In Figure 6.3, we see the degree distributions of the different datasets. All networks exhibit power law like degree distributions at least for the tail. To get an initial idea of the navigability of these networks, we apply the method presented by Boguña *et al.* [BKC09] who studied navigability of networks by looking at their clustering coefficients and power law exponents. In Table 6.2, we see that the values of the clustering coefficient of all networks are in the range defined in [BKC09]. Additionally, we determine the power law exponent of the degree distributions with the methods presented in [CSN09, ABP14]. We see that if we try to fit the power law distribution for the whole range of data points ($x_{min} = 1$), all networks are navigable according to Boguña *et al.* [BKC09]. This is not the case, if we try to find the best power law fit and let the method estimate the best x_{min}. In this case, only the Facebook network is efficiently navigable.

[1] http://snap.stanford.edu/data/index.html

Table 6.1: Datasets collection. The table shows the network type and the number of nodes and edges. Two networks are directed and two undirected. For each network type there is a small and a big network regarding the nodes and the edges.

	Type	Nodes	Edges
Wikifs	directed	4,604	119,882
Facebook	undirected	4,039	88,234
Twitter	directed	81,306	1,768,149
DBLP	undirected	317,080	1,049,866

Inequality of degree distributions. The Gini index is a metric that reviews the inequality in the degree distributions. A Gini index of zero means that the degree is equally distributed over the network, whereas a Gini index of one means that one node of the network possesses all links. In Table 6.3, we highlight the Gini index and the corresponding functions generating distributions with such inequality for the four datasets at hand. The corresponding generating functions support the results of the estimated first data point. We see that Wikipedia for schools, Facebook and DBLP possess Gini indices of 0.54. The inequality in the degree distribution is more explicit in the Twitter network. Inequality in the degree distribution is important for achieving good results with greedy navigation since it assures easy decision making.

Pareto principle. Since all networks possess power law like degree distributions, the Pareto principle suggests that we will need at least 20% of the nodes to achieve similar success rates and stretches for the networks with the popularity strategy and partially informed decentralized search as with a fully informed decentralized search. Additionally, we see that only one network is navigable according to the classification of Boguña *et al.* [BKC09] (if we use higher x_{min} values), thus, we cannot necessarily expect the popularity strategy with smaller amounts of nodes to perform well in these networks. The results presented in Section 6.5 contradict this intuition. We believe that this is tightly related to the clustering coefficient distributions for the four networks.

Differences in clustering coefficient distributions. In Figure 6.4, we see that the networks possess very different profiles regarding the clustering coefficient distributions. We see that the Facebook and Twitter networks exhibit similar clustering coefficient distributions, despite the different network size. In these networks, most of the nodes have a clustering coefficient between 0.3 and 0.7. Nodes in DBLP exhibit very high clustering coefficients and most of the nodes in Wikipedia for schools have a clustering coefficient between 0.1 and 0.5. Thus, we also expect to see differences in the results produced by the clustering strategy for node selection.

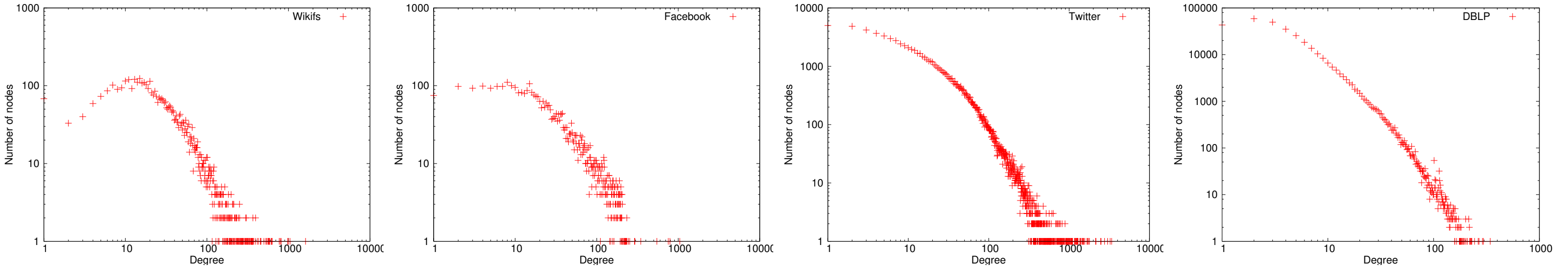
Shortest path distributions. Beside the clustering coefficient and the degree distribution of a network, the shortest distance distribution is also important for the emergence of the small world property of a network [WS98] which significantly increases its navigability. The shortest distance distribution also provides insight into how difficult it generally is to navigate a network. In Figure 6.5, we can see the shortest distance distributions of studied networks. For the Wikipedia for schools network, we see that most of the node pairs have a shortest distance of three. The Facebook and Twitter network exhibit a bit longer shortest distance, whereas DBLP has the longest shortest distance distribution.

Table 6.2: Small world classification of the datasets. Depending on the point from where we try to fit the power law in the distribution (from the first data point or x_{min} estimated automatically), we see that either all of the networks are efficiently navigable (the clustering coefficient C and the power law exponent α are in the range defined by Boguña *et al.* [BKC09]) or just the Facebook network. Table 6.3 suggests that we have higher trust in the results of the second row where the x_{min} is placed automatically.

	C	α, x_{min}	SW?	α, x_{min}	SW?
Wikifs	0.27	1.25, 1	✓	3.05, 142	✗
Facebook	0.61	1.26, 1	✓	2.51, 47	✓
Twitter	0.57	1.30, 1	✓	3.27, 188	✗
DBLP	0.63	1.48, 1	✓	3.26, 29	✗

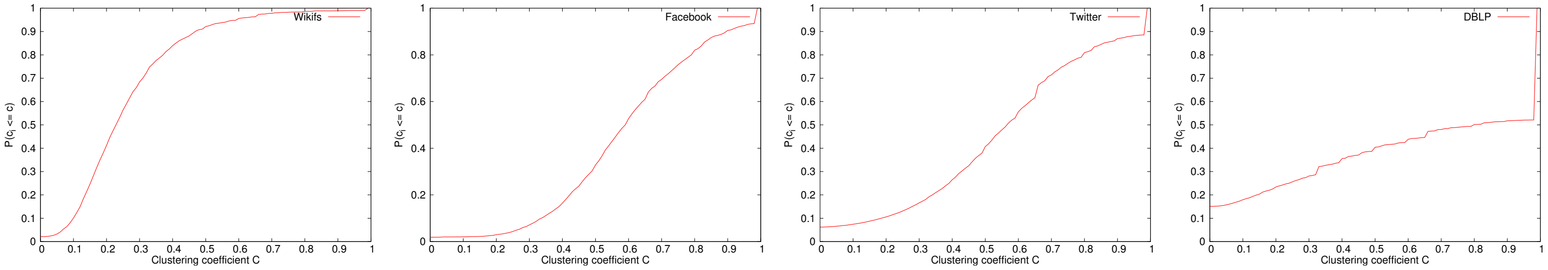
Table 6.3: Gini Index. The table shows the Gini index of the used networks and the corresponding distribution functions.

	Wikifs	Facebook	Twitter	DBLP
Gini Index	0.54	0.54	0.64	0.54
$f(x)$	x^2	x^2	x^3	x^2



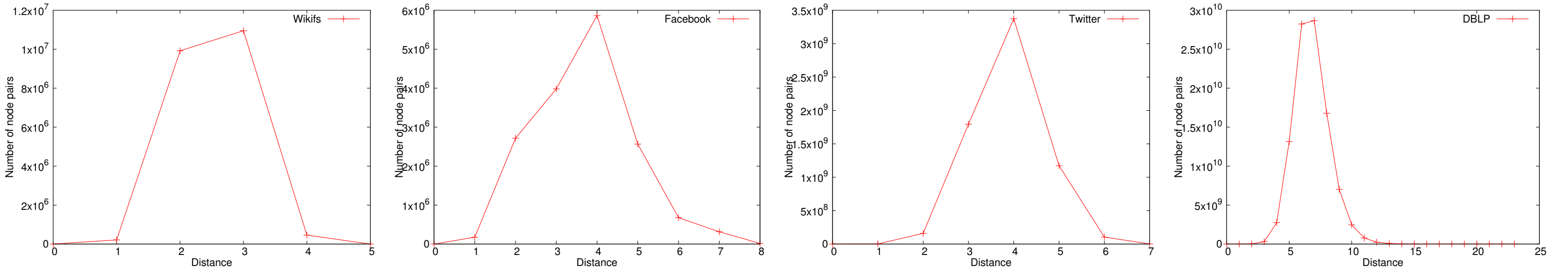
(a) Wikifs (b) Facebook (c) Twitter (d) DBLP

Figure 6.3: Degree distributions on log scaled axes. We see that all networks have a power law like degree distributions. This implies that degree greedy navigation will be very successful. For α-values *cf.* Table 6.2.



(a) Wikifs (b) Facebook (c) Twitter (d) DBLP

Figure 6.4: Clustering coefficient distribution. Most nodes in Wikipedia for school have clustering around 0.2 meaning that the network has no clearly defined clusters. Facebook and Twitter exhibit similar distributions despite the different network size; there is a fraction of nodes with clustering near zero and a bigger fraction of nodes with very high clustering near one. All other nodes have clustering coefficient nearly uniformly distributed between zero and one. Very characteristic for the DBLP network is the high clustering coefficient; around half of the nodes have a clustering coefficient of around one.



(a) Wikifs (b) Facebook (c) Twitter (d) DBLP

Figure 6.5: Shortest distance distributions. Most of the node pairs in Wikipedia for schools have very short shortest paths; this makes this network very efficiently navigable. We see that the Facebook and Twitter networks have very similar distributions despite the different network size. Also, in these networks most of the node pairs have very short shortest paths. DBLP is the most difficult to navigate considering the fraction of node pairs with relatively long shortest paths.

6.5 Results

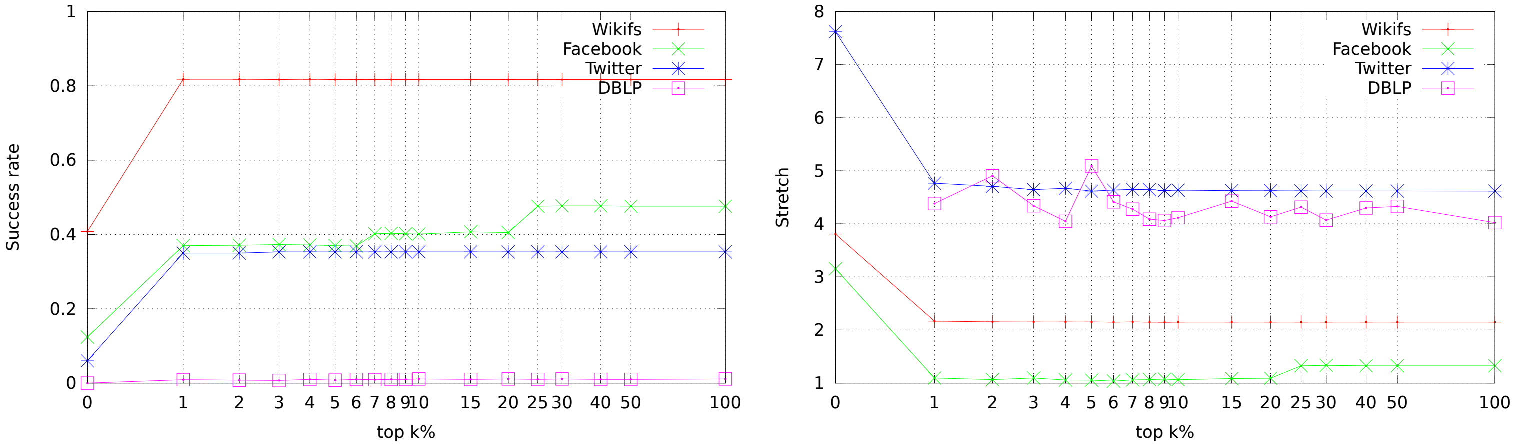
In the following, we provide the results of our empirical evaluation. For both node selection strategies presented in Section 6.3.2, the decentralized search algorithm navigates greedy by degree. The amount of information needed for efficient navigation depends on the type of the structural information used and differs in the distinct networks.

Popularity strategy results. First, the popularity strategy tries to identify important nodes based on their popularity. Figure 6.6 shows the success rate and stretch for this strategy in all networks. In this case, the algorithm achieves with just 1% of the nodes similar efficiency results as with 100%. For Facebook, the partially informed decentralized search achieves slightly worse results than the fully informed search already with 2% for navigation by degree and the same or even a bit better results with 25% of the nodes. For this setting, the algorithm achieves similar performance as the fully informed decentralized search for Wikipedia for schools and Twitter also already with 1% of the nodes. We see that navigation in DBLP is very difficult in general. The best results in this network are realized with 2-3%.

Clustering strategy results. The clustering strategy tries to identify structurally important nodes based on their clustering coefficient. Figure 6.7 shows the success rate and stretch for greedy navigation by degree. We see that for Wikipedia for schools and Twitter the success rate initially falls with increasing amount of information, and then it jumps to the level of the fully informed search at 2% and 6% for Wikipedia for schools and Twitter, respectively. For Facebook, we observe very interesting success rate values since we are able to achieve considerably better results with less structural information. The success rate grows from 1% to 6% of the nodes to a value higher than the value achieved by the fully informed search (100%). After a drawback between 6% and 9% of the nodes, the success rate achieves even better results than for 6% with 15% of the nodes. Using more than the top 15% of the nodes worsens the success rate to the level of fully informed search. As before, we can see that navigation in DBLP is also very difficult with this strategy. The best results in this network are realized with 30% of the structural information.

Summary and findings. Next, we summarize the results in the following two main findings answering the research questions tackled throughout this work as proposed in Section 6.1.

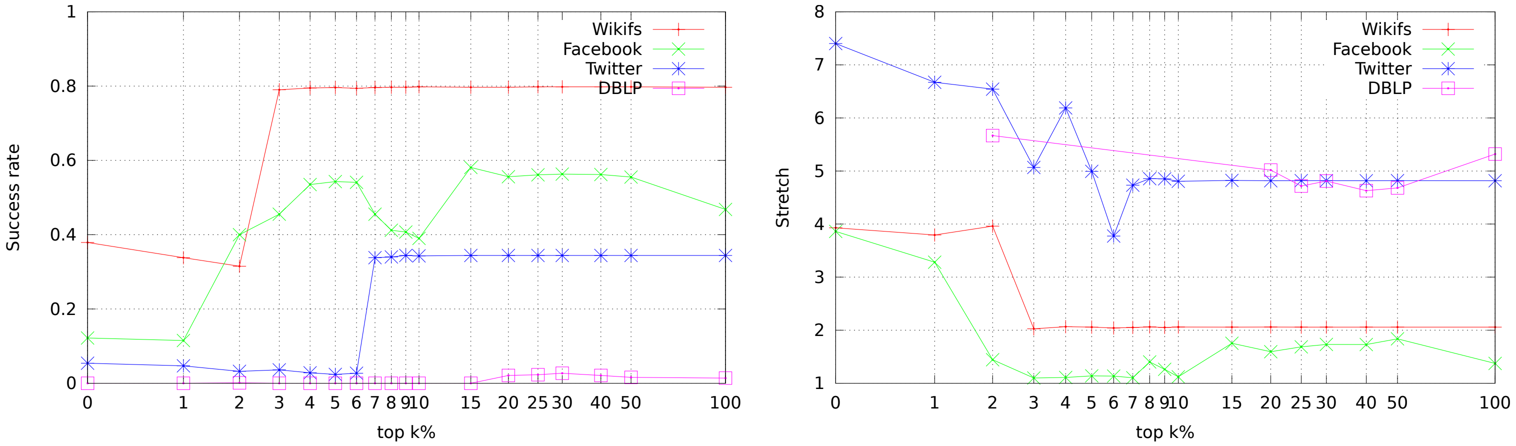
(i) What kind of structural information is needed for efficient navigation? Strongly outperforming the fully informed search with the popularity strategy is not possible. With increasing amount of information about the popularity, the success rate and the stretch improves continuously. With the clustering strategy, it is partly possible to substantially outperform the fully informed search. There is an initial drawback in success rate and stretch



(a) Success rate

(b) Stretch

Figure 6.6: Success rate (s) and stretch (τ) for popularity strategy for different amount of information. Left (a): The success rate achieved for the popularity strategy and degree as popularity metric—the higher the better. To improve readability, we added one to all values and logarithmically scaled the x axis which shows the amount of information used. Right (a): The stretch achieved for the popularity strategy and degree as popularity metric—the lower the better. To improve readability, we added one to all values and scaled the axes logarithmically. We can see that we can achieve the success rate and stretch levels of fully informed search already with very small amount of information—about 1-2%. Strongly outperforming the fully informed search is not possible with this strategy.



(a) Success rate

(b) Stretch

Figure 6.7: Success rate (s) and stretch (τ) for clustering strategy for different amount of information. Left (a): The success rate achieved for the clustering strategy and degree as popularity metric—the higher the better. To improve readability, we added one to all values and logarithmically scaled the x axis which shows the amount of information used. Right (b): The stretch achieved for the clustering strategy and degree as popularity metric—the lower the better. To improve readability, we added one to all values and scaled the axes logarithmically. We can see that for achieving the success rate and stretch levels of fully informed search (100%) we need slightly more information with the clustering strategy compared to the popularity strategy presented in Figure 6.6. Nonetheless, with this strategy, we are also able to outperform the fully informed search in some networks by only utilizing a low amount of clustering information.

in all networks with the clustering strategy. After this initial drawback the success rate and the stretch increase until the levels of the fully informed search or even outperform the fully informed search.

Our results suggest that nodes with high popularity and low clustering are very important and can guide navigation very well and thus, should be exposed to the user through the interface.

(ii) How much structural information is needed for efficient navigation? With the popularity strategy, the levels of success rate and stretch produced by the fully informed decentralized search are achieved already with 1% of the popularity information. With the clustering strategy, the levels of success rate and stretch produced by the fully informed search are achieved with a bit more information than with the popularity strategy, depending on the network.

Our results suggest that with intelligent selection of nodes based on their structural properties, we can significantly reduce the amount of information that is needed to be presented to the user in navigational interfaces without reducing the efficiency of navigation.

6.6 Discussion

In Section 6.6.1, we start with a discussion and interpretation of our results (*cf.* Section 6.5) tailored around the research questions posed in Section 6.1. In Section 6.6.2, we discuss the implications followed by an elaboration of the advantages and limitations of our approach in Section 6.6.3.

6.6.1 Discussion and Interpretation of Results

Quality of Structural Information—Popularity vs. Clustering. Ranking the nodes by popularity and clustering is a good way to identify structurally important nodes. Furthermore, if the popularity information is combined with small amounts of clustering information which is a local metric, we can navigate even more efficiently. Nodes with high popularity and low clustering are very important and can guide navigation very well and should be exposed to the user through the interface. Knowing the important nodes on the local level regarding popularity and clustering can result in reducing the amount of nodes that need to be exposed to the user. This way we would be able to relax constraints of the screen size [HS11]. The initial drawback in the performance of the algorithm for this strategy can be explained by the degree distributions of the informed set of nodes. If the set is too small, there are not enough nodes with high popularity. Once the informed set has a sufficient amount of nodes for which the user has an intuition not only about the popularity but also about the clustering coefficient of the nodes, the user can navigate more confidently to the target.

Amount of Structural Information—Partially vs. Fully Informed Search. Surprisingly low amount of structural information is needed to achieve the same or even better results than with all information. This finding is really surprising if we consider the level of inequality in the degree distributions suggested by the Gini index and the exponent of the power law degree distribution (*cf.* Table 6.2 and Table 6.3). For the popularity strategy, outperforming the fully informed search is not possible, whereas for the clustering strategy we are able to top the results produced by the fully informed decentralized search.

6.6.2 Implications

Navigation in online networks is supported by smart user interfaces like tag clouds, breadcrumbs, subcategories and related categories. Normally, these navigational user interfaces make use of algorithmically preprocessed information about the content of the network. Our results have direct implications for these algorithms and for the ways data is presented to the user through the navigational interfaces.

Rethinking algorithms. Our findings suggest to reorganize the way we build hierarchies and to rethink algorithms creating hierarchies like [HGM06b, HS11, PLG10b, DFG01b, Zho05]. In [HS11], the authors showed that the ability of hierarchies to guide navigation is significantly reduced through the restrictions introduced by the user interfaces. The main problem identified by the authors was that the top level of the hierarchies produced by the algorithms have too many subcategories—*i.e.*, a too high *branching factor*. To tackle this problem, they adapted one of the best known algorithms for hierarchy induction proposed by Heymann and Garcia-Molina [HGM06b]. This algorithm creates a hierarchy by producing a similarity network. The hierarchy is then developed by ranking the nodes in the similarity network by popularity. Nodes are then placed in the hierarchy in a descending order of their popularity and their similarity to nodes that are already in the hierarchy. This way, nodes with high popularity are placed in the top of the hierarchy and nodes with low popularity at the bottom. Our results suggest an alternative ranking for the nodes of the similarity network. The results of the popularity strategy suggest that we should concentrate on the top 1% of the nodes in the network and try to produce a hierarchy with a well structured top. In contrast, the order of the bottom levels of the hierarchy is not really important, since we are able to achieve the same efficiency in navigation with only 1% of the nodes. This result also suggests that even if we break the semantics in the low levels of the hierarchy, we still will be able to navigate efficiently. Hints of how we should reorganize the top levels of the hierarchy are given by the clustering strategy we presented. We can re-rank the nodes of the similarity network considering not only their popularity, but also their clustering coefficient.

Our result could also be applied to the adapted version of the algorithm by Heymann and Garcia-Molina [HGM06b] proposed by Helic and Strohmaier [HS11] which generates a hierarchy in two stages. First, it produces hierarchies with a given branching factor. The largest hierarchy is called the main tree and all other hierarchies are then added to the main tree. After sorting the hierarchies by size, they are attached to the main tree in a way that preserves the branching factor of the hierarchy. Here, we could again try to re-rank the most popular nodes also by their clustering and put them in the main tree as suggested by the clustering strategy.

Presentation and information scent. Our results suggest that for efficient navigation, only a very small amount of local popularity and clustering information is necessary. Thus, we can derive that for efficient navigation, the user needs to have a good intuition only about the most important nodes in the network. Exposing the nodes with high structural importance through the user interface does not ensure that the user is going to select them. If the user has no sufficient knowledge and understanding of the most important nodes, the system has to deliver the explanation and in this way strengthen the *information scent* of the user for these specific nodes [Pir97, PC99]. By providing additional information about the important nodes regarding popularity and clustering, the information system would help the user to create an intuition about the presented choices. Assuring that a user has a high understanding about the topology of the information space—high exposure to the structurally most important nodes—would allow us to reduce the actual amount of nodes that are presented to the user through the interfaces. Without such information, random navigation performs well, which is consistent with previous results for navigation by popularity in power law networks [ALPH01].

Helic *et al.* [HTSA10] studied the navigability of social tagging systems and showed that the tagging networks are power law networks. They showed that limiting the tag cloud size to practically feasible sizes (*e.g.*, 5%,10%) does not affect the navigability. Our results suggest that we can reduce the tag cloud size even further to 1% of the nodes, according to the popularity strategy. In the same work, the authors also provided theoretical and empirical arguments against existing approaches of tag cloud construction. Possible improvements of these approaches can be achieved for instance with alternative rankings considering the clustering of the tagging network as the results of the clustering strategy presented.

6.6.3 Advantages and Limitations

In the following, we would like to address some limitations and advantages of our work.

Correlation between strategies. It has been shown that networks might exhibit a negative correlation between the degree and the clustering co-

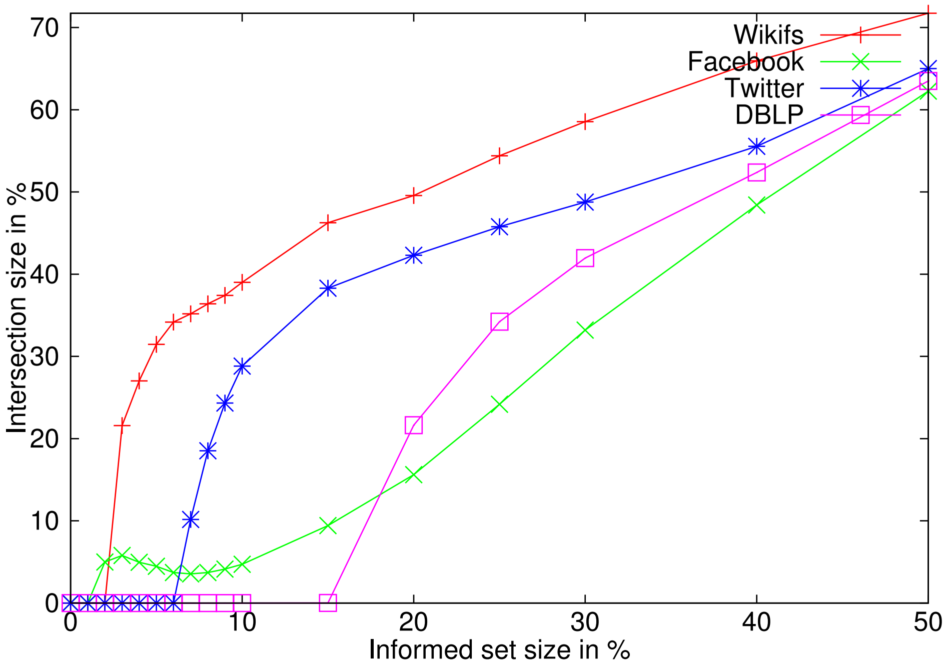


Figure 6.8: Informed sets overlap. The intersection size in percentage for the popularity and clustering strategy for the different informed sets sizes (except for 100%, $I = V$) of nodes used in the experiments in Section 6.3.3. For the datasets at hand, there is no big overlap in the informed sets selected by the two strategies. Additionally, with increasing set size the overlap does not necessarily increase as so the efficiency of navigation (*cf.* Figure 6.6 and Figure 6.7).

efficient of nodes based on the formal definition of the clustering coefficient [SV05]. Due to this negative correlation, it is possible that there is big overlap in the informed sets created by the popularity and clustering strategy in this work. That is why it is important to quantify up to which extent the two strategies for important node selection differ in the experiments conducted in Section 6.3.3. In Figure 6.8, we illustrate the size of the intersection of the popularity and clustering strategies for all datasets for different sizes of the informed sets of nodes. Overall, we can see that the overlap of nodes between the two strategies is considerably low for smaller set sizes. Not surprisingly, with increasing set size, the overlap is generally rising as the chance of overlapping node selection increases. However, as it can be seen in Figure 6.6 and Figure 6.7, an increasing overlap does not reflect an increase in the performance of the partially informed decentralized search. By and large, these observations support the importance of the findings from Section 6.5 and give a confirmation that both strategies select mostly different nodes and structurally important nodes that could support navigation.

However, there might exist some few nodes that are highly beneficial to be included in an informed set for efficient navigation. Both strategies might select them early on and as soon as they include these nodes, efficiency increases drastically. Thus, in future work, we plan on further investigating the overlaps between both strategies which might also help us to find even better (potentially smaller) informed sets that can guide navigation well.

Alternative strategies for node selection. With our experiments, we have concentrated on the degree and clustering coefficient as metrics for measuring the structural importance of nodes. Above, we have discussed the potential correlation between both strategies but have also shown that the overlap is low for small informed sets. Nonetheless, other strategies might be amendable. For example, previous work [SV05] has suggested an alternative way to calculate the clustering coefficient by removing the degree bias (*cf.* Equation 6.1). By utilizing this method, we might be able to further investigate the differences of both strategies for finding important nodes for navigation. Also, we could simply try to implement strategies for node selection that produce mostly distinct sets of nodes. By doing so, we might be able to further improve our approach potentially leading to even better results in terms of success rate and stretch. Nodes of the distinct sets selected by the strategies can then be exposed to the user through the navigational interfaces.

Also, there exist other thinkable metrics (*e.g.*, k-core and link irregularity) describing the structure of a network that can be applied in straightforward fashion [DGM06, Est10]. We leave these investigations open for future work.

Alternative user models. In our experiments we utilized a greedy neighbor selection if at least one of the candidate nodes is in the informed set of nodes otherwise we selected one at random. This models a user who always follows her intuition if it has one. Although this is a valid user model, it is a very simple one. In future work, we plan to experiment with alternative neighbor selection mechanisms that model a user who is greedy or stochastic to different extents in following her intuition [HSGS13]. Additionally, it is also thinkable to use different informed sets at different stages of the search *e.g.*, the informed set created with the popularity strategy can be used in the beginning of the search where the user is interested in exploring the information space, whereas the informed set created with the clustering strategy can be applied in stages of the search where bridging a gap between two clusters is needed.

Global information. One limitation of the decentralized search is the amount of global information used for navigation. The models presented by Watts *et al.* [WDN02], Kleinberg [Kle00b, Kle00a] and Boguña *et al.* [BKC09] make use of the global position of the target node. One could argue that partially informed decentralized search is using to much global information in the sense that it uses the information about the distribution of the degree and clustering coefficient. A way to tackle the problem would be to make a random sample of n% (*i.e.*, 30%) of the nodes and apply the popularity and the clustering strategies only at these n% of the nodes in the network.

6.7 Conclusion

Navigational interfaces are only useful for augmenting navigation to the extent to which they are able to expose the underlying structure of the information space. In this chapter, we have been interested in studying (i) which and (ii) how much structural information is necessary for properly exposing the hidden structure of the information space. To that end, we have utilized an adapted version—*i.e.*, partially informed—of the message passing decentralized search algorithm. This adaption allows to model a user who is limited in her exposure to the structure of the information space having only limited knowledge about the topology of the information space. In detail, we have focused on two strategies for selecting important nodes based on their (i) popularity and (ii) clustering coefficient.

With simulations on four distinct datasets, we have observed that a surprisingly low amount of structural information is needed by the partially informed version of decentralized search in order to achieve the same or even better performance than the fully informed decentralized search.

Besides the popularity, for choosing structurally important nodes, also the clustering coefficient has turned out to be a good indicator for this task. The clustering strategy would expose nodes of high structural importance to users, and thus reduce the amount of information that has to be presented to users and relax screen size constraints. Our results have implications on the algorithms used for the structuring of the information space. These algorithms should take into account the supportive properties of the clustering coefficient for navigation.

In future work, we would like to propose and evaluate another version of decentralized search that models exploitation on the local level. In this version, we plan on combining centrality metrics as proxies for popularity and clustering information as a proxy of homophily. With this extended version, we would like to study how the clustering coefficient can be used to jump from one network region to another or to stay in the same cluster and explore it.

Chapter 7

Conclusions

Before the digital age, information was predominantly consumed in a linear manner. The introduction of technologies such as Hypertext and the World Wide Web has drastically changed our reading behavior. The key feature of these technologies—the possibility to interlink text passages— allows readers to decide how they would like to experience connections between individual pieces of information, leading to the emergence of new information consumption patterns. Consequently, consuming information by navigating through large information networks such as Wikipedia and the Web in general has been investigated with excitement by the research community. This interest is in part due to the topic's broad application spectrum, ranging from improving network structure through ranking and recommendations to user interface design and website layout enhancement. However, the majority of previous studies and proposed navigational hypotheses are limited to the amount and nature of the data. Moreover, these consider navigation independently from search—the other dominant information seeking strategy—and not as a building block of the search-vs-navigation ecosystem. Therefore, the aim of this thesis has been to advance towards a comprehensive *understanding and modeling of navigational user behavior in information networks* with the goal of paving the way for improvement and maintenance of web-based information systems such as Wikipedia. This work approaches this objective by *analyzing large-scale openly available data capturing information consumption* within the English version of Wikipedia. First, I focus on the preferred user information access form, *i.e.*, search or navigation, with respect to Wikipedia article characteristics. By introducing searchshare and resistance, this thesis *illustrates how articles attract external traffic and transform it into internal navigation on Wikipedia.* Further, it shows how both search and navigation shape the Wikipedia article popularity distribution, which is dominated by search for articles at the top of the distribution and by navigation for articles at the tail of the distribution. Second, this work *validates and builds upon existing navigational hypotheses* *to create a "reasonable surfer" version of the PageRank algorithm.* This version is based on empirical evidence that users tend to select links leading to the periphery of the Wikipedia network and located at the top and left side of the screen. Finally, this thesis introduces a *partially informed version of decentralized search, and applies it in order to simulate agent-based navigation.* By analyzing the success of agents that navigate with various levels of topological network knowledge, this work highlights the amount and type of structural knowledge needed for efficient navigation. To conclude, I provide an overview of the results and contributions of the present work, address its limitations, and briefly elaborate on promising directions for future research.

7.1 Results and Contributions

By addressing each research question as posed in Section 1.4, the following section summarizes the contributions made by this thesis.

RQ1: How do user reading preferences shape the external and internal traffic on Wikipedia? The content diversity and richness of web-based information systems requires users to access content by both search and navigation. While both search and navigation have been thoroughly studied in related work, they have been thus far mostly considered independently. Consequently, little knowledge has been gained regarding which elements and content types of any specific website (inter)act in which structural roles, resulting in different information access patterns. Understanding the user access patters and interplay between search and navigation can highlight situations in which web-based systems fail to support users. To this end, this thesis asks the question: How are user reading preferences shaping the external and internal traffic on Wikipedia? This question can be divided into two sub-questions, as posed in Chapter 3: (i) How do search and navigation interplay to shape the article traffic on Wikipedia? Given any article, this thesis investigates how its role as a search entry point is related to (not) relaying navigation traffic into the Wikipedia network, and vice versa. This also addresses the issue of how search and navigation contribute to an article's popularity. Beyond these characteristics of a web-based information system in general, this work also examines which specific properties of articles influence their roles in the search-vs-navigation ecosystem. In other words: (ii) Which article features (*i.e.*, topic, network, content and edit features) are indicative of specific information access behavior? To answer these questions, this work introduces two metrics that capture individual article traffic behavior, *i.e.*, (i) searchshare—the amount of views an article received by search—, and (ii) resistance—the ability of an article to channel traffic into and through Wikipedia. The main contributions can be summarized as follows: (i) Regarding general (collective) access behavior on Wikipedia, this work provides empirical evidence that for the most-viewed articles, search dominates navigation in both the number of articles accessed and the number of views. However, navigation appears to become more and more important for the tail of the views distribution. (ii) This thesis links article properties such as position in the Wikipedia network, number of article revisions, and topics to preferred access behavior, *i.e.*, search or navigation. Finally, (iii) this work has also quantified the strength of the relationship between article properties and preferred access behavior. This work argues that (i) search and navigation are both used to access and explore different articles, and are thus both crucial information access types for Wikipedia. (ii) Articles representing exit points of navigation sessions are located at the periphery of the network, whereas entry/relay points articles are located an the core. Further, (iii) edit activity is strongly related to the ability of an article to relay traffic. These results have various potential applications, *e.g.*, improving and maintaining the visual appearance and hyperlink structure of articles and identifying articles exhibiting changes in access behavior patterns, for example due to vandalism or other online misbehavior. This analysis represents a first step towards a better understanding of how search and navigation interplay to shape user access behavior on platforms such as Wikipedia and other web-based systems. Additionally, it provides helpful insights into the development of new navigational hypotheses, an issue addressed with the next research question.

RQ2: What makes a link successful on Wikipedia? The main goal of web-based systems such as Wikipedia is to provide appropriate information access to users. This depends on proper content presentation, efficient link structures, as well as ranking algorithms such as PageRank. Understanding which link features affect user click behavior can be beneficial not only for improving PageRank, which disregards potential user click preferences and assumes random click behavior, but also for maintaining the constantly growing link structures of Wikipedia's network. To this end, this work questioned what makes a link successful on Wikipedia. This question has two dimensions, studied using large-scale link transitions data. The first dimension addresses competition between links with different target articles (*cf.* Chapter 4), whereas the second dimension deals with competition between links with the same target article (*cf.* Chapter 5). The main findings can be summarized as follows: This thesis provides empirical evidence that Wikipedia users (i) select links leading to the periphery of the underlying topological link network, (ii) prefer links leading to semantically similar articles, and (iii) are more likely to traverse links positioned at the top and left side of the screen. Additionally, this thesis integrates these findings with first-order Markov chain models to demonstrate improvement of respective model fits by incorporating the assumed behavior of human navigation into the inference process. Finally, the utility of these findings is demonstrated by adapting PageRank to better reflect human navigation behavior. The methodological framework provided in this work can easily be applied to transitional click data from any web-based information system.

RQ3: What kind of and how much network structural information is needed for efficient navigation? After studying information access at both the link and article level, the third and final research question addresses the network as a whole with reference to what kind of and how much network structural information is needed for efficient navigation. To tackle this question, I introduced partially informed decentralized search, allowing for the modeling of users with limited knowledge of the information network structure (*cf.* Chapter 6). The role of the available structural information is then evaluated by varying the number (amount of information used for navigation) and type (structural properties, *i.e.*, degree and clustering coefficient) of nodes for which a navigating agent can take an informed decision. The results obtained from an agent-based simulation of four different networks suggest that efficient navigation is possible even with a minimal amount of knowledge of the network topology. Overall, this finding indicates that the information scent of high degree and low clustering nodes must be strengthened when designing navigational interfaces. Moreover, this finding suggests the redesign of hierarchical clustering algorithms for structuring an information space in such a way that both popular and low clustering nodes are placed an the top hierarchy levels to ensure high navigability.

7.2 Implications and Applications

Gaining a better understanding of users' navigational behavior in information networks is indispensable for the improvement of numerous aspects of web-based information systems. With this work, I have addressed this challenge by building navigational models using data pertaining to information consumption in the Wikipedia context, one of the largest and most visited information systems on the Web. I am convinced that the findings of the thesis, which have mainly been obtained through empirical analyses, could be beneficial to the Wikimedia Foundation as well as inspirational for future research on Web navigation in general. In this section, I provide a short discussion of the potential implications and applications of the insights presented in this work.

Interface design and webpage layout. Several empirical findings have been made throughout this thesis with implications for user interface design and page layout. For example, by highlighting links to articles with low resistance values, the appearance of Wikipedia pages acting as entry points can be modified to provide further navigational guidance within the system, and thus retain visitors (*cf.* Section 3.3). Exit point pages, on the other hand, can be assigned to editors for improvement. Shortcomings in their layout can be identified by creating article heatmaps showing the discrepancy between the positions of links and clicks. Moreover, as different topics exhibit varying access behavior, the Wikipedia community might consider developing topic-specific article layouts.

Results from Section 6.5 suggest that only a modest amount of local popularity and clustering information is necessary for efficient navigation. This result has implications for user interface design, as it suggests that users must possess a good intuition concerning only the most important nodes in the network. Identifying and exposing nodes with high structural importance through the user interface does not, however, ensure that a user will select them. Therefore, the information system must deliver an explanation, and thus strengthen users' information scent for these specific nodes. For example, by providing additional information about important nodes or by highlighting them to indicate their popularity and clustering coefficients, an information system would support users in exercising their intuition regarding the presented choices and, eventually, match them to his/her motivation and information needs. Assuring that users sufficiently understand the information network topology—have a high exposure to the structurally most important nodes—would help reduce the actual amount of nodes presented to a user through the interface to ultimately improve user experience.

Re-thinking clustering algorithms. The empirical results obtained in Section 6.5 provide an intuition about how to build hierarchies suitable for guiding navigation (*cf.* Section 2.4). For example, the popularity strategy suggests that we try to produce hierarchies with a well-structured top containing the top 1% of the most popular nodes in the network. In contrast, the order of the bottom levels of a hierarchy is less important. This has implications for clustering algorithms accounting for semantic relationships between the nodes in a hierarchy. For such algorithms, these findings suggest that even if we deviate from the semantics in the lower levels of the hierarchy, it will still efficiently guide navigation. Further intuitions about the population of the upper levels of a hierarchy are given by the presented clustering strategy. The goal here could be to place nodes with lower clustering coefficients into the upper hierarchy levels when building a hierarchy, ensuring lower hierarchy distances between network clusters.

Content ranking and search personalization. The navigational hypotheses introduced in this thesis have applications related to content ranking and search personalization. As shown in Section 4.5.2, one could improve the ranking of Wikipedia content by creating a "reasonable surfer" version of PageRank in a straightforward manner by weighting the network edges with the best-performing hypothesis (*kcore+visual*). However, other scenarios are also possible. For example, by analyzing individual user sessions, one can create a personalized PageRank version for each user. While this approach would provide search personalization, it could also have negative consequences, such as a potential filter bubble effect. The hypotheses presented in this thesis follow certain general intuitions and represent *collective* navigational behavior. As such, they are suitable for achieving personalization while reducing filter bubble effects. To this end, one can first cluster user sessions by using the numerous navigational hypotheses presented in this thesis as cluster centroids. A version of PageRank based on the centroid hypotheses can then be assigned to each user in the respective cluster in order to improve Wikipedia search results.

7.3 Limitations

The results and insights of this thesis exhibit the following limitations that I would like to discuss here.

- **Generality of results and findings.** This thesis has concentrated on Wikipedia due to its central role in the Internet's infrastructure, its tremendous amount of users, and openness, making it an excellent learning and vibrant research environment. Although these characteristics establish Wikipedia as a small, idealized example of the World Wide Web, the empirical results and insights presented in this thesis are not necessarily directly transferable to other web-based information systems, or to the Web in general. Nonetheless, I still hope that the overall approach of this thesis will be further utilized with other datasets, *e.g.*, on other language editions of Wikipedia or datasets from other web-based information systems, in order to enrich our understanding of users' navigational behavior in information networks.

- **Data limitations.** The clickstream data used in this work capture information consumption on Wikipedia on a large scale; however, only at an aggregated level. Any results obtained by analyzing this data are therefore only indicative of *collective* user behavior on Wikipedia. While this was suited for validating the presented set of hypotheses, it is not possible to differentiate between potentially varying phases of navigation sessions, *i.e.*, zoom-out and zoom-in phases. Moreover, due to privacy restrictions, the data only contain (referrer, resource) pairs that occurred at least ten times during the respective data collection period. This could lead to a skewed view of user access behavior. For example, if an article is navigated in total much more than ten times over different links, but each individual link is transitioned less than ten times, these transitions will not be included in the data. This data limitation is especially relevant when contrasting search and navigation, as the searchshare for an article might be substantially overestimated (*cf.* Section 3.3). Moreover, this might occur specifically for articles with overall meager page views and explain some findings suggesting that, *e.g.*, articles in the periphery of the link network show a stronger search prevalence. At the same time, we observe that users tend to select links leading to the periphery of the information network, suggesting that in general, peripheral articles (potentially articles about specific concepts that are not common knowledge) are interesting to users.

- **Definition of link success.** In this thesis, the success of a link has been defined as the click frequency derived from user transitions. However, other forms of link success can also be studied. For example, instead of explaining the transition preference of outgoing links, one could modify the research question to observe incoming links. Consequently, we might ask: Given a set of incoming links to a Wikipedia article, which are the most popular, and how can we explain this? The success of a link could also be defined in light of the degree to which a single link contributes to each type of navigation. Link success could also be examined in terms of a link's ability to enable and encourage further navigation, and thus expose users to richer content or to the strongly connected network component. Further research could also concentrate on studying the relationship between the strength of the information scent and the success of a link.

- **Better combinations of existing, and development of further, hypotheses.** The results of Section 4.5.2 depend on how we code the hypotheses as priors. While the presented hand-crafted hypotheses are in-line with the insights gained from Section 4.4.2, better-engineered combinations of favorable priors and weighting schemes that could even further improve the results may exist. For example, the narrow textual context of a link is strongly related to the strength of the information scent, and can thus also be used to propose navigational hypotheses. Additionally, one might argue that some results from Section 4.4.2 could have been expected given our previous findings in a different context and with a different methodology. However, the main goal was to demonstrate the utility of incorporating assumptions about navigational user behavior into existing models of navigation (*i.e.*, Markov chain and PageRank) that predominantly assume uniform behavior.

7.4 Future Work

Finally, to conclude this thesis I would like to give an overview of some potential future research directions worth pursuing, which have been inspired in part by the limitations presented in the previous section.

- **Analyzing individual user sessions.** In order to better understand individual user behavior, one could analyze entire user sessions. This would allow for the testing of hypotheses regarding zoom-out and zoom-in phases of navigation (*cf.* Section 2.5.2). The findings of the present work suggest that, at least for some topics, the zoom-out phase might have been completely eliminated, *e.g.*, through search (*cf.* Section 3.4.2), which raises the question as to whether the zoom-out/zoom-in navigation pattern observed in navigational games data is also present in real Wikipedia user sessions.

- **Cultural differences in information consumption.** The results and findings of this work are based on analyses of data pertaining to user behavior on the English language Wikipedia. However, behavior may vary across language editions; consequently, studying cultural differences across language editions of Wikipedia would be a natural continuation of this line of research. For example, our results concerning the visual position of clicks (*cf.* Section 4.4.2), *i.e.*, users predominantly clicking on links located on left side of the screen, may not be valid for the Arabic or Farsi Wikipedia versions, as both languages have an inverted reading/writing direction (right to left) compared to English (left to right). Further cultural differences may also be related to the preferred presentation of information, *i.e.*, tabular organization of content, running text, and the number of pictures per article. Moreover, the network structure of different language editions may vary substantially, resulting in disparate traffic patterns within the Wikipedia network. Another interesting direction would be to examine to what extent transitions between different language editions occur. Is there an exceptional behavior in that regard, *i.e.*, do unexpectedly high/low number of transitions between two language editions occur over a low/high number of cross-language links?

- **Approximating eye tracking data using click data.** Eye tracking has been extensively applied in the field of human-computer interaction and usability research to study how a given webpage is perceived by users. Although eye tracking studies provide valuable, fine-grained data, they are costly to conduct (w.r.t. time), complicated in their execution, and it often proves hard to recruit participants. Navigational click data could be used as a substitute for eye tracking data in numerous contexts. Moreover, such data is relatively inexpensive to collect while it captures "real world" user behavior. Knowing how well click data can approximate eye tracking data and in which cases such data is a valid substitute would accelerate research of both human-computer interaction and user interface design.

- **Vandalism detection and navigational user behavior over time.**

Detecting vandalism on Web platforms such as Wikipedia is an important topic that is currently gaining more and more attention due to the rising amount of misbehavior on the Web. Wikipedia represents a valuable and trustworthy information source that has been recognized by other online platforms such as YouTube, which recently announced to link Wikipedia content to conspiracy theory videos in order to fight misinformation[1]. Although this is a noble initiative, such direct linkage could potentially lead to an increase of vandalism targeted at Wikipedia articles. Identifying disruption in the navigability of an information network by observing changes in users' navigational behavior over time is another line of research that shows potential. This is especially valid for the case of sensitive articles and topics such as conspiracy theories, as vandalism might not be apparent at first glance, *i.e.*, not through aggressive deletion of text and links, but rather through subtle textual and re-linking changes that attempt to violate Wikipedia's article neutrality principle.

- **Mobile usage.** During the composition of this thesis, Wikimedia Foundation launched a mobile Wikipedia App. Complementing the analyses of this work using transitional data that capture mobile Wikipedia user behavior could be another issue of special interest. The following questions arise in this context: How do search and navigation interplay on mobile devices? Is mobile usage different from desktop usage, *e.g.*, in terms of topic popularity? Can desktop eye tracking data be even better approximated by mobile usage data? Are there cultural differences between the mobile usage of different Wikipedia language editions?

Research on navigation in information networks has made significant progress since the breakthrough of Hypertext and the Web. With this thesis, I hope to have highlighted promising avenues and directions for future research in this exciting field.

[1] `https://www.wired.com/story/youtube-will-link-directly-to-wikipedia-to-fight-conspiracies/` 2018-04-17

List of Figures

List of Tables

Bibliography

[AA05] Lada Adamic and Eytan Adar. How to search a social network. *Social Networks*, 27, 2005.

[ABD06] Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *Int. Conference on Research and Development in Information Retrieval*, 2006.

[ABP14] Jeff Alstott, Ed Bullmore, and Dietmar Plenz. powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE*, 9, 2014.

[AdR05] Sisay Fissaha Adafre and Maarten de Rijke. Discovering missing links in wikipedia. In *Proceedings of the 3rd international workshop on Link discovery*, pages 90–97. ACM, 2005.

[AH00] Lada A Adamic and Bernardo A Huberman. Power-law distribution of the world wide web. *Science*, 287(5461):2115–2115, 2000.

[AHCSZ06] Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*, 2006.

[AL08] Richard Atterer and Philip Lorenzi. A heatmap-based visualization for navigation within large web pages. In *Proceedings of the 5th Nordic conference on Human-computer interaction: building bridges*, pages 407–410, 2008.

[ALPH01] Lada Adamic, Rajan Lukose, Amit Puniyani, and Bernardo Huberman. Search in power-law networks. *Physical Review E*, 64, 2001.

[ATD08] Eytan Adar, Jaime Teevan, and Susan T Dumais. Large scale analysis of web revisitation patterns. In *Proceedings of the*

SIGCHI conference on Human Factors in Computing Systems, pages 1197–1206. ACM, 2008.

[ATD09] Eytan Adar, Jaime Teevan, and Susan T Dumais. Resonance on the web: web dynamics and revisitation patterns. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1381–1390. ACM, 2009.

[AZN99] David W Albrecht, Ingrid Zukerman, and Ann E Nicholson. Pre-sending documents on the www: A comparative study. In *IJCAI*, pages 1274–1279, 1999.

[BA99] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[Bat89] Marcia J Bates. The design of browsing and berrypicking techniques for the online search interface. *Online review*, 13(5):407–424, 1989.

[BB08] Arindam Banerjee and Sugato Basu. A social query model for decentralized search. In *Proceedings of the 2nd Workshop on Social Network Mining and Analysiss. ACM, New York*, volume 124, 2008.

[BC06] Luca Becchetti and Carlos Castillo. The distribution of pagerank follows a power-law only for particular values of the damping factor. In *Proceedings of the 15th international conference on World Wide Web*, pages 941–942. ACM, 2006.

[BCM09] Georg Buscher, Edward Cutrell, and Meredith Ringel Morris. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 21–30. ACM, 2009.

[Bes96] Azer Bestavros. Speculative data dissemination and service to reduce server load, network traffic and service time in distributed information systems. In *Data Engineering, 1996. Proceedings of the Twelfth International Conference on*, pages 180–187. IEEE, 1996.

[BGN08] Scott Bateman, Carl Gutwin, and Miguel Nacenta. Seeing things in the clouds: The effect of visual features on tag cloud selections. In *Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia*, HT '08, pages 193–202, New York, NY, USA, 2008. ACM.

[BGS05] Monica Bianchini, Marco Gori, and Franco Scarselli. Inside pagerank. *ACM Transactions on Internet Technology (TOIT)*, 5(1):92–128, 2005.

[BH03] Gary D Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):1, 2003.

[BHSS10] Dominik Benz, Andreas Hotho, Gerd Stumme, and Stefan Stützer. Semantics made by you and me: Self-emerging ontologies can capture the diversity of shared knowledge. In *Proceedings of the 2nd Web Science Conference*, WebSci '10, 2010.

[BKC09] Marian Boguna, Dmitri Krioukov, and Kimberly C Claffy. Navigability of complex networks. *Nature Physics*, 5(1):74–80, 2009.

[BKM$^+$00] Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. Graph structure in the web. *Computer networks*, 33(1):309–320, 2000.

[BL00] Jose Borges and Mark Levene. Data mining of user navigation patterns. In *Web usage analysis and user profiling*, pages 92–112. Springer, 2000.

[BL11] Lars Backstrom and Jure Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 635–644. ACM, 2011.

[BLFFBD00] Tim Berners-Lee, Mark Fischetti, and Michael L Foreword By-Dertouzos. *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. Harper-Information, 2000.

[BLS$^+$16] Martin Becker, Florian Lemmerich, Philipp Singer, Markus Strohmaier, and Andreas Hotho. Mixedtrails: Bayesian hypotheses comparison on heterogeneous sequential data. *arXiv:1612.07612*, 2016.

[BMB$^+$14] Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, et al. lme4: Linear mixed-effects models using eigen and s4. *R Package Version*, 1(7), 2014.

[BMH$^+$16] Martin Becker, Hauke Mewes, Andreas Hotho, Dimitar Dimitrov, Florian Lemmerich, and Markus Strohmaier. Sparktrails: a mapreduce implementation of hyptrails for comparing hypotheses about human trails. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 17–18. International World Wide Web Conferences Steering Committee, 2016.

[BMK97] Rob Barrett, Paul P Maglio, and Daniel C Kellem. How to personalize the web. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 75–82. ACM, 1997.

[BNJ03] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[BP98] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Computer networks and ISDN systems*, 30(1-7):107–117, 1998.

[BP12] Sergey Brin and Lawrence Page. Reprint of: The anatomy of a large-scale hypertextual web search engine. *Computer networks*, 56(18):3825–3833, 2012.

[BRCA09] Fabrício Benevenuto, Tiago Rodrigues, Meeyoung Cha, and Virgílio Almeida. Characterizing user behavior in online social networks. In *Internet Measurement Conference*, 2009.

[Bri90] John S Bridle. Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer, 1990.

[BS03] Randolph E Bucklin and Catarina Sismeiro. A model of web site browsing behavior estimated on clickstream data. *Journal of Marketing Research*, 40(3):249–267, 2003.

[BS09] Randolph E Bucklin and Catarina Sismeiro. Click here for internet insight: Advances in clickstream data analysis in marketing. *Journal of Interactive Marketing*, 23(1):35–48, 2009.

[BSV05] Paolo Boldi, Massimo Santini, and Sebastiano Vigna. Pagerank as a function of the damping factor. In *Proceedings of the 14th international conference on World Wide Web*, pages 557–566. ACM, 2005.

[Bur09] Ronald S Burt. *Structural holes: The social structure of competition.* Harvard University Press, 2009.

[BZ02] Vladimir Batagelj and Matjaž Zaveršnik. Generalized cores. *arXiv preprint cs/0202039*, 2002.

[CD07] Julie Coiro and Elizabeth Dobler. Exploring the online reading comprehension strategies used by sixth-grade skilled readers to search for and locate information on the internet. *Reading research quarterly*, 42(2):214–257, 2007.

[CGM00] Junghoo Cho and Hector Garcia-Molina. Synchronizing a database to improve freshness. In *ACM Sigmod Record*, volume 29, pages 117–128. ACM, 2000.

[CHN03] Patrali Chatterjee, Donna L Hoffman, and Thomas P Novak. Modeling the clickstream: Implications for web-based advertising efforts. *Marketing Science*, 22(4):520–541, 2003.

[CKRS12] Flavio Chierichetti, Ravi Kumar, Prabhakar Raghavan, and Tamas Sarlos. Are web users really markovian? In *Proceedings of the 21st international conference on World Wide Web*, pages 609–618. ACM, 2012.

[CM99] S CAMPBELL CHRISTOPHER and PAUL P MAGLIO. Facilitating navigation in information spaces. *International Journal of Human-Computer Studies*, 50(4):309–327, 1999.

[CMN08] Aaron Clauset, Cristopher Moore, and Mark EJ Newman. Hierarchical structure and the prediction of missing links in networks. *Nature*, 453(7191):98–101, 2008.

[CP95] Lara D Catledge and James E Pitkow. Characterizing browsing strategies in the world-wide web. *Computer Networks and ISDN systems*, 27(6):1065–1073, 1995.

[CPCP01] Ed H Chi, Peter Pirolli, Kim Chen, and James Pitkow. Using information scent to model user information needs and actions and the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 490–497. ACM, 2001.

[CPP00] Ed H Chi, Peter Pirolli, and James Pitkow. The scent of a site: A system for analyzing and predicting information scent, usage, and usability of a web site. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 161–168. ACM, 2000.

[CRS$^+$03] Ed H Chi, Adam Rosien, Gesara Supattanasiri, Amanda Williams, Christiaan Royer, Celia Chow, Erica Robles, Brinda Dalal, Julie Chen, and Steve Cousins. The bloodhound project: automating discovery of web usability issues using the infoscentπ simulator. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 505–512. ACM, 2003.

[CSC$^+$06] Andrea Capocci, Vito DP Servedio, Francesca Colaiori, Luciana S Buriol, Debora Donato, Stefano Leonardi, and Guido Caldarelli. Preferential attachment in the growth of social networks: The internet encyclopedia wikipedia. *Physical review E*, 74(3):036116, 2006.

[CSN09] Aaron Clauset, Cosma Rohilla Shalizi, and Mark EJ Newman. Power-law distributions in empirical data. *SIAM Review*, 51, 2009.

[CW88] Joanne Francis Cove and Brian Colin Walsh. Online text retrieval via browsing. *Information Processing & Management*, 24(1):31–37, 1988.

[CZTR08] Nick Craswell, Onno Zoeter, Michael Taylor, and Bill Ramsey. An experimental comparison of click position-bias models. In *Int. Conference on Web Search and Data Mining*, 2008.

[DAB10] Jeffrey A Dean, Corin Anderson, and Alexis Battle. Ranking documents based on user behavior and/or feature data, 2010. US Patent 7,716,225.

[Dav04] Brian D Davison. Learning web request patterns. In *Web dynamics*, pages 435–459. Springer, 2004.

[DBW13] Dimitar Dimitrov, Erdal Baran, and Dennis Wegener. Making data citable-a web-based system for the registration of social and economics science data. In *WEBIST*, pages 155–159, 2013.

[DFG01a] Inderjit S Dhillon, James Fan, and Yuqiang Guan. Efficient clustering of very large document collections. In *Data mining for scientific and engineering applications*, pages 357–381. Springer, 2001.

[DFG01b] Inderjit S. Dhillon, James Fan, and Yuqiang Guan. *Efficient Clustering of Very Large Document Collections*, pages 357–381. Springer US, Boston, MA, 2001.

[DGM06] Sergey N Dorogovtsev, Alexander V Goltsev, and Jose Ferreira F Mendes. K-core organization of complex networks. *Physical Review Letters*, 96, 2006.

[DHBW13] Dimitar Dimitrov, Daniel Hienert, Katarina Boland, and Dennis Wegener. Linking research data and literature: Integration of da|ra and sowiport based on link information from infolis. In *IASSIST Conference*, 2013.

[DHS18] Dimitar Dimitrov, Denis Helic, and Markus Strohmaier. Tag-based navigation and visualization. In Peter Brusilovsky and Daqing He, editors, *Social Information Access: Systems and Technologies*, volume 10100 of *LNCS*, pages 181–212. Springer International Publishing, Cham, 2018.

[Dim17] Dimitar Dimitrov. Modeling navigation in information networks. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 845–845. ACM, 2017.

[Din11] Ying Ding. Topic-based pagerank on author cocitation networks. *Journal of the Association for Information Science and Technology*, 62(3):449–466, 2011.

[DK04] Mukund Deshpande and George Karypis. Selective markov models for predicting web page accesses. *ACM Transactions on Internet Technology (TOIT)*, 4(2):163–184, 2004.

[DLFS18] Dimitar Dimitrov, Florian Lemmerich, Fabian Flöck, and Markus Strohmaier. Query for architecture, click through military: Comparing the roles of search and navigation on wikipedia. In *Proceedings of the 10th ACM Conference on Web Science*, pages 371–380. ACM, 2018.

[DLWS17] Dimitar Dimitrov, Daniel Lamprecht, Robert West, and Markus Strohmaier. Link disambiguation in wikipedia: From transition counts between articles to click probabilities of individual links. 2017.

[DNLH16] Alexander Dallmann, Thomas Niebler, Florian Lemmerich, and Andreas Hotho. Extracting semantics from random walks on wikipedia: Comparing learning and counting methods. In *Wiki@ ICWSM*, 2016.

[DOD$^+$06] Nathaniel D Daw, John P O'Doherty, Peter Dayan, Ben Seymour, and Raymond J Dolan. Cortical substrates for exploratory decisions in humans. *Nature*, 441(7095):876–879, 2006.

[DSHS15] Dimitar Dimitrov, Philipp Singer, Denis Helic, and Markus Strohmaier. The role of structural information for designing navigational user interfaces. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*, pages 59–68. ACM, 2015.

[DSLS16] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. Visual positions of links and clicks on wikipedia. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 27–28. International World Wide Web Conferences Steering Committee, 2016.

[DSLS17] Dimitar Dimitrov, Philipp Singer, Florian Lemmerich, and Markus Strohmaier. What makes a link successful on wikipedia? In *Proceedings of the 26th International Conference on World Wide Web*, pages 917–926. International World Wide Web Conferences Steering Committee, 2017.

[ERG$^+$89] Dennis E Egan, Joel R Remde, Louis M Gomez, Thomas K Landauer, Jennifer Eberhardt, and Carol C Lochbaum. Formative design evaluation of superbook. *ACM Transactions on Information Systems (TOIS)*, 7(1):30–57, 1989.

[Est10] Ernesto Estrada. Quantifying network heterogeneity. *Physical Review E*, 82, 2010.

[FD07] Brendan J Frey and Delbert Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[FEA17] Fabian Flöck, Kenan Erdogan, and Maribel Acosta. Toktrack: A complete token provenance and change tracking dataset for the english wikipedia. In *Proceedings of the Eleventh International AAAI Conference on Web an Social Media (ICWSM 2017)*, 2017.

[FLGD87] George W. Furnas, Thomas K. Landauer, Louis M. Gomez, and Susan T. Dumais. The vocabulary problem in human-system communication. *Communications of the ACM*, 30(11):964–971, 1987.

[For65] Edward W Forgy. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. *Biometrics*, 21:768–769, 1965.

[FP07] Wai-Tat Fu and Peter Pirolli. Snif-act: A cognitive model of user navigation on the world wide web. *Hum.-Comput. Interact.*, 22(4):355–412, November 2007.

[FR98] George W Furnas and Samuel J Rauch. Considerations for information environments and the navique workspace. In *Proceedings of the third ACM conference on Digital libraries*, pages 79–88. ACM, 1998.

[Fur97] George W Furnas. Effective view navigation. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 367–374. ACM, 1997.

[Gol97] Gene Golovchinsky. Queries? links? is there a difference? In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 407–414. ACM, 1997.

[GY17] Patrick Gilderslave and Taha Yasseri. Inspiration, captivation, and misdirection: Emergent properties in networks of online navigation. *arXiv:1710.03326*, 2017.

[Har14] Xavier A Harrison. Using observation-level random effects to model overdispersion in count data in ecology and evolution. *PeerJ*, 2:e616, 2014.

[Hav99] Taher Haveliwala. Efficient computation of pagerank. Technical report, Stanford, 1999.

[Hav02] Taher H Haveliwala. Topic-sensitive pagerank. In *Proceedings of the 11th international conference on World Wide Web*, pages 517–526. ACM, 2002.

[Hav03] Taher H Haveliwala. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *IEEE transactions on knowledge and data engineering*, 15(4):784–796, 2003.

[HdH84] HP Hogeweg-de Haart. Characteristics of social science information: A selective review of the literature. part ii. *Social Science Information Studies*, 4(1):15–30, 1984.

[Hea09] Marti Hearst. *Search user interfaces.* Cambridge University Press, 2009.

[Hel12] Denis Helic. Analyzing user click paths in a wikipedia navigation game. In *MIPRO, 2012 Proceedings of the 35th International Convention*, pages 374–379. IEEE, 2012.

[HGM06a] Paul. Heymann and Hector. Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Stanford University, April 2006.

[HGM06b] Paul Heymann and Hector Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Stanford, 2006.

[HK03] Taher Haveliwala and Sepandar Kamvar. The second eigenvalue of the google matrix. Technical report, Stanford, 2003.

[HKG$^+$12] Denis Helic, Christian Körner, Michael Granitzer, Markus Strohmaier, and Christoph Trattner. Navigational efficiency of broad vs. narrow folksonomies. In *Proceedings of the Conference on Hypertext and Social Media*. ACM, 2012.

[HKJ03] Taher Haveliwala, Sepandar Kamvar, and Glen Jeh. An analytical comparison of approaches to personalizing pagerank. Technical report, Stanford, 2003.

[HMvdS10] Joop J Hox, Mirjam Moerbeek, and Rens van de Schoot. *Multilevel analysis: Techniques and applications*. Routledge, 2010.

[HPN11] Mei-Chen Hu, Martina Pavlicova, and Edward V Nunes. Zero-inflated and hurdle models of count data with extra zeros: examples from an hiv-risk reduction intervention trial. *The American Journal of Drug and Alcohol Abuse*, 37(5):367–375, 2011.

[HPPL98] Bernardo A Huberman, Peter LT Pirolli, James E Pitkow, and Rajan M Lukose. Strong regularities in world wide web surfing. *Science*, 280(5360):95–97, 1998.

[HS11] Denis Helic and Markus Strohmaier. Building directories for social tagging systems. In *Proceedings of the International Conference on Information and Knowledge Management*. ACM, 2011.

[HSGS13] Denis Helic, Markus Strohmaier, Michael Granitzer, and Reinhold Scherer. Models of human navigation in information networks based on decentralized search. In *Conference on Hypertext and Social Media*, 2013.

[HST$^+$11] Denis Helic, Markus Strohmaier, Christoph Trattner, Markus Muhr, and Kristina Lerman. Pragmatic evaluation of folksonomies. In *Proceedings of the 20th international conference on World wide web*, pages 417–426. ACM, 2011.

[HTSA10] Denis Helic, Christoph Trattner, Markus Strohmaier, and Keith Andrews. On the navigability of social tagging systems. In *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pages 161–168. IEEE, 2010.

[JFM$^+$97] Thorsten Joachims, Dayne Freitag, Tom Mitchell, et al. Webwatcher: A tour guide for the world wide web. In *IJCAI (1)*, pages 770–777, 1997.

[JHW07] Seikyung Jung, Jonathan L Herlocker, and Janet Webster. Click data as implicit relevance feedback in web search. *Information Processing & Management*, 43(3):791–807, 2007.

[Joa02] Thorsten Joachims. Optimizing search engines using clickthrough data. In *Int. Conference on Knowledge Discovery and Data Mining*, 2002.

[Joa03] Thorsten Joachims. Evaluating retrieval performance using clickthrough data. Technical report, Cornell University, 2003.

[JW03] Glen Jeh and Jennifer Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, pages 271–279. ACM, 2003.

[KBP00] Muneo Kitajima, Marilyn H Blackmon, and Peter G Polson. A comprehension-based model of web navigation and its application to web usability analysis. *People and Computers*, pages 357–374, 2000.

[KCN06] Christian Kohlschütter, Paul-Alexandru Chirita, and Wolfgang Nejdl. Efficient parallel computation of pagerank. In *ECIR*, volume 3936, pages 241–252. Springer, 2006.

[KHMG03] Sepandar D Kamvar, Taher H Haveliwala, Christopher D Manning, and Gene H Golub. Extrapolation methods for accelerating pagerank computations. In *Proceedings of the 12th international conference on World Wide Web*, pages 261–270. ACM, 2003.

[KK09] Jaap Kamps and Marijn Koolen. Is wikipedia link structure different? In *Proceedings of the second ACM international conference on Web search and data mining*, pages 232–241. ACM, 2009.

[Kle99] Jon M Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

[Kle00a] Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the Symposium on Theory of Computing*. ACM, 2000.

[Kle00b] Jon M Kleinberg. Navigation in a small world. *Nature*, 406(6798):845–845, 2000.

[Kle02] Jon Kleinberg. Small-world phenomena and the dynamics of information. *Advances in neural information processing systems*, 1, 2002.

[Kle06] Jon Kleinberg. Complex networks and decentralized search algorithms. In *Proceedings of the International Congress of Mathematicians: invited lectures*, 2006.

[KM09] Weimao Ke and Javed Mostafa. Strong ties vs. weak ties: Studying the clustering paradox for decentralized search, 2009.

[KM10] Weimao Ke and Javed Mostafa. Scalability of findability: Effective and efficient ir operations in large information networks. In *Proceedings of the International Conference on Research and Development in Information Retrieval.* ACM, 2010.

[KNB$^+$15] Kasper Kristensen, Anders Nielsen, Casper W Berg, Hans Skaug, and Brad Bell. Tmb: automatic differentiation and laplace approximation. *arXiv:1509.00660*, 2015.

[KR95] Robert E Kass and Adrian E Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995.

[KRRT99] Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, and Andrew Tomkins. Trawling the web for emerging cyber-communities. *Computer networks*, 31(11):1481–1493, 1999.

[KT09] Ravi Kumar and Andrew Tomkins. A characterization of online search behaviour. *Data Engineering Bullettin*, 32(2):2009, 2009.

[KT10] Ravi Kumar and Andrew Tomkins. A characterization of online browsing behavior. In *Proceedings of the 19th international conference on World wide web*, pages 561–570. ACM, 2010.

[Law00] Steve Lawrence. Context in web search. *IEEE Data Eng. Bull.*, 23(3):25–32, 2000.

[LBL01] Mark Levene, José Borges, and George Loizou. Zipf's law for web surfers. *Knowledge and Information Systems*, 3(1):120–129, 2001.

[LBS$^+$16] Florian Lemmerich, Martin Becker, Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. Mining subgroups with exceptional transition behavior. In *Int. Conference on Knowledge Discovery and Data Mining*, 2016.

[LCH$^+$05] Donald J Leu, Jill Castek, D Hartman, Julie Coiro, L Henry, J Kulikowich, and Stacy Lyver. Evaluating the development of scientific knowledge and new forms of reading comprehension during online learning. *Final report presented to the North Central Regional Educational Laboratory/Learning Point Associates. Retrieved May*, 15:2006, 2005.

[LDHS16] Daniel Lamprecht, Dimitar Dimitrov, Denis Helic, and Markus Strohmaier. Evaluating and improving navigability of wikipedia: A comparative study of eight language editions. In *Proceedings of the 12th International Symposium on Open Collaboration*, pages 17:1–17:10. ACM, 2016.

[LECZ$^+$12] Donald J Leu, Heidi Everett-Cacopardo, Lisa Zawilinski, Greg McVerry, and W Ian O'Byrne. New literacies of online reading comprehension. *The Encyclopedia of Applied Linguistics*, 2012.

[LFK09] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[LHK10] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Predicting positive and negative links in online social networks. In *Proceedings of the 19th international conference on World wide web*, pages 641–650. ACM, 2010.

[LLHS16] Daniel Lamprecht, Kristina Lerman, Denis Helic, and Markus Strohmaier. How the structure of wikipedia articles influences user navigation. *New Review of Hypermedia and Multimedia*, pages 1–22, 2016.

[LLHS17] Daniel Lamprecht, Kristina Lerman, Denis Helic, and Markus Strohmaier. How the structure of wikipedia articles influences user navigation. *New Review of Hypermedia and Multimedia*, 23(1):29–50, 2017.

[Llo82] Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[LM04] Amy N Langville and Carl D Meyer. Deeper inside pagerank. *Internet Mathematics*, 1(3):335–380, 2004.

[LMBL$^+$14] Janette Lehmann, Claudia Müller-Birn, David Laniado, Mounia Lalmas, and Andreas Kaltenbrunner. Reader preferences and behavior on wikipedia. In *Conference on Hypertext and Social Media*, 2014.

[LNK07] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *journal of the Association for Information Science and Technology*, 58(7):1019–1031, 2007.

[LSB$^+$17] Florian Lemmerich, Philipp Singer, Martin Becker, Lisette Espin-Noboa, Dimitar Dimitrov, Denis Helic, Andreas Hotho, and Markus Strohmaier. Comparing hypotheses about sequential data: A bayesian approach and its applications. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 354–357. Springer, 2017.

[LSH$^+$15] Daniel Lamprecht, Markus Strohmaier, Denis Helic, Csongor Nyulas, Tania Tudorache, Natalya F Noy, and Mark A Musen. Using ontologies to model human navigation behavior in information networks: A study based on wikipedia. *Semantic web*, 6(4):403–422, 2015.

[LW04] Mark Levene and Richard Wheeldon. Navigating the world wide web. In *Web Dynamics*, pages 117–151. Springer, 2004.

[Man08] Anne Mangen. Hypertext fiction reading: haptics and immersion. *Journal of research in reading*, 31(4):404–419, 2008.

[Mar97] Stephen M Marson. A selective history of internet technology and social work. *Computers in Human Services*, 14(2):35–49, 1997.

[May12] V Valli Mayil. Web navigation path pattern prediction using first order markov model and depth first evaluation. *International Journal of Computer Applications (0975-8887)*, 45(16), 2012.

[MC07] Rada Mihalcea and Andras Csomai. Wikify!: linking documents to encyclopedic knowledge. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242. ACM, 2007.

[McS05] Frank McSherry. A uniform approach to accelerated pagerank computation. In *Proceedings of the 14th international conference on World Wide Web*, pages 575–582. ACM, 2005.

[Mil67] Stanley Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

[MISL07] Lev Muchnik, Royi Itzhack, Sorin Solomon, and Yoram Louzoun. Self-emergence of knowledge trees: Extraction of the wikipedia hierarchies. *Physical Review E*, 76(1):016106, 2007.

[MJ02] Naomi Mandel and Eric J Johnson. When web pages influence choice: Effects of visual primes on experts and novices. *Journal of consumer research*, 29(2):235–245, 2002.

[MJH17] Connor McMahon, Isaac Johnson, and Brent Hecht. The substantial interdependence of wikipedia and google: A case study on the relationship between peer production communities and information technologies. In *Proceedings of the Eleventh International AAAI Conference on Web an Social Media (ICWSM 2017)*, 2017.

[MLSL04] Alan L Montgomery, Shibo Li, Kannan Srinivasan, and John C Liechty. Modeling online browsing and path analysis using clickstream data. *Marketing Science*, 23(4):579–595, 2004.

[Moe03] Wendy W Moe. Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of consumer psychology*, 13(1-2):29–39, 2003.

[MTC$^+$12] Gengxin Miao, Shu Tao, Winnie Cheng, Randy Moulic, Louise E. Moser, David Lo, and Xifeng Yan. Understanding task-driven information flow in collaborative networks. In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 849–858, New York, NY, USA, 2012. ACM.

[MW07] Gary Marchionini and Ryen White. Find what you need, understand what you find. *International Journal of Human-Computer Interaction*, 23(3):205–237, 2007.

[MW08] David Milne and Ian H Witten. Learning to link with wikipedia. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 509–518. ACM, 2008.

[MWDR12] Edgar Meij, Wouter Weerkamp, and Maarten De Rijke. Adding semantics to microblog posts. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 563–572. ACM, 2012.

[Nel65] Theodor H Nelson. Complex information processing: a file structure for the complex, the changing and the indeterminate. In *Proceedings of the 1965 20th national conference*, pages 84–100. ACM, 1965.

[New01] Mark EJ Newman. Clustering and preferential attachment in growing networks. *Physical review E*, 64(2):025102, 2001.

[New10] Mark Newman. *Networks: An Introduction.* Oxford University Press, 2010.

[Nie06] Jakob Nielsen. F-shaped pattern for reading web content. https://www.nngroup.com/articles/f-shaped-pattern-reading-web-content, 2006. Accessed: 2016-07-12.

[OC03] Christopher Olston and Ed H Chi. Scenttrails: Integrating browsing and searching on the web. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 10(3):177–197, 2003.

[OJ93] Vicki L O'Day and Robin Jeffries. Orienteering in an information landscape: how information seekers get from here to there. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*, pages 438–445. ACM, 1993.

[OKU$^+$08] Jong-Hoon Oh, Daisuke Kawahara, Kiyotaka Uchimoto, Jun'ichi Kazama, and Kentaro Torisawa. Enriching multilingual language resources by discovering missing cross-language links in wikipedia. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology-Volume 01*, pages 322–328. IEEE Computer Society, 2008.

[PB06] Jose Pinheiro and Douglas Bates. *Mixed-effects models in S and S-PLUS.* Springer Science & Business Media, 2006.

[PBMW99] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web., 1999.

[PC99] Peter Pirolli and Stuart Card. Information foraging. *Psychological Review*, 106, 1999.

[PE00] Mike Perkowitz and Oren Etzioni. Towards adaptive web sites: Conceptual framework and case study. *Artificial intelligence*, 118(1-2):245–275, 2000.

[PF03] Peter Pirolli and Wai-Tat Fu. Snif-act: A model of information foraging on the world wide web. In *International Conference on User Modeling*, pages 45–54. Springer, 2003.

[Pir97] Peter Pirolli. Computational models of information scent-following in a very large browsable text collection. In *Proceedings of the Conference on Human Factors in Computing Systems*. ACM, 1997.

[Pir07] Peter Pirolli. *Information foraging theory: Adaptive interaction with information*. Oxford University Press, 2007.

[Pir09] Peter Pirolli. An elementary social information foraging model. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 605–614, New York, NY, USA, 2009. ACM.

[PLG10a] Anon Plangprasopchok, Kristina Lerman, and Lise Getoor. From saplings to a tree: Integrating structured metadata via relational affinity propagation. In *Proceedings of the AAAI workshop on Statistical Relational AI*, July 2010.

[PLG10b] Anon Plangprasopchok, Kristina Lerman, and Lise Getoor. Growing a tree in the forest: Constructing folksonomies by integrating structured metadata. In *Proceedings of the International Conference on Knowledge Discovery and Data Mining*. ACM, 2010.

[PP99] Peter LT Pirolli and James E Pitkow. Distributions of Surfers' Paths through the World Wide Web: Empirical Characterizations. *World Wide Web*, 2(1-2):29–45, 1999.

[PPC77] Graham H Pyke, H Ronald Pulliam, and Eric L Charnov. Optimal foraging: a selective review of theory and tests. *The quarterly review of biology*, 52(2):137–154, 1977.

[PWZL16] Ashwin Paranjape, Robert West, Leila Zia, and Jure Leskovec. Improving website hyperlink structure using server logs. In *Int. Conference on Web Search and Data Mining*, 2016.

[RFF$^+$10] Jacob Ratkiewicz, Santo Fortunato, Alessandro Flammini, Filippo Menczer, and Alessandro Vespignani. Characterizing and modeling the dynamics of online popularity. *Physical review letters*, 105(15):158701, 2010.

[ŘS10] Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the*

LREC 2010 Workshop on New Challenges for NLP Frameworks, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

[RTG00] Yossi Rubner, Carlo Tomasi, and Leonidas J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[Sar00] Ramesh R Sarukkai. Link prediction and path analysis using markov chains. *Computer Networks*, 33(1):377–386, 2000.

[SBC97] Ben Shneiderman, Don Byrd, and W Bruce Croft. Clarifying search: A user-interface framework for text searches. *D-lib magazine*, 3(1):18–20, 1997.

[Sch06] Patrick Schmitz. Inducing ontology from flickr tags. In *Collaborative Web Tagging Workshop at WWW2006, Edinburgh, Scotland*, volume 50, 2006.

[SE98] Alistair Sutcliffe and Mark Ennis. Towards a cognitive theory of information retrieval. *Interacting with computers*, 10(3):321–351, 1998.

[SHHS15] Philipp Singer, Denis Helic, Andreas Hotho, and Markus Strohmaier. Hyptrails: A bayesian approach for comparing hypotheses about human trails on the web. In *Int. Conference on World Wide Web*, 2015.

[SHTS14] Philipp Singer, Denis Helic, Behnam Taraghi, and Markus Strohmaier. Detecting memory and structure in human navigation patterns using markov chain models of varying order. *PloS One*, 9(7):e102070, 2014.

[SJ$^+$05] Ozgur Simsek, David Jensen, et al. Decentralized search in networks using homophily and degree disparity. In *Proceedings of the 19th international joint conference on Artificial intelligence*, pages 304–310. Morgan Kaufmann Publishers Inc., 2005.

[SKK$^+$08] Christin Seifert, Barbara Kump, Wolfgang Kienreich, Gisela Granitzer, and Michael Granitzer. On the beauty and usability of tag clouds. In *Proceedings of International Conference on Information Visualisation.* IEEE, 2008.

[SLW$^+$17a] Anna Samoilenko, Florian Lemmerich, Katrin Weller, Maria Zens, and Markus Strohmaier. Analysing timelines of national histories across wikipedia editions: A comparative computational approach. In *Proceedings of the Eleventh International*

AAAI Conference on Web an Social Media (ICWSM 2017), pages 210–219, 2017.

[SLW$^+$17b] Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, and Jure Leskovec. Why we read wikipedia. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1591–1600. International World Wide Web Conferences Steering Committee, 2017.

[SNSH13] Philipp Singer, Thomas Niebler, Markus Strohmaier, and Andreas Hotho. Computing semantic relatedness from human navigational paths on wikipedia. In *Int. Conference on World Wide Web*, 2013.

[Spo07] Anselm Spoerri. What is popular on Wikipedia and why? *First Monday*, 12(4), 2007.

[SPWL14] Aju Thalappillil Scaria, Rose Marie Philip, Robert West, and Jure Leskovec. The last click: Why users give up information network navigation. In *Int. Conference on Web Search and Data Mining*, 2014.

[Ste80] James H Steiger. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245, 1980.

[Sto82] Sue Stone. Humanities scholars: information needs and uses. *Journal of documentation*, 38(4):292–313, 1982.

[Sto84] Stephen K Stoan. Research and library skills: An analysis and interpretation. *College & research libraries*, 45(2):99–109, 1984.

[SV05] Sara Nadiv Soffer and Alexei Vazquez. Network clustering coefficient without degree-correlation biases. *Physical Review E*, 71, 2005.

[TAAK04] Jaime Teevan, Christine Alvarado, Mark S Ackerman, and David R Karger. The perfect search engine is not enough: a study of orienteering behavior in directed search. In *Conference on Human Factors in Computing Systems*, 2004.

[TSHS12] Christoph Trattner, Philipp Singer, Denis Helic, and Markus Strohmaier. Exploring the differences and similarities between hierarchical decentralized search and human navigation in information networks. In *Proceedings of the International Conference on Knowledge Management and Knowledge Technologies*. ACM, 2012.

[tTVLK12] Marijn ten Thij, Yana Volkovich, David Laniado, and Andreas Kaltenbrunner. Modeling and predicting page-view dynamics on wikipedia. *CoRR*, abs/1212.5943, 2012.

[WADZM12] Andias Wira-Alam, Dimitar Dimitrov, and Wolfgang Zenk-Möltgen. Extending basic dublin core elements for an open research data archive. In *International Conference on Dublin Core and Metadata Applications*, pages 56–61, 2012.

[Wal11] Vivienne Waller. The search queries that took australian internet users to wikipedia. *Information Research*, 16(2), 2011.

[WASB16] Chao-Yuan Wu, Christopher V Alvino, Alexander J Smola, and Justin Basilico. Using navigation to improve recommendations in real-time. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 341–348. ACM, 2016.

[WDN02] Duncan J Watts, Peter Sheridan Dodds, and Mark EJ Newman. Identity and search in social networks. *science*, 296(5571):1302–1305, 2002.

[WDY08] Youwei Wang, Weihui Dai, and Yufei Yuan. Website browsing aid: A navigation graph-based recommendation system. *Decision Support Systems*, 45(3):387–400, 2008.

[Wes15] Robert West. Electronic cigarette heatmap, 2015. Accessed: 2017-4-13.

[WJ11] Ingmar Weber and Alejandro Jaimes. Who uses web search for what: and how. In *International Conference on Web Search and Data Mining*, 2011.

[WKW$^+$13] Gang Wang, Tristan Konolige, Christo Wilson, Xiao Wang, Haitao Zheng, and Ben Y Zhao. You are how you click: Clickstream analysis for sybil detection. In *USENIX Security Symposium*, volume 9, pages 1–008, 2013.

[WL03] Richard Wheeldon and Mark Levene. The best trail algorithm for assisted navigation of web sites. In *Web Congress, 2003. Proceedings. First Latin American*, pages 166–178. IEEE, 2003.

[WL12] Robert West and Jure Leskovec. Human wayfinding in information networks. In *Int. Conference on World Wide Web*, 2012.

[WLW04] Ting-Fan Wu, Chih-Jen Lin, and Ruby C Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5(Aug):975–1005, 2004.

[WM08] Ian Witten and David Milne. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy*, pages 25–30, 2008.

[WMZ10] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):20, 2010.

[WPL15] Robert West, Ashwin Paranjape, and Jure Leskovec. Mining missing hyperlinks from human navigation traces: A case study of wikipedia. In *Int. Conference on World Wide Web*, 2015.

[WPP09a] Robert West, Joelle Pineau, and Doina Precup. Wikispeedia: An online game for inferring semantic distances between concepts. In *IJCAI*, pages 1598–1603, 2009.

[WPP09b] Robert West, Doina Precup, and Joelle Pineau. Completing wikipedia's hyperlink structure through dimensionality reduction. In *Proceedings of the 18th ACM conference on Information and knowledge management*, pages 1097–1106. ACM, 2009.

[WPP10] Robert West, Doina Precup, and Joelle Pineau. Automatically suggesting topics for augmenting text documents. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, pages 929–938. ACM, 2010.

[WS98] Duncan J Watts and Steven H Strogatz. Collective dynamics of 'small-world'networks. *Nature*, 393(6684):440–442, 1998.

[WS04] Feng-Hsu Wang and Hsiu-Mei Shao. Effective personalized recommendation based on time-framed navigation clustering and association mining. *Expert systems with applications*, 27(3):365–377, 2004.

[WT] Ellery Wulczyn and Dario Taraborelli. Wikipedia clickstream. figshare. doi:10.6084/m9.figshare.1305770. Accessed: 2015-12-13.

[WWZL16] Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. Growing wikipedia across languages via recommendation. In *Proceedings of the 25th International Conference on World Wide Web*, pages 975–985. International World Wide Web Conferences Steering Committee, 2016.

[XZC$^+$04] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi, and WeiGuo Fan. Optimizing web search using web click-through data. In *Int. Conference on Information and Knowledge Management*, 2004.

[YJGMD96] Tak Woon Yan, Matthew Jacobsen, Hector Garcia-Molina, and Umeshwar Dayal. From user access patterns to dynamic hypertext linking. *Computer Networks and ISDN Systems*, 28(7-11):1007–1014, 1996.

[YRM$^+$09] Eric Yeh, Daniel Ramage, Christopher D Manning, Eneko Agirre, and Aitor Soroa. Wikiwalk: random walks on wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 41–49. Association for Computational Linguistics, 2009.

[ZAN99] Ingrid Zukerman, David W Albrecht, and Ann E Nicholson. Predicting users' requests on the www. In *UM99 User Modeling*, pages 275–284. Springer, 1999.

[ZBŠD06] Vinko Zlatić, Miran Božičević, Hrvoje Štefančić, and Mladen Domazet. Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74(1):016115, 2006.

[Zho05] Shi Zhong. Efficient online spherical k-means clustering. In *Proceedings of the International Joint Conference on Neural Networks*, pages 3180–3185. IEEE, 2005.

[ZLZ09] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting missing links via local information. *The European Physical Journal B-Condensed Matter and Complex Systems*, 71(4):623–630, 2009.

Appendix A

Further Publications

During the course of this dissertation, I co-authored several publications listed here that are not part of this manuscript. The list order represents the relationship strength of each list entry to the topic of this thesis.

- Article 1 [LDHS16]: Daniel Lamprecht, Dimitar Dimitrov, Denis Helic and Markus Strohmaier. Evaluating and Improving Navigability of Wikipedia: A Comparative Study of Eight Language Editions. In *Proceedings of the 12th International Symposium on Open Collaboration*, pages 17:1–17:10. ACM, 2016.

- Article 2 [LSB^{+}17]: Florian Lemmerich, Philipp Singer, Martin Becker, Lisette Espín-Noboa, Dimitar Dimitrov, Denis Helic, Andreas Hotho and Markus Strohmaier. Comparing Hypotheses on Sequential Behavior: A Bayesian Approach and its Applications. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 354–357. Springer, 2017.

- Article 3 [BMH^{+}16]: Martin Becker, Hauke Mewes, Andreas Hotho, Dimi- tar Dimitrov, Florian Lemmerich and Markus Strohmaier. SparkTrails: A MapReduce Implementation of HypTrails for Comparing Hypotheses About Human Trails. In *Proceedings of the 25th International Conference Companion on World Wide Web*, pages 17–18. International World Wide Web Conferences Steering Committee, 2016.

- Article 4 [DHBW13]: Dimitar Dimitrov, Daniel Hienert, Katarina Boland and Dennis Wegener. Linking Research Data and Literature: Integration of da|ra and Sowiport Based on Link Information from InFoLiS. In *IASSIST Conference.* 2013.

- Article 5 [DBW13]: Dimitar Dimitrov, Erdal Baran and Dennis Wegener. Making Data Citable - A Web-based System for the Registration of Social and Economics Science Data. In *Proceedings of the 9th*

International Conference on Web Information Systems and Technologies. 2013.

- Article 6 [WADZM12]: Andias Wira-Alam, Dimitar Dimitrov and Wolfgang Zenk-Möltgen. Extending Basic Dublin Core Elements for an Open Research Data Archive. In *International Conference on Dublin Core and Metadata Applications*. 2012.

Dimitar Dimitrov

Curriculum Vitae

Waisenhausgasse 7
50676 Cologne
Germany
☎ *+49 (179) 1475505*
✉ *dimitar.dimitrov@gmx.net*
🖱 *www.dimitardimitrov.info*

Citizenship	Bulgaria
Place of birth	Plovdiv, Bulgaria
Date of birth	09.02.1984
Interests	Data Science, Machine Learning, Data Mining, Network Science

Experience

2011 to current **Research assistant** in Knowledge Technologies for the Social Sciences at GESIS — Leibniz Institute for the Social Science, Cologne, Germany

2014–2014 **Visiting researcher** at the Knowledge Technologies Institute at Graz University of Technology, Graz, Austria; Advisor: Prof. Dr. Denis Helic

2008–2011 **Software developer** at InnoSysTec GmbH, Salem, Germany

2006–2007 **Software development Internship** in Reading/Coding Department at Siemens Postal Automation, Konstanz, Germany

2005–2006 **Software development Internship** in Software Applications Department at Siemens Postal Automation, Konstanz, Germany

Education

2013 to current **Ph.D.** in Computer Science at University of Koblenz-Landau, Koblenz, Germany; Dissertation: Aspects of Information Access: Modeling Navigation on Wikipedia; Advisors: Prof. Dr. Markus Strohmaier and JProf. Dr. Claudia Wagner

2009–2011 **Master of Science** in Computer Science at HTWG Konstanz, Konstanz, Germany; Thesis: Investigations of Pivot Tightening for the Interval Cholesky Method and of the Solution of Toeplitz Systems of Linear Interval Equations; Advisor: Prof. Dr. Jürgen Garloff

2004–2009 **Diplom** in Software Engineering at HTWG Konstanz, Konstanz, Germany; Thesis: Construction of Totally Non-negative Interval Matrices and Experimental Studies of Interval-Neville-Elimination (in German); Advisor: Prof. Dr. Jürgen Garloff

2003–2004 **German Matura** at HTWG Konstanz, Konstanz, Germany

2002–2003 **Intensive German language course** at Unilingua, Göttingen, Germany

1997–2002 **Bulgarian Matura** at TEE Plovdiv, Plovdiv, Bulgaria

Personal and Technical Skills

Data Analysis Statistical hypothesis testing, regression analysis, feature engineering and outlier detection for further machine learning and data mining activities, e.g., classification, clustering, pattern mining, and information retrieval

Network Analysis Analysis of large-scale information and social networks, e.g., identifying and understanding collective behavior phenomena shaping the network structure, i.e., degree distribution, clustering coefficients, k-core and bow-tie structures

User Modeling Data driven development and evaluation of agent-based models capturing user behavior in information and social networks, e.g., Markov chains- and decentralized search-based click models

Web-based Systems Design and architecture of web-based systems for parallel processing of large-scale and high frequency data requests

Programming Skills **Python:** SciKit Learn, Pandas, NumPy; **Java:** Spring, Tomcat, Solr, ActiveMQ; further knowledge in R, SQL, Matlab, Grails, Groovy, C/C++, LaTeX, HTML

Databases MySQL, PostgreSQL

Project- and Team-leading Team and project leadership and coordination experience throughout various academic and industry projects

Languages

Bulgarian **Native:** my native language

German **Fluent:** written and spoken

English **Fluent:** written and spoken

Publications

Proceedings

2018 **D. Dimitrov**, F. Lemmerich, F. Flöck, and M. Strohmaier. *Query for Architecture, Click through Military: Comparing the Roles of Search and Navigation on Wikipedia.* In 10th ACM Conference on Web Science (WebSci18), 27-30 May, Amsterdam, Netherlands, 2018.

2017 F. Lemmerich, P. Singer, M. Becker, L. Espín-Noboa, **D. Dimitrov**, D. Helic, A. Hotho, and M. Strohmaier. *Comparing hypotheses on sequential behavior: A Bayesian approach and its applications.* In European Conference on Machine Learning and Knowledge Discovery in Databases (ECML/PKDD), 2017.

2017 **D. Dimitrov**, P. Singer, F. Lemmerich and M. Strohmaier. *What Makes a Link Successful on Wikipedia?* In 26th International World Wide Web Conference (WWW'2017), 3-7 April, Perth, Australia, 2017. (CORE A* conference, acceptance rate 164/966, 17% quota)

2017 **D. Dimitrov**. *Modeling Navigation in Information Networks.* In Doctoral Consortium at the 10th ACM International Conference on Web Search and Data Mining (WSDM'2017), 6-10 February, Cambridge, United Kingdom, 2017. (CORE A* conference)

2016 D. Lamprecht, **D. Dimitrov**, D. Helic and M. Strohmaier. *Evaluating and Improving Navigability of Wikipedia: A Comparative Study of Eight Language Editions.* In 12th International Symposium on Open Collaboration (OpenSym'2016), Berlin, Germany, 2016.

2015 **D. Dimitrov**, P. Singer, D. Helic and M. Strohmaier. *The Role of Structural Information for Designing Navigational User Interfaces.* In 26th ACM Conference on Hypertext and Social Media (HT'2015), 1-4 September, Middle East Technical University Northern Cyprus Campus, Cyprus, 2015. (CORE A conference)

2013 **D. Dimitrov**, E. Baran and D. Wegener. *Making Data Citable - A Web-based System for the Registration of Social and Economics Science Data.* In 9th International Conference on Web Information Systems and Technologies (WEBIST'2013) 8-10 May, Aachen, Germany, 2013.

2012 A. Wira-Alam, **D. Dimitrov** and W. Zenk-Möltgen. *Extending Basic Dublin Core Elements for an Open Research Data Archive.* In International Conference on Dublin Core and Metadata Applications (DCMI'2012), 3-7 September, Malaysia, 2012.

Posters

2016 **D. Dimitrov**, P. Singer, F. Lemmerich and M. Strohmaier. *Visual Positions of Links and Clicks on Wikipedia.* In 25th International World Wide Web Conference (WWW'2016), 11-15 April, Montreal, Canada, 2016. (CORE A* conference)

2016 M. Becker, H. Mewes, A. Hotho, **D. Dimitrov**, F. Lemmerich and M. Strohmaier. *SparkTrails: A MapReduce Implementation of HypTrails for Comparing Hypotheses About Human Trails.* In 25th International World Wide Web Conference (WWW'2016), 11-15 April, Montreal, Canada, 2016. (CORE A* conference)

2013 **D. Dimitrov**, D. Hienert, K. Boland and D. Wegener. *Linking Research Data and Literature: Integration of da|ra and Sowiport based on Link Information from InFoLiS.* In IASSIST Conference (IASSIST'2013) 28-31 May, Cologne, Germany, 2013.

Book Chapters

2018 **D. Dimitrov**, D. Helic and M. Strohmaier. *Tag-based Navigation and Visualization.* In: Peter Brusilovsky, Daqing He (eds) Social Information Access: Systems and Technologies, LNCS, vol 10100, pages 181–212, Springer International Publishing, 2018.

Working Papers

2017 **D. Dimitrov**, D. Lamprecht, R. West and M. Strohmeier. *Link Disambiguation in Wikipedia: From Transitions Counts Between Articles to Click Probabilities of Individual Links.* (under submission)