Rainer Maria Hofmann-Wellenhof BSc

Modelling User Navigation with biased Random Surfers

Master's Thesis

Graz University of Technology

Knowledge Technologies Institute
Head: Univ.-Prof. Dr. Stefanie Lindstaedt

Supervisor: Assoc.Prof. Dipl.-Ing. Dr.techn. Denis Helic
Advisor: Dipl.-Ing. Daniel Lamprecht BSc

Graz, October 2015

Statutory Declaration

I declare that I have authored this thesis independently, that I have not used other than the declared sources/resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.

Graz, _____ _____

Date Signature

Eidesstattliche Erklärung[1]

Ich erkläre an Eides statt, dass ich die vorliegende Arbeit selbstständig verfasst, andere als die angegebenen Quellen/Hilfsmittel nicht benutzt, und die den benutzten Quellen wörtlich und inhaltlich entnommenen Stellen als solche kenntlich gemacht habe.

Graz, am _____ _____

Datum Unterschrift

[1] Beschluss der Curricula-Kommission für Bachelor-, Master- und Diplomstudien vom 10.11.2008; Genehmigung des Senates am 1.12.2008

Acknowledgements

I would like to thank Denis Helic[2] for providing me the opportunity to conclude my Master's studies with research at the Knowledge Technologies Institute (KTI).

This research included the publication of a research paper. Thus, I would further like to thank my collaborators of the paper: Florian Geigl, Daniel Lamprecht, Simon Walk, Markus Strohmaier, and (again) Denis Helic. The in-depth insights into the process of scientific publishing was compelling and the work ethic admirable.

To all of my colleagues who regularly participated in the research meetings and vividly discussed recent publications: thank you.

My advisor for this thesis, who provided valuable guidance in weekly meetings, responded to E-Mails on Sunday nights, and never lost focus deserves my deepest gratitude: Thank you Daniel for bearing with me.

Lastly, and most importantly I would like to thank my parents and family for their continuous support not just during this thesis, but throughout my time as a student and my entire life.

Rainer Hofmann-Wellenhof

Graz, October 2015

[2] I am not including any titles in these acknowledgements to represent the considerate and friendly environment at the Knowledge Technologies Institute.

Abstract

When users navigate the Web they tend to navigate differently depending on the domain and the underlying graph structure. Several researchers analysed user navigation in information networks and created extensive models based on random walks in order to simulate it. These models were enhanced by accounting for backtracking, page topicality, or sophisticated measures which determined whether the next click brings the user closer to the target or if it is a click in a random direction. Some of the used data sets were created through games played on Wikipedia. This thesis had the privileged access to real user click data of Austria-Forum, an online encyclopedia. Previous research came to the conclusion that a new kind of random surfer needs to be created in order to simulate user behaviour in Austria-Forum more accurately. Based on an analysis of the click data, new random surfers are introduced to improve previous results. Furthermore, the influence of a varying jump probability, which depends on the amount of past clicks, is investigated in such models. This thesis shows that a simple bias in random surfers enhances the capabilities to model user navigation in information networks. Including a varying jump probability further approximates the navigation behaviour of users. These findings can be used to test the accuracy of such models on other domains as well as predict the influence on user navigation when design changes are made on encyclopedias.

Kurzfassung

Benutzer navigieren im Web unterschiedlich, abhängig von der Domäne und der zugrundeliegenden Graph-Struktur. Mehrere Forscher haben Benutzerverhalten in Informationsnetzwerken analysiert und umfangreiche Modelle basierend auf *Random Walks* erstellt. Diese Modelle wurden um Fähigkeiten wie Zurück-klicken, themenbezogenes surfen, oder raffinierte Verfahren die ermitteln ob der nächste Klick den Benutzer näher ans Ziel bringt oder ein Klick in eine zufällige Richtung ist, erweitert. Manche der verwendeten Datensätze wurden durch Spiele die in Wikipedia gespielt werden erstellt. Diese Arbeit hatte den privilegierten Zugang zu echten Benutzer-Klick-Daten von Austria-Forum, einer Online Enzyklopädie. Durch vorhergehende Forschung wurde festgestellt, dass eine neue Art von *Random Surfer* erstellt werden muss um Benutzerverhalten im Austria-Forum akkurat zu modellieren. Daher wurden Random Surfer erstellt die auf der Analyse der Klick-Daten beruhten, um die existierenden Resultate zu verbessern. Des Weiteren wurde der Einfluss von dynamischen Sprung-Wahrscheinlichkeiten in solchen Modellen untersucht; diese hängen von der Anzahl der getätigten Klicks ab. Durch diese Arbeit wird gezeigt, dass Modelle durch einfache Beeinflussungen Benutzerverhalten besser modellieren können. Die Einbindung von dynamischen Sprung-Wahrscheinlichkeiten verbessert die Annäherung an das echte Benutzerverhalten. Die Resultate der Arbeit können sowohl verwendet werden um die Präzision der erstellten Modelle auf anderen Domänen zu testen, als auch um den Einfluss auf die Benutzer Navigation vorherzusagen wenn Änderungen in Enzyklopädien vorgenommen werden.

Contents

List of Figures

1 Introduction

1.1 Motivation

The *World Wide Web* consists of *Web sites*, which are often split into *Web pages*. Today, sites and pages are highly interlinked with *hyperlinks*, which are the clickable paths between pages, hidden behind text. Pages can be seen as *nodes*, and links as *edges*: the Web is then an immense *graph* with directed edges, meaning, often one node can be reached from another, but can not reach the other node: page A having a link leading to page B does not imply there is a link from page B to A as well. Furthermore, there can be multiple links on page A leading to page B: the page *Graz*[1] on *Wikipedia* has multiple links to *Schloss Eggenberg* on it, since one is already included in the abstract.

In encyclopedias, links are of great importance: articles can often only be understood with some context, provided by related articles, reachable via these links. A general goal of Web sites is to keep users engaged; to keep them navigating on the site. In order to measure how well this goal is met, the behaviour of users needs to be analysed. User behaviour analysis can have several other benefits as well. The effect of newly implemented features or an adapted design can be measured. The information about (un)popular pages or features can be used to better connect these pages or improve these features. Or, as Suneetha and Krishnamoorthi showed, it can be used to find errors of Web pages [SK09].

User behaviour is split into sessions, which are instances of humans navigating a site. Sessions can thus be extracted from user click data. The duration of sessions is particularly useful for banking sites [Kap+14]. It needs to be

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

analysed what timeframe between two clicks is still realistic in order to automatically log-out users who forgot to close their browsers, possibly on a public device. Sessions are also used to determine how often users jump to different pages by clicking on bookmarks or entering an address. Research showed that the probability of such a jump depends on the domain. Users on Wikipedia are less likely to follow links than users of movie platforms [Gle+10]. When users are seeking a specific article on Wikipedia (given an arbitrary starting article), the chances of a random click to a different category are high in the beginning but decay over time [Hel+13]. Researchers were already aware of the fact that some pages are more likely to trigger a jump than others [Gon+09]. The research on these jump probabilities is plentiful. Yet, there have not been any investigations about the probabilities of jumps when users advanced in their sessions.

Having explored several aspects of user behaviour, models can be created which incorporate properties of user behaviour. These models are often simulations performed on a specific domain. A model which perfectly simulates user behaviour on a recommender system might be inaccurate on an information platform. These models can be used to predict user navigation on a system where no data is available. As an example: a new online store would benefit greatly from a model which simulates user traffic on a well-established online store to see which parts of the system will be used the most. Similarly, a model can be used to test a new feature on a site, predicting how many clicks will be made based on automated user navigation simulations.

The data of user navigation is needed in order to create a model which simulates user navigation. Such data is seldom available to the public. This thesis uses the data provided by Austria-Forum[2]. Previous research created two simple random surfers in order to simulate user behaviour on Austria-Forum. The research found that search engines have a big influence on the click data of encyclopedias, it stated: "To capture the lateral access to a website from a search engine we need a new kind of random surfer model" [Gei+15].

[2] http://austria-forum.org

1.2 Research Questions and Contributions

This thesis was driven by two research questions. These questions and the findings thereof are described in this section. The detailed results to these research questions can be found in section 4.2. The questions are further examined in chapter 5.

Research Question 1: How well can a simple biased random surfer simulate the lateral user navigation behaviour in Austria-Forum?

The analysis on the data set of Austria-Forum showed that search engines do indeed have an impact on page views in Austria-Forum. The correlations between the simulated click data and the reference click data (of real users) was improved with the development of new random surfers, compared to the results achieved in previous research [Gei+15]. The new random surfer models are using a simple personalised vector (sec. 2.2.2) in order to mitigate the lateral access raised by search engines. The random surfer models do not possess any knowledge about past decisions (sec. 2.2.1).

Research Question 2: What impact does a dynamic jump probability have on the results of random surfers?

Dynamic jump probabilities have not been investigated so far. Usually, an approximate average jump probability is used. For Austria-Forum, the average jump probability for each click is 69%. All random surfers with dynamic jump probabilities modelled user navigation in Austria-Forum better than the corresponding random surfers with the static jump probability of Austria-Forum.

This thesis is the first to perform an extensive analysis on user click data in Austria-Forum (sec. 4.1). Several of the performed methods were inspired by the analysis of other data sets (sec. 2.1) which were performed in the past. User behaviour and the Web have since evolved with the introduction of new Web development techniques and handheld devices. Thus, this thesis provides insight into the current usage of an information platform. Such insight can be taken as reference for other encyclopedias, given that similar

data is available for analysis. For most encyclopedias (e.g. Wikipedia) it is not feasible to be given access to this data, which enhances the importance of this research. Furthermore, the finding that dynamic jump probabilities increase the accuracy when modelling user navigation might be an incentive for future research to investigate whether this finding holds true for other domains as well.

1.3 Overview

This work is split into 5 further chapters. The next chapter, chapter 2, extensively describes research related to the research questions (presented in sec. 1.2). Chapter 3 goes into detail about the available materials and methods applied upon these materials in order to answer the research questions. Chapter 4 then presents and describes the results which are discussed in detail in chapter 5 with a focus on the research questions. Finally, chapter 6 concludes this thesis and provides suggestions on how to further advance the topic of modelling user navigation in Austria-Forum or information networks in general.

2 Related Work

This chapter is split into two parts. The first part focuses on the analysis of and findings from human click data and click trails (sessions) on encyclopedias. Following that, practicalities on how to reconstruct and model user behaviour, or concepts appearing in the practical world, are presented.

2.1 Click Analysis

There are plenty of possibilities for analysing click data. It can be used to determine whether pages, which are ranked highly in search engines, are also used most often by users and, thus, measuring the influence search engines have on encyclopedias (*Wikipedia*) or the Web overall [QLC05]. Through click trails, it can be studied which links are the most important, if link placement or emphasising makes a difference, which links are redundant or broken, and similarly more [SK09; MHM06].

In more general terms: understanding how users navigate the Web and use its features can help improving the Web (and features).

2.1.1 Sessions

Analysing sessions is used in order to [Mei+09]:

- reproduce real world user navigation via agent-based models;
- detect abnormal user behaviour created by machines (bots);
- determine the importance of search engines and measure the influence they have on browsing behaviour.

One of the main questions, when trying to define a session, is *when does a session end*.

- Does it end when the user stopped using the device?
- Does it end when the browser is closed?
- Are sessions bound to a specific task/topic?
- Are separate browser windows/tabs part of the same or a different session?

No matter the definition, what all sessions have in common is that they *always* consist of at least one page visit and an arbitrary amount of clicks. The difference lies in how those clicks are grouped together.

Not long after the spread of the *stateless random surfer* (sec. 2.2.2), which naively follows links within a graph, Qiu et al. found that real user sessions are better represented through trees, and not lines, with the beginning of the session as the root of the tree [QLC05].

When users only follow links, without backtracking or surfing with multiple tabs, the ratio of

$$\frac{\text{number of clicks}}{\text{depth of the tree}} \tag{2.1}$$

equals one. This is the assumption the *stateless random surfer model* makes. A visualisation of this behaviour would resemble a line. This ratio is called the *branching factor*.

In order to reconstruct sessions, some form of logging needs to be performed. This can either be done on the server side (by logging HTTP requests), or on the client side (via browser add-ons like toolbars) which requires the cooperation (or ignorance) of the user. Even though logging on the server side is more common, the information contained in such logs is often limited and always unstructured due to HTTP being a connectionless and stateless protocol. Depending on the amount of information logged from one request, session reconstruction can be a lengthy task. Bayir et al. first extracted the graph behind the web page they had log files on. For every series of requests $s = [r_1, r_2, r_3, ..., r_n]$, grouped by users, they checked whether there exists an

in-link into the node corresponding to r_i with $i > 1$ from another node of the series r_j with $1 <= j < i$ [Bay+12].

When log files contain the information (referrer) about the last page visited by the user, sessions can be reconstructed without the corresponding graph. Instead of checking whether links from previously visited nodes exist to the currently investigated node, one can check if the referrer page was the target of one of the previous requests. The resulting structure is called a *referrer tree* [QLC05].

Most often, sessions are split via timeouts [DM12]. Determining a feasible value finds application in real world scenarios: *online banking* solutions contain sensible data and should not be displayed longer than necessary (in case the user forgot to log out and left the device). Therefore security measures are often implemented which automatically terminate a session after a certain period of inactivity. This period should not be chosen too brief either, since the user should not be rushed into making clicks in order to keep the session active [Kap+14].

More closely related to the contributions of this work, the authors of "What's in a session: tracking individual behavior on the web" investigated the concept of creating sessions solely based on timeouts. Varying the timeout heavily influences the *mean node count* (session length) and thus *mean depth* and *sessions per user*. The concept of logical sessions best resembles that of user behaviour. Logical sessions end when the user stops clicking on links (makes a jump). By modelling referrer trees they found that revisitation patterns were higher than anticipated. They attributed the reason for that to users utilising the back button and then branching out to a different page, as well as the introduction of multi-tab browsing [Mei+09; Gon+09].

Once users start to use the back button, and then click a different link on a page which was already visited in this session, another branch is created. Consequently, the *branching factor* from 2.1 grows. Meiss et al. found an average value of $\mu = 1.94$ ($\sigma = 0.25$) for this ratio, meaning the number of nodes is almost twice the depth of a session and the *branching factor* of sessions is higher than commonly assumed. Therefore the stateless random surfer does not accurately imitate user behaviour [Mei+09]. An average node-based branching factor was calculated at 2.95 [QLC05].

Even though various Web traffic analyses follow a log-normal fit at user level, the cumulative of this data often results in a power-law distribution [Mei+09].

2.1.2 Analysis of Wiki-Games

The Wiki Game data set

When it comes to analysing user behaviour and navigation patterns, *The Wiki Game*[1] data set is commonly used and will be referenced several times throughout this thesis[2]. As the name suggests, the Wiki Game operates on Wikipedia. There are several versions within the Wiki Game, but the common goal is to reach a certain target (article), given a random starting article. The different versions and its goals are:

- *speed race:* reach the target in minimal time;
- *least clicks:* reach the target with minimal clicks;
- *six degrees of Wikipedia:* reach the target with exactly six clicks;
- *five clicks to Jesus:* reach the "Jesus" article with five clicks or less;
- *no United States:* since the United States article is a *hub* that links to many different categories, which makes finding things much easier, it is restricted in this game.

Even though the collected click data is created by users (humans), their behaviour outside of such games might differ due to the following:

- the concept of starting at a random article is not realistic for the real world;
- the use of the search function is prohibited in the game, even though it is probably used extensively in real world scenarios;
- in games where speed is important, but the number of clicks is not, users are presumably making clicks they would not make when reading articles for information;

[1] http://thewikigame.com

[2] Sometimes the data set used is from *Wikispeedia* (http://www.wikispeedia.net), an equivalent data set. For simplicity, only the *Wiki Game*-phrasing will be used throughout this thesis, referring to either *The Wiki Game* or *Wikispeedia*

- not all users have specific target articles;
- in reality, users are often interested in a certain topic and navigate within the related category [Mei+10]. In the game, chances are that start and target article are not within the same category.

The optimal data set for analysing user click trails in encyclopedias, or on the Web, would be the click trails gathered by Wikipedia or users themselves. Getting access to this data is difficult due to privacy concerns. Although the Wiki Game data set does have its limitations, it has one tremendous advantage over usual log files: *the target is always known*.

The following findings have come from the data set of the Wiki Game.

Findings

Since the target is always known, Scaria et al. analysed why users would abort their search and determined differences between users who found what they were looking for, and those who did not. They found that users often gave up even when they were close to their target information. Those users, who did not reach the target, were three times more likely to backtrack (click the back button in the browser) even though the click they made brought them closer to the target [Sca+14].

Takes and Kosters measured the difficulty of finding a path between two pages. They noticed a strong correlation between distance (shortest path length) and difficulty (path length of users): the longer the shortest possible path to the target node, the more trouble users had finding that node. Understandably, users were not always navigating the shortest path, but were overall doing well in finding quick ways to reach the target [TK12].

West and Leskovec confirmed these findings: even though humans do not know the graph behind *information networks*, they tend to navigate it efficiently. The authors attribute that humans make use of their general knowledge when navigating in information networks. They further investigated, whether general knowledge is *necessary* in order to find targets quickly: they created an agent which did not possess *any* additional information about the network. The agent found paths quicker than humans do. That was partly due to the fact that humans missed the possibility of

reaching the target with just one more click went another route instead. They attributed this circumstance to users having—and sticking to—a plan on how to reach the target, so that they missed better opportunities arising along the way [WL12a]. In follow up research, they discovered the last click to be the most difficult one: In longer games, users tended to circle around the target, until they reached it. In shorter games they made quick and direct progress in almost every step [WL12b].

No matter the length of a game, a phenomenon observed by most researchers was that click trails could be split into two phases: initially, users tried to get away from the starting node, in order to reach a *Hub* (a page with many links) due to the many possibilities it provides. Users then started approaching towards the target (category) [WL12b; TK12]. In *2013*, Helic et al. made a similar discovery: users either clicked to *explore* or to *exploit*. When exploring, users chose links almost arbitrarily until they reached topics they were familiar with. Afterwards users started *exploiting*, which is equivalent to the "narrowing in"-phase of West and Leskovec. 15-20% of clicks were found to be *exploring*, which notably matches the *teleportation probability* in (sec. 2.2.2) PageRank [Hel+13].

Hubs are of great importance due to their centrality: users tend to make quick progress until they reach a hub, then the probability of making a good click remains stable, until the deciding clicks are made. On the way towards the target, users not only decrease the distance in the graph, but also the conceptual distance to the target article in a steady manner: articles with high textual similarity are also close (shortest path) to each other in the Wikipedia graph [WL12b].

Links are not always where they should be: in recent research, West et al. have engaged in the problem of finding missing links between articles: creating links to related articles can be considered an easy task, since potential articles can be derived from the written text. The problem of finding sources (besides the obvious) is much harder, since *all articles* are candidates. Instead of scanning every article for keywords which correlate to the new article, they made use of the click-paths. When users navigate from article A to B and then C, where C is the target article, they concluded that A should be searched for (possible) links to C, since users were looking for such a link [WPL15].

2.2 Click Models

Click models simulate the behaviour on a system, such as the Web or a section thereof. The behaviour to simulate is the click stream data of users. Accurate models can be used to simulate traffic and measure the influence of a potential change in link structure as well as make predictions about the next click of a user. This section explains the development of click models and the underlying heuristic.

2.2.1 Random Walk

Suppose $G = (V, E)$ is a directed graph where $V = v_1, v_2, ..., v_n$ are the nodes and E is the set of edges. Let $E(v_i)$ be the set of out-edges of node v_i. A random walk is then a stochastic process in which an agent walks the graph by selecting a successor node of the current node v_i out of $E(v_i)$.

A random walk which fulfils these properties and acts on a *finite* graph is called *Markov chain*. A Markov chain consists of a set of possible states, called *state-space*. In the above example, the set of nodes V represents the state-space of the random walk. For every state, the next state is never dependent on past states. The random walk is therefore *memoryless*. This property of memorylessness is called the *Markov property*. The future states only depend on the current state [ES02; Lov93].

Random walks have played a substantial role in mathematics and physics for decades. In mathematics, they were used for probability theories. In physics, one of the earlier findings showed that "The Brownian motion of a dust particle is random walk in the room" [Lov93]. Later, studies found strong connections between electric networks and random walks [Flo12; DS00]. Floryance showed that, with the graph mapped to the electric network and given some starting nodes, the voltages per node within said electric network closely resemble the probability that a random walk terminates [Flo12].

In the field of computer science, Grady used random walks on graphs to create a new algorithm for image segmentation: initially some pixels of an image are labeled, which will later be parts of the different segments

of the output image. By using a random walk, the probabilities for every other pixel reaching one of the labeled pixels is calculated. Each pixel is then assigned the label of the pixel it has the highest probability of reaching first [Gra06]. This, and similar algorithms, are often used to improve the readability of medical images [AHS10; GF04; Mai+08].

2.2.2 PageRank

PageRank is a stochastic process which extends a simple random walk with a teleportation probability: the Web is a directed graph where several web pages do not contain any links. The random surfer would get stuck on such pages, having no way to continue navigating. Furthermore, the random surfer can get stuck in a cycle, a small group of pages. PageRank accounts for those cases, it enhances the basic random surfer with a probability to jump to a new page[3]. Thus, the random surfer includes a probability to follow a link (α), and the probability of a jump ($1 - \alpha$). With probability α, the next page will be chosen uniformly at random from the available connected pages, with probability $1 - \alpha$ the next page will be chosen uniformly at random from all pages [Pag+99].

Page et al. found 0.15 to be a good fit for the jump probability. Consequently, the random surfer keeps clicking links with a probability of 0.85. The Markov property is still given. PageRank was created to rank the pages of the World Wide Web based on their incoming and outgoing links. It served as the basis for ranking the search results of Google[4].

Page et al. proposed the possibility of a *personalised PageRank*, where pages have different probabilities of being the *landing page*. This was accomplished through the vector E, which is a vector over web pages, holding the probability distribution for the web page being the one the random surfer jumps to [Pag+99].

[3] This probability is often called *teleportation factor* or *jump probability*
[4] http://www.google.com/about/company/

2.2.3 PageRank Improvements

The branching factor of users surfing the Web is not anticipated by PageRank (see sec. 2.1.1) and not all pages are equally likely as entry page for sessions. Therefore, Meiss et al. argue that a *stateless random surfer* does not properly model the browsing behaviour of users [Mei+09].

Gonçalves et al. created an agent-based model called *BookRank*. The new model counters PageRanks' weaknesses: *BookRank* owes its name to the fact that users tend to have bookmarks, which are often the entry points of a session. Furthermore, pages which were already visited by the user are also more likely an entry page than other pages. Therefore BookRank includes a memory-like feature that every page visited by a user becomes a candidate for being the entry page of new sessions, with higher chances for pages with more visits. BookRank also includes the possibility of backtracking to increase the similarity to real user behaviour. These additions violate the Markov property [Gon+09].

Meiss et al. extended BookRank further by adding *page topicality*: they attributed the lateral distribution of traffic to many users with focused interest, instead of few users with many interests. They found that sessions tend to get longer (more clicks) when users are staying within a thematic area. The new model incorporates this topicality-based surfing. The results came substantially closer to the aggregated click data of users than the results of BookRank. They concluded that topicality, alongside of bookmarks and multiple tabs, are key aspects when modelling user behaviour.

In exchange for being more sophisticated than PageRank, those new models better simulate user traffic. Although, technically, PageRank is for ranking pages by importance and not traffic, it can be argued that traffic is (often) the main goal of web pages and the observable measurement of importance [Mei+10].

Teleportation (Targets) in Random Surfers

In PageRank, when the random surfer jumps, a vector v holds the probabilities for all pages to be the entry point of the next session. Those prob-

abilities are mostly chosen to be uniform and static. Gyöngyi et al. used different probabilities among pages, creating a biased random surfer alongside *TrustRank*, for ranking pages by trustworthiness and fighting spam [GGP04].

Throughout time, popularities of web sites fluctuate. This holds true for encyclopedia articles as well. Often these rises (and falls) in popularity are due to events on a global scale: after earthquakes or other natural disasters, the associated searches[5] and articles record spikes in popularity. In the scope of an encyclopedia, different parts of the underlying graph gain more importance at certain times [GR14].

This lead Gleich and Rossi to extensively research the effect of *time-dependent teleportation*, choosing v not only non-uniform, but also as a function $v(t)$, depending on time t. Incorporating these advancements highlights pages of momentary importance, which might be used in monitoring applications [GR14].

In earlier work, Gleich et al. measured teleportation parameters of PageRank: based on empirical data, extracted from toolbar logs, they calculated that users jump with a probability of $0.275 - 0.4$ when surfing the Web in general; on *Wikipedia*, these chances increase to $0.575 - 0.675$. In order to conclude those values, they calculated jump-probabilities on the user level, followed by an estimation of a density function [Gle+10].

The power of different jump probabilities was shown by Bressan and Peserico: on some graphs, the top n results generated by PageRank can be arranged in all possible $(n!)$ sequences, even when adjusting the jump probability only within a marginally small $[0.14999, 0.15001]$ interval [BP10].

2.2.4 Decentralised Search

Multiple findings from sec. 2.1.2 reported efficient navigational behaviour of users in an *information network*. This phenomenon was first observed in *social networks*, when Milgram published "The small world problem" in 1967. He showed that arbitrary people, living in the United States, can reach each

[5] https://www.google.com/trends/explore#q=earthquake

other within six steps. Apart from the short distance between two "nodes", the efficiency of finding paths was striking and and caused subsequent research [Mil67].

Small world graphs are present in every day life: the power grid of the United States can be classified as such, as well as the graph extracted from collaborations between actors in the film industry. Watts and Strogatz investigated the attributes which identify a small world graph: by randomly replacing edges in a *regular graph*, the ultimate outcome is a *random graph*. Small world graphs lie somewhere between these two extremes; high clustering and a small diameter are noticeable characteristics. Even in sparse graphs, by introducing a small amount of "short cuts" (additional edges), they were able to reproduce the *small world phenomenon*. The presence of this phenomenon in society, and a vast variety in nature, enables information to spread fast through the underlying networks; on the other hand, diseases spread just as fast [WS98].

Following up, Kleinberg developed the first decentralised algorithm which effectively finds shortest paths in small world graphs. A decentralised algorithm has access to information from its local context and a heuristic. Kleinberg used an *agent based-model*—located at some node on a graph—with a target node t: the knowledge was limited to the locations of the current neighbour nodes, as well as the relative location of t. The heuristic was the underlying grid of the graph. Thus, the agent knew which direction to take in order to reach t, even though the path to t is not known. The algorithm was able to find shortest paths reliably by making the right decision step by step, in a decentralised manner [Kle00].

In more recent research, Helic et al. modelled human navigation behaviour in information networks based on decentralised search. They developed multiple algorithms, one of them called ϵ-*greedy*: the algorithm is given a start and a target node, just like users are in a *wiki game* (sec. 2.1.2). In every step, the next click is chosen at random (from the available links) with probability ϵ, and with probability $1 - \epsilon$ it follows the link which brings it the closest to the target page, making the best choice within the local context. Therefore, the structure of the network has to be known.

One would argue that users do not know the structure of navigation networks like Wikipedia, and consequently, an algorithm which knows the

structure, would navigate differently. This is true for the *ϵ-greedy* algorithm, when applied with $\epsilon = 0$ (making full use of the structural knowledge), since it always finds the shortest path to the target node: those results differed from the ones produced of real users. This confirms the findings in sec. 2.1.2, that users do not navigate along the shortest path, but close to it. Thus, with $\epsilon = 0.15$, Helic et al. were able to yield better results. Since users are more erratic than machines, they concluded the presence of a stochastic component in human navigation trails.

Even though *ϵ-greedy* modelled user navigation equally well as the other two—more sophisticated—algorithms, Helic et al. took it one step further: due to the *exploration* and *exploitation* phase, they included a *decay* functionality. Instead of having a constant ϵ, it is reduced in every step until the algorithm reaches the target or ϵ equals zero. This imitates the fact that users tend to navigate towards known topics, where the probability of a random click is substantially smaller than in the beginning. The result of this approach mapped human behaviour in navigation networks almost perfectly [Hel+13].

The success of this *decaying factor* is the inspiration for this work. Instead of decaying link selection, this thesis investigated the effect of varying jump probabilities which were empirically measured through session extraction. It then determines the influence of applying those dynamic probabilities to random surfers which are not based on decentralised algorithms. In random surfers, the probability of a session ending is commonly chosen to be static and 0.15. This results in an exponential distribution of session lengths. Session length distributions of click data are frequently classified as log-normal distributions (see chapter 4) [Gon+09; Mei+09].

Gonçalves et al. stated further that "Node dependent jump-probabilities" as well as "node dependent entry-points" could be used to enhance these models to better match the heterogeneity of user traffic [Gon+09]. Meiss et al. suggested a non-uniform out-link selection as an additional improvement to their (non-markovian) random surfers [Mei+10].

2.2.5 Random Surfer on an Encyclopedia

In a joint effort with fellow researchers Geigl et al. of the *Knowledge Technologies Institute*[6] *(at Graz University of Technology)*, the Paper "Random Surfers on a Web Encyclopedia" was published[7]. The paper investigated the following two questions:

1. **Comparison of two random surfer models with real user click data:** to what extent do random surfers with teleportation imitate user navigation behaviour in information networks?
2. **Influence of search engines:** how do search engines affect how users access and navigate websites?

In order to find answers to those questions, the following steps were taken: first, one of the peers crawled the link structure of Austria-Forum. The crawler was given the home page as the entry point. It traversed the underlying graph in a breadth-first manner where all neighbours are visited and scanned for further links, which were then followed. This automated procedure was performed recursively until all reachable pages were visited. All binary files (videos, audio-files, and others) which did not contain information about user behaviour were removed. The result was a graph with more than 426 000 nodes (pages) and more than 16.5 million edges (links).

Next, the research used a subset (59 days) of the real world click data of *Austria-Forum* (details in sec. 3.2). Out of this data, an adjacency matrix A was created. An adjacency matrix is the matrix representation of a graph. In this case, the values of the matrix corresponded to the weights of the edges in the graph. In other words, the transition count between all pages was stored. Therefore, the value w_{ij} is the number of times the link was clicked, which originates at the page corresponding to index i and leads to the page corresponding to index j. The matrix was thus a square n times n matrix, where n is the number of pages.

[6] http://kti.tugraz.at/en/

[7] The Paper was accepted for the i-KNOW 2015 Conference and was presented on October 21-22 in Graz

A page view vector v was created as well. The vector also contains n elements where the values are the number of visits for the associated page. The vector was normalised and is called *reference distribution*[8]. The page visits and link transitions were then mapped to the graph.

Another peer created two random surfers: one random surfer was the PageRank random surfer called *uniform random surfer (URS)*. The second random surfer was called *pragmatic random surfer (PRS)* and was a slight modification of the URS: instead of choosing edges at random, the previously described weights of the empiric click data were used. Hence, the pragmatic random surfer should model link selection behaviour similar to users. Since link positions are biased, the weights (w_{ij}) were scaled using sub linear scaling:

$$scaled(w_{i,j}) = \begin{cases} 1 + \ln w_{i,j} & \text{if } w_{i,j} > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

The distributions of click count probabilities of the two random surfers were compared to the *reference distribution* using the Pearson correlation coefficient p. With a jump probability of 15% for both random surfers, it yielded the following results: the URS and the PRS have a positive correlation of $p = 0.98$, making them quasi equal. The conclusion was drawn, that the URS models user click navigation very well since the PRS symbolises click behaviour.

Compared to the reference distribution, the correlations were substantially lower. The URS reached $p = 0.38$ and the PRS $p = 0.47$. This can be attributed to the following cause: several poorly ranked pages in the URS/PRS are more important in the reference distribution, diversifying it in a lateral way. This phenomenon of a lateral distribution was already observed [Mei+10]. These lateral accesses are often related to pages situated deep in the hierarchical structure and can be connected to search engines. Users often enter on a specific page and leave without making a single click. They either found the information they were looking for, or went back to the search results to consult another page.

[8] The paper calls this distribution *lateral random surfer*. It will be called *reference (distribution)* throughout the thesis for clarity.

The following conclusion was made: "This result suggests that the lateral access to a website can not be solely captured by a random surfer with teleportation. [...] To capture the lateral access to a website from a search engine we need a new kind of random surfer model." [Gei+15]

This thesis builds upon these results and open issues. Several random surfer models were developed in hopes of outperforming the uniform and pragmatic random surfer. Then, the influence of variable jump probabilities on these random surfers is measured. Further analysis on click data, the importance of certain pages, and other structural components was performed as well.

My contributions to the paper [Gei+15] were the data set, its description, and plots. This thesis uses the following parts of it: the extracted graph of Austria-Forum, the calculations of the URS and PRS. These calculations were repeated in this thesis with an extended version of the data set. The framework for these calculations was provided by one of the fellow researchers.

3 Materials and Methods

This section is split into three parts. First it describes Austria-Forum which provided several data sets for this thesis. It then explains the choice of data set and how it was processed in order to extract sessions from it. The last part explains several random surfers in detail.

3.1 Austria-Forum Background

Austria-Forum[1] is an online encyclopedia. Its content is limited to the geographic scope of *Austria*. The concept of *Austria-Forum* as an information platform (encyclopedia) has been present in the World Wide Web for more than 20 years [Mau]. In 2008 it has been adapted to the growing *Web 2.0* by changing its format to a wiki concept [HMW08]. Some of the basic principles of Web 2.0 are:

- user generated content;
- power of the crowd;
- data on a large scale;
- architecture of participation;

which can be summarised to: knowledge created by the collective wisdom of many people, who simultaneously inspect and control the provided information, on a global scale.

Wikis, such as *Austria-Forum* (and *Wikipedia*), are among the core services of the Web 2.0. The quality of the contents of *Wikis* has been in discussion since their beginnings. Users can create or edit arbitrary articles, however

[1] http://austria-forum.org

they see fit. This results in unchecked facts and content of doubtful accuracy [And07; Stv+05].

3.1.1 Quality in Encyclopedias

Information on the Web was and still is of varying quality, not just when comparing unrelated web sites, but even within domains or encyclopedias [Mau; Stv+05]. Due to this varying quality, the user needs to decide whether the author or the information can be trusted [HMW08].

On encyclopedias, articles are often created by a collaboration of multiple authors, and, due to the scope of the Web, mostly by someone unknown to the reader. Often authors are editing articles anonymously. More authors working on a single article does not necessarily mean a higher quality outcome. Authors have to overcome the difficulty of collaborating over the Web. Coordination is a key aspect in that matter. Kittur and Kraut found that editors, who show commitment to Wikipedia, and contribute often on multiple articles, raise the quality of articles by contributing to them, due to their experience; as well as users, who only contribute to a single article multiple times, showing commitment to a very specific topic, often raise the quality indirectly, by participating and encouraging the discussion concerning that article [KK08]. Anthony et al. found that "Good Samaritans" (anonymous users with sole contributions) positively impact on quality as well [ASW07].

In case the author is not known, a decision regarding the trustworthiness of the provided information has to be made. Users having partial knowledge about the topic can check if the found information extends and fits their previous knowledge [HMW08]. Since users mostly look for information they do not possess, it is difficult to determine the correctness of it without checking another source, which contradicts the purpose of an information platform.

Information on the Web can be used for personal or professional knowledge gain, but is also used to spread knowledge through scientific publications and journalism. Problems which occur due to the influx and ease of access to information can be plagiarism, copy and paste, fragmented knowledge

(incomplete articles), copyright violations, biased information, and more [HMW08].

In recent studies, biases of information depending on cultural similarities have been shown in Wikipedia articles: even though editors of Wikipedia are urged to follow Wikipedias "Ethical principles for editors"[2] and write in an unbiased, objective way, Laufer showed that the (supposedly) same article in different languages had different emphases depending on geographic proximity and, thus, cultural similarity [Lau14].

3.1.2 Austria-Forum Revision

The problems of anonymous, potentially incorrect or fragmented information can be countered by only accepting publications from named, known, and trusted editors. Applying this principle to an information platform throws it back to its infancy, having to reevaluate every article to determine its quality. Reaching the tremendous amount of information currently available on Wikipedia would be an immense task, only solvable with many academics. Finding enough editors is, on account of the limited number of experts and funding, difficult if not impossible. Taking this step means losing quantity of content, but gaining quality. Few are willing to make this trade-off since having many articles, and thus links in between those articles, increases the chances of users finding relevant pages in search engines or visiting the web site in general, due to a larger variety in content [HMW08].

In 2008, based on the conditions at the time, Austria-Forum made the decision in favour of quality, allowing content to be published only by known and renowned contributors (editors). The approach taken should include the community—in form of critical discussions—as well as experts. The biggest downside of the novel approach, which is the limited number of experts and consequently content, was countered by reducing the scope of the encyclopedia geographically to Austria and by allowing links to quality-assessed external resources. The main objective was "an encyclopedia that

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

is citable and can be used in educational institutions such as schools or universities, as a starting point in many scientific inquiry, or simple as a fast lookup tool" [HMW08].

Even back then, Wikipedia was already the focal point of user-created information, with highly interlinked articles and knowledge collections. Google[3] was dominating the search-engine market almost exclusively. Kappe et al. showed that Google was the primary search engine in Austria also and that "Google is clearly privileging Wikipedia sites in its ranking", where the German version of Google ranked results from the German Wikipedia even higher than the international version of Google ranked English Wikipedia results. Kappe et al. formed the hypothesis that the high link count in between encyclopedia pages, which was even higher in the German Wikipedia than in the English version at that time, leads to Wikipedia entries being ranked so high [Kap+07]. Research published in the same year strengthens this hypothesis, showing that PageRank, and therefore highly interlinked pages, were still substantially contributing to higher ranks in Googles results [FE07].

Research showed that the link position plays a significant role when it comes to user behaviour on the Web: the *Primacy and Recency Effect* states that users are more likely to click the first or last item in a list, followed by the second, third, and its successors, each with fewer possibility than the one before. This is the case, even when the content behind the links is irrelevant [MHM06]. Weinreich et al. found that 76% of clicks are made without scrolling. Figure 3.1 shows a heat map, which highlights the regions where users are most likely to click, primarily at the top (left) [Wei+06].

As a result, having a high ranking in search results should increase the percentage of visitors coming from search engines. In this work, this phenomenon was observed in *Austria-Forum* as well.

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

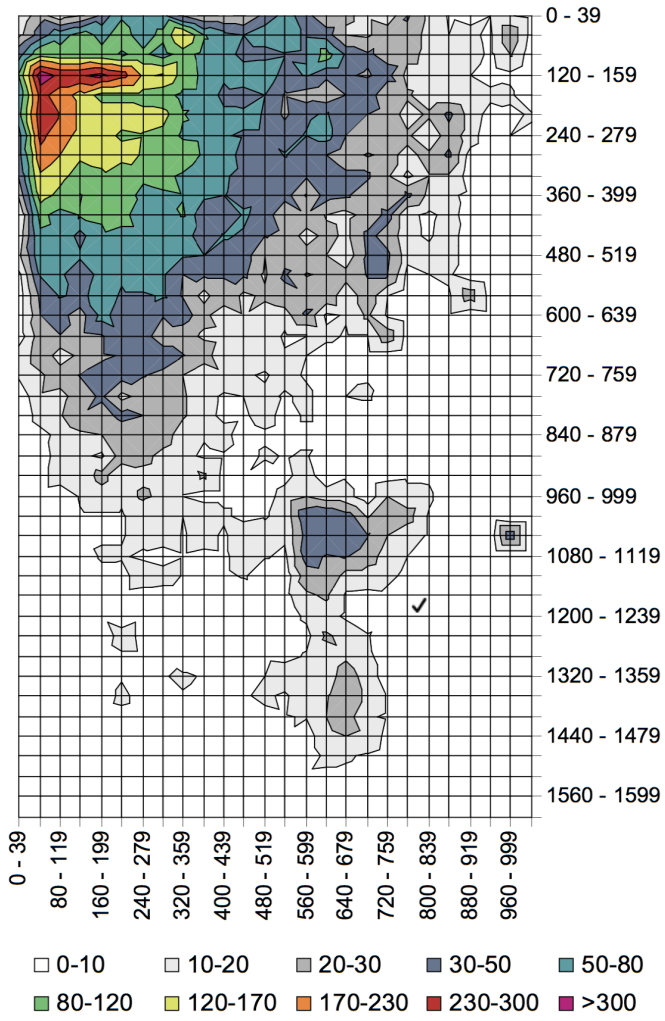


Figure 3.1: Click heat map: the position on screen where users are most likely to make a click. Users are most likely to follow links situated at the top of the site. The clustering at $\approx$ 1000 on the vertical axis is due to users often clicking on the "next" link in googles search results [Wei+06]. (reprinted from [Wei+06])

3.2 Data Set Description

Even though logging user request and activity of sites is—especially in recent times and primarily due to privacy concerns—often debated, valuable information can be gained from it. At the very least it can be used to analyse the growth of a web site.

For this thesis, three different log data sets of Austria-Forum were available, expanding over several years (see table 3.1). Even though the period of data set Z is the shortest of the three, it was chosen over the others for several reasons:

- The data set contains more details per request.
- Austria-Forum went through several updates, heavily influencing the structure of the site. Thus, recent log files provide the best match for the graph used in this thesis.
- Creating the graph which would map to the older logs is improbable, due to availability issues.
- Mobile devices have changed user behaviour since the beginning of logs.

Table 3.1: **Available data sets X, Y, and Z: data set Z was chosen over the other data sets. It contains the most recent and most detailed information about user behaviour on Austria-Forum.**

Month	2009	2010	2011	2012	2013	2014	2015
January		X	X	X, Y	Y	Y	
February		X	X	X, Y	Y	Y	
March		X	X	X, Y	Y	Y	
April		X	X	X, Y	Y	Y	Z
May		X	X	X, Y	Y	Y	Z
June		X	X	X, Y	Y	Y	Z
July		X	X	X, Y	Y	Y	Z
August		X	X	X, Y	Y	Y	Z
September		X	X	X, Y	Y	Y	
October	X	X	X	Y	Y	Y	
November	X	X	X	Y	Y	Y	
December	X	X	X	Y	Y	Y	

- Article collection[4] and user base grow steadily. The system has a different scale compared to several years ago.
- Findings from current logs best portray the current situation, providing the developers of Austria-Forum with potentially valuable information.

The chosen data set consists of 144 daily log files, collected from the middle of April through August 2015, resulting in slightly more than 38 million logged HTTP requests.

3.2.1 HTTP Request

Browsers are tools for displaying web pages. They communicate with servers hosting sites by sending requests and rendering the responses. These requests commonly use the *Hypertext Transfer Protocol (HTTP)*, and are either triggered when clicking on a hyperlink, or when entering a *uniform resource locator*, better known as *URL*, into the browser. If successful, the servers respond to the requests and transmit the requested pages.

An HTTP request consists of multiple fields which can be used to extract sessions or perform general analyses. Table 3.2 shows a sample request, the more important fields are:

- **Target**: holds the identification of the requested page, in this case the biography page of *Waltraud Klasnic* (former governess of Styria).
- **Content-type**: shows that a HyperText Markup Language (HTML) page was requested.
- **Remote-IP** and **session-ID**: can be used to identify users, where *session-ID* is often derived from cookies and *remote-IP* is the public IP-address of the requesting instance.
- **Referrer**: is a field of special importance. If the user clicks a link, the *referrer* is filled with the *URL* of the page where the link was clicked. If the user enters an *URL* or clicks a bookmark, the *referrer* is empty.

[4] http://austria-forum.org/af/Infos_zum_AF/Grunds%C3%A4tze#section-Infos_zum_AF_2FGrunds_C3_A4tze-4.Projektverlauf

- **Response code**: is part of the logged response of the server (and not part of the actual request). The code 200 implies *OK* which resulted in the transmitted page.
- **User-agent:** provides information about browser and device, but not enough to make distinctions between users.
- **User-name-hash**: is used for identifying logged-in users without disclosing their identity.

Clicking a link often triggers multiple requests, since web pages usually consist of multiple segments which need to be transmitted to the user. This thesis focuses primarily on user navigation, therefore the logs were filtered extensively to extract only the necessary requests.

3.2.2 Log Cleaning

First, in order to reduce the amount of (unnecessary) requests quickly, those created by machines (bots) were filtered. Bots are often called crawlers. They are automated processes. A crawler is a bot that follows links on the web, often used to create a layout of the web. More than 550 bots were identified through bot-related keywords in the *user-agent* field. Since bots tend to crawl as many pages as possible, a substantial amount of

Table 3.2: **HTTP request log entry**: the table shows the HTTP parameters which were logged and an example query entry where the user came from Google and visited the page of *Waltraud Klasnic* which was successfully transmitted.

Date	2015-04-12 23:22:13,893
Method	GET
Response Code	200
Host	austria-forum.org
Target	[...]/Biographien/Klasnic,_Waltraud
Referrer	https://www.google.at/
Content-Type	text/html;charset=UTF-8
Session-ID	DCFBFBECFDE865896890674085346441
Remote-IP	1.1.1.1 (for anonymity)
User-Agent	Mozilla/5.0 (iPad; CPU OS 8_2 like Mac OS [...]
User-Name-Hash	−15832049

requests was removed. In general, bots have good intentions (e.g. indexing all sites so they can be ranked in search engines), but not all do. Some bots do not identify themselves as such (through the *content-type* field). Therefore, another approach was taken: sessions with more than four clicks, where more than half of these clicks had no entry in the *referrer* field, were filtered. Bots rarely set the *referrer*. Even though it is also possible for users to have multiple empty referrers, within this thesis it was found that the above parameters are properly set, keeping shorter sessions which could be created by real users: clicking on the bookmark of Austria-Forum twice (or refreshing the page) and making one more click afterwards, results in more than half of this sessions' referrers empty (with a session length of three). Even though similar scenarios are possible for sessions with length of four and more, they become increasingly unlikely.

Several requests were *self loops*: the *target* and the *referrer* field contained the same information. These requests were removed as well since it was considered abnormal behaviour, probably by bots, which does not represent user behaviour and could not be reproduced.

Next, all requests which yielded templates, xml files, attachments or other structural information were removed via the *content-type* field, since they do not contain any navigational information. Only requests yielding HTML pages remained.

The *request method* field was used to only keep those requests, which would try to "GET" something from the server. The resulting requests were then further diminished: only those with the positive *response code* (200) were kept.

Lastly, all sessions which contained requests indicating admin, editor, or staff behaviour were removed as well: requests generated from staff were identified via a proxy IP-address; admins requested pages which were not reachable by normal users. In order to reach those pages, they needed to be logged in. Thus, the *user-name-hash* field had a value which could then be classified as an admin user. All sessions by admins were removed, even those which did not access any restricted areas. The equivalent was done for editors, since they do not portray standard user behaviour: sessions which contained requests alike *pressing the upload button* to attach files to an article or triggering the preview window for a new article.

Table 3.3 structures the reduction into its steps[5].

Out of the initial (approximately) 38 million requests, 1 568 422 were associated with navigational behaviours of users, or simply put, *page views*. The data set consists of 144 days, resulting in an average of approximately 10 892 daily page views.

Pages can be reached via (slightly different) URLs. The following links all lead to the home page of Austria-Forum:

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

In order to correctly assess the page view count for each page, to map those pages to the pages of the graph, and to compare the target of one request to the referrer of the next, all URLs were normalised.

Table 3.3: **Extensive log data filtering: methods and removed requests. The data set contained slightly more than 38 million requests. After applying several filter methods, about 1.5 million requests remained. These requests contain user behaviour and were the basis for the analysis performed in this thesis.**

Method	Requests	additional information
user-agent	16 829 393	keywords: *bot, slurp, crawler, spider*
content-type	14 048 089	non-HTML: *audio-/video, XML, templates*
request method	3 340 129	keep GET requests
status codes	312 487	requests not yielding status *OK*
spoofed referrers	128 223	conspicuous text in referrers
empty referrers	1 525 848	prevailing empty referrers in session
self loops	126 868	referrer equals target, indicating bots
editor sessions	295 395	sessions with editorial privileges
staff & admin	72 810	IP-Address, user names, privileges
Total	36 679 242	

[5]Depending on the order of execution, removed requests per method vary.
[6]This is the outdated domain of Austria-Forum.

3.2.3 Session Reconstruction

The *session-ID* parameter of the HTTP requests is used as the initial assessment of sessions. The timeout value of those session-IDs is set to 30 minutes[7]. Meiss et al. argue that sessions should not be split solely based on timeouts but rather logical affiliation; 15 minutes was found to be the best value, since the investigated user sessions tended to last a little over ten minutes. Those studies were conducted on the entire Web where users browse everything from news to social networks, online shops, and more [Mei+09].

The data set of this thesis is restricted to one logical domain: an information platform. Information platforms help users gain knowledge. Pages on information platforms often contain long texts whereas social networks are generally dominated by pictures. Since it takes substantially longer to read informative articles than consume other available media, a timeout of 30 minutes was considered appropriate for encyclopedias.

The logs consisted of 916 182 unique session-IDs, resulting in an average session length of 1.71 clicks and a site jump probability of

$$\frac{\text{number of sessions}}{\text{number of requests}} = \frac{916\,182}{1\,568\,422} \approx 0.58 \tag{3.1}$$

meaning that on average, the probability of users making a click is 42%.

Qiu et al. introduced the concept of *referrer trees*: by following the chains of referrers, one can construct a tree-like structure. Only an empty referrer would indicate the beginning of a new session, setting that request as the root of the tree [QLC05]. The policies used in this thesis are stricter: a new session starts when the referrer is one or more of the following:

- *empty*: the user either clicked on a bookmark, or entered the URL into the address bar.

[7] According to the staff of Austria-Forum it should be the default value of *apache tomcat*. See: https://tomcat.apache.org/tomcat-7.0-doc/config/manager.html#Common_Attributes

- *non-empty, external*: this is the most common case. It appears whenever the user reached Austria-Forum through an external link, for example a Google search result.
- *search result, internal*: using the search functionality of the site enables the user to "jump" to different places in the underlying graph. This splits several sessions, making them shorter. Having a random surfer which jumps equally often in the graph as users do, enables for better modelling of user behaviour. Therefore, this step is necessary.

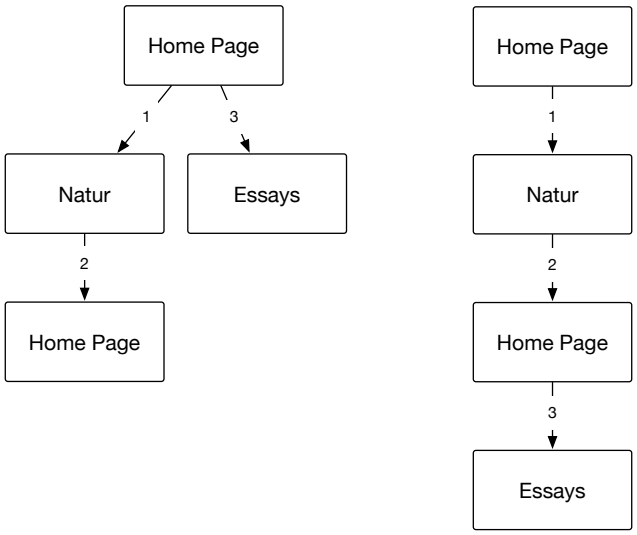
Applying these policies increases the session count to 1 089 763 sessions, which elevates the site jump probability of equation 3.1 to $\approx 69\%$. The average session length drops to ≈ 1.44.

Table 3.4 shows a simplified example session, consisting of four requests and thus, four referrer/target pairs with the initial referrer empty. When

Time t	Referrer	Target
0		Home Page
1	Home Page	Natur
2	Natur	Home Page
3	Home Page	Essays

Table 3.4: Four consecutive requests of pages (targets). The initial referrer is empty, the home page was most likely opened trough a bookmark. Recreating the trees from this sequence of requests yields two different possibilities, shown in figure 3.2.

pages occur multiple times as a target page ($t(0)$ and $t(2)$) and at some later point of the session in the referrer ($t(3)$), the resulting referrer tree is not distinct. Figure 3.2 shows the two possible trees for this sequence of requests. This thesis uses the option portrayed in figure 3.2b, mapping referrers to the targets which appeared last (by time).

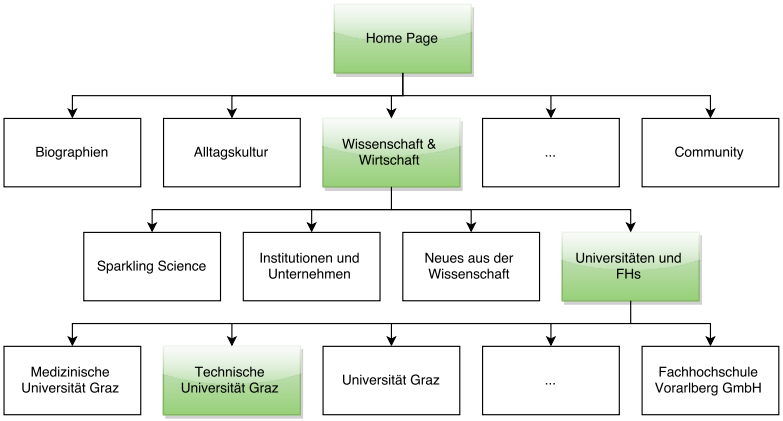


(a) Possibility 1 (b) Possibility 2

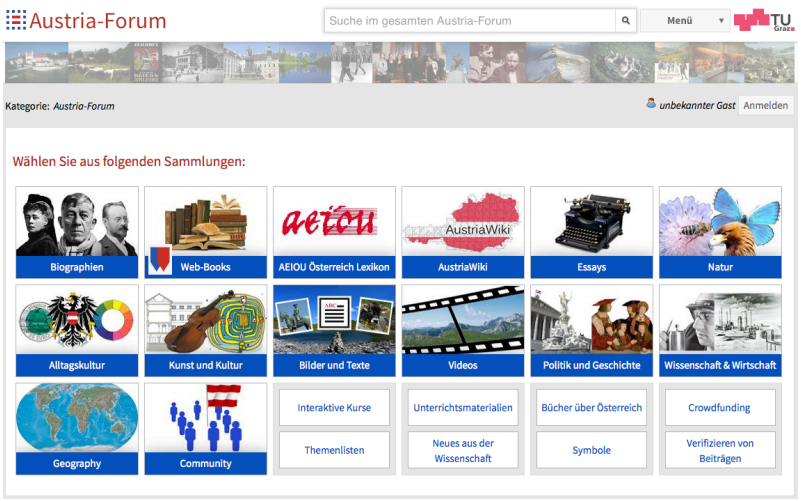
Figure 3.2: Multiple possible referrer trees: one sequence of clicks can produce multiple referrer trees. These trees are the possible representations of the sequence of requests displayed in table 3.4. The request at t(3) has the *home page* specified as the referrer page. Due to the home page being requested twice before (indicated by the target field), the succession of clicks is not uniquely identifiable. Figure 3.2a shows the tree produced if the user opened the "Natur" page in a new tab (window), clicked on a link back to the home page, and closed the tab (window). The request of the "Essays" page was then performed from the original tab (window) which still showed the home page. Figure 3.2b shows the tree which is produced when the referrer of a request is linked to the last time it was the target. This method is used in this thesis.

3.2.4 Austria-Forum Structure

Austria-Forum is built in a taxonomic way, which can be seen as a tree-like structure, shown in figure 3.3a: The home page (figure 3.3b) is the root of the tree which is the top level of the hierarchy. The content is split into several categories, often followed by sub-categories. As an example, "Graz University of Technology" as well as its parent categories are highlighted in the figure. The current position within this taxonomy is always visible to the user. Links to the parent categories are available on the page. This leads to high in-link counts and PageRank in category pages and the home page.



(a) Austria-Forum hierarchy



(b) Austria-Forum home page

Figure 3.3: Austria-Forum structure: Figure 3.3a shows the taxonomic structure of Austria-Forum. The home page is the root of the tree, followed by its categories and subcategories. *Graz University of Technology* and its affiliations are highlighted. Figure 3.3b shows the home page. The 22 categories are visible in the body of the page. Below the search bar and the logo of Austria-Forum (top left) is the banner. It spans horizontally across the site and contains variable links.

3.3 Random Surfer Models

Random surfer models, also called random surfers or random walkers, are agent-based models of a random walk (sec. 2.2.1). They are used for simulations in order to perform measures on the importance of nodes in a graph (sec. 2.2.2). Random surfers use probabilities when selecting the next node or edge. A biased random surfer uses probabilities which are not uniform, thus, making it biased towards a preferred set of nodes or edges. The addition of a bias enhances the capabilities of a random surfer in order to approximate user behaviour and user choices.

Geigl et al. created a biased random surfer which does not choose edges (links) uniformly, instead it privileges edges which were used more frequently by real users. Neither the *uniform random surfer*, nor the biased version were capable of accurately reconstructing real user data (see sec. 2.2.5 for more details). The click data of real users was determined more *lateral* than the simulated results of the random surfers. A lateral access includes pages which are far away (several clicks) from *hubs* of the underlying graph of web sites. In encyclopedias hubs are often the home page or category pages (sec. 3.2.4) [Gei+15].

This work extends the research by Geigl et al. [Gei+15], testing several random surfers with the intentions of best approximating the user click data. These random surfers are not biased when choosing edges. Some of them are choosing the starting pages of the simulated sessions in a biased manner. These random surfers were created in order to answer the research question (sec. 1.2) of *whether it is possible to create an agent-based model with simple measures, which simulates the lateral access of users on an encyclopedia better than previously created models*. The created random surfers are memoryless, basing the next decision only on the current state of the surfer. They can thus be seen as stochastic processes which comply with the Markov property (sec. 2.2.1).

3.3.1 Output of Random Surfers

A probability distribution created by random surfers can be seen as a vector v with n elements, where n is the number of nodes in the graph. Then v_i (with $i = 1, 2, 3, ..n$) is the probability a random surfer is at node i when stopped at an arbitrary time during the random walk. Thus, the sum of the vector adds up to one:

$$\sum_{i=1}^{n} v_i = 1. \tag{3.2}$$

Such a probability distribution is called *stationary distribution* in the context of converging stochastic processes (sec. 2.2.1 and 2.2.2).

When simulating a random walk, there are several possible termination conditions:

- n steps are performed;
- the starting node v_1 is re-visited (n times);
- every node or edge of the graph is visited (at least n times);
- the changes in the probability distribution are insignificant when performing another n steps.

Two possible termination conditions were used: a random surfer either created four million sessions, or the probability distribution within the last 100 000 steps only changed marginally. All random surfers reached the former condition first and stopped after four million sessions. Figure 3.4 shows the relative error of a typical random surfer simulation. The formula of the relative error was provided by Geigl et al. [Gei+15]. It was calculated between the distribution of page views at step x, and the distribution at step $x - 100\,000$. Due to the flattening of the curve, four million sessions seemed appropriate in order to minimise the error and keep the calculations feasible.

A fixed amount of sessions does not result in the same amount of clicks for every random surfer model. Random surfers with lower (average) jump

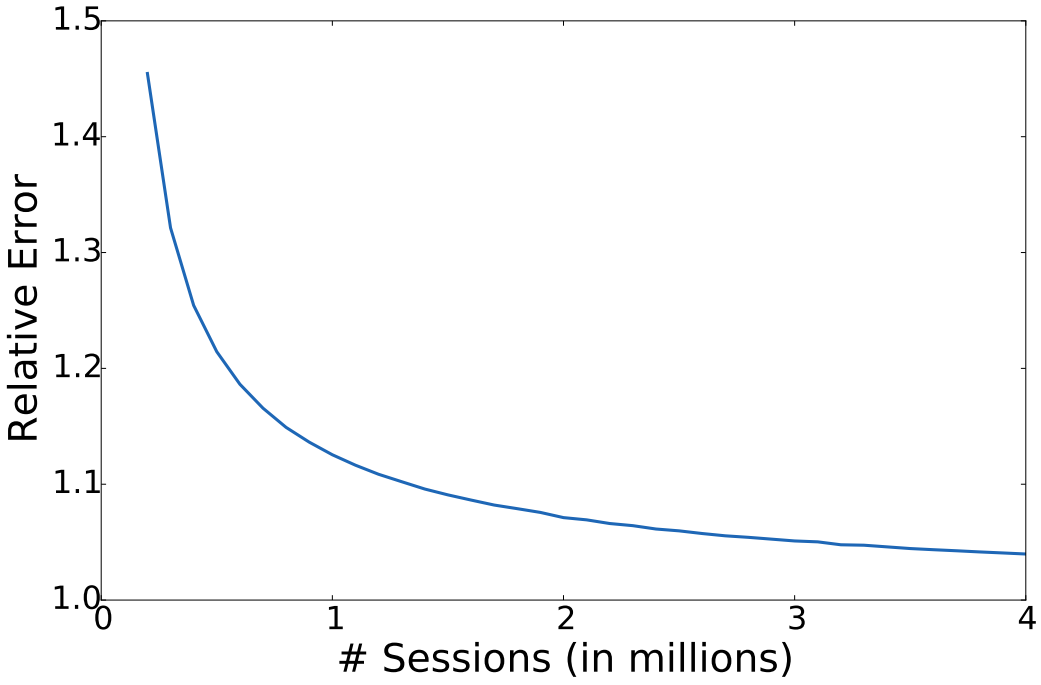


Figure 3.4: Random surfer relative error: the relative error between the distribution created by the random surfer after x sessions (horizontal axis), and the distribution from $x - 100\,000$. Due to the flattening of the curve and the infeasibility of the calculation, every random surfer simulated four million sessions.

probabilities created longer sessions (more clicks) and thus more page views than those with larger average jump probabilities.

Accuracy of Random Surfers Now two distribution vectors can be compared with different correlation methods. In this work, the *Pearson Product-Moment Correlation Coefficient* (often only called *Pearson Correlation*) was used.

The resulting distributions of the random surfers were compared to two reference distributions:

1. *empiric user click data*: the probability distribution of page visits by real users, extracted from the log files. This distribution is used to measure

how well (biased) random surfers can emulate a lateral distribution. In this distribution many page views are influence by search engines.

2. *empiric user click data without external sessions of length one (zero clicks)*: this distribution was created in order to best measure how well random surfers model page views created through *navigation*. Due to many users leaving the site without making a click, these page views were not created through *navigation within Austria-Forum*, but rather a targeted jump guided by search engines or other sources (Wikipedia). Thus, all sessions of length one (zero clicks) which originated at an external site were removed.

3.3.2 Properties of Random Surfers

The developed random surfers possess two properties which can be set:

1. the selection of entry pages, once the random surfer jumps;
2. the frequency of jumps.

Both of these properties are split into three possibilities. For the selection of entry pages (the starting pages of sessions) the following options are available:

- *uniform probability distribution* (abbreviation: *(U)*): this is a common choice for random surfers. Geigl et al. showed that this uniform selection does not properly cover the lateral access of real users [Gei+15].
- *landing pages (LP)*: unpredictable events like earthquakes can lead to users visiting different pages at different times (sec. 2.2.3), possibly pages they would not have visited otherwise [GR14]. In order to minimise this stochastic component of user navigation, the entry pages of user sessions (called landing pages) are used as the entry pages for random surfers.
- *preferential pages (PP)*: these preferential pages are proportional to the combined view counts of the real user click data. More important pages, those with many views, have a higher chance of being the starting page of a session.

The options for *frequency of jumps* are the following:

- the commonly chosen probability of 15%;
- the average jump probability of Austria-Forum: 69%;
- the dynamic jump probabilities (see chapter 4).

In order to answer the second research question (sec. 1.2), namely, *the influence of dynamic jump probabilities on the results produced by random surfers*, some of the random surfers include such dynamic jump probabilities.

The probability of 69% was chosen in order to best measure the influence of dynamic jump probabilities. Random surfers with a jump probability of 69% are producing sessions of the same average length as those random surfers with dynamic jump probabilities. Therefore they produce an equal click count, and take advantage of the landing pages and preferential pages to the same degree. The only difference is a static versus a dynamic jump probability.

Table 3.5 summarises of the developed random surfers.

Name	Page Selection	Jump Probability
U15	uniform	static 15%
U69	uniform	static 69%
UDyn	uniform	dynamic
LP15	landing pages	static 15%
LP69	landing pages	static 69%
LPDyn	landing pages	dynamic
PP15	preferential pages	static 15%
PP69	preferential pages	static 69%
PPDyn	preferential pages	dynamic

Table 3.5: Random surfers: this table shows the random surfers which were used in the simulation of user navigation behaviour. Random surfers with uniform page selection are true random surfers. Random surfers with non-uniform page selection are biased random surfers. The jump probability of 15% is used most commonly by random surfers in related research (chapter 2); 69% is the average jump probability on Austria-Forum; and *dynamic* are the jump probabilities which vary in every step based on the extraction from real user click data. Every type of random surfer (U, LP, PP) was performed with all three possible jump probabilities in order to determine whether a different jump probability than the commonly chosen 15% increases the capabilities of a random surfer to model user navigation.

4 Results

This chapter is split into two parts. The first part presents the results of the click analysis with a focus on sessions. The second part breaks down the results of random surfers and shows their correlations which determine how well they model user navigation in Austria-Forum.

4.1 Click Data Analysis

The size of the encyclopedia Austria-Forum is smaller compared to other encyclopedias (e.g. Wikipedia) but the underlying wiki structure with many inter-linked nodes is similar. This section analyses several aspects of user navigation in Austria-Forum in order to compare it to previously found attributes of user navigation on the entire Web and on other encyclopedias. First it investigates the influence of timeout values in order to split sessions; followed by the influence of search engines on session lengths and jump probabilities. It further provides an estimation of session times as well as calculations on the amount of consecutive clicks and the usage of multiple tabs and the back button. Lastly, it shows the influence of search engines on all nodes while introducing the two distributions of click counts (per page) which will be used by the random surfer in the next section.

Session timeout intervals The most fitting timeout value between two requests is often debated and measured (sec. 2.1.1). In this thesis, sessions are split when the path of a user is not continuous and can thus not be cleanly reconstructed. A user making use of the search functionality generates a custom page which only fits the specific search requests. Such a session can not be mapped to the underlying graph, since such nodes

are also generated. Thus, a new session would start in this scenario (sec. 3.2.3).

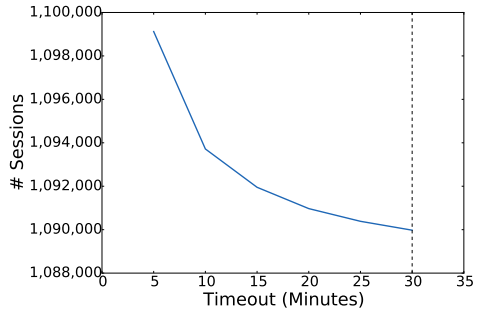
By making use of the logical linking of requests, namely matching referrers to previous target pages, sessions are split when the paths can not be reconstructed. Thus, the need of a timeout value is reduced to sessions including clicks which were several minutes apart. Still, it should not be chosen arbitrarily, otherwise sessions could last for days if the user forgot to close the browser window. The logged *session-ID* had a timeout of 30 minutes which seems plausible considering the length of articles in information platforms.

In order to measure the effects of a shorter timeout, to see how long users are actually taking between clicks, sessions were split with timeouts of 5, 10, 15, 20, and 25 minutes. Figure 4.1 shows what happens when reducing this timeout.

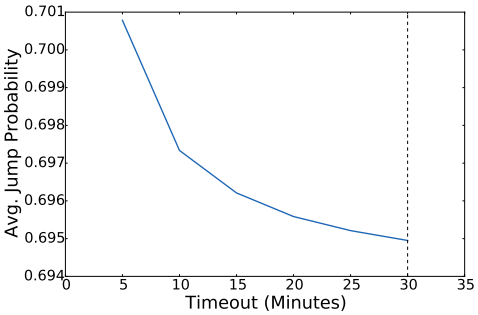
A shorter timeout results in more sessions (some are split), consequently in fewer clicks per session (on average), and a higher average jump probability. Note in Figure 4.1a the *vertical axis*: the amount of split sessions is marginal compared to the overall number of sessions. Assuming a timeout of 5 minutes, fewer than 10 000 sessions are split, that is less than 1% of sessions. On the other hand, about 90% of sessions consist of only one page view (zero clicks). These sessions cannot be split. Hence, there are roughly 100 000 sessions with more than one click where potentially 10% would be split, still assuming a timeout of 5 minutes.

For all three plots in figure 4.1 the biggest drop/incline (vertical axis) is between five and ten minutes (horizontal axis). This means that out of the sessions where two clicks were more than five minutes apart, only half also took longer than ten minutes. Considering this and the flattening of the curve in figure 4.1a, it suggests that the majority of users fall within the selected 30 minute timeout and only a few sessions were split by the preexisting timeout of 30 minutes. Still, out of the sessions with an inter-click time of more than 5 minutes, approximately 25% have at least 15 minutes between two clicks. This strengthens the hypothesis that a timeout value of 30 minutes is appropriate for an information platform and should not be reduced further even though other researchers calculated a timeout of 15 minutes for the entire Web (sec. 3.2.3). One should keep in mind that the

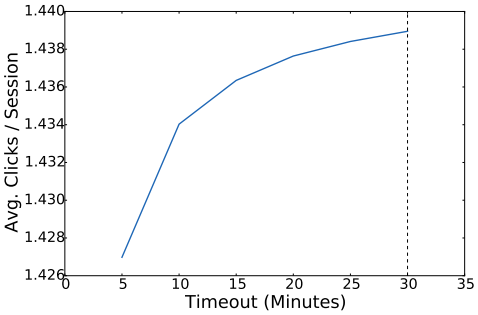
time spent on the last page is not included, which likely makes sessions even longer than the log files show.



(a) Session count



(b) Site jump probability



(c) Average session length

Figure 4.1: Impact of different timeout values: timeout values are often used to split sessions when two consecutive requests take longer than the specified timeout. This figure shows the minimal effect of shorter timeout values on *session-count* and *-length* as well as the average jump probability. For this thesis, a timeout value of 30 minutes was chosen.

Session lengths The session length distribution for lengths less or equal ten (up to nine clicks) can be seen in figure 4.2a. The distribution is highly skewed, with many users leaving the site without making a single click. This suggests that users were already steered towards the exact page they were looking for. In order to determine who provided them with the perfect information, the domains included in the referrer were extracted. These were extracted *only* from the sessions of length one (zero clicks). Table 4.1 shows the segmentation of these domains. The vast majority were directed to Austria-Forum by search engines. Several users came from *AEIOU*[1], a deprecated cultural lexicon which is now integrated into Austria-Forum[2].

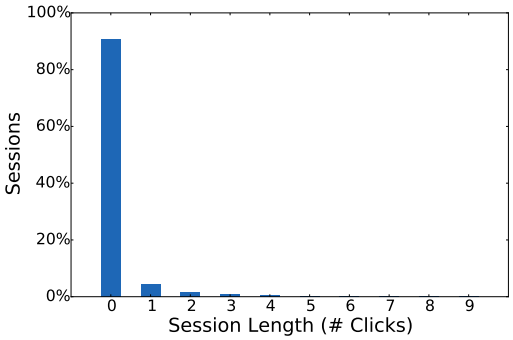
Source	Count	% (rounded)
Google	489 653	89
AEIOU	11 591	2
Wikipedia	9 942	2
Yahoo	6 111	1
Bing	6 143	1
other search engines	8 563	2
other	16 822	3
Total	548 825	

Table 4.1: Segmentation of domains: more than half (548 825) of all logged sessions (1 089 763) were of length one (zero clicks) and originated from an external site (mostly search engines). This table shows the predominant domain(s). AEIOU still seems to have several active users.

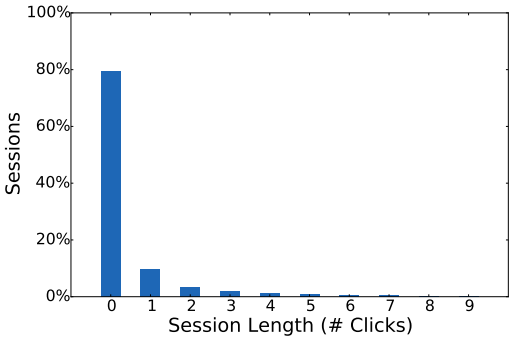
Figure 4.2b then shows the distribution with all these sessions of length one (zero clicks) removed where the user *came from another site* (referrer not empty) and left without making a click. Even though a large portion of sessions ($\approx$ 50%, $\approx$ 550000) was removed, the distribution changed only to some extent. The zero-click sessions were still the dominant force. These sessions include bookmark clicks (without a follow-up click), users leaving after entering a search, possible left-over bots, and others.

In order to put this into perspective, figures 4.2c and 4.2d show session

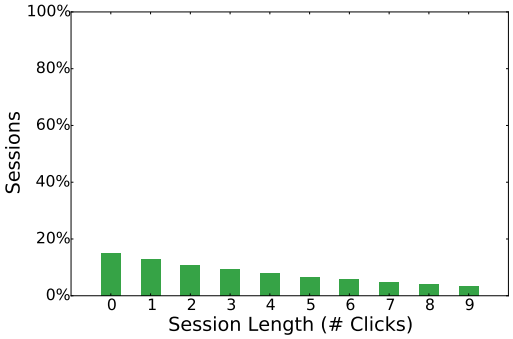
[1] http://www.aeiou.at
[2] http://austria-forum.org/af/AEIOU



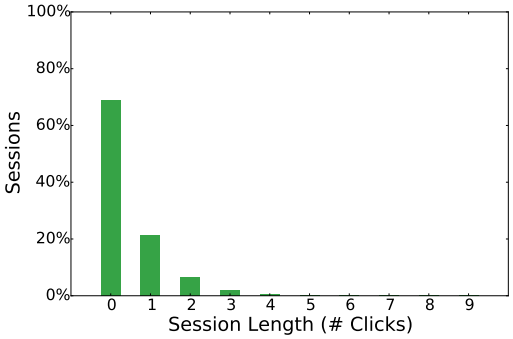
(a) Austria-Forum session distribution



(b) Austria-Forum session distribution without sessions of length one (zero clicks) originating from an external site.



(c) Common random surfer session distribution: the resulting distribution when the jump probability is set to 15%.



(d) Session distribution with jump probability 69%, the average jump probability of Austria-Forum.

Figure 4.2: Initial ten session lengths: this figure shows the session length distribution of all data in 4.2a. About half of the sessions do not include any clicks and originate at other domains. Even when these sessions are removed, the distribution is still strongly skewed left (4.2b). The figures in green (4.2c and 4.2d) illustrate the resulting session lengths when the jump probability is the same at every click. This suggest that constant jump probabilities, especially the standard of 15%, are not resulting in proper session lengths and thus link-usage.

length distributions for constant jump probabilities of 15% and 69%. These jump probabilities are the *common random surfer jump probability* and the *average Austria-Forum jump probability*, respectively. The figures show the session length distribution random surfers with these probabilities generate. The difference between the distributions of empirical data (figures 4.2a and 4.2b), and those generated with a static jump probability (figures 4.2c and 4.2d) strengthens the drive to further investigate the influence a dynamic jump probability has on the outcomes of random surfers.

Due to dominant sessions of length one (zero clicks), the session length distribution of Austria-Forum (figure 4.2a) is difficult to interpret based on that representation. It is skewed left with a long tail. It just fits a log-normal distribution with Log-Likelihood ratio $R = 0.69$ and significance $p = 0.31$ compared to a power-law distribution. In order to see the distribution of session lengths beyond ten clicks, figure 4.3 shows the session length distribution on a log-normal scale.

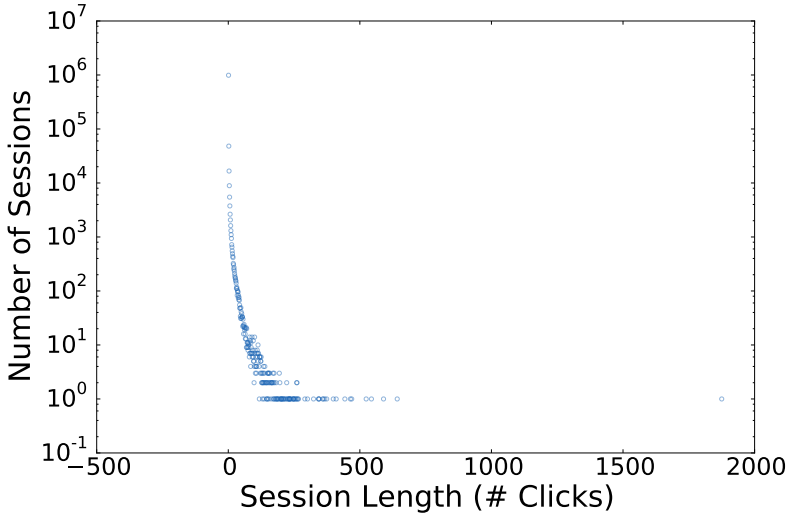
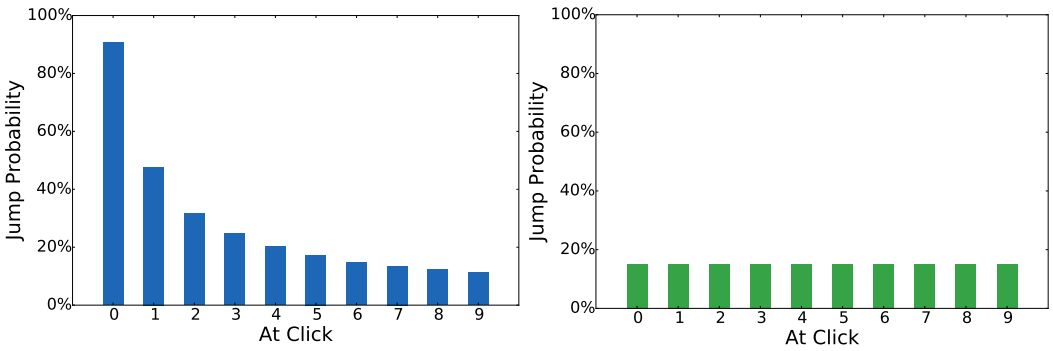


Figure 4.3: Session length distribution (log-normal): this plot shows the trend of session lengths. The shorter the sessions, the more there are. Still, several long sessions (large values on the horizontal axis) are visible. The longest sessions were investigated and were all verified to originate from human users, not from machines. The distribution is on the verge between a log-normal and a power-law distribution. It leans more towards log-normal with a log-Likelihood ratio of $R = 0.69$ (significance $p = 0.31$).

Two extreme values are apparent: at the value one on the horizontal axis the number of sessions is substantially larger than any other, as mentioned before. The longest logged session is visible at session length ≈ 1900. Thus, random surfers with the empiric and dynamic jump probabilities would have a probability 100% to terminate a session if they reached ≈ 1900 clicks. From this perspective it becomes clear that shorter sessions are more dominant. The longer the sessions get, the fewer sessions of that length exist.

Jump probabilities Due to the session distribution plot (figure 4.2a), the percentage of sessions with length one (zero clicks) is known. This percentage value equals the jump probability in the first step. It is now of interest, how many of the remaining users are leaving instead of making another click. Figure 4.4a shows the trend for the first nine clicks. With every click, the probability of a jump decreases. Figure 4.4b visualises the uniform jump probability of common random surfers. This contrast discloses the potential for improvement in terms of choosing the proper parameters for modelling user navigation.



(a) Austria-Forum (b) Common random surfer

Figure 4.4: Comparison of jump probabilities: even though the majority of users left Austria-Forum without making a click (figure 4.4a), it is still interesting to see the percentage of users that leave before the following click, and the one after that. It becomes clear that the probability of another click increases with every click and jumps becoming less likely (for the first few clicks). Figure 4.4b shows the contrast to the commonly assumed uniform jump probability of 15%.

At some point, the jump probability is extracted from the longest session. At this threshold, no user continued clicking. This results in a jump probability of 100%. Figure 4.5 shows the fall and rise of all jump probabilities. The jump probability of 100% is reached after $\approx$ 1900 clicks, matching the longest logged session. For missing values (e.g. between $x \approx 700$ and $x \approx 1900$), the last known value was taken.

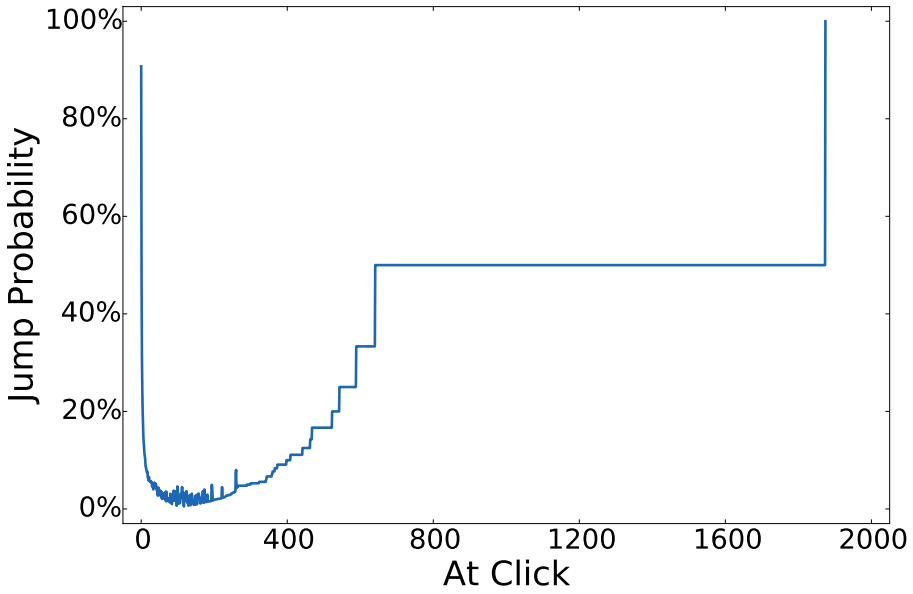


Figure 4.5: Austria-Forum jump probabilities: this plot shows the entire progression of the jump probabilities, empirically measured through its click data. Starting at $\approx$ 90% the probability drops rapidly. For now it most resembles a parabola in the beginning, but due to several missing and small values no accurate assessment can be made (yet). Missing values were filled with the previous value (creating horizontal lines). Once more data is gathered, I predict that the jump probabilities initially rise slowly after the global minimum, but with increasing values on the horizontal axis the rise quickens. A jump probability of 100% will always be the result once the length of the longest logged session is reached.

Session times Figure 4.6 shows the appearances of session lengths *by time*. These values are estimates. The times users spent on the last page could not be extracted through HTTP request logs (sec. 3.2.3). The times for all but

the last page were calculated by comparing the difference of the timestamps between two requests. The total session time was calculated by adding the average time the user spent on the previous pages of the session. This is, especially for shorter sessions, an approximation at best: a user who opened a page and made one click after five seconds would be classified at ten seconds seconds on the horizontal axis (two page views with five seconds each). This explains the sparse population between zero and ten seconds session time. It can be argued that the last page is the page where the user found the information he was looking for, and thus the page they stayed the longest on (studying the information). On the other hand, the last page might be the page where the user spent only a short period, if the information was simple or the user gave up looking for it.

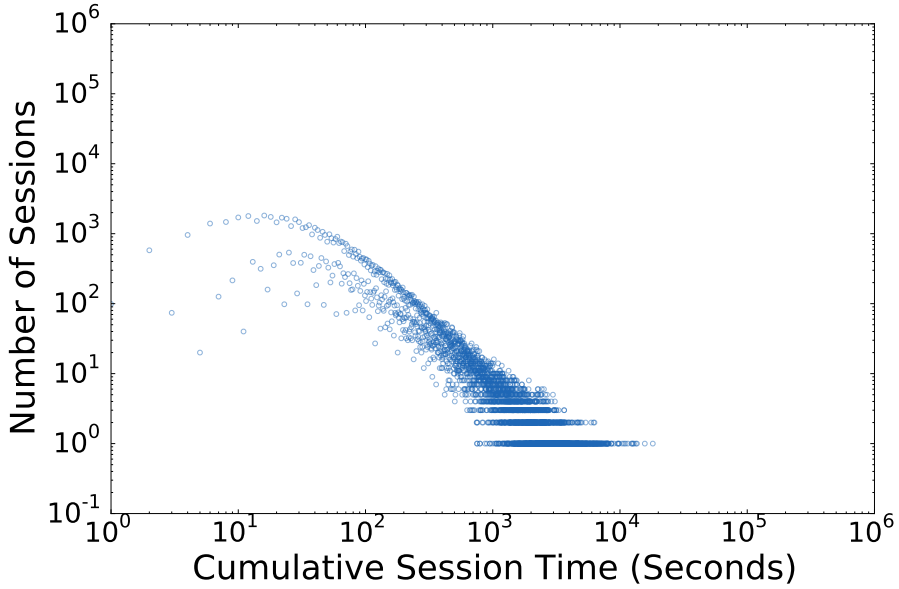


Figure 4.6: Session time distribution (log-log): the time spent on the last (arguably most important) page is not logged, thus this distribution is an approximation of session durations. The average page-view-time of the session was added to compensate for the missing value. Still, there are several sessions taking 15 minutes (900 seconds) and more; indicating that several users spend longer periods navigating within Austria-Forum.

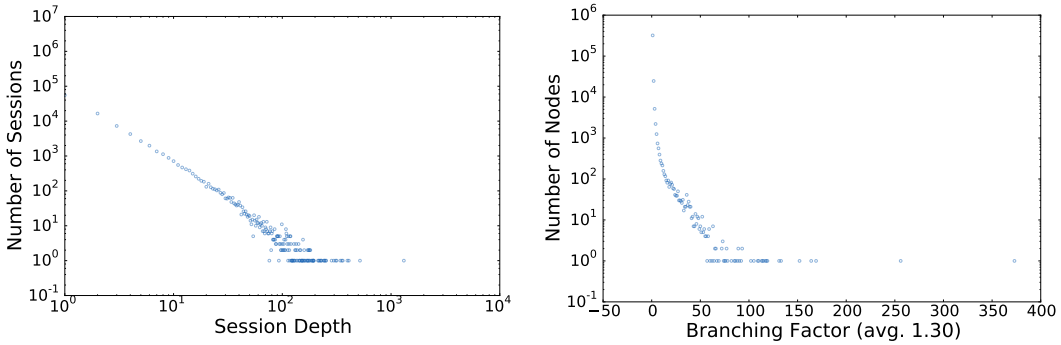
It is still interesting to see that several sessions are longer than 15 minutes (approximately the 10^3 tick on the horizontal axis); they might be even

longer in reality. It is an indication that users are generally interested in the content. It will be interesting to see the session time distribution calculated on other domains compared to figure 4.6. Qiu et al. first used this method of session time calculation. They found the distribution follows the power-law. The average session length was two hours. The average session length of two hours was calculated on the entire Web. Users browsed on approximately 21 pages on 5 different web sites (on average). The average time spent per page was five minutes. The research was conducted in 2005 [QLC05]. The average time per web site approximates the time distribution of Austria-Forum. The average sessions lengths as well as the average session lengths per site do not resemble the findings of this thesis. This might be due to the rise of mobile devices, the evolvement of the Web, or the influence of search engines on the domain of Austria-Forum.

Referrer tree analysis In order to analyse how often users go back to the previous site, open multiple tabs, or how many consecutive links they follow, the referrer trees were analysed. Figure 4.7 provides information about the session depths (figure 4.7a) and the branching factor (figure 4.7b).

The average session depth is 0.31 due to the dominant zero-click sessions. Yet, there are several trees of greater depth, meaning sessions with matching referrer-target pairs. One session had a depth of more than 1 000 clicks. Figure 4.7a shows the distribution of these session tree depths on a log-log scale. It shows that there are still several occurrences where users followed link after link.

Figure 4.7b shows the branching factor of trees in a plot with log-normal scale: the value for 50 on the horizontal axis amounts to about 10. This means that there were ≈ 10 occurrences where a session had a node which had 50 children. These nodes with abnormally large branching factors are due to some sites keeping the referrer, even when the target changes (due to the implementation of Austria-Forum, see sec. 5.3 *limitations*). They should thus be discarded. On average, a node has 1.3 children. This shows that users are backtracking and opening multiple tabs. In order to perfectly model this behaviour with random surfers, one needs to take this into account. This has been discussed in related research (sec. 2.1.2) and is beyond the scope of this thesis.



(a) Session depth (log-log) (b) Node branching factor (log-normal)

Figure 4.7: Referrer tree analysis: referrer trees provide information about the link-usage patterns of users. The session depth (4.7a) portrays the number of consecutively followed links which equals the depth of the referrer tree. There are several sessions of significant depth, most probably created at the cause of interesting image galleries. The average session depth is—due to the vast amount of sessions without clicks— only 0.31. The node branching factor 4.7b provides information about the usage of the back button and multiple tabs. The plot shows the occurrences (vertical axis) of a node in a tree having children as displayed on the horizontal axis. On average, every node has 1.3 children. In order to perfectly imitate user behaviour, one needs to account for this.

Page views The large proportion of users coming from search engines—and leaving immediately without making a click—was already pointed out (table 4.1). This is oftentimes named as the cause for a lateral distribution: a distribution with pages which are far away from hub nodes which are being viewed without ever being navigated to through internal links. Therefore, figure 4.8 shows two distributions of page views: in *blue* is the actual click data containing all page views. The data set consists of $\approx$ 1 568 000 page views on $\approx$ 124 000 pages. Red is the click data with all external sessions of length one (zero clicks) removed. After removal, $\approx$ 1 020 000 page views on $\approx$ 112 000 pages are still left. This means that only a small portion ($\approx$ 12 000) of pages is only accessed without follow-up or previous clicks. Of the remaining nodes, the vast majority is influenced by removing the external sessions of length one (zero) clicks: the entire red distribution shifts towards the origin. This suggests that users are lead to navigational pages as well as pages deeper in the hierarchy. Longer sessions which might

have started out from search engines are still left in the red distribution, since they include link-following and thus navigational behaviour. For both distributions, there are many pages with few clicks and a few pages with many clicks. This is common in several systems, for example: on twitter there are few users with many followers compared to the many users with few followers.

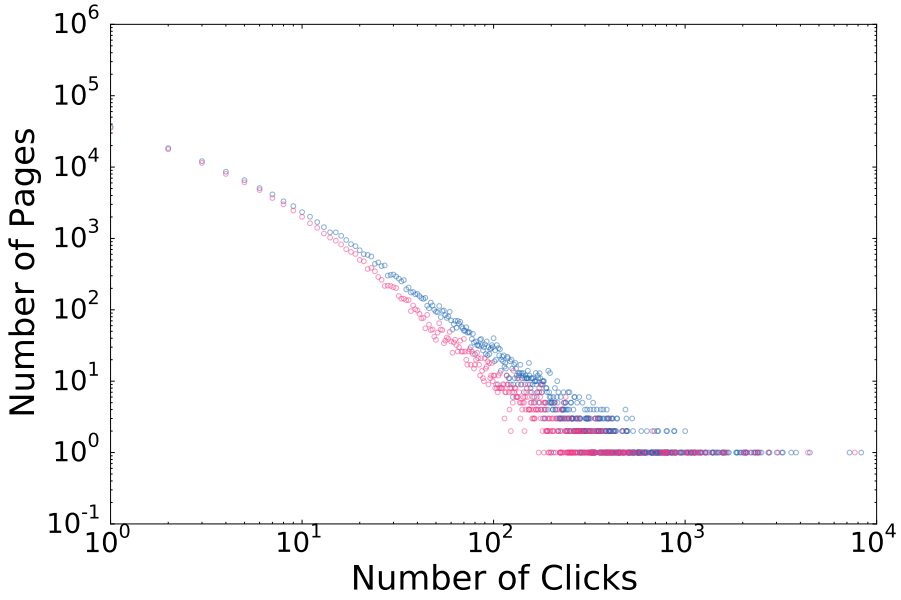


Figure 4.8: Page view distributions (log-log): the click data distribution (in blue) is said to be highly influenced by search engines. It consists of approximately 1 570 000 page views on 125 000 different pages. When removing all sessions of length one (zero clicks) which originated at an external domain, the page views drop to approximately 1 000 000 visits on 112 000 pages. This means that the $\approx$ 500 000 sessions from remote pages of length one (zero clicks) are only visiting $\approx$ 13 000 pages which are not visited otherwise. These sessions do have an influence on the entire distribution: it shifts towards the origin, suggesting that pages with more views are influences as well as pages with fewer views. For both distributions there are many pages with few views and few pages with many views.

Click data analysis - conclusion This section analysed several aspects of user navigation in an encyclopedia. It showed that timeout values should be

chosen carefully, in order to not disrupt sessions of users who take their time between clicks. It then explained the impact of search engines on session lengths and consecutively on jump probabilities. It showed that both differ substantially from those created by random surfers which asks for further investigations on the contributions of dynamic jump probabilities.

The section further showed that most pages were visited through search engines. Approximately 10% of the visited pages were only reached from external domains and the user left immediately afterwards. Qiu et al. found that $\approx$ 14% of web traffic was influenced by search engines. They further stated: "[...] search engines help users reach 20% more sites by presenting them in search results, that may be otherwise unreachable by the users". The research was conducted in 2005 [QLC05].

4.2 Random surfer

Random surfer models are based on a random walk and used in several fields: math, physics, computer science, and more (sec. 2.2.1). In computer science, the most well-known random surfer model is PageRank. It ranks pages by importance based on the proportion of visits the random surfer pays to them (sec. 2.2.2). Following PageRank, several researchers worked on improving simple random surfers in order to model user behaviour on the entire Web or specific domains more accurately. Therefore, these surfers were often enhanced by biasing the random aspects of the models to better match the decisions of users (sec. 2.2.3 and 2.2.5). In order to best model user behaviour on a specific domain (or the Web in general), an analysis needs to be performed to extract the properties of the navigation behaviour on that domain (sec. 2.1.2 and 4.1). A random surfer model can then be extended with one of these properties, creating a biased random surfer.

In this thesis two biased random surfers were introduced (summarisation in table 3.5). The goal was to create an agent which accurately models user navigation on an information platform, namely Austria-Forum. Such a model might then be used to simulate the effects new designs or features have on user behaviour in Austria-Forum. It might be used to compare user navigation on an information platform with user navigation on a platform where users are more likely to click on links. Currently, users tend to make fewer clicks (on average) in encyclopedias than they do in other domains [Gle+10]. Comparing models of these domains might lead to knowledge on how to keep users engaged on information platformss.

4.2.1 Click data distributions

In total, the click data consists of 1 568 000 page views on 124 000 different pages. The sessions extracted from log files contain several sessions starting from external domains (e.g. Google) of length one. This means the user did not click on any links in Austria-Forum. There are 549 000 of these *external zero click (EZC)* sessions. Since they consist of length one, this results in an equal amount of page views. Without these sessions, the data set thus consists of approximately one million page views on 112 000 different pages.

This means that approximately 10% (12 000) of the visited pages are viewed within Austria-Forum as a result of the EZC sessions, which would not be viewed otherwise. Thus, these external domains are the source for a more *lateral* distribution. The first research question (sec. 1.2) focuses on simulating this lateral behaviour.

On the other hand, the EZC page views do not express user navigation behaviour within Austria-Forum. They are only influenced by outside sources. This thesis also compares the results of the random surfers to the distribution extracted from click data without the EZC sessions. This measures how well the random surfers apply to page views which were not only created through links from external sources. Sessions where the user came from an external domain and stayed in Austria-Forum for at least one click are still included.

Furthermore, the data set contains 1 090 000 logged landing pages which equals the amount of sessions since every session has a landing page. These sessions started on 97 000 different pages. Without the external zero click (EZC) sessions, the landing page count is reduced to 79 000 distinct pages. This means that 18 000 landing pages were the starting point solely due to external sites. Note that some of these 18 000 sites were visited (through internal clicks) by other session, since only 12 000 pages were only visited by EZC sessions (as explained above). Table 4.2 gives a summarisation of the segmentation.

Table 4.3 shows the five pages with the most visits for the:

- reference distribution including all clicks;
- reference distribution without external sessions of length one (zero click);
- landing pages;
- and landing pages without external sessions of length one (zero clicks), respectively.

The panels in the table show the following:

- Landing pages strongly influence the view count of pages. The home page was the starting page of 58 655 sessions and was viewed 64 165

Name	Page Views	Pages
external zero click (EZC)	549 000	58 000
reference	1 568 000	124 000
reference (-EZC)	1 020 000	112 000
landing pages	1 090 000	97 000
landing pages (-EZC)	541 000	79 000

Table 4.2: This table shows the amount of sessions with length one (zero clicks) originating at an external site (EZC). It further shows the page view count and the amount of distinct pages of the user click data (reference) and the influence EZC sessions have on this data set. 12 000 pages are only reached from EZC sessions. The landing pages equal the number of sessions. The EZC sessions include 18 000 landing pages which are not landing pages in the rest of the sessions. Out of these 18 000 pages, 6 000 are reached through other clicks (not landing page clicks).

times in total. Thus, more than 90% of these page views came from session starts.

- The home page of Austria-Forum is likely bookmarked by several people. Even without the EZC sessions, the home page is still dominating the general view count, as well as the landing page count.
- The interests of users vary: the majority of users start at the home page but there is no dominant category the majority of users is interested in. The *AEIOU* lexicon leads the race, but is again a collection of many categories. The old AEIOU domain links directly to the */af/AEIOU* page (sec. 4.1). This explains the higher view count. Biographies and Symbols seem to be well received, but the many clicks of the home page are fairly well distributed between the categories.
- The "U-Bahn in Wien" page seems to be a popular target of users from external sites, many leaving without clicking on a link on that page. The page which occurred most often in the referrer of the "U-Bahn in Wien" page was a Google image search which produced the image of the vienna underground map from the domain of Austria-Forum.

Page	Count
Home Page	64 165
/af/AEIOU	7 222
/af/Wissenssammlungen/Symbole	4 483
/af/Heimatlexikon/U-Bahn_in_Wien	3 796
/af/Wissenssammlungen/Briefmarken	3 581

Panel A: top five page views

Page	Count
Home Page	51 152
/af/AEIOU	6 067
/af/Wissenssammlungen/Briefmarken	3 350
/af/AustriaWiki	2 726
/af/Wissenssammlungen/Symbole	2 480

Panel B: top five page views without EZC sessions

Page	Count
Home Page	58 655
/af/AEIOU	6 729
/af/Wissenssammlungen/Symbole	3 893
/af/Heimatlexikon/U-Bahn_in_Wien	3 795
/af/Wissenssammlungen/Symbole/Runen	3 226

Panel C: top five landing pages

Page	Count
Home Page	45 642
/af/AEIOU	5 574
/af/Heimatlexikon/Birnensorten_im_Überblick	2 201
/af/Wissenssammlungen/Biographien	1 919
/af/Wissenssammlungen/Symbole	1 890

Panel D: top five landing pages without EZC sessions.

Table 4.3: This table shows the top five pages (panel A). The home page has by far the most page views followed primarily by category pages and some outlier pages like the "U-Bahn in Wien". By removing the EZC sessions, the "U-Bahn in Wien" page is not in the top five anymore. Several users reach it from an external site and leave without a click. The difference between panels A and B further show that 13 000 user were directed to the home page and did not make a single click. The home page is dominant in all panels. Panels C and D suggest that several users have it bookmarked and follow links on it.

4.2.2 Random surfer models

The previous section explained the two reference distributions as well as the distribution of landing pages (table 4.2). In order to compare two distributions, the Pearson correlation coefficient (p) was calculated. Values in the range $-1 <= p < 0$ represent a negative correlation, and $0 < p <= 1$ a positive correlation. The correlation between the two reference distributions is 0.99. The high correlation suggests that page views of all pages are influenced in a proportionally similar way, confirming the interpretation of page view distributions in figure 4.8.

These reference distributions were then compared to the distributions created by the random surfer models. The random surfer models were described in sec. 3.3 (which contains a summarisation in table 3.5). The abbreviations used represent:

- uniform *(U)*;
- landing page *(LP)*;
- preferential page *(PP)*.

These define the distribution used as the starting nodes for new sessions. Thus, random surfers using the uniform distribution are *real random* surfers, whereas random surfers using LP or PP are *biased* random surfers. The landing pages were used as starting points to imitate the fact that some pages are more likely to be the entry point than others (see table 4.3 panel C for top five landing pages). The preferential pages are the normalised vector of page views (see table 4.3 panel A for the top five pages). Thus, pages with more views have higher chances of being the starting point of sessions.

Furthermore, the values *15* and *69* contained in the names of random surfers represent the commonly chosen jump probability and the average jump probability of Austria-Forum, respectively. *Dyn* means the random surfer used the dynamic jump probabilities.

Reference Distribution including EZC sessions Table 4.4 displays the correlations between the reference distribution (including all real user click data) and the random surfers. The correlation of $p = 0.51$ between the

reference distribution and the U15 random surfer confirms that the common uniform random surfer can model user navigation in Austria-Forum only to some extent (sec. 2.2.5).

The introduction of simple biases (which can be easily measured by system administrators) increased the result of the U15 random surfer. The random surfers using landing pages and page views as starting pages reached $p = 0.57$ and $p = 0.56$, both with constant jump probabilities of 15%. Due to the *low* jump probability for all random surfers in every step, the resulting sessions are of considerable length. Therefore, the entry pages of sessions constitute only a small amount of the total page views. Thus, the biased random surfers perform only slightly better than the uniform random surfer; the influence if the bias is limited.

By raising these jump probabilities to 69% (the average jump probability in Austria-Forum) for all three random surfer types, the correlation between the uniform random surfer and the reference distribution decreased ($p = 0.48$ compared to $p = 0.51$ before). The correlations between the reference distribution and the other two random surfers increased substantially. The LP69 reached $p = 0.87$ and the PP69 $p = 0.81$ (compared to $p = 0.57$ and $p = 0.56$ before). These random surfers follow fewer links and jump more compared to the random surfers with 15% jump probability. This means that the average session length drops (to 1.44 clicks). The uniform random surfer jumps to all pages with equal probability. Thus, it frequently jumps to those pages, which are further away from the main page and other hubs (e.g. category pages). Due to the shorter average session length, category pages which are reachable with one or two clicks from the main page (and have many visits in the reference distribution) cannot be reached regularly by the uniform random surfer. This explains the drop in correlation compared to the uniform random surfer with longer sessions (15% jump probability) where the chances of sessions reaching hubs are substantially higher. On the other hand, the LP and PP random surfers are positively affected by the increase in jump probability. These random surfers jump to pages which have many visits in the reference distribution. Thus, shorter sessions (higher jump probabilities) have a positive impact.

The differences between random surfers using a constant jump probability of 69% and those using a dynamic jump probability are of particular interest

to this thesis. The simulation used the same number of sessions for all calculations. Sessions with jump probability of 69% as well as sessions created with dynamic jump probabilities result in the *same average session length*. Thus, these simulations had the same session count and click count, meaning that the bias influenced both random surfers to the same extent. The results were still different. Random surfers with dynamic jump probabilities performed better than random surfers with a constant jump probability of 69% in every instance. The dynamic jump probabilities produce more sessions of length 1 (zero click), due to the jump probability being $\approx$ 90% in the first step. This means that $\approx$ 10% of sessions make at least one click, compared to the 31% of sessions which make at least one click when the random surfer jumps with constant probability of 69%. Both of these subsets (10% and 31%) include an equal amount of clicks, since *dynamic* and static 69% jump probabilities both create the same average session length. Thus, these remaining sessions created with dynamic jump probabilities are on average longer due to the many zero-click sessions. These longer sessions have higher chances of reaching hubs and other pages with many page views multiple times. Random surfers with jump probability 69% produce more sessions of length two (one click); if the starting page of a session is several clicks away from hubs and other pages with many page views, these are likely not reached. Thus, the random surfer with dynamic jump probabilities performs better.

Table 4.4 shows the improvements in detail.

Reference Distribution without EZC sessions Research showed that search engines lead to a more lateral distribution of the user click data [Gei+15]. Pages are accessed which otherwise would not have been accessed through click trails not originating from search engines. In more than 50% of sessions users came from an external site, visited a page in Austria-Forum and left immediately without making a single click (called EZC sessions, table 4.2). In order to measure how well the presented random surfers model user navigation, these sessions were removed from the reference distribution since they do not include navigation behaviour in Austria-Forum.

The biased random surfers (using landing pages and preferential pages as entry points of sessions) were affected by removing these page views as

well. The distributions of landing pages and preferential pages changed. Every random surfer improved its correlation to the adapted reference distribution compared to the original correlations on the real click data (table 4.4). This shows that random surfers are having trouble modelling pages where users do not make any clicks. Thus, by removing these pages (clicks) all correlations increased.

The Pearson correlations between the uniform random surfers and the reference distribution without EZC sessions had the smallest improvements (0.02) compared to their correlations with the reference distribution which includes all click data (table 4.4). This shows that the results of the uniform random surfer are hardly influenced by a more lateral distribution. Due to the uniform selection of starting pages, even the most remote pages are reached. It will be interesting to see whether this holds true once the user data includes more visited pages. The removal of the sessions with length one (zero clicks) which were only accessed from external sources has a positive affect on the LP and PP random surfers. The landing page random surfers improved between 0.04 and 0.05 Pearson correlation units; the preferential page between 0.03 and 0.05. The remaining pages in the reference distribution are all reached through navigation (those sessions without without navigational elements were removed). Due to the bias of these random surfers, these pages are more likely to be reached since the starting pages are very likely connected to them.

Equivalent to the comparison of the random surfers to the real user click data, all random surfers achieved higher correlations with dynamic jump probabilities, compared to the average jump probability of Austria-Forum (69%).

Thus, the peak values of all random surfers improved as well compared to the peak levels of the distribution containing all clicks (table 4.4). For the uniform from 0.51 (U15, UDyn) to 0.53 (U15, UDyn), the landing page random surfers from 0.91 (LPDyn) to 0.95 (LPDyn), and the preferential page random surfers from 0.87 (PPDyn) to 0.90 (PPDyn).

The correlations of the landing page random surfer with dynamic jump probabilities and the two reference distributions are substantially higher than the correlations of previous research on random surfers in Austria-Forum [Gei+15]. The LP random surfer can be considered a simple random

surfer. It is a random surfer with the bias of a personalised vector (sec. 2.2.2) for the starting pages of sessions. This bias mitigates the—possibly stochastic (sec. 2.2.3)—component, that a subset of pages are more likely to be the starting pages of sessions. This elevates the accuracy of the random surfer.

Nonetheless, all random surfers have problems modelling user navigation behaviour on pages where users do not follow any links. In order to incorporate this behaviour, a random surfer needs to be enhanced by providing information about the percentage of users that leave a page instead of making a click.

Table 4.5 shows the correlations between the random surfers and the reference distribution without the EZC sessions.

Conclusion This section presented the results of the random surfers. It showed that selecting a subset of pages as the starting pages for sessions enhanced the correlation to the real user data. With a higher jump probability (69% compared to the common 15%) the uniform random surfer produced worse results when modelling user navigation due to the shorter average session length. The other random surfers both substantially improved the modelling of user behaviour when adjusting the constant probability to the average jump probability of Austria-Forum (69%) due to the introduced bias. The introduction of dynamic jump probabilities yielded better results for all random surfers compared to the corresponding random surfers with a static jump probability of 69%. The random surfer with dynamic jump probabilities which starts at the same pages as users do reached a Pearson correlation of 0.91 (table 4.4). By removing the sessions of length one (zero clicks) which came from an external source (e.g. search engines), this correlation improved to 0.95 (table 4.5).

Name	p
U15	0.51
U69	0.48
UDyn	0.51
LP15	0.57
LP69	0.87
LPDyn	0.91
PP15	0.56
PP69	0.81
PPDyn	0.87

Table 4.4: Reference Distribution

Name	p
U15	0.53
U69	0.50
UDyn	0.53
LP15	0.62
LP69	0.93
LPDyn	0.95
PP15	0.59
PP69	0.86
PPDyn	0.90

Table 4.5: Reference Distribution -EZC

Correlations between click data (see table 4.2) and random surfers: these tables show the random surfers with uniform (U), landing page (LP), and preferential page (PP) session starting nodes; with the static jump probabilities of 15% and 69% as well as dynamic jump probabilities. The biased random surfers (LP and PP) perform better (higher correlation) than the uniform random surfers with corresponding jump probabilities. Random surfers with the average jump probability of Austria-Forum (69%) perform better than with 15% jump probability; except for the uniform random surfer, which performs worse (details in text). The dynamic jump probabilities always increase the correlation to the real user click data, compared to the same random surfer with the static probability of Austria-Forum (69%). Using the reference distribution without ECZ sessions increases all correlations of the random surfers. This shows that pages which were only reached from external domains are difficult to simulate with these simple random surfers and decrease their correlations.

5 Discussion

This section is split into three parts. First, it analyses the results of sec. 4.1 and puts them into perspective with the results of previous research. Following up, the results of the random surfers, presented in sec. 4.2, are interpreted. Lastly, this section explains the limitations.

5.1 Click Data Analysis

This thesis had the privilege of having access to real user click data of an online encyclopedia. Previous studies on user behaviour in online encyclopedias were primarily conducted on data sets restricted to Wiki-Games (sec. 2.1.2), or data sets of the entire web with only a small portion of clicks in encyclopedias [Mei+09; Gle+10]. Gleich et al. estimated the average jump probability on Wikipedia between $57.5\% - 67.5\%$ [Gle+10]. For Austria-Forum, this probability is even higher, at 69%. The reason for this might be Austria-Forum being lesser known, with a smaller active user base, and containing less information than Wikipedia. The majority of visitors are stumbling onto Austria-Forum via search engines. Due to the smaller scope of Austria-Forum, it contains fewer links which would keep users engaged in the site. Thus, users leave quickly and the jump probability rises.

Meiss et al. further found the value of 15 minutes the best fitting to split sessions via timeout [Mei+09]. The findings in this thesis support the fact, that most sessions with large gaps between two clicks are within the timeout of 15 minutes. Half of them are even within 10 minutes. Since users are possibly reading long articles, summarising them for research, or keeping them displayed as a reference while researching, it can be argued that a lager timeout value fits for encyclopedias. A close approximation of continuous

navigation on the underlying graph can be achieved with a larger timeout in combination with referrer trees and splitting sessions at search requests.

Web traffic is often associated with skewed distributions which follow the power-law. Research showed that user navigation is log-normally distributed at user-level. The aggregate of these log-normal distributions results in power-law distributions at Web- or website-level. Meiss et al. attributed this to many users with different behaviour and interests [Mei+09]. Several findings of this thesis on click data are on the verge between power-law and log-normal distributions (figures 4.3, 4.6, 4.7a, 4.7b, 4.8). The majority of these figures tend towards log-normal.

The extracted referrer trees show less depth (average 0.31) compared to previous research (average 2.52 [QLC05]). This value is highly influenced by sessions without clicks (depth = 0). Without these sessions, the average depth increases to 3.46. Qiu et al. further calculated a node-based branching factor of 2.95 (sessions were on the entire web) [QLC05]. The branching factor of Austria-Forum is merely 1.3 (sessions of length 1 do not have any influence here). This means users are not backtracking or multi-tab browsing as much as expected (but still more than the common random surfer accounts for). The cause for that is not straightforward to interpret. On the one hand, users might not see the need to go back since they found the information they were looking for. They might not need to open multiple tabs because it is clear to them what lies behind links, or where the information they are looking for is located. On the other hand, the low branching factor might arise from missing links and users not having the opportunity to explore related articles.

In general, the attributes of Austria-Forum match the expectations for an online encyclopedia. The large impact of search engines suppresses the real user navigation consisting of several clicks. Investigating these longer sessions is left to future research, when the data set will be substantially larger.

5.2 Random Surfer Results

The random surfer models in this thesis were created in order to answer the research questions presented in sec. 1.2.

Research Question 1: How well can a simple biased random surfer simulate the lateral user navigation behaviour in Austria-Forum?

Previous research showed that neither the common *uniform random surfer (URS)* nor an enhanced model—called *pragmatic random surfer (PRS)*—with the capability of biased edge selection were accurate when modelling user navigation on Austria-Forum [Gei+15] (sec. 2.2.5). The URS and the PRS reached Pearson correlations of $p = 0.38$ and $p = 0.47$ compared to the distribution created by real user click data [Gei+15]. Therefore the first research question was created. The goal was to create a simple model relying on minimal heuristics.

The results of the uniform random surfer ($p = 0.38$) were improved with the pragmatic random surfer ($p = 0.47$) [Gei+15]. The expense of this improvement is the calculation of link transitions. Thus, every click needs to be logged and mapped to the graph (distinctions between two links with the same source and target cannot easily be made). These calculations were performed on a subset of the current data set. On the current data set, the uniform random surfer reached a correlation of $p = 0.51$.

These results were improved with two different methods which need less data compared to the PRS where every transition (click on a link) needs to be logged. First with the introduction of landing pages (LP), then with preferential pages (PP). The concept of a random surfer with preferential pages is easily reproducible, the page view count of the investigated domain should suffice. At the commonly chosen jump probability of 15%, the results improved to $p = 0.57$ and $p = 0.56$ of the LP- and PP random surfer respectively. By adjusting the jump probability for all three types of random surfers, the uniform random surfer modelled user navigation even worse ($p = 0.48$) than before. LP random surfer ($p = 0.87$) and PP random surfer ($p = 0.81$) both substantially increased the correlation to the real user click data.

This shows that the lateral access of search engines can be compensated by steering a random surfer onto pages which are either more likely to be clicked, or more likely to be the starting pages of a new session. Either way, preferring these pages as the starting pages for sessions produces more accurate results than those created by other random surfers where one of them included a more sophisticated bias [Gei+15]. The simplicity of such a random surfer is given since these preferred pages can be easily identified by a site administrator. Due to many of the top pages (sec. 4.2, table 4.3) including category pages of Austria-Forum, a random surfer which prefers these pages over others might also model user behaviour to a sufficient extent. Investigating this claim is left to future research.

Research Question 2: What impact does a dynamic jump probability have on the results of random surfers?

The second research question is based on the finding of a decaying factor when users navigated towards a certain article on Wikipedia. The results of the $\epsilon - greedy$ algorithm, using the advanced heuristic of decentralised search, inspired the objective of this thesis. In the beginning, users explored by making several random clicks. The closer they got towards the article, the lower was the probability of a random click [Hel+13] (sec. 2.2.4). The analysis on jump probabilities showed a similar pattern. The probability of a jump is high in the beginning, but decays with every step (at least for the first few steps; see sec. 4.1).

The results showed that random surfers with dynamic jump probabilities *always* performed better than the corresponding random surfers with the jump probability of 69% (the average jump probability in Austria-Forum). The improvements were between 0.03 and 0.06 points on the Pearson correlation scale. This is interesting due to the following facts:

- All simulations were conducted with 4 million sessions, therefore including 4 million jumps. The distribution created due to jumps solely can thus be assumed equal for all random surfers using the same subset of starting pages.
- The jump probability of 69% is the site jump probability in Austria-Forum. The average session length (1.44 clicks) of random surfers

using this probability approximately equals the average session length of random surfers using dynamic jump probabilities.

Thus, these two groups of random surfers performed *the same number of jumps* and *the same number of clicks*. Still, random surfers with dynamic jump probabilities *always* yielded better results than those with the static jump probability (of 69%). The only difference, between these two groups of random surfers, is the distribution of session lengths. Dynamic jump probabilities result in more sessions of length 1, and more longer sessions. Random surfers using the jump probability of 69% result in more sessions of length 2, and 3. In other words, the average session length is the same, but the variance is higher with dynamic jump probabilities.

When modelling user navigation behaviour, agent-based models commonly use static jump probabilities. This thesis showed that the inclusion of dynamic jump probabilities in these models improves the accuracy of the modelling. Dynamic jump probabilities depend on the domain; this thesis provided the first empirically measured jump probabilities of Austria-Forum which should be applicable on other encyclopedias as well.

5.3 Limitations

Several limitations were presupposed or had to be made in order to make this work feasible.

5.3.1 The Graph

The crawled graph is a snapshot. It only represents a temporal limited state of Austria-Forum. Content in encyclopedias is added continuously: new pages are generated, other pages updated (deleted), links are inserted (deleted), and parts of the graph structure change. This results in the following limitations:

- *Unmapped edges:* Austria-Forum provides dynamic content in its banners. These banners are present on several pages and display a set

of pages. Since the crawler crawled every page once, only the links which were in the banner at the time of the crawl are present in the graph. The majority of user clicks in banners will thus not be present as link transitions (but they are as page views) and random surfers are only able to use the links which were crawled. Due to encoding standards, several links could not be mapped to the crawled links. In total, approximately 4% of links could not be mapped.

- *Unmapped nodes:* the graph as well as the log files only contain HTML pages. Austria-Forum has several features written in JavaScript, unreached by the crawler, not visible to the logger. Several key pages trigger JavaScript requests: *search*, *login*, *upload*, *comment*, *user preferences*, and more. These are also not present as transitions. They were adapted in order to enable the referrer tree construction without discontinuities. They are thus considered in session length and depth, as well as node branching factor. In total less than 1% of pages were not mapped. This results in approximately 7% of unmapped clicks, due to many of these being key pages.
- The graph was limited to the strongly connected component. This had only minor impacts since less than 0.2% of nodes were removed.

5.3.2 Log Data

The log data was limited in the following ways:

- It does not include the E-Books section.
- When clicking through specific image galleries, the referrer field does not change. This results in some nodes having an abnormally large branching factor.

5.3.3 Random Surfer

The click data (reference distribution) only visited approximately 30% (125 000) of the available $\approx$ 420 000 pages (nodes).

The result of the mathematical calculation of the uniform random surfer with 15% jump probability (provided by the fellow researchers from [Gei+15]) had a variation of approximately 4% on the Pearson correlation, compared to the simulation. For consistency, the values created by the simulation were used.

6 Conclusion

This thesis explored several aspects of Austria-Forum, an online encyclopedia limited to Austrian context. It had firsthand access to empirical, unbiased log data. In related research, several calculations of user behaviour in online information platforms (primarily Wikipedia) have been conducted. The data of these researches was often limited or biased by target groups. This thesis thus reevaluated several preexisting findings: the *influence of search engines* [QLC05; HMW08], *distributions in web traffic* [Mei+09], and *user navigation behaviour* (session- length, -depth, -time, and branching factor) [QLC05; Mei+09; MHM06; SK09]. These calculations were extended through empirically measured jump probabilities on the entire encyclopedia, as well as at each click (of a session).

Thereafter, modelling of user behaviour in Austria-Forum was explored. The results of a preexisting biased random surfer [Gei+15] were used as a reference. These results modelled user behaviour with a Pearson correlation of $p = 0.47$. New random surfers with simple enhancements were introduced yielding results up to $p = 0.87$. These random surfers still comply to the Markov property of memorylessness. By using the novel approach of dynamic jump probabilities, the results were further enhanced with peak correlations to real user click data of $p = 0.91$.

This shows that—at least on the encyclopedia Austria-Forum—the common assumption of a uniform jump probability is inaccurate. It will be interesting to see dynamic jump probabilities on other domains, especially Wikipedia. The incorporation of these dynamic probabilities in stochastic processes and their mathematical calculations would be an extension of this work and improvement of existing methods. These dynamic jump probabilities clearly enhanced the capabilities of random surfers to model user navigation in information networks and should be considered in future research.

6.1 Future Work

The click analysis investigated several aspects of user behaviour and compared them to previous results. There were some slight deviations, possibly due to changing user behaviour (comparisons are made with research being several years old), related research based on games instead of unbiased data, and the size of the data set used in this thesis (see sec. 5.3). In future research, it will be interesting to see how statistics on Austria-Forum will evolve, with the data set containing more months or even years.

Another open issue is whether the jump probability curve (figure 4.5) can be mapped to a function. For now, only speculations are possible due to the sparse values. With more data available, one could try to map different functions to the curve. The best fit might then be used for random walk simulations on encyclopedias. Mathematically, the PageRank model could be extended in order to incorporate this function of dynamic jump probabilities. Gleich and Rossi already included a function for time-dependent teleportation [GR14]. Their research might be used as incentive to extend PageRank mathematically with dynamic jump probabilities.

In the future, it will be interesting to see empirically measured jump probabilities (for each click) on *Wikipedia*, and how they compare to those of Austria-Forum or domains with other purposes and link structures.

Concerning further development of random surfers: it will be interesting to measure the performance of a random surfer which takes *exit probabilities* of pages into account. Currently, the data set does not include enough data to adequately portray these probabilities. A model using *landing pages*, *biased link selection* (PRS), and *page dependent exit probabilities* in combination with *dynamic jump probabilities* can further improve results.

Bibliography

[AHS10] Shawn Andrews, Ghassan Hamarneh, and Ahmed Saad. "Fast random walker with priors using precomputation for interactive medical image segmentation." In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2010*. Springer, 2010, pp. 9–16 (cit. on p. 12).

[And07] Per Andersen. *What is Web 2.0?: ideas, technologies and implications for education*. Vol. 1. 1. JISC Bristol, UK, 2007 (cit. on p. 22).

[ASW07] Denise Anthony, Sean W. Smith, and Tim Williamson. *The Quality of Open Source Production: Zealots and Good Samaritans in the Case of Wikipedia*. Tech. rep. TR2007-606. Hanover, NH: Dartmouth College, Computer Science, Sept. 2007. URL: http://www.cs.dartmouth.edu/reports/TR2007-606.pdf (cit. on p. 22).

[Bay+12] Murat Ali Bayir et al. "Discovering better navigation sequences for the session construction problem." In: *Data & Knowledge Engineering* 73 (2012), pp. 58–72 (cit. on pp. 6, 7).

[BP10] Marco Bressan and Enoch Peserico. "Choose the damping, choose the ranking?" In: *Journal of Discrete Algorithms* 8.2 (2010), pp. 199–213 (cit. on p. 14).

[DM12] Martin Drlik and Michal Munk. "Session Timeout Thresholds Impact on Quality and Quantity of Extracted Sequence Rules." In: *International Journal of New Computer Architectures and their Applications (IJNCAA)* 2.1 (2012), pp. 34–51 (cit. on p. 7).

[DS00] Peter G. Doyle and J. Laurie Snell. "Random walks and electric networks." In: *Carus mathematical monographs* 22 (2000) (cit. on p. 11).

[ES02] Brian S. Everitt and Anders Skrondal. "The Cambridge dictionary of statistics." In: *Cambridge: Cambridge* (2002) (cit. on p. 11).

[FE07] Steven Furnell and Michael P. Evans. "Analysing Google rankings through search engine optimization data." In: *Internet research* 17.1 (2007), pp. 21–37 (cit. on p. 24).

[Flo12] Patrick J. Floryance. "Random Walks, Electrical Networks, and Perfect Squares." PhD thesis. University of Minnesota-Twin Cities, 2012 (cit. on p. 11).

[Gei+15] Florian Geigl et al. "Random Surfers on a Web Encyclopedia." In: *arXiv preprint arXiv:1507.04489* (2015) (cit. on pp. 2, 3, 17, 19, 35, 36, 38, 60, 61, 67, 68, 71, 73).

[GF04] Leo Grady and Gareth Funka-Lea. "Multi-label image segmentation for medical applications based on graph-theoretic electrical potentials." In: *Computer Vision and Mathematical Methods in Medical and Biomedical Image Analysis*. Springer, 2004, pp. 230–245 (cit. on p. 12).

[GGP04] Zoltán Gyöngyi, Hector Garcia-Molina, and Jan Pedersen. "Combating web spam with trustrank." In: *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*. VLDB Endowment. 2004, pp. 576–587 (cit. on p. 14).

[Gle+10] David F. Gleich et al. "Tracking the random surfer: empirically measured teleportation parameters in PageRank." In: *Proceedings of the 19th international conference on World wide web*. ACM. 2010, pp. 381–390 (cit. on pp. 2, 14, 54, 65).

[Gon+09] Bruno Gonçalves et al. "Remembering what we like: Toward an agent-based model of Web traffic." In: *arXiv preprint arXiv:0901.3839* (2009) (cit. on pp. 2, 7, 13, 16).

[GR14] David F. Gleich and Ryan A. Rossi. "A Dynamical System for PageRank with Time-Dependent Teleportation." In: *Internet Mathematics* 10.1-2 (2014), pp. 188–217 (cit. on pp. 14, 38, 74).

[Gra06] Leo Grady. "Random walks for image segmentation." In: *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 28.11 (2006), pp. 1768–1783 (cit. on pp. 11, 12).

[Hel+13] Denis Helic et al. "Models of human navigation in information networks based on decentralized search." In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM. 2013, pp. 89–98 (cit. on pp. 2, 10, 15, 16, 68).

[HMW08] Denis Helic, Hermann Maurer, and Bebo White. "AUSTRIA-FORUM: A CITABLE WEB ENCYCLOPEDIA." In: *IADIS International Conference WWW/Internet*. 2008, p. 1 (cit. on pp. 21–24, 73).

[Kap+07] Frank Kappe et al. "Report on dangers and opportunities posed by large search engines, particularly Google." In: (2007) (cit. on p. 24).

[Kap+14] Jozef Kapusta et al. "Determining the Time Window Threshold to Identify User Sessions of Stakeholders of a Commercial Bank Portal." In: *Procedia Computer Science* 29 (2014), pp. 1779–1790 (cit. on pp. 1, 7).

[KK08] Aniket Kittur and Robert E. Kraut. "Harnessing the Wisdom of Crowds in Wikipedia: Quality Through Coordination." In: *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*. CSCW '08. San Diego, CA, USA: ACM, 2008, pp. 37–46. ISBN: 978-1-60558-007-4. DOI: 10.1145/1460563.1460572. URL: http://doi.acm.org/10.1145/1460563.1460572 (cit. on p. 22).

[Kle00] Jon Kleinberg. "The small-world phenomenon: An algorithmic perspective." In: *Proceedings of the thirty-second annual ACM symposium on Theory of computing*. ACM. 2000, pp. 163–170 (cit. on p. 15).

[Lau14] Paul Laufer. "Cultural Similarity, Understanding and Affinity on Wikipedia Cuisine Pages." MA thesis. University of Technology Graz, 2014 (cit. on p. 23).

[Lov93] László Lovász. "Random walks on graphs: A survey." In: *Combinatorics, Paul erdos is eighty* 2.1 (1993), pp. 1–46 (cit. on p. 11).

[Mai+08] Florian Maier et al. "Automatic liver segmentation using the random walker algorithm." In: *Bildverarbeitung für die Medizin 2008*. Springer, 2008, pp. 56–61 (cit. on p. 12).

[Mau] Hermann Maurer. "Austria-Forum and Beyond." In: () (cit. on pp. 21, 22).

[Mei+09] Mark Meiss et al. "What's in a session: tracking individual behavior on the web." In: *Proceedings of the 20th ACM conference on Hypertext and hypermedia*. ACM. 2009, pp. 173–182 (cit. on pp. 5, 7, 8, 13, 16, 31, 65, 66, 73).

[Mei+10] Mark R. Meiss et al. "Agents, bookmarks and clicks: a topical model of web navigation." In: *Proceedings of the 21st ACM conference on Hypertext and hypermedia*. ACM. 2010, pp. 229–234 (cit. on pp. 9, 13, 16, 18).

[MHM06] Jamie Murphy, Charles Hofacker, and Richard Mizerski. "Primacy and recency effects on clicking behavior." In: *Journal of Computer-Mediated Communication* 11.2 (2006), pp. 522–535 (cit. on pp. 5, 24, 73).

[Mil67] Stanley Milgram. "The small world problem." In: *Psychology today* 2.1 (1967), pp. 60–67 (cit. on pp. 14, 15).

[Pag+99] Lawrence Page et al. "The PageRank citation ranking: bringing order to the Web." In: (1999) (cit. on p. 12).

[QLC05] Feng Qiu, Zhenyu Liu, and Junghoo Cho. "Analysis of User Web Traffic with A Focus on Search Activities." In: *WebDB*. Citeseer. 2005, pp. 103–108 (cit. on pp. 5–7, 31, 50, 53, 66, 73).

[Sca+14] Aju Thalappillil Scaria et al. "The last click: Why users give up information network navigation." In: *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM. 2014, pp. 213–222 (cit. on p. 9).

[SK09] K.R. Suneetha and Raghuraman Krishnamoorthi. "Identifying user behavior by analyzing web server access log file." In: *IJCSNS International Journal of Computer Science and Network Security* 9.4 (2009), pp. 327–332 (cit. on pp. 1, 5, 73).

[Stv+05] Besiki Stvilia et al. "Information quality discussions in Wikipedia." In: *Proceedings of the 2005 international conference on knowledge management*. Citeseer. 2005, pp. 101–113 (cit. on p. 22).

[TK12] Frank W. Takes and Walter A. Kosters. "The Difficulty of Path Traversal in Information Networks." In: *KDIR*. 2012, pp. 138–144 (cit. on pp. 9, 10).

[Wei+06] Harald Weinreich et al. "Off the beaten tracks: exploring three aspects of web navigation." In: *Proceedings of the 15th international conference on World Wide Web*. ACM. 2006, pp. 133–142 (cit. on pp. 24, 25).

[WL12a] Robert West and Jure Leskovec. "Automatic Versus Human Navigation in Information Networks." In: 2012 (cit. on pp. 9, 10).

[WL12b] Robert West and Jure Leskovec. "Human wayfinding in information networks." In: *Proceedings of the 21st international conference on World Wide Web*. ACM. 2012, pp. 619–628 (cit. on p. 10).

[WPL15] Robert West, Ashwin Paranjape, and Jure Leskovec. "Mining missing hyperlinks from human navigation traces: A case study of wikipedia." In: *Proceedings of the 24th international conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2015, pp. 1242–1252 (cit. on p. 10).

[WS98] Duncan J. Watts and Steven H. Strogatz. "Collective dynamics of 'small-world' networks." In: *nature* 393.6684 (1998), pp. 440–442 (cit. on p. 15).