Dietmar Paulus

Trends in User Navigation Behavior on Wikipedia

Master's Thesis

Graz University of Technology

Institute for Knowledge Technologies
Head: Univ.Prof. Dipl-Ing. Dr.techn. Stefanie Lindstaedt

Supervisor: Assoc.Prof. Dipl-Ing. Dr.techn. Denis Helic
Advisor: Dipl-Ing. Florian Geigl, BSc

Graz, April 2016

This document is set in Palatino, compiled with pdfLaTeX2e and Biber.

The LaTeX template from Karl Voit is based on KOMA script and can be found online: https://github.com/novoid/LaTeX-KOMA-template

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

Abstract

Wikipedia is one of the most frequently used websites worldwide and is an influential source of knowledge in the World Wide Web. Though Wikipedia aims to offer information in a neutral way, it is known that the community behind Wikipedia is biased and the network structure topologically prefers articles about men to articles about women. Until now it is not known to what extent user navigation behaviour on Wikipedia is biased.

This thesis focuses on analyzing user navigation behaviour and making tendencies in this behaviour visible. By using the random surfer model it is possible to calculate the probability of visiting a Wikipedia article based on provided user click data.

The obtained results show that there is a gender bias in the Wikipedia network structure, which is already known, but additionally, the user navigation behaviour has gender tendencies too. Whereas it is already proven that the Wikipedia network structure prefers articles about men, the user navigation behaviour tends to visiting articles about females. Regardless whether Wikipedia articles about persons are analyzed or whether the articles are grouped by certain criteria, a gender tendency is obvious throughout all results.

Kurzfassung

Die Onlineenzyklopädie Wikipedia ist eine der am häufigsten genutzten Webseiten weltweit und ist weiters eine einflussreiche Quelle von Wissen innerhalb des World Wide Web. Obwohl es das Ziel von Wikipedia ist, neutrale Informationen bereitzustellen, ist es bekannt, dass die Gemeinschaft der Autoren, die Artikel in die Wikipedia stellt, einen Bias aufweist und auch die Struktur der verlinkten Seiten Artikel über Männer gegenüber Artikel über Frauen bevorzugt. Bis jetzt ist aber noch nicht bekannt, in wie weit das Navigationsverhalten von Benutzern der Wikipedia gewissen Navigationstrends aufweist.

Der Fokus dieser Arbeit liegt auf der Analyse des Navigationsverhaltens von Benutzern der Wikipedia und auf dem Erkennen und Aufzeigen solcher Tendenzen. Basierend auf Klickdaten von Wikipediaartikeln ist es mit Hilfe des *Random Surfer* Modelles möglich, die Besuchswahrscheinlichkeit von Wikipediaartikeln zu berechnen.

Die gefundenen Resultate zeigen, dass ein Geschlechterbias sowohl in der Struktur der verlinkten Seiten der Wikipedia als auch geschlechtsspezifische Tendenzen im Navigationsverhalten von Benutzern existiert. Der bereits bekannte Geschlechterbias in der Linkstruktur der Wikipedia bevorzugt Wikipediaartikel über Männer, wohingegen das Navigationsverhalten der Benutzer zu Artikeln über Frauen tendiert. Diese Ungleichheit gilt sowohl im Generellen für Wikipediaartikel über Personen als auch in einzelnen Gruppierungen von Artikeln, wie zum Beispiel einer Gruppierung der Artikel nach dem Geburtsland der Person.

Acknowledgements

First of all I want to thank my supervisor Prof. Denis Helic for his great support throughout my master's thesis. Whenever it was necessary I was able to talk to him and always got valuable feedback.

Special thanks to my advisor Dipl.-Ing. Florian Geigl for his effort during my master's thesis. He always responded very fast to my questions and gave me constructive comments and feedback.

Thanks for all the discussion and inspiring moments to the staff of the Knowledge Technologies Institute of the Technical University Graz.

I also want to thank Dimitar Dimitrov from GESIS Leibniz - Institute for the Social Sciences in Cologne for providing the Wikipedia network dataset, which was a fundamental part of this thesis.

Finally I want to thank my parents Margaretha and Wilhelm Paulus and especially my beloved girlfriend Mag. Bettina Korb for always supporting me during my studies.

Contents

List of Figures

List of Tables

1. Introduction

This chapter provides an introduction to this master's thesis. Section 1.1 illustrates the motivation behind this work. In Section 1.2 the main contributions are listed and Section 1.3 describes the outline of this thesis.

1.1. Motivation

Gender inequalities and gender differences are a common phenomenon of mankind. Whether there are findings that show that men and women should be treated differently within organizations (Acker, 2006) or there are differences in healthcare for women and men (Annandale and Hunt, 2000), gender inequalities occur in different spheres of life. These gender gaps are not restricted to the *off-line* world itself, they are also transferred into the World Wide Web. A study in 1991 (Selfe and Meyer, 1991) showed that even in anonymous discussions women are treated differently. Recent studies (Lam et al., 2011), (Collier and Bear, 2012), (Wagner et al., 2015) also illustrate that the online encyclopedia Wikipedia is not immune to gender inequalities.

The Wikipedia project claims to be a massive collection of mankind's knowledge, that the content is free for everybody and that every person can contribute to the online encyclopedia. Regarding the guidelines[1], Wikipedia should not have biased sources in anyway. As described above this is not the case. To begin with the whole network structure favours articles about men (Wagner et al., 2015) and furthermore the community that is responsible for creating and editing these articles, are mainly male and white (Lam et al., 2011). Despite the fact that articles about males are preferred within the

[1] https://en.wikipedia.org/wiki/Wikipedia:Policies_and_guidelines

network structure of Wikipedia, it is not known whether the navigation behaviour of users shows the same tendency.

The navigation behaviour of users is an important field of interest for various target groups. Website owners could adapt their systems to the needs of the users or information scientists could investigate in new navigation methods. Simply knowing about the existence of certain biases and tendencies in the user navigation behaviour, could have numerous benefits. Trends of user visited articles, in particular articles about notable persons, would give insight into user navigation behaviour.

This thesis focuses on analyzing tendencies and trends of user navigational behaviour concerning Wikipedia articles that are about persons. Furthermore, it is of interest, how the user navigation differ regarding Wikipedia articles about persons that are born in different decades or countries. Another motivation for this thesis was to get insight into how Wikipedia articles about persons are perceived within the network structure compared to the user navigated articles.

These analysis are new as they rely on the Wikipedia network structure and combine this topology information with the actual user navigational behaviour in form of a dataset that contains the click data of links between Wikipedia articles.

1.2. Contribution

This thesis was driven by the following questions:

- ***When users navigate Wikipedia articles about persons, does their navigational behaviour show any tendency in general?*** This question addresses the user navigational behaviour in a general way. If there exists such a general gender tendency, this finding can be interesting for a wide field of research and for further activity. One the one hand, researchers can investigate in finding out why there is such a trend or what facts lead to a general gender tendency. On the other hand, website owners or administrators can modify their systems so that users are able to navigate information systems in a better way.

- ***Does the user navigational behaviour show any tendency regarding the year of birth or country of birth, when they navigate Wikipedia articles about persons?*** The results of this question also address the user navigational behaviour, like the question before, but not only in a general way. If the Wikipedia articles about persons are grouped by birth year and birth country, does the user navigational behaviour restricted to these groups also show a gender trend? By analyzing these groups and comparing the results to the findings of the general user navigational behaviour, it can be concluded that user navigation behaviour is different for certain Wikipedia article groups compared to their general navigation behaviour. Furthermore comparisons between the different groups can be made and with these findings it is possible to gain a better understanding of the user navigation process.

- ***Who benefits most from the user navigational behaviour on Wikipedia? Are female or male articles more preferred by user clicks?*** A comparison between the Wikipedia network structure, which is the predefined source for navigation, and the actual user navigation provides additional information on finding a trend in the user navigational behaviour. These results illustrate what genders benefit most from the user clicks and for example, website owners can adapt their information systems to improve navigation.

All these questions have in common, that they try to gain insight into the user navigation process regarding navigating Wikipedia articles about persons. With these findings it could be possible to get a better understanding of the user navigational behaviour on complex information networks such as Wikipedia and they could be a basis for adapting the structure of information networks to better fit the needs of navigating users.

To answer these questions, this thesis uses the Wikipedia network link structure as well as a user click dataset that reflects which Wikipedia articles were visited during user navigation sessions. Furthermore the Random Surfer Model - a model for random walks in weighted directed graphs - is used to interpret the user navigational behaviour.

Moreover, this thesis provides the following contributions:

- A method is shown to analyze the Wikipedia network structure as well as the user navigation data and to compare these results.

- It is shown that a grouping of Wikipedia articles about persons can be done and that it is possible to automatically label the gender of Wikipedia articles about persons.

- This study gives some insight in the user navigation behaviour and answers the above mentioned questions to analyze tendencies and trends regarding user navigational behaviour on Wikipedia.

1.3. Thesis Outline

This thesis is divided into seven chapters. The introduction is followed by Chapter 2 in which related work is discussed. Chapter 3 provides the background that is necessary for the context of the thesis as well as the metrics and the Random Surfer Model - the main model used in this thesis The three used datasets (Wikipedia network, Wikipedia Clickstream, DBpedia) are illustrated in Chapter 4. Chapter 5 describes the experimental setup used in this thesis and in Chapter 6 the results are discussed. Finally, in Chapter 7 the thesis is concluded, limitations are pointed out and suggestions for future work are made. The Appendix includes detailed quartile result tables of the calculated stationary distributions (Appendix A) as well as the bibliography.

2. Related Work

This thesis concentrates on analyzing biases in user navigation behaviour. Therefore the research areas of navigation models and gender inequalities and biases in information and social networks are of particular importance. Section 2.1 discusses related works of other scientists in the field of navigation and navigation behaviour in complex networks such as Wikipedia. In Section 2.2 related works regarding gender inequalities and biases are illustrated.

2.1. Navigation in Complex Networks

Wikipedia - a complex information network - is one of the most frequently used websites world wide and thus an interesting platform for scientists to analyze user navigation behaviour as well as the underlying network structure.

In 2010, Ratkievicz et al. investigated the traffic of social media including information networks such as Wikipedia (Ratkiewicz, Flammini, and Menczer, 2010). Their study shows that Wikipedia itself takes a central position within the World Wide Web and *attracts* users from other websites. Out of a dataset with more than $5,000,000$ requests they categorized the network traffic into four groups. These requests solely consist of referrers and targets that are Wikipedia articles. **Directory** means that the traffic mostly comes from inside Wikipedia and immediately leaves to other websites whereas **Search** is interpreted as page visits in order to find external resources. The third group **Encyclopedia** is defined as page visits from outside and resulting in visiting internal resources and the last group **Browsing** refers to request from one Wikipedia resource to another. The results show that Browsing and Encyclopedia are the dominant groups. Moreover, Ratkievicz et al. investigated the navigation behaviour inside the Wikipedia. For that purpose they analyzed a Wikipedia proxy server log dataset that includes the hit counts for the Wikipedia articles. They concluded that the user visits of neighbour articles are correlated and that content correlation outperforms traffic correlation.

Helic (D. Helic, 2012) analyzed the user navigation behaviour of users based on the navigation game *Wikigame* as well as the provided Wikipedia network structure. The structure analysis shows that there are some Wikipedia articles which are connected very well - these are called hubs - and that there are a lot of articles that do not have a high in-degree and out-degree - referred to as peripheral hubs. Another finding is that the Wikipedia network structure provides a strongly connected component that consists of 55 % of all articles and the weakly connected component covers 99 % of all articles. Furthermore, Helic concludes that users are very efficient at navigating. On average, they are able to navigate from one random Wikipedia article to another Wikipedia article with 6.27 clicks, which is just half a click more than the global shortest path with the value of 5.70. The results of the average navigation path distribution of the users are similar to the findings of the well known small-world paper from Milgram (Milgram, 1967). Finally Helic deduces that the user navigation process consists of two phases. The first phase is called *Zoom-Out* and the second phase is the so called *Zoom-In* phase.

In 2009, Boguna et al. (Boguna, Krioukov, and Claffy, 2009) also found these two phases when users navigate complex networks. They showed that the *Zoom-Out* phase is a coarse-grained search which leads the user towards the central network core. In general, this network core is a hub that connects many objects, for example, Wikipedia articles. So in the first phase the user navigates from the starting point to a hub (zooming out) and switches to the second phase. The *Zoom-In* phase - a fine-grained search - then leads the user from the network core towards the target (zooming in) in the periphery of the network. Another finding of Boguna et al. was that this two phase navigation process has two preconditions: The underlying network structure has to have a certain amount of hubs and a strong clustering. Networks that are scale-free and that have a degree distribution with small exponents, such as Wikipedia, fulfill these requirements.

West and Leskovec (West and Leskovec, 2012) questioned, if it was necessary to have structured knowledge and use high-level reasoning to navigate information networks. Therefore they compared the shortest paths of a large scale human navigational dataset from an online game *Wikispeedia*, a game similar to *Wikigame*, to the shortest paths computed by an automatic method, using agents for the latter. They concluded that, in general, high-level reasoning is not required and even simple agents outperform humans. They also argued that humans have a fixed plan in mind while navigating from a starting point to a target. This leads to missing good link opportunities for humans. Finally their feature analysis for agents stated that the influence of the degree should be less important on later path positions and similarity should be more weighted.

In 2012, Ghosh and Lerman (Ghosh and Lerman, 2012) studied dynamical processes in social networks, such as information diffusion and epidemics. Therefore they classified two types of dynamical processes (*conservative* and *non-conservative*) and related them to the centrality measures *PageRank* and *Alpha-Centrality*. A stochastic process is conservative if the weights in a graph are distributed among all the nodes of a network and that the total weight of the nodes remains constant. For example, the nodes of a network can be seen as a group of connected people each having a certain amount of money. In every iteration each person retains some of the initial money and distributes the remaining money among its neighbours. A non-conservative process was described as a stochastic process where the total weight of the network can change over time. To illustrate a non-conservative process they gave the following example. Imagine a group of connected persons each with a certain amount of money and a money minting machine. In contrast to the previous example each person in the network now can make some money and distribute additional money. Furthermore they showed that non-conservative processes have an effective transmissibility threshold and that Alpha-Centrality serves as a better centrality measure for social networks than PageRank. They showed for conservative processes that the quantity that is distributed on the network is constant and that the transfer matrix is a stochastic matrix, since every row sums up to one. The steady distribution of conservative processes can be described as an information diffusion on the network and Ghosh and Lerman showed that PageRank, which is based on random walks like the random surfer model, reflects this diffusion process.

Helic et al (Denis Helic et al., 2013) investigated in human navigation models of information networks such as Wikipedia based on decentralized search, a well established navigation model for social networks. The action selection mechanism for decentralized search can be divided into two groups. The first group includes deterministic methods like greedy methods and the second group consists of probabilistic approaches like ϵ-greedy, softmax-rule and inverse distance rule. In their study they used the greedy method as a baseline which is compared to the results of the three previously mentioned methods. They applied the different decentralized search methods to a dataset that contained the click paths of $250,000$ successfully completed games of TheWikiGame. The results showed that greedy decentralized search almost has the same success rates than human navigation paths but that the navigation behaviour of humans in information networks cannot be fully explained by decentralized search. They concluded that human navigation in information networks seems to be rather a stochastic process where deterministic greedy actions and random probability actions are combined. Furthermore they stated that these two actions correspond to the two navigation phases *exploitation* and *exploration*. Deterministic greedy actions are used in the exploitation phase when the user thinks they know the network whereas the exploration phase is carried out when links are followed randomly because the user does not know enough about the network.

2.2. Gender Inequalities and Biases

Gender inequalities are very common in today's society. This phenomenon is well known in many fields of mankind. For example, in the following spheres:

- **Education**: (Buchmann, DiPrete, and McDaniel, 2008), (Colclough, Rose, and Tembon, 2000)
- **Health**: (Annandale and Hunt, 2000), (Artazcoz, Borrell, and Benach, 2001), (Okojie, 1994)
- **Psychology**: (Barreto and Ellemers, 2005)

- **Economy**: (Forsythe, Korzeniewicz, and Durrant, 2000), (Beneria and Sen, 1982)

As the World Wide Web can be seen as a mapping of the world itself, aspects such as gender inequalities and gender biases are also transferred from the real world into the virtual world. The following researches are related to gender differences in Wikipedia.

In 2011 Reagle and Rhue (Reagle and Rhue, 2011) compared bibliographies on the basis of the English Wikipedia and the online encyclopedia *Britannica*. One of their findings was that the coverage of articles about persons is way higher in Wikipedia than in Britannica, regardless of the gender. As a baseline for their coverage measure they used six different sources that provide information about persons and their corresponding bibliographies. Another result was that they found an imbalance in gender representation. They used a dataset consisting of 18, 495 entries as a baseline and analyzed the entries published on Wikipedia. The results show that only 12 % of the biographies are about females and 88 % are about men. Finally they also stated that the perception of women is changing over time. For instance, in 2006 a top 100 list of the most influential people in American history covered 10 % females, whereas in 2008 the Time magazine's list of the most influential people included 23 % females.

Wagner et al. (Wagner et al., 2015) also investigated the coverage of Wikipedia articles about women and men and furthermore illustrated how different genders are perceived. They showed that Wikipedia does not have a significant difference in coverage of females and males. Also the visibility of women and men is nearly equal. They concluded this by finding no gender bias for featured articles on Wikipedia's start page. However, they also showed that women and men are portrayed differently. Therefore, Wagner et al. analyzed the network structure of Wikipedia and the articles about well-known persons which led to a finding of a structural as well as a lexical gender bias. The structural bias describes the centrality of the articles. In three of six language editions of Wikipedia, men are significantly more central than women, regardless of the used centrality measure. The lexical gender bias of the English Wikipedia shows that female biographies focus more on their romantic relationship and family-related issues. Their calculations are based on three different datasets that consist of around 125, 000 articles of notable people.

The gender inequality in the network structure of Wikipedia was analyzed in 2012 (Aragón et al., 2012). For that purpose Aragón et al. used the biographies of notable persons of the 15 biggest Wikipedia language editions where the list of the English Wikipedia is the base line for comparisons. For the centrality measure, *PageRank* (Brin and Page, 2012) and *Betweenness* were used. Both measuring methods showed that women are significantly less central than men regardless of the language edition. For example, only four persons out of the top five lists of all 15 language editions are female whereas 71 are male. Further results illustrated that men also have a higher in-degree. For instance, in the English Wikipedia the in-degree of the top female article is 665 whereas the in-degree for the top male article is 2, 123.

Collier and Bear (Collier and Bear, 2012) investigated the Wikipedia community concerning contributors, editors and readers. For this purpose they analyzed survey data from 176, 192 people who are or were interacting with Wikipedia contributions. Those contributions cover 22 different Wikipedia language editions including English, German, Spanish, French and Russian. In their study they formulated the following four hypotheses:

- **H1**: "Female Wikipedia users are less likely to contribute to Wikipedia due to the high level of conflict involved in the editing, debating, and defending process."
- **H2**: "Female Wikipedia users are less likely to contribute to Wikipedia due to gender differences in confidence in expertise to contribute and lower confidence in the value of their contribution."
- **H3**: "Female contributors are less likely to contribute to Wikipedia because they prefer to share and collaborate rather than delete and change other's work."
- **H4**: "Female contributors are less likely to contribute to Wikipedia because they have less discretionary time available to spend contributing."

Finally they concluded that their results support hypotheses H1, H2 and H3 whereas H4 is not supported. They also stated that there is a gender gap between females and males who are contributing to Wikipedia. The proportion of females is significantly lower than the proportion of men. (Hill and Shaw, 2013) and (Lam et al., 2011) show similar results.

3. Technical Background

This chapter outlines the background, technical methods and measures used in this thesis. The Sections 3.1 and 3.2 describe the online encyclopedia Wikipedia and the community project DBPedia. In Section 3.3 the process and background knowledge of web navigation is defined. Section 3.4 covers the creation and representation of the graphs. Section 3.5 presents the problem of categorising Wikipedia articles. Finally Section 3.6 describes the process of the energy calculation and its basic information.

3.1. Wikipedia

The online encyclopedia Wikipedia[1] is a free to use, collaborative internet encyclopedia. Launched on January 15, 2001 by Larry Sanger and Jimmy Wales, as an English encyclopedia, it is now one of the most frequently visited websites worldwide. Later on, this project became multilingual and now there are more than 290 different editions in over 250 different languages. The English Wikipedia [2] is the edition with the far most articles amongst all editions. Altogether the English Wikipedia combines more than $5,100,000$ articles. In contrast the second largest edition - the Swedish one - has about $2,880,000$ articles. On the whole the encyclopedia consists of over $38,000,000$ articles.

The detailed structure of a Wikipedia article varies depending on the topic. For example, articles about sport persons often have a career section whereas articles about countries or cities generally have a history or tourism section. However, a common pattern for articles is, that they have a title, followed by

[1] https://www.wikipedia.org

[2] https://en.wikipedia.org

an short definition called abstract and that statements must be referenced. Figure 3.1 illustrates the English Wikipedia article of Roger Federer. The first part of the article is the abstract and is located right before the content list. On the right upper corner, a so called infobox that is optional, is placed. An infobox is a table that represent a summary of some unifying aspects that are shared amongst articles. For instance, articles about tennis players should contain, amongst other things, the following:

- Name
- Residence
- Year of Birth
- Turned Pro
- Career Titles
- Highest Ranking

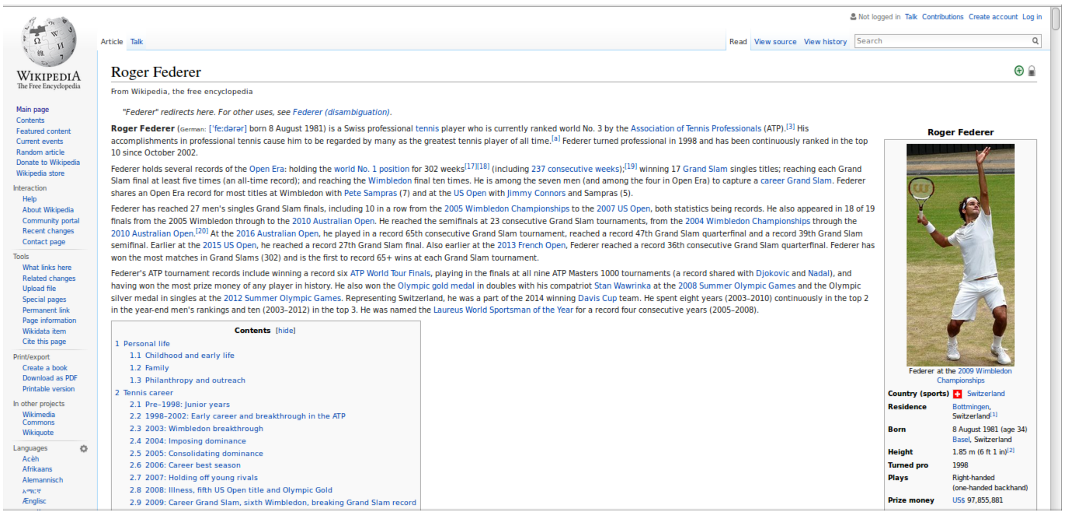
In opposition to that, the infobox of articles about countries should contain:

- Conventional Long Name
- Common Name
- Longitude and Latitude
- Area information
- Currency

The infobox should help the user to get a quick impression of important facts of the article. In the Wikipedia Help section, the following attributes are mentioned, that an infobox should have:

- **Comparable**: For common types of articles it is useful to compare their corresponding attributes that are shared between the articles.
- **Concise**: Succinct information gives a quick impression of the article.
- **Relevant**: Only relevant material should be included.
- **Cited elsewhere**: The provided data in the infobox should have cited entries to reliable sources.

The Wikipedia uses 17 different subject namespaces within the encyclopedia. The purpose of a namespace is to define a scope of websites with common topics. For example, all articles and encyclopedia redirects are combined in the main or article namespace, whereas templates for articles or infoboxes



Roger Federer

From Wikipedia, the free encyclopedia

"Federer" redirects here. For other uses, see Federer (disambiguation).

Roger Federer (German: ['fe:dərər] born 8 August 1981) is a Swiss professional tennis player who is currently ranked world No. 3 by the Association of Tennis Professionals (ATP).[3] His accomplishments in professional tennis cause him to be regarded by many as the greatest tennis player of all time.[a] Federer turned professional in 1998 and has been continuously ranked in the top 10 since October 2002.

Federer holds several records of the Open Era: holding the world No. 1 position for 302 weeks[17][18] (including 237 consecutive weeks);[19] winning 17 Grand Slam singles titles; reaching each Grand Slam final at least five times (an all-time record); and reaching the Wimbledon final ten times. He is among the seven men (and among the four in Open Era) to capture a career Grand Slam. Federer shares an Open Era record for most titles at Wimbledon with Pete Sampras (7) and at the US Open with Jimmy Connors and Sampras (5).

Federer has reached 27 men's singles Grand Slam finals, including 10 in a row from the 2005 Wimbledon Championships to the 2007 US Open, both statistics being records. He also appeared in 18 of 19 finals from the 2005 Wimbledon through to the 2010 Australian Open. He reached the semifinals at 23 consecutive Grand Slam tournaments, from the 2004 Wimbledon Championships through the 2010 Australian Open.[20] At the 2016 Australian Open, he played in a record 65th consecutive Grand Slam tournament, reached a record 47th Grand Slam quarterfinal and a record 39th Grand Slam semifinal. Earlier at the 2015 US Open, he reached a record 27th Grand Slam final. Also earlier at the 2013 French Open, Federer reached a record 36th consecutive Grand Slam quarterfinal. Federer has won the most matches in Grand Slams (302) and is the first to record 65+ wins at each Grand Slam tournament.

Federer's ATP tournament records include winning a record six ATP World Tour Finals, playing in the finals at all nine ATP Masters 1000 tournaments (a record shared with Djokovic and Nadal), and having won the most prize money of any player in history. He also won the Olympic gold medal in doubles with his compatriot Stan Wawrinka at the 2008 Summer Olympic Games and the Olympic silver medal in singles at the 2012 Summer Olympic Games. Representing Switzerland, he was a part of the 2014 winning Davis Cup team. He spent eight years (2003–2010) continuously in the top 2 in the year-end men's rankings and ten (2003–2012) in the top 3. He was named the Laureus World Sportsman of the Year for a record four consecutive years (2005–2008).

Contents [hide]

1 Personal life
- 1.1 Childhood and early life
- 1.2 Family
- 1.3 Philanthropy and outreach

2 Tennis career
- 2.1 Pre–1998: Junior years
- 2.2 1998–2002: Early career and breakthrough in the ATP
- 2.3 2003: Wimbledon breakthrough
- 2.4 2004: Imposing dominance
- 2.5 2005: Consolidating dominance
- 2.6 2006: Career best season
- 2.7 2007: Holding off young rivals
- 2.8 2008: Illness, fifth US Open title and Olympic Gold
- 2.9 2009: Career Grand Slam, sixth Wimbledon, breaking Grand Slam record

Roger Federer

Federer at the 2009 Wimbledon Championships

Country (sports)	Switzerland
Residence	Bottmingen, Switzerland[1]
Born	8 August 1981 (age 34) Basel, Switzerland
Height	1.85 m (6 ft 1 in)[2]
Turned pro	1998
Plays	Right-handed (one-handed backhand)
Prize money	US$ 97,855,881

Figure 3.1.: **Wikipedia article of Roger Federer:** This is an example for a Wikipedia article. The title is placed in the left upper corner followed by an abstract. The content list is directly after the abstract. On the right upper side is the infobox with certain properties such as residence, date of birth and place of birth.

are located in the template namespace. The following list contains the different namespaces:

- Main
- User
- Wikipedia
- File
- MediaWiki
- Template
- Help
- Category
- Portal
- Book
- Draft
- Education Program
- TimedText
- Module
- Gadget
- Gadget definition

- Topic

As mentioned before the Wikipedia encyclopedia is a collaborative project. This means that everybody can read these articles but also has the right to create and edit one. These privileges lead to a bunch problems when using Wikipedia. Since everybody can edit the content, Wikipedia should not be used as a reference for scientific work. Moreover there are different kinds of biases within the Wikipedia. A full list of Criticism is also available as a Wikipedia article (Wikipedia, 2016).

3.2. DBpedia

The DBPedia is a community project started as a cooperation between the Free University of Berlin[3], Leipzig University[4] and OpenLink Software[5]. The aim of this project is to provide structured information of the Wikipedia encyclopedia and make this information publicly available on the World Wide Web. Wikipedia articles itself mostly consist of free text but also provide the in the previous section described infobox information. All gathered structured data then is represented with the Resource Description Framework (RDF)[6] to allow users to ask sophisticated queries against Wikipedia. With the use of RDF, DBpedia is connected with other Linked Datasets such as Freebase[7] or GeoNames[8] by more than 50 million RDF links. Figure 3.2 shows the DBpedia within the Linking Open Data Cloud of August 2014. This figure shows that the DBpedia dataset is a central concept within the Linking Open Data Cloud. It is connected with many other datasets (which is indicated with the huge amount of black lines around the DBpedia dataset) and thus is a good source for retrieving information.

[3] http://www.fu-berlin.de/en

[4] https://www.zv.uni-leipzig.de/en/

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

[6] RDF is cumulation of specifications of the World Wide Web Consortium. The RDF is used as a general method for conceptual description or modelling of information.

[7] http://www.freebase.com/

[8] http://www.geonames.org/

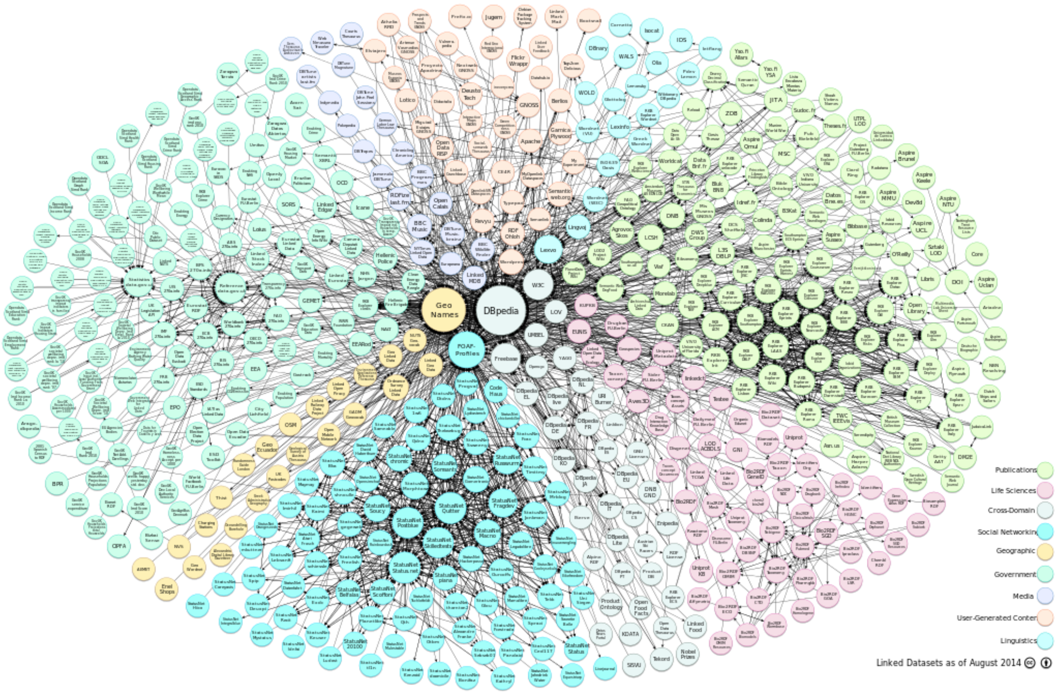


Figure 3.2.: **DBpedia in the Linking Open Data Cloud 2014:** This figure shows the Linking Open Data Cloud of 2014 and DBpedia is placed in the center of this cloud. The size of the circle describes the size of the linked dataset whereas the lines between the circles illustrates links between the different datasets. It can be seen that the DBpedia dataset is a central concept within the Linking Open Data Cloud and that it is connected to many other open datasets and thus is a good source for retrieving information. (CC-BY-SA-3.0 from `https://commons.wikimedia.org/wiki/File:LOD_Cloud_2014.svg`)

This project provides structured information in localized versions in 125 different languages. The biggest version is the English one and the corresponding DBpedia knowledge base describes $4,580,000$ things. For instance, more than $1,445,000$ persons, $735,000$ places or $411,000$ creative works are included in the knowledge base. Beside the sophisticated queries against Wikipedia, DBpedia also offers the datasets for download. The following enumeration lists some provided localized datasets:

- **Titles**: Titles of all Wikipedia articles
- **Short Abstracts**: Short abstracts of Wikipedia articles
- **Images**: Main image and corresponding thumbnail from Wikipedia articles
- **Geographic Coordinates**: Extracted coordinates from Wikipedia
- **Raw Infobox Properties**: The raw information that has been extracted from Wikipedia infoboxes
- **Persondata**: Information about persons, such as place of birth or date of birth
- **Page IDs**: Linking the Wikipedia page ID to the corresponding DBpedia resource

3.3. Web Navigation

Web navigation is a special type of navigation an refers to a process of navigating networks in the World Wide Web. These networks are organized as hypertext or hypermedia and their structure can be described with a graph or a matrix. In context of websites, such as Wikipedia, navigation means following links on these websites, typing an URL in the address bar of a web browser or using the internal search functionality. As mentioned before, networks can be described with graphs where articles are the vertices and links are the edges. Following a link of a website here means going from one edge - the article that has the link or also called source - to another edge - the article that is the target of the link. If a user types an URL in the address bar of the web browser or uses the internal search functionality, the source page does not have a link to the target page in most cases. With this in mind it would not be possible to directly navigate to the target page

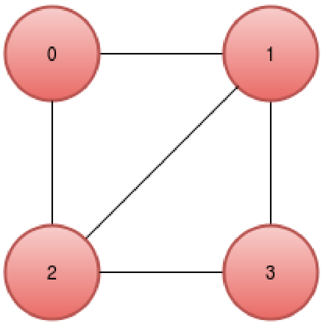


Figure 3.3.: **Sample graph:** This figure illustrates a sample graph with four vertices and five edges. The vertices are marked as red circles and the edges are the connections between the vertices.

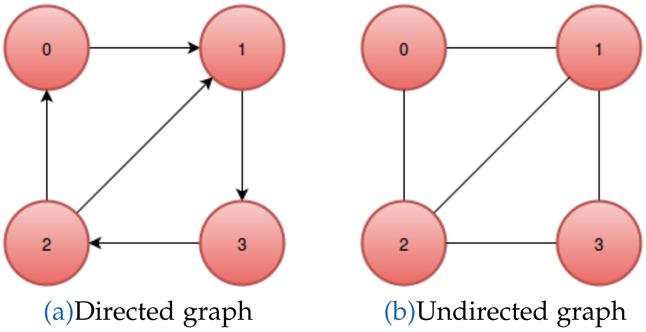
with the available link structure of the graph. This navigation step is called teleportation. Web navigation without teleportation is named strict web navigation and this is the basis for the later used Wikipedia Clickstream Dataset which is described in Section 4.1.

3.4. Graph Theory

Graph theory deals with graphs and its attributes. A graph consists of vertices and edges and a common description is

$$G = (V, E) \quad , \tag{3.1}$$

where G is the graph, V is the set of vertices and E the set of edges. A vertex often symbolises objects - in this thesis Wikipedia articles - and an edge describes a connection between edges - in this thesis a link from one Wikipedia article to another. Figure 3.3 shows a graph with four vertices and five edges where the vertices are drawn as circles with numbers in it and the edges are lines which connect the vertices.



(a)Directed graph (b)Undirected graph

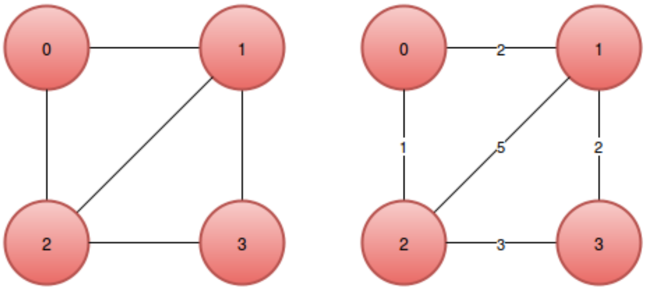
Figure 3.4.: **Directed graph:** The connection between edges can have an orientation. If so, this graph is a directed graph as seen in 3.4a. The direction of the edge is marked with an arrow. For example, the vertex 2 is connected with vertex 0 but not vice versa. 3.4b shows an undirected graph an thus the edges do not have arrows. The connection between the vertices is bidirectional. For example, the vertex 3 is connected with vertex 1 as well as vertex 1 with vertex 3.

3.4.1. Directed and Undirected Graphs

Graphs can be divided into directed and undirected graphs. A graph attribute reflects the type of connection between vertices. Undirected graphs do not have an orientation for their edges and thus connection between the vertices are bilateral. A directed graph in contrast consists of directed edges between vertices. Typical applications of graphs are maps for street navigation where crossovers are represented as vertices and the street between crossings are edges. If it is possible to use the streets in both directions the map can be represented as an undirected graph, however if the map has one ways - streets which can be used just in one direction - the more suitable graph is a directed one. Figure 3.4a is a directed graph whereas Figure 3.4b shows an undirected graph.

3.4.2. Weighted and Unweighted Graphs

Another attribute of a graph is the possibility to weight the edges. If it is necessary to distinguish between the connections between vertices a weighted graph can be used to assign a number (weight) to the connection.



(a)Unweighted graph (b)Weighted graph

Figure 3.5.: **Weighted graph:** Vertices in a graph are connected with edges and these connections can be weighted or unweighted. In 3.5a the edges are unweighted and thus all edges between vertices are the same. Compared to this, in 3.5b the edges are weighted. For example, vertex 0 is connected with vertex 1 an this connection has the weight 2 whereas the connection between vertices 3 and 2 are weighted with value 3.

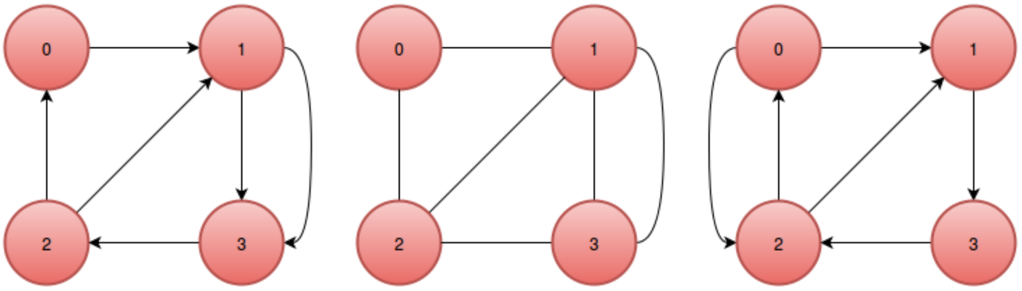
A street map for example models a weighted graph where the distance between two crossovers can represent the weight for the edge. Figure 3.5a shows an unweighted graph whereas Figure 3.5b is a weighted graph.

3.4.3. Multi- and non Multigraphs

Another categorisation is the necessity of multiple edges. In multigraphs two vertices can be connected with several edges whereas non multigraphs just have one edge between two vertices. There is an exception in directed graphs concerning multiple edges, namely if two vertices are connected with two edges, it only counts as a multigraph if the edges have the same orientation. Figure 3.6a shows a directed multigraph, Figure 3.6b is an undirected multigraph and Figure 3.6c represents a directed non multigraph.

3.4.4. Connectivity

Furthermore undirected graphs can be connected or disconnected. An undirected graph is connected when every vertex is reachable within the



(a)Directed multigraph (b)Undirected multigraph (c)Directed non multigraph

Figure 3.6.: **Mulitgraphs:** Mulitgraphs are graphs that have multiple connections between two or more vertices. In 3.6a a directed multigraph is illustrated. There vertex 1 is connected with vertex 3 with two edges with the same direction. 3.6b shows an undirected mulitgraph. Vertex 3 and vertex 1 are connected with two edges. 3.6c shows a directed non mulitgraph. Although the vertices 0 and 2 are connected with two edges this graph is a non mulitgraph, because the edges do not have the same direction or orientation.

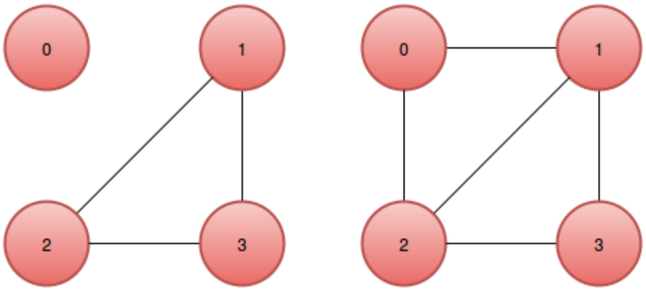
graph. Figure 3.7a is an undirected disconnected graph and Figure 3.7b shows an undirected connected graph.

In general directed graphs have three connectivity statuses:

- *Disconnected*
  A directed graph is disconnected when not every vertex is reachable.
- *Weakly connected*
  If the replacement of all directed edges with undirected edges produces a connected (undirected) graph, the directed graph is weakly connected.
- *Strongly connected*
  A directed graph is strongly connected if there exists a directed connection between every pair of vertices.

Figure 3.8a shows a directed disconnected graph, Figure 3.8b is a directed weakly connected graph and Figure3.8c represents a directed strongly connected graph.

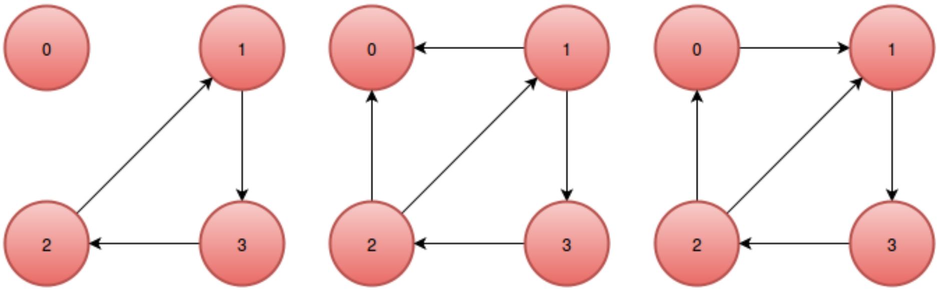
In this thesis just strongly connected, directed, weighted, non multigraphs are used for further analysis.



(a)Undirected disconnected graph

(b)Undirected connected graph

Figure 3.7.: **Undirected connected graph:** A connected graph is a graph where all vertices are connected. 3.7a shows an undirected disconnected graph because vertex 0 is not connected to any other vertex of the Graph. In 3.7b a undirected connected graph is shown. All four vertices are reachable within the graph.



(a)Directed disconnected graph

(b)Directed weakly connected graph

(c)Directed strongly connected graph

Figure 3.8.: **Directed connected graph:** Directed graphs can be disconnected, weakly connected or strongly connected. 3.8a shows a directed disconnected graph, because vertex 0 is not connected to any other vertex of the graph. In 3.8b a directed weakly connected graph is shown. This graph is weakly connected because every vertex is reachable but it is not possible to reach every vertex from any other vertex. For example, it is not possible to go from vertex 0 to vertex 3. Furthermore if every directed connection is replaced with a undirected connection, the graph would be a undirected connected graph. 3.8c shows a directed strongly connected graph. For every pair of vertices, a directed connection exits.

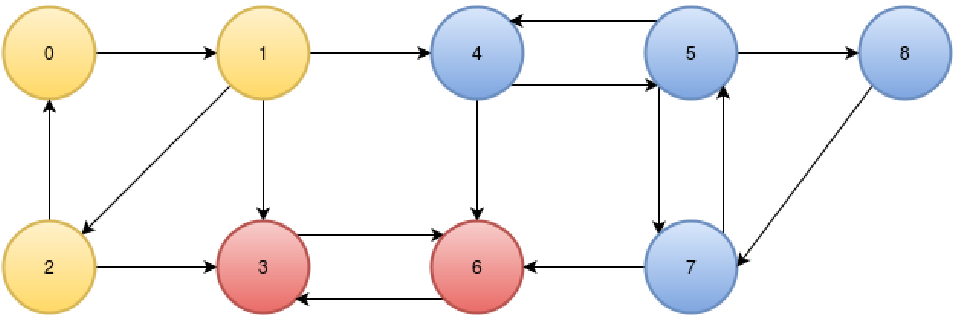


Figure 3.9.: **Strongly connected components**: A graph can be decomposed into several disjoint SCCs. If the graph just consists of one SCC the graph itself is a strongly connected graph. In this figure the graph has three SCCs that are marked with different colours. The yellow SCC consists of vertices 0, 1 and 2 and the red SCC includes the vertices 3 and 6. Finally the biggest SCC is the blue one with the vertices 4, 5, 7, and 8.

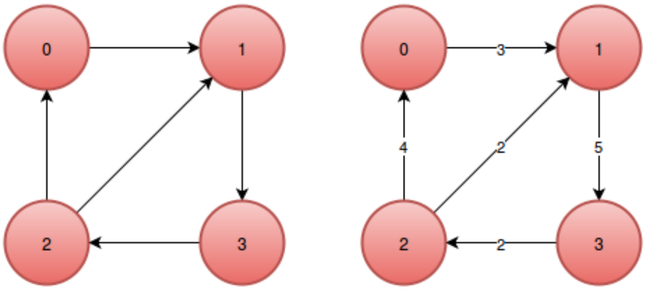
3.4.5. Strongly Connected Component

A strongly connected component (SCC) of a directed graph G is the maximal subgraph G′ which is strongly connected. Every directed graph can be decomposed into several disjoint SCCs which is represented in Figure 3.9. The biggest SCC is the subgraph with the highest amount of vertices. In Figure 3.9 this is the subgraph with the blue vertices. In this thesis the biggest SCC of directed graphs is used several times.

3.4.6. Adjacency Matrix

An adjacency matrix $\boldsymbol{A}$ is a matrix representation of a graph which indicates which pairs of vertices are adjacent. $\boldsymbol{A} = [a_{ij}]$ for an unweighted non multigraph is defined as

$$a_{ij} = \begin{cases} 1 & \text{if } (j,i) \in E, \\ 2 & \text{if } i = j \text{ and } (i,i) \in E, \\ 0 & \text{otherwise} \end{cases} . \tag{3.2}$$



(a)Unweighted non multigraph (b)Weighted non multigraph

Figure 3.10.: **Unweighted non multigraph and weighted non multigraph:** 3.10a illustrates an unweighted non multigraph and 3.10b is a weighted non multigraph.

$\boldsymbol{A} = [a_{ij}]$ for a weighted non multigraph with edge weight c is defined as

$$a_{ij} = \begin{cases} c_{ij} & \text{if } (j,i) \in E, \\ 0 & \text{otherwise} \end{cases} . \tag{3.3}$$

Figure 3.10a shows an unweighted non multigraph and Figure 3.10b is a weighted non multigraph. The corresponding adjacency matrices are as follows:

$$\boldsymbol{A_A} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} \boldsymbol{A_B} = \begin{pmatrix} 0 & 0 & 4 & 0 \\ 3 & 0 & 2 & 0 \\ 0 & 0 & 0 & 2 \\ 0 & 5 & 0 & 0 \end{pmatrix}$$

3.5. Article Categorisation

The categorisation of articles is an important part for analysing tendencies. Later on analysis can be granular detailed based on the categories of the Wikipedia articles. In this thesis there are three different categories defined which are all inferred from personal attributes and so the categorisation is focused on articles about persons. The category for each Wikipedia article with page identifier *id* is defined as:

- *Gender*

$$gender_{id} = \begin{cases} female & \text{if } id \text{ refers to a female person,} \\ male & \text{if } id \text{ refers to a male person,} \\ no\_gender & \text{otherwise} \end{cases} \tag{3.4}$$

- *Birth Year by*

$$birthyear_{id} = \begin{cases} by & \text{if } id \text{ refers to a person and} \\ & \text{infobox properties includes} \\ & \text{birth year,} \\ no\_birthyear & \text{otherwise} \end{cases} \tag{3.5}$$

- *Birth Country bc*

$$birthcountry_{id} = \begin{cases} bc & \text{if } id \text{ refers to a person and} \\ & \text{infobox properties includes} \\ & \text{birth country,} \\ no\_birthcountry & \text{otherwise} \end{cases} \tag{3.6}$$

3.5.1. Term Frequency and Inverse Document Frequency

In the field of information retrieval there are two common weighting schemes: Term Frequency (TF) and Inverse Document Frequency (IDF). These methods measure how important a word in a collection of documents is and both schemes are working with vectors. Documents, such as Wikipedia articles, can be represented as such vectors. The dimension of these vectors is defined through the amount of different terms used in the documents. Counting the occurrences of each term in the documents and using these counts as a weighting scheme is called Term Frequency (TF). Table 3.1 provides five example documents and Table 3.2 shows the TF of these example documents.

Document ID	**Document**
D1	Vienna Vienna
D2	Vienna Vienna Austria Austria Germany Germany Germany
D3	Vienna Austria Austria Austria Germany Germany
D4	Austria Austria Austria
D5	Austria

Table 3.1.: **Example documents for TF and IDF:** This table contains five different example documents with the three terms *Austria*, *Vienna* and *Germany*.

Term **Document**	Germany	Vienna	Austria
D1	0	2	0
D2	3	2	2
D3	2	1	3
D4	0	0	3
D5	0	0	1

Table 3.2.: **TF for example documents:** This table shows the TF of every term for the five example documents. For example, document 2 has 3 occurrences of term *Germany*, 2 of term *Vienna* and of term *Austria*.

When TF is used for measuring relevance of a document in a query, this weighting scheme will prioritize documents which have a high occurrence of such words. For instance, a query with the terms *Austria* and *Germany* will retrieve two documents (D2, D3). D2 has three occurrences of *Germany* and two of *Austria* and D3 has two occurrences of *Germany* and three of *Austria*. Since the combination of both TFs have the same value, both documents get the same relevance. One approach to penalize words which appear to often is too use an IDF factor. Therefore the Document Frequency (DF) must be calculated. The DF is the number of documents that contain a specified term. So a discriminative word will have a low DF and the IDF will be high. A common definition of the IDF is the following:

$$IDF_{term} = log(\frac{N}{DF_{term}}) \quad , \tag{3.7}$$

where N is the number of documents and DF_{term} is the document frequency of the given term. Table 3.3 shows the DF and IDF of the example documents provided in 3.1. Finally the combination of TF and IDF

$$TFxIDF_{iterm} = TF_{iterm} log(\frac{N}{DF_{term}}) \tag{3.8}$$

will have a high result if the *term* occurs just a couple of times in the document collection and most appearances are in the document i. In table 3.4 the TFxIDF weighting is shown. Now the previous query for the terms *Austria* and *Germany* will also retrieve D2 and D3, but D2 has a higher relevance since the term *Germany* is more discriminative.

3.5.2. Support Vector Machine

Support Vector Machines (SVMs) are a supervised learning model in the field of machine learning. SVMs support binary classification of data as well as regression. In this thesis SVM is used as a binary classifier.

The fundamental function of a SVM is to train a model with labelled data

$$\{(\vec{x}_i, y_i) | i = 1, ..., n; y_i \in \{-1, 1\}\} \quad , \tag{3.9}$$

Term	Germany	Vienna	Austria
DF	2	3	4
IDF	0.3979	0.2218	0.0969

Table 3.3.: **DF and IDF for example documents:** This table shows the DF of the three terms *Austria*, *Vienna* and *Germany* as well as the calculated IDF for every term. For example, the term *Germany* appears in 2 documents, less occurrences than any other term and thus has the highest IDF of all terms.

Term **Document**	Germany	Vienna	Austria
D1	0	0.4436	0
D2	1.1937	0.4436	0.1938
D3	0.7958	0.2218	0.2907
D4	0	0	0.2907
D5	0	0	0.0969

Table 3.4.: **TFxIDF weighting for example documents:** This tables shows the TFxIDF weighting for every term with every document. For example, the term *Germany* has the highest weights whereas the term *Austria* has smallest values.

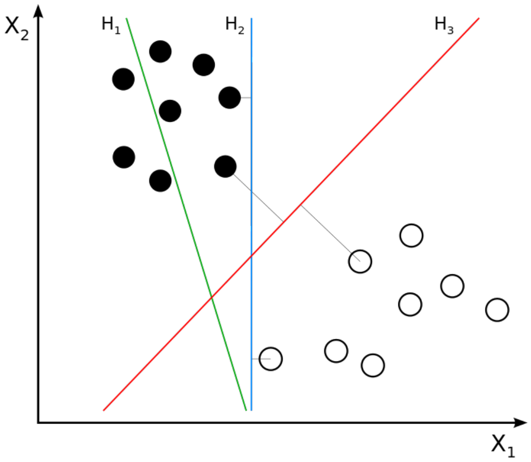


Figure 3.11.: **SVM hyperplanes**: This figure illustrates data points for two classes. One class is marked with black circles and the other class are the white circles. The classes are not separated with H1, only H2 and H3 separate them. H3 has a higher margin than H2 and in this case is the maximum-margin hyperplane. (CC-BY-SA-3.0 from `https://commons.wikimedia.org/wiki/File:Svm_separating_hyperplanes_(SVG).svg`)

where $\vec{x}_i$ is a training sample and y_i the corresponding class, and construct a hyperplane in a high-dimensional vector space to separate the data into two classes. This placement of the hyperplane should maximize the minimal margin between the two classes and thus create a so called maximum-margin hyperplane. Figure 3.11 shows different hyperplanes and the maximum-margin hyperplane for a two dimensional space.

Many classifications can be done with a linear separable SVM. If two classes can be linearly separated, the hyperplane can be represented as

$$bmvecw \cdot \vec{x} - b = 0 \quad , \tag{3.10}$$

where $\vec{x}$ is the set of training data, $\vec{w}$ is the normal vector of the hyperplane and b is the offset of the hyperplane from the origin along $\vec{w}$. Figure 3.12 illustrates the maximum-margin hyperplane with the maximal margin and support vectors. The margin is defined through two hyperplanes

$$\vec{w} \cdot \vec{x} - b = 1 \tag{3.11}$$

and

$$\vec{w} \cdot \vec{x} - b = -1 \quad , \tag{3.12}$$

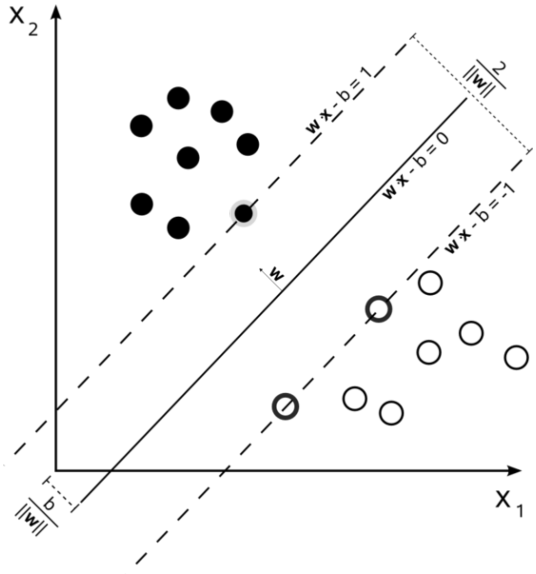


Figure 3.12.: **Maximum-margin hyperplane**: The maximum-margin hyperplane is located in the center of the margins. These margins are solely defined through the samples. These samples are on these margins and are called support vectors. (Public Domain from `https://commons.wikimedia.org/wiki/File:Svm_max_sep_hyperplane_with_margin.png`)

where the maximum-margin hyperplane is in the middle of them. Geometrically the distance between these two hyperplanes or the length of the margin is $\frac{2}{||\vec{w}||}$. In order to maximize the margin, $\vec{w}$ must be minimized. Another restriction is that each sample of the training data must be on the margin or outside the margin. Each training sample i must either fulfill

$$\vec{w} \cdot \vec{x}_i - b \geq 1, if y_i = 1 \tag{3.13}$$

or

$$\vec{w} \cdot \vec{x}_i - b \leq 1, if y_i = -1 \quad . \tag{3.14}$$

In case of linear SVM the decision function (classifier) is the solution of following optimisation problem:

$$Minimize\, ||\vec{w}||\, subject\, to\, y_i(\vec{w} \cdot \vec{x}_i - b) \geq 1, \forall 1 \leq i \leq n \quad . \tag{3.15}$$

As seen in Figure 3.12 the margin is solely defined by those data points which are closest to it. These are specially marked and are called support vectors.

In this thesis only a linearly separable SVM is used, however there are several different parameters for SVM. For all tried parameters see Subsection 4.3.1.

3.5.3. Evaluating Binary Classifiers

To evaluate the performance of binary classifiers different measures are used. Three of them are Precision (P), Recall (R) and F_1 score (F_1). P describes the fraction of retrieved samples which are relevant whereas R is defined as the fraction of relevant samples which are retrieved. F_1 is a trade-off between P and R and can be defined as harmonic mean of P and R. Table 3.5 shows a contingency table of a binary classification. With this classification the measures are calculated as follows:

$$P = \frac{tp}{tp + fp} \quad , \tag{3.16}$$

$$R = \frac{tp}{tp + fn} \quad , \tag{3.17}$$

$$F_1 = 2\frac{P \cdot R}{P + R} \quad . \tag{3.18}$$

The P and R measures have a range between zero and one and thus resulting in a F_1 range between zero - when P or R is zero - and one - when P and R are one. In general the F_1 measure is a good trade off and is used as the significant measure for binary classification in this thesis.

Real class **Prediction**	Positiv condition	Negative condition
Positive predicted condition	True positive (tp)	False positive (fp)
Negative predicted condition	False negative (fn)	True negative (tn)

Table 3.5.: **Contingency table for binary classification:** This table illustrates the four cases for binary classification: *True positives* are those cases where the positive condition was predicted positive and *false negative* where the positive condition was predicted negative. The other cases are *false positive* (negative condition was predicted positive) and *true negative* (negative condition was predicted negative).

3.6. Random Surfer Model and Stationary Distribution

The random surfer model is an example explanation for navigating graphs. As the name states, the navigation behaviour is described by random walks through the graph. It is possible to calculate the stationary distribution π for a strongly connected weighted directed graph - which solely is used in this thesis. The stationary distribution includes the energy values of each vertex and these energy values define the probability that a random surfer visits a certain vertex of the graph in the limit of infinitely many steps. Besides the previously mentioned graph properties, another restriction applies to the graph so that such an energy exists: The adjacency matrix of the graph must not allow only periodic returns to a given state (Geigl et al., 2015, p.4). The mathematical definition of π is:

$\boldsymbol{W} = [w_{ij}]$ for a strongly connected weighted graph $G = (V, E)$ with edge weight c is defined as

$$\boldsymbol{W}_{ij} = \begin{cases} c_{ij} & \text{if } (j,i) \in E, \\ 0 & \text{otherwise} \end{cases} . \tag{3.19}$$

The weighted out-degree k_i^+ defines the weight sum of the outgoing links of node i:

$$k_i^+ = \sum_{j=1}^{n} \boldsymbol{W}_{ji} \tag{3.20}$$

$\boldsymbol{D} = [d_{ii}]$ is a diagonal matrix of weighted out-degrees:

$$\boldsymbol{D}_{ii} = \begin{cases} k_i^+ & \text{if } k_i^+ > 0, \\ 1 & \text{otherwise} \end{cases} \tag{3.21}$$

$\boldsymbol{P} = [p_{ij}]$ is the transition matrix of the graph for the random surfer. The element $[p_{ij}]$ is the probability of the random surfer going from vertex j to vertex i.

$$\boldsymbol{P} = \boldsymbol{W}\boldsymbol{D}^{-1} \tag{3.22}$$

The solution for π is the right eigenvector of the appropriate largest eigenvalue, mathematically defined as follows:

$$\boldsymbol{\pi} = \boldsymbol{P}\boldsymbol{\pi} \tag{3.23}$$

Figure 3.13 shows an example graph which is strongly connected and consists of five vertices. The corresponding adjacency matrix, weighted out-degree diagonal matrix, probability matrix and stationary distribution are as follows:

$$\boldsymbol{W} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \end{pmatrix} \boldsymbol{D} = \begin{pmatrix} 3 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

$$\boldsymbol{P} = \begin{pmatrix} 0 & 0 & 0.5 & 0 \\ 0.333 & 0 & 0 & 1 \\ 0.333 & 1 & 0 & 0 \\ 0.333 & 0 & 0.5 & 0 \end{pmatrix} \boldsymbol{\pi} = \begin{pmatrix} 0.167 \\ 0.278 \\ 0.333 \\ 0.222 \end{pmatrix}$$

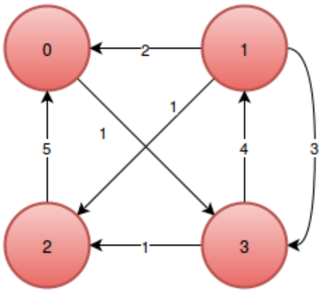


Figure 3.13.: **Strongly connected weighted directed graph:** This figure illustrated a graph with four vertices and five edges. The connections have an orientation and are weighted. Vertex 1 and vertex 3 are connected with two edges but they do not have the same direction. Furthermore all pairs of vertices are directed connected and thus this figure shows a strongly connected weighted directed graph.

4. Datasets & Data Preprocessing

This chapter describes the used datasets in this thesis as well as the preliminary data preprocessing of them. Section 4.1 illustrates a dataset of Wikipedia user clicks. In Section 4.2 the deployed Wikipedia network is introduced. Finally Section 4.3 outlines the used data of DBpedia.

4.1. Wikipedia Clickstream

The Wikipedia Clickstream (Wulczyn and Taraborelli, 2015) project analyses the HTTP request logs of the English Wikipedia and provides the prepared data. The main aim of this data preparation is to create pairs of referer and resource and the corresponding amount of clicks. The first part of the pair - the referer - points to the source of the request and the resource to the target. Both, referer and resource, are part of the HTTP header and are identified with a URL. Altogether this dataset contains $22,509,896$ different referer-resource pairs with a total of 3.2 billions requests which were made in February 2015.

Wikipedia defines 35 different namespaces and this dataset just includes the main namespace of the desktop version of the English Wikipedia which indicates a Wikipedia article. The referers were mapped to a set of values based on the following scheme:

Referer is

- an article that has a main namespace of English Wikipedia → *title of the article*
- another Wikipedia page which does not have a main namespace of English Wikipedia → *other-wikipedia*

- empty → *other-empty*
- a page from any other Wikimedia project → *other-internal*
- from Google → *other-google*
- from Yahoo → *other-yahoo*
- from Bing → *other-bing*
- from Facebook → *other-facebook*
- from Twitter → *other-twitter*
- anything else → *other-other*

Redirects of Wikipedia pages were also resolved. So if a Wikipedia article refers to several pages, this dataset includes the resolved page based on the Wikipedia redirects table. Another data preparation of this project is the attempt to exclude spider traffic as well as traffic generated by bots. Finally just referer-resource pairs which were clicked at least eleven times were included in this dataset.

Table 4.1 shows an example of this dataset which has the following format:

- *prev_id*: This contains the MediaWiki page ID of the article in the main namespace of the referer. Otherwise it will be emtpy.
- *curr_id*: This containt the MediaWiki page ID of the article in the main namespace of the resource.
- n: The number of occurrences of the referer-resource pair
- *prev_title*: The value of the above described referer mapping
- *curr_title*: The requested Wikipedia article in the main namespace
- *type*: A type definition of the referer-resource pair which can be:
  - *link*: If referer and resource are both Wikipedia articles in the main namespace
  - *redlink*: If referer is a Wikipedia article in the main namespace and the resource points to a Wikipedia article which does not exist yet.
  - *other*: If not *link* or *redlink*

Since this thesis is focusing on Wikipedia articles, this dataset is reduced to data with the type value **link**. Finally the used dataset consists of $12,194,530$ different referer-resource pairs which cover $972,643,099$ HTTP requests.

prev_id	**curr_id**	**n**	**prev_title**	**curr_title**	**type**
	331586	274	other-yahoo	Crocodile_Dundee	other
39737124		23	Yoon_So-hee	EXO_Music_Video	redlink
2371832	1261557	16	Carlos_Alomar	Heroes	link

Table 4.1.: **Example data of the Wikipedia Clickstream dataset:** This table illustrates all three types of Wikipedia Clickstream data. The first row does not have a *prev_id* (referer of HTTP request has not in the Wikipedia main namespace) and thus has type *other*. The second row is a *redlink* and so the *curr_id* field is empty (Wikipedia page for resource of HTTP request does not exist). The last type is *link* and this indicates a HTTP request from one Wikipedia article to another, in this example from Page ID 2371832 to Page ID 1261557. The *n* field describes the amount of clicked links. For example, from the Wikipedia article *Carlos_Alomar* the link to article *Heroes* was clicked 16 times.

4.2. Wikipedia

The online encyclopedia Wikipedia is a huge collection of knowledge represented as articles on web pages. The Wikipedia dataset is divided into two parts:

- Network structure
- Article data

4.2.1. Network Structure

Altogether the used network structure dataset which was provided by Dimitar Dimitrov includes 340, 131, 071 links from articles within the English Wikipedia main namespace. The links represented in this dataset are unique. If one Wikipedia article has two or more links to another Wikipedia article - which should not be the case due to Wikipedia rules - this dataset includes this link just once.

The data preparation for this dataset is based on the previously described Wikipedia Clickstream data. Links from and to Wikipedia articles which are not part of the Clickstream dataset are removed. In the end the resulting

Wikipedia dataset had 165, 383, 230 links which cover 2, 190, 218 different Wikipedia articles.

For further processing the links are used to create a network (graph) with graph-tool (Peixoto, 2014) and to calculate the SCC of the network. This SCC covers 164, 707, 695 (99, 59 %) links and 2, 140, 423 (97, 73 %) articles of the reduced Wikipedia dataset and is used for further analysis.

4.2.2. Article Data

The second type of Wikipedia dataset consists of the article content itself. This is used for classifying the gender of Wikipedia articles which are about persons. For crawling the articles content the Python *wikipedia*[1] package is used.

4.3. DBpedia

DBpedia's project aim is to extract structured information from Wikipedia and Wikidata and provide this information on the web. The used datasets are from April 2015 and can be downloaded from DBpedia's dataset section[2]. For this thesis the following sub datasets are analysed:

- **Persondata**: Reference that a Wikipedia article is about a person - containing 1, 176, 762 different persons
- **Raw Infobox Properties**: Values of the infobox from a Wikipedia article - containing 73, 686, 499 pieces of information for 3, 487, 062 different Wikipedia articles
- **Short Abstracts**: Short abstracts of Wikipedia articles - containing 4, 305, 028 different Wikipedia articles

The persondata dataset then is filtered so that it only contains Wikipedia articles which are also present in the Wikipedia Clickstream dataset. Also the infobox dataset and short abstract dataset are reduced so that these

[1] https://pypi.python.org/pypi/wikipedia

[2] http://wiki.dbpedia.org/Downloads2015-04

datasets solely contain Wikipedia Clickstream articles which are about persons resulting in information on $473,910$ different persons.

All three datasets are used to label or categorize Wikipedia articles for their *gender*, *birth year* and *birth country* if possible.

4.3.1. Gender Categorization

One main data preparation step was to categorize the gender of Wikipedia articles that are about people. Therefore a supervised learning approach was applied. A pipeline of CountTokenizer, TfidfTransformer and SVM was used to train this machine learning model. The input for the gender classification was the downloaded Wikipedia article set described in Subsection 4.2.2. Out of this set $2,000$ articles were manually labelled as female or male and used for training.

Every single step of the pipeline supports several parameters. GridSearchCV[3] is a method for trying every kind of possible combination of parameters and, as a result, providing the best one. The scoring method for choosing the best combination is the F_1 measure described in Subsection 3.5.3. After every training iteration (one combination of parameters) a 5-fold cross validation of the training set was performed. The following parameters of the pipeline steps are used:

- **CountVectorizer**:
  - *Stemming*: Use of Porter English Stemmer (Porter, 1997). Stemming is the process of putting words to their word stem. A stem is a part of the word that describes the root of a word. For example, the word stem of the words waiting and waited is wait and thus count as the same word.
    - Parameter Values: Yes, No
  - *Stop Words*: Use of English Stop Words (except *she*, *her*, *herself*, *he*, *his*, *hisself*. Stop words are common words like *and*, *it* or *the* and are filtered out.
    - Parameter Values: Yes, No

[3] http://scikit-learn.org/stable/modules/grid_search.html

- **TfidfTransformer**:
  - *IDF*: Enable IDF reweighting. Parameter if IDF should be used as a weighting scheme.
    - Parameter Values: Yes, No
  - *Smooth IDF*: Use smooth IDF weights by adding one to DF, as if an extra document was seen containing every term in the collection exactly once. Prevents zero divisions.
    - Parameter Values: Yes, No
- **SVM**:
  - *Kernel*: The used kernel type for calculating the hyperplane that is used as class separator. this parameter either can be a radial basis function like a Gaussian function, polynomial with different kinds of degree or a linear function.
    - Parameter Values: Radial basis functions, polynomial or linear
  - *Shrinking*: Enable shrinking heuristic. Shrinking is used to speed up the optimization process.
    - Parameter Values: Yes, No
  - *C*: Penalty parameter C of the error term (Cortes and Vapnik, 1995). This penalty parameter is used to prevent over-fitting.
    - Parameter Values: 0.1, 0.5, 1, 3.33, 10

A schematic tree of all used parameter combinations is shown in Figure 4.1. The best combination is marked with the green rectangle and has the following parameters:

- **Stemming:** no
- **Stop words:** yes
- **IDF:** no
- **Smooth IDF:** all
- **Shrinking:** all
- **C error term:** 3.33
- **Kernel:** linear

The parameter value *all* means that every parameter value leads to the same result. For example, it does not matter if the value for the parameter

Shrinking is set to *yes* or *no*, it always leads to the same result. The above used parameter combination has a F_1 score of 0.98 and a standard deviation of 0.020. As mentioned above, 2,000 manually labelled (872 female and 1,128 male) Wikipedia articles are used for training the SVM. For the remaining articles about persons the content of its page was used for the gender prediction if possible. If the wikipedia article for a page does not exist any more the short abstract of the DBpedia dataset was used. This could be due to the fact that the content of the Wikipedia articles was crawled in February 2016 and the persondata dataset - which indicates if a Wikipedia article is about a person - is from April 2015 and so it could be that an article was deleted. The result of the gender classification is: 87,413 females and 386,497 males.

4.3.2. Birth Country and Birth Year Categorization

The infobox properties dataset provides information such as birth year and birth country. If such information for a person is provided in this dataset it is parsed out of it. Finally it was possible to get the birth year for 428,852 persons and the birth country for 294,501 persons. People from countries which do non exist any more, such as the *German Democratic Republic*, are categorized to the successor country under international law. If this was not possible the birth country was not categorized. Figure 4.2 shows the result of the birth year categorization. About 87 % of the people were born in the 20th century and in this century there is a steady increase of born persons until the 9th decade. In Figure 4.3 the histogram of the persons' birth country is illustrated. About 46 % of the people are born in North America followed by Europe with about 32 %. Table 4.2 lists the top 25 birth countries with the number of persons and the ranking within the continent.

4.3.3. Categorization Summary

The whole categorization process is based on all three used datasets. Table 4.3 lists the summary of this operation. It was possible to infer the gender for 473,910, the birth country for 294,501 and the birth year for 428,852 out of

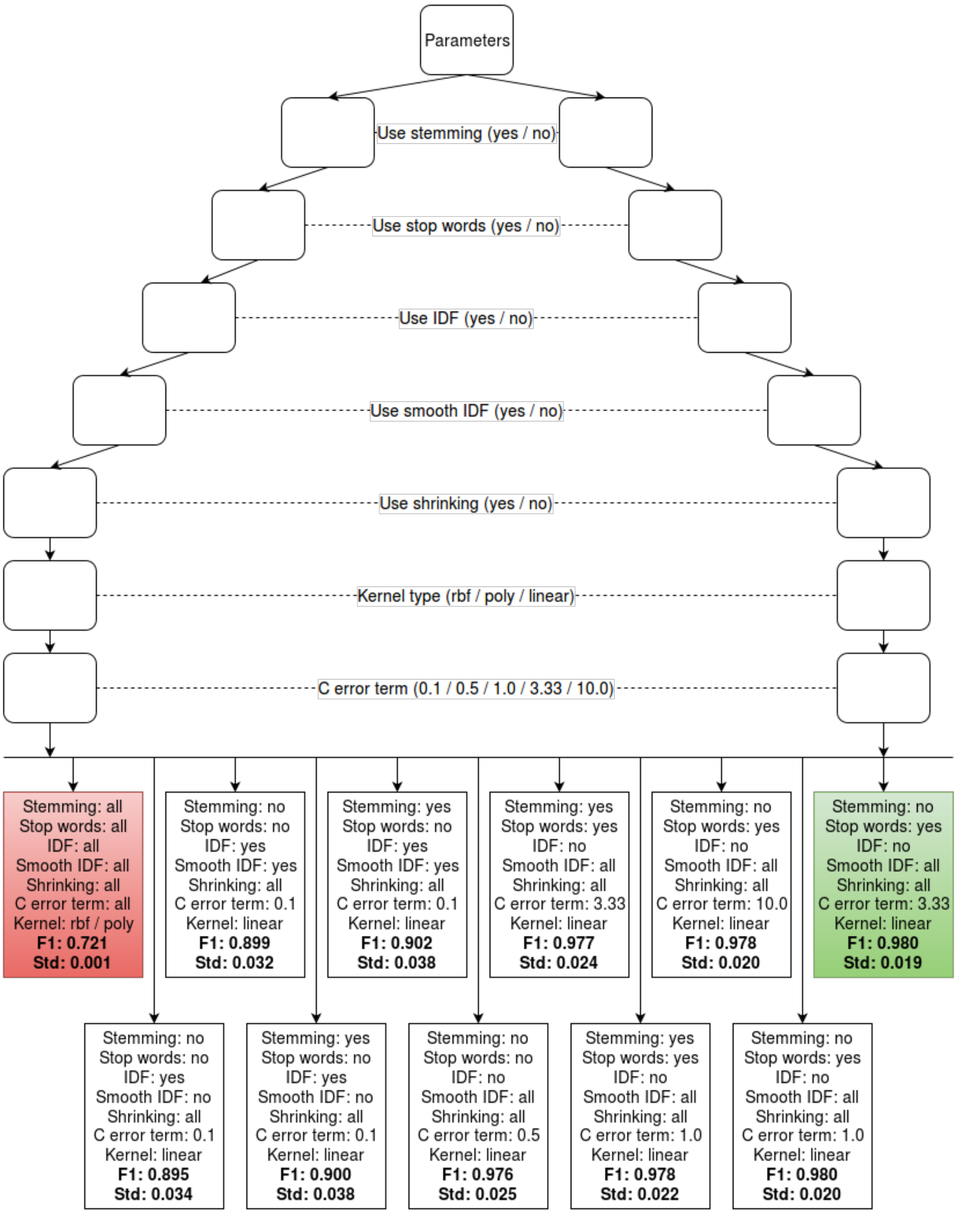
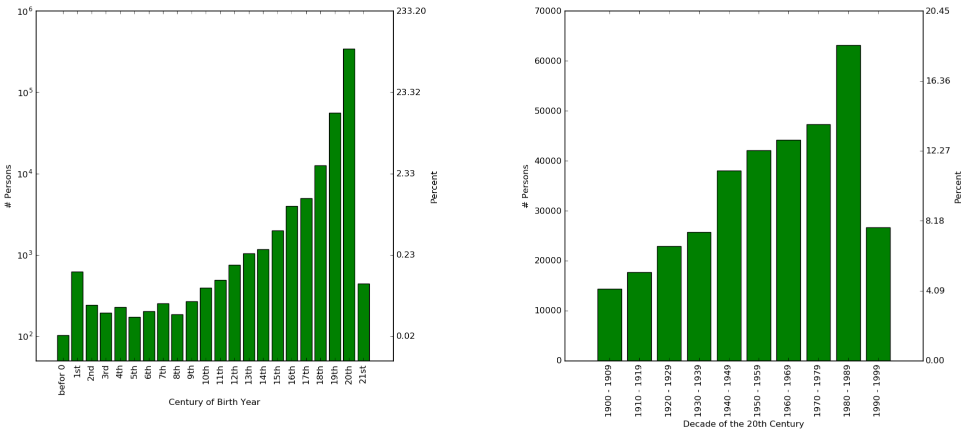

Parameters
Use stemming (yes / no)
Use stop words (yes / no)
Use IDF (yes / no)
Use smooth IDF (yes / no)
Use shrinking (yes / no)
Kernel type (rbf / poly / linear)
C error term (0.1 / 0.5 / 1.0 / 3.33 / 10.0)
Stemming: all Stop words: all IDF: all Smooth IDF: all Shrinking: all C error term: all Kernel: rbf / poly F1: 0.721 Std: 0.001
Stemming: no Stop words: no IDF: yes Smooth IDF: yes Shrinking: all C error term: 0.1 Kernel: linear F1: 0.899 Std: 0.032
Stemming: yes Stop words: no IDF: yes Smooth IDF: yes Shrinking: all C error term: 0.1 Kernel: linear F1: 0.902 Std: 0.038
Stemming: yes Stop words: yes IDF: no Smooth IDF: all Shrinking: all C error term: 3.33 Kernel: linear F1: 0.977 Std: 0.024
Stemming: no Stop words: yes IDF: no Smooth IDF: all Shrinking: all C error term: 10.0 Kernel: linear F1: 0.978 Std: 0.020
Stemming: no Stop words: yes IDF: no Smooth IDF: all Shrinking: all C error term: 3.33 Kernel: linear F1: 0.980 Std: 0.019
Stemming: no Stop words: no IDF: yes Smooth IDF: no Shrinking: all C error term: 0.1 Kernel: linear F1: 0.895 Std: 0.034
Stemming: yes Stop words: no IDF: yes Smooth IDF: no Shrinking: all C error term: 0.1 Kernel: linear F1: 0.900 Std: 0.038
Stemming: no Stop words: no IDF: no Smooth IDF: all Shrinking: all C error term: 0.5 Kernel: linear F1: 0.976 Std: 0.025
Stemming: yes Stop words: yes IDF: no Smooth IDF: all Shrinking: all C error term: 1.0 Kernel: linear F1: 0.978 Std: 0.022
Stemming: no Stop words: yes IDF: no Smooth IDF: all Shrinking: all C error term: 1.0 Kernel: linear F1: 0.980 Std: 0.020

Figure 4.1.: **A schematic tree with several parameter combinations for gender prediction:** Every level in the tree illustrates a new used parameter with its valid values. The last level indicates the used parameter combination with the resulting average F_1 score of the 5-fold cross validation and standard derivation. The left side of the last level shows the five worst combinations whereas the right side shows the six best combinations and the used combination for gender prediction is marked with a green rectangle. The combinations between are skiped due to space restriction.



(a)Persons per birth century

(b)Persons per birth decade of the 20th century

Figure 4.2.: **Birth year histogram per century and for decades of the 20th century:** 4.2a shows the amount of born persons per century. Most people (87 %) were born in the 19th century. This century was also divided into decades and the results are shown in 4.2b. The amount of born persons steadily increases until the 9th decade.

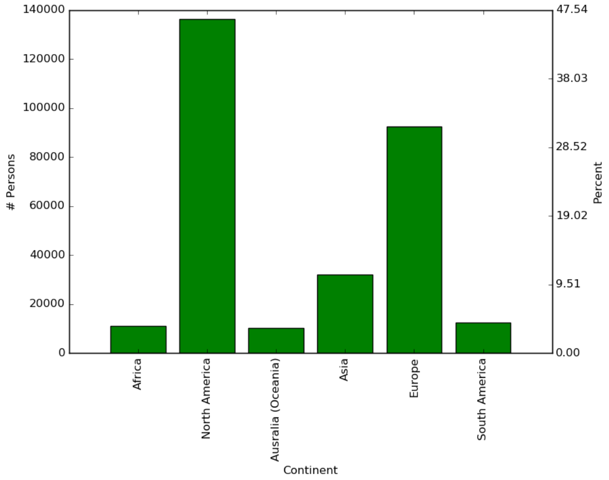


Figure 4.3.: **Birth country histogram per continent:** This plot shows the amount of born people per continent. About 46 % of all people were born in North America followed by Europe with around 25 %.

Country	Amount Persons	Ranking within Continent
United States	116,898	1
United Kingdom	34,442	1
Canada	10,372	2
Germany	7,674	2
France	7,596	3
India	7,434	1
Australia	7,261	1
Italy	5,160	4
Japan	5,121	2
Argentina	4,219	1
Spain	3,857	5
Brazil	3,179	2
Netherlands	2,926	6
Russia	2,828	7
Indonesia	2,730	3
Ireland	2,497	8
Mexico	2,327	3
South Korea	2,113	4
Philippines	2,015	5
New Zealand	1,870	2
Austria	1,794	9
Jamaica	1,648	4
South Africa	1,622	1
Sweden	1,530	10
Liberia	1,517	2

Table 4.2.: **Top 25 birth countries:** This table shows the top 25 birth countries. Each with the amount of born people as well as the ranking of the country within its continent. For example, in *Australia* 7,261 people were born and thus is ranked 7th. Limited to the continent where *Australia* belongs, it is ranked first.

Description	Amount
Articles	2, 140, 423
Articles about persons with gender	473, 910
Articles about persons with birth year	428, 852
Articles about persons with birth country	294, 501
Articles about persons with birth year and gender	412, 513
Articles about persons with birth country and gender	280, 812
Articles about persons with birth year and birth country	281, 048
Articles about persons with birth year, birth country and gender	267, 460

Table 4.3.: **Summary of the categorization of Wikipedia articles:** This table illustrates the summary of the categorization of Wikipedia articles. Altogether this dataset consists of 2, 140, 423 articles It was possible for 473, 910 articles to predict the gender and 428, 852 articles to parse the year of birth and for 294, 501 articles to parse the country of birth.

2, 140, 423 articles. Furthermore it was possible to match 267, 460 Wikipedia articles to all three categories.

5. Experimental Setup

In this chapter the application of the methods and materials used in this thesis are described. Section 5.1 provides the details for calculating the energy values of the Wikipedia network influenced by the Wikipedia Clickstream data. Section 5.2 describes the methods for analysing different stationary distributions.

5.1. Stationary Distribution of Wikipedia Clickstream Network

For analysing the Wikipedia Clickstream dataset the stationary distribution is used. Therefore a Clickstream matrix C is created and this matrix is a square matrix with dimension $2,140,423$, which is the amount of used Wikipedia articles. For every different referer-resource pair the amount of clicked links is inserted into $\boldsymbol{C}$ so that in the end this matrix has $12,194,530$ non zero entries.

In order to calculate the energy it is necessary to have a strongly connected graph or matrix. Matrix $\boldsymbol{C}$ does not fulfill this requirement and therefore this matrix is added to the adjacency matrix $\boldsymbol{A}$ of the Wikipedia network structure. Since $\boldsymbol{A}$ is a SCC, the addition of another matrix with the same dimension also leads to a SCC. With this newly created matrix it is possible to create its stationary distribution. However, the use of the Wikipedia network structure - which is mandatory - creates some *noise* in the data, since these links influence the resulting energy. Hence a β-factor is introduced which increases the influence of the Wikipedia Clickstream data or decreases the influence of the Wikipedia network structure and so reduces the *noise* of it. In this study this factor is between (excluding) zero and (including) one and is multiplied with $\boldsymbol{A}$. For example a β-factor of 0.1 means that the influence of the Wikipedia Clickstream data is increased ten times.

Additionally to the above mentioned calculation, another variation is used where the values of the matrix are logarithmized. The reason for this is to smooth the weights of the Wikipedia Clickstream matrix. If a referer-resource pair was clicked, for example, 100 times and another pair 50 times, it does not automatically mean that the amount of clicks for the first pair is worth twice as much as the second pair. This process is called sublinear scaling.

In conclusion, three different types of stationary distributions are calculated:

- π_{WNS}: Stationary distribution of the Wikipedia network structure
- π_i: Stationary distribution of the Wikipedia network structure and Wikipedia Clickstream data with different β-factors
- π_{i_log}: Stationary distribution of the Wikipedia network structure and Wikipedia Clickstream data with different β-factors and logarithmic weights

These stationary distributions are calculated with the previous described equations 3.19, 3.20, 3.21, 3.22 and 3.23, where for the weighted adjacency matrix $\boldsymbol{W}$ the following values are substituted:

- **For** π_{WNS}: $\boldsymbol{W} = \boldsymbol{A}$
- **For** π_i: $\boldsymbol{W} = \beta \cdot \boldsymbol{A} + \boldsymbol{C}$
- **For** π_{i_log}: $\boldsymbol{W} = log_{10}(1 + (\beta \cdot \boldsymbol{A} + \boldsymbol{C})$,

where i is the β-factor between zero and one.

Finally π_i and π_{i_log} are compared with π_{WNS} on the basis of the two correlation measures Pearson measure ρ and Spearman measure r_s, which leads to the following four different correlation values:

$$\rho_i = \rho(\pi_{WNS}, \pi_i) \tag{5.1}$$

$$\rho_{i\_log} = \rho(\pi_{WNS}, \pi_{i\_log}) \tag{5.2}$$

$$r_{s\_i} = r_s(\pi_{WNS}, \pi_i) \tag{5.3}$$

$$r_{s\_i\_log} = r_s(\pi_{WNS}, \pi_{i\_log}) \tag{5.4}$$

The Pearson correlation value describes the linear relationship between two variables - in this study two stationary distributions. The value of this correlation can be between -1 and 1, where -1 means a total negative linear correlation, 1 a total positive linear correlation and 0 no linear correlation. Since the stationary distributions can have other correlations than linear, the Spearman correlation is also calculated. The Spearman correlation ranks the values of the stationary distribution and calculates the Pearson correlation of the ranked values.

In this study the results of the analyses of the sub-linear scaled stationary distributions and non sub-linear scaled stationary distributions showed similar results. Because of this and space limitations within this study, only the results of the non sub-linear scaled stationary distributions were used.

5.2. Stationary Distribution Analysis

This section concentrates on describing the methods of analysing the previously illustrated stationary distributions with different β-factors. Every stationary distribution contains 2.140.423 values, each defining the probability that a random surfer visits the corresponding Wikipedia article in the limit of infinitely many steps. For further analysis which solely focuses on Wikipedia articles about persons, only the energy of the predicted $87,413$ females and $386,497$ males are used.

5.2.1. Gender Tendency

π_{WNS} contains how likely it is for a random surfer to visit a Wikipedia article solely based on the Wikipedia network structure. For this energy values, a frequency distribution was plotted as a histogram-like plot and as a boxplot. This energy is the basis for a comparison to other distributions. Additionally, the corresponding frequency distributions are calculated for π_i and compared to the basis in order to evaluate if there is a gender tendency. Provided there is a gender tendency, it is analyzed how this changes with increasing influence of the Wikipedia Clickstream dataset.

5.2.2. Gender Tendency grouped by Birth Year

Another analysing step is to group the Wikipedia articles about persons by birth year. In this experiment all persons who were born between 1900 and 1999 are included. Altogether ten groups are defined. These groups correspond to the decades of the 20th century. The energy values then are compared as described in Subsection 5.2.1.

5.2.3. Gender Tendency grouped by Birth Country

A second grouping is defined by the country of birth. Every Wikipedia article about a person who had a categorized birth country is included in this experiment. The energy values then are compared as described in Subsection 5.2.1.

5.2.4. Gender Factor

This experiment focuses on gender factors for Wikipedia articles about persons. The gender factor describes the relative increase or decrease of the energy values of a Wikipedia article about a gender from $\pi_{0.01}$ and π_{WNS} and is calculated as follows:

$$gender\,factor = \frac{\pi_{0.01}}{\pi_{WNS}} \tag{5.5}$$

The frequency distribution of the different gender factors then is analysed to obtain information who profits most from the user clicks defined by the Wikipedia Clickstream dataset.

5.2.5. Top Genders

The last experiment analyses 100 persons - the top 50 females and the top 50 males according to their energy - of each stationary distribution. The corresponding energy value of these persons then are compared to see if there is a preference towards females or males. Also the change factors of the top 50 females and top 50 males are compared to evaluate which gender in general benefits most from the user click dataset.

6. Results & Discussion

This chapter describes the results of this thesis. Furthermore the impact and meaning of the findings will be discussed. Section 6.1 illustrates the impact of the β-factor on the energy values of the different stationary distributions. In Section 6.2 the focus lies on describing the gender tendency and trend. Finally Section 6.3 concludes the findings.

6.1. Stationary Distribution of Wikipedia Clickstream Network

Since the weighted matrix of the Wikipedia Clickstream dataset could not be used to calculate a stationary distribution, it was necessary to add this matrix to the adjacency matrix of the Wikipedia network structure. To increase the influence of the user clicks which were made during navigating Wikipedia, a β-factor was multiplied with the adjacency matrix and the resulting energy values are compared to the energy values of the Wikipedia network structure which is not influenced by user clicks. With a steady decrease of the β-factor, the correlation values between the network structure and the user influenced structure are declining. For instance, the Pearson correlation value of π_{WNS} and $\pi_{1.0}$ is 0.82 and the Spearman correlation value is 0.76. In comparison, the Pearson value of π_{WNS} and $\pi_{0.01}$ is 0.18 and the Spearman correlation value is 0.55. All results are shown in Figure 6.2.

Until β-factor 0.1 the Pearson correlation values are higher than the Spearman correlation values. On the contrary, between β-factor 0.1 and 0.05 the opposite is true. That is the case because with a β-factor lower than 0.1, the resulting stationary distributions are more skewed. This skewness leads to a lower Pearson correlation value because of a non-linear correlation between

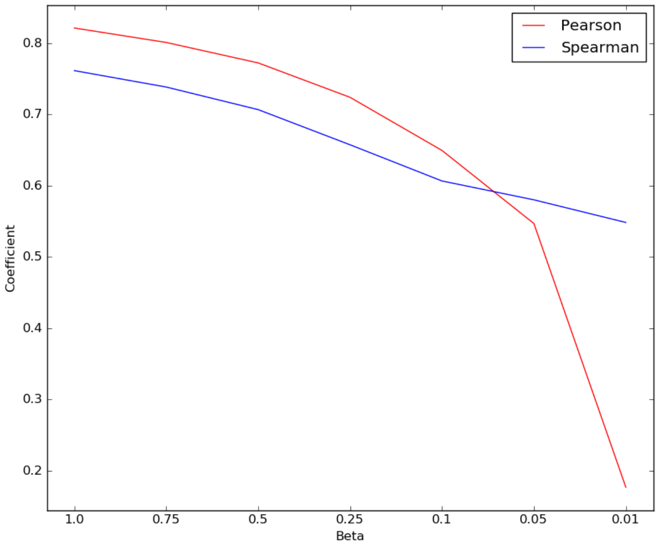


Figure 6.1.: **Stationary distribution β-factor correlation:** This plot shows the Pearson (red line) and Spearman (blue line) correlation between the stationary distribution of the Wikipedia network structure and the stationary distributions with different β-factors. The results show that with decreasing β-factor both correlation values are decreasing too. This indicates that the stationary distribution with lower β-factor increases the influence of the Wikipedia Clickstream data.

the stationary distributions and thus the Pearson correlation value is lower than the Spearman correlation value.

When the weights for the user click influenced stationary distributions are sub-linear scaled, the correlation values are dropping with decreasing β-factor. For example, the Pearson correlation value of π_{WNS} and $\pi_{1.0\_log}$ is 0.98 and the Spearman correlation value is 0.97. In comparison, the Pearson value of π_{WNS} and $\pi_{0.01\_log}$ is 0.80 and the Spearman correlation value is 0.71. All results for sub-linear scaled weights are illustrated in Figure 6.2.

The decreasing correlation values indicate that the use of the β-factor indeed increases the influence of the user clicks defined by the Wikipedia Clickstream dataset. Hence the use of energy values calculated with a lower β-factor for the remaining experiments focuses more on user clicks than on data noise - which was introduced through the usage of the Wikipedia network structure.

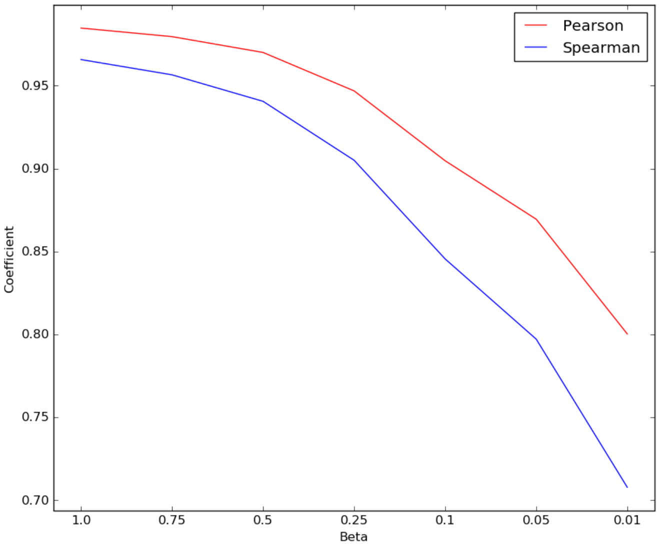


Figure 6.2.: **Stationary distribution β-factor correlation with sub-linear scaled weights:** This plot shows the Pearson (red line) and Spearman (blue line) correlation between the stationary distribution of the Wikipedia network structure and the stationary distributions with different β-factors and sub-linear scaled weights. The results show that with decreasing β-factor both correlation values are decreasing too. This indicates that the stationary distribution with lower β-factor increases the influence of the Wikipedia Clickstream data.

	Quartile 1	Quartile 2 or Median	Quartile 3
Females	$1.57 \cdot e^{-06}$	$4.51 \cdot e^{-06}$	$1.44 \cdot e^{-05}$
Males	$2.26 \cdot e^{-06}$	$6.96 \cdot e^{-06}$	$2.32 \cdot e^{-05}$

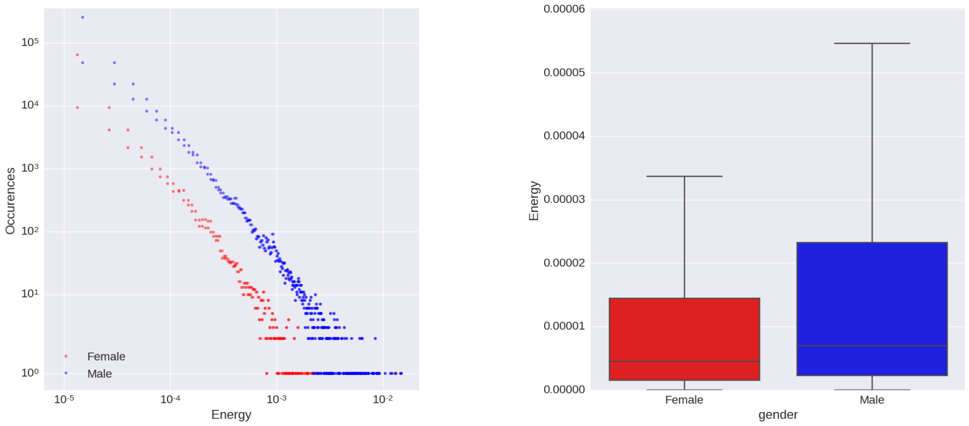
Table 6.1.: **Quartiles for genders of the energy values of the base line:** This table shows the quartiles of the base line energy values. For example, the median for female articles is $4.51 \cdot e^{-06}$ whereas the median for male articles is $6.96 \cdot e^{-06}$. These results show that the Wikipedia network structure on average prefers articles about men.

6.2. Stationary Distribution Analysis

This section focuses on the results of the stationary distribution concerning only Wikipedia articles about people. The stationary distribution of the Wikipedia network structure π_{WNS} defines the base line for further comparisons. The quartile values of the female and male energy values for the base line can be seen in Table 6.1. For example, the median for female persons is $4.51 \cdot e^{-06}$ and for male persons $6.96 \cdot e^{-06}$. The results are also illustrated in Figure 6.3.

The stationary distribution of the base line clearly shows that the Wikipedia network structure prefers Wikipedia articles that are about male persons. On average, males have significantly higher energy values and for the random surfer it is more likely to visit a male Wikipedia article than a female one.

The distribution of the energy values follows a power law. There are many articles about persons, regardless of the gender, with lower energy values and there are articles with a higher energy value that just occur once. This power law is also valid for the whole Wikipedia network structure itself and the stationary distribution of it also states that this is the case for the used subset (articles about persons) of the network. In Figure 6.3b the boxplot also shows the power law of the distribution. Both boxes, red for females and blue for males, have the median in the lower half of the boxes and the start of the boxes are also located near zero. These facts indicate that most of the energy values have a low value and only a few have higher energy values, thus resulting in a power law.



(a)Persons per birth century

(b)Distribution of the energy values per gender

Figure 6.3.: 6.3a shows the distribution of the base line stationary distribution per gender. The results show that male articles (blue dots) have significantly higher energy values than articles about women.
6.3b is the boxplot of the base line stationary distribution per gender. A boxplot is a diagram to visualize a series of numerical data. The plot consists of a number line, a box, whiskers or antennas and outliers. The first quartile (Q_1) of the represented data marks the beginning of the box, here $1.57 \cdot e^{-06}$ for female, and the end of the box is the third quartile (Q_3), here $1.44 \cdot e^{-05}$ for female. Between (Q_1) and (Q_3) lie fifty percent of the data and this area is called interquartile range (IQR). Within the box there is a vertical line which defines the second quartile (Q_2) or median, here $4.51 \cdot e^{-06}$ for female. Both ends of the box are extended with whiskers and a common definition for the length of the antennas is $1,5 \cdot IQR$ but the maximal length of the whiskers is bound to a value of the data series which is between $Q_1 - IQR$ and $Q_3 + IQR$. Finally outliers are marked as dots or rectangles to indicate data points which are beyond the whiskers. In this thesis outliers are not plotted.
This boxplot shows that on average male articles have significantly higher energy values that female articles.

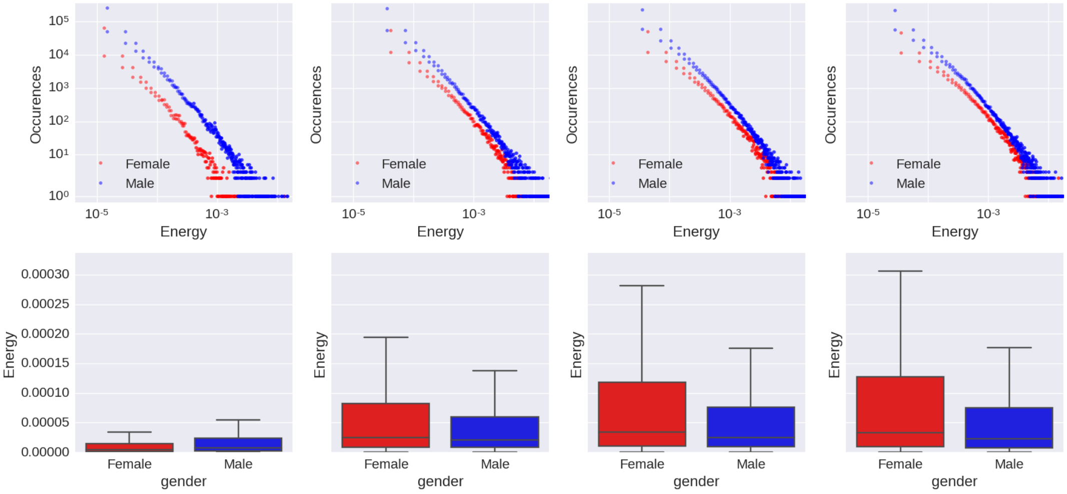
	Quartile 1	**Quartile 2** or **Median**	**Quartile 3**
Females (π_{WNS})	$1.57 \cdot e^{-06}$	$4.51 \cdot e^{-06}$	$1.44 \cdot e^{-05}$
Males (π_{WNS})	$2.26 \cdot e^{-06}$	$6.96 \cdot e^{-06}$	$2.32 \cdot e^{-05}$
Females ($\pi_{1.0}$)	$8.03 \cdot e^{-06}$	$2.50 \cdot e^{-05}$	$8.27 \cdot e^{-05}$
Males ($\pi_{1.0}$)	$7.80 \cdot e^{-06}$	$2.10 \cdot e^{-05}$	$5.97 \cdot e^{-05}$
Females ($\pi_{0.25}$)	$1.00 \cdot e^{-05}$	$3.36 \cdot e^{-05}$	$1.17 \cdot e^{-04}$
Males ($\pi_{0.25}$)	$9.02 \cdot e^{-06}$	$2.51 \cdot e^{-05}$	$7.57 \cdot e^{-05}$
Females ($\pi_{0.05}$)	$8.96 \cdot e^{-06}$	$3.31 \cdot e^{-05}$	$1.28 \cdot e^{-04}$
Males ($\pi_{0.05}$)	$7.73 \cdot e^{-06}$	$2.28 \cdot e^{-05}$	$7.52 \cdot e^{-05}$

Table 6.2.: **Quartiles for genders of the following stationary distributions: π_{WNS}, $\pi_{1.0}$, $\pi_{0.25}$ and $\pi_{0.05}$:** This table illustrates the quartiles for the calculated energy values. These values show that with increasing influence of the Wikipedia Clickstream data (decreasing β-factor) the gender tendency shifts from male favoured articles (π_{WNS}) towards female preferred articles ($\pi_{1.0}$, $\pi_{0.25}$ and $\pi_{0.05}$). For example, the median for females is less for the Wikipedia network structure than for men whereas the median is higher for females than for males for all stationary distributions calculated with different β-factors.

6.2.1. Gender Tendency

Compared to the base line, all quartiles of the stationary distributions have a higher value. Table 6.2 shows the results for π_{WNS}, $\pi_{1.0}$, $\pi_{0.25}$ and $\pi_{0.05}$. For instance, the female median of $\pi_{1.0}$ is $2.50 \cdot e^{-05}$ whereas the corresponding value of the base line is $4.51 \cdot e^{-06}$.

Comparing the median values between genders of the base line and the energy values calculated with β-factors clearly shows that there is a shift from male preferred Wikipedia articles of the Wikipedia network structure to a more likely probability to visit a female Wikipedia article. The boxplots of Figure 6.4 also illustrate this finding. Moreover, the stationary distribution plots of Figure 6.4 show that the gap between the genders of the stationary distribution base line is getting closer compared to the energy values derived with different β-factor. These results show that although the Wikipedia network structure on average prefers male articles, the users who navigated



(a)WNS (b)$\beta = 1.0$ (c)$\beta = 0.25$ (d)$\beta = 0.05$

Figure 6.4.: **Energy values of genders for base line and stationary distributions derived with different β-factors:** The first row illustrates the distribution of the energy values and the second row shows the boxplot of the stationary distribution. Each column represents a different stationary distribution (from left to right): π_{WNS}, $\pi_{1.0}$, $\pi_{0.25}$ and $\pi_{0.05}$.
The left column illustrates the base line and it can be seen that male articles have higher energy values than female articles. The other columns illustrate the stationary distributions calculated with different β-factors.
All these energy values show that there is a shift towards female favoured articles. Also the plots in the first row show that the gaps between the genders are getting closer from the base line compared to the stationary distributions calculated with different β-factors. Also the top position changes from a male article (blue dot) to a female article (red dot).

Wikipedia articles are more likely to visit a female Wikipedia article.

The power law which was observed for the stationary distribution of the Wikipedia network structure, is also valid for $\pi_{1.0}$, $\pi_{0.25}$ and $\pi_{0.05}$. All three stationary distributions calculated with different β-factors show that they have many Wikipedia articles about persons with low energy values and just a few articles about females and males with high energy values. The same observations are illustrated in the boxplots of figure 6.4. The medians are placed in the lower half of the boxes and Q_1 of the boxes is located near zero.

Another finding is that, in general, the energy values of articles about persons are increasing. For example, the median for females is $4.51 \cdot e^{-06}$ and for males $6.96 \cdot e^{-06}$ for the base line whereas the energy values derived with β-factor 0.05 are $3.36 \cdot e^{-05}$ for females and $2.51 \cdot e^{-05}$ for males. This indicates that it is more likely for the random surfer to visit an article about a person with increasing influence of the Wikipedia Clickstream data than other Wikipedia articles.

6.2.2. Gender Tendency grouped by Birth Year

The following results are influenced by grouping the genders according to their birth decade limited to the 20th century. Here, the base line on average prefers male articles too. All values for the three quartiles of the Wikipedia network structure stationary distribution for males in every decade are higher than for females. The stationary distribution derived with β-factor 1.0 shows that in some cases the quartile values are higher for females than for males. With increasing influence of the Wikipedia Clickstream data, there is also an increase in the number of female quartile values exceeding their corresponding male quartile values. For example, $\pi_{0.05}$ shows that for every decade every female quartile except one is higher than the male ones. Only from the last decade the female Q1 value $5.67 \cdot e^{-06}$ is lower than the male value $7.17 \cdot e^{-06}$.

Comparing the decades of the base line, females who were born in the 5th decade of the 20th century have the highest quartile values, whereas the highest quartile values for male articles are from men who were born in the

1930s. This shows that the Wikipedia network structure on average prefers articles about persons who were born in these two decades. For female articles this fact is also valid for the stationary distributions calculated with different β-factors. All three distributions have the highest median values for female articles for the 5th decade. In contrast, articles about males do not have the highest median values for the 4th decade for $\pi_{1.00}$, $\pi_{0.25}$ and $\pi_{0.05}$, as it is true for the base line. They also show the highest median values for the 1940s.

The results are listed in Table A.1 and illustrated in Figure 6.5.

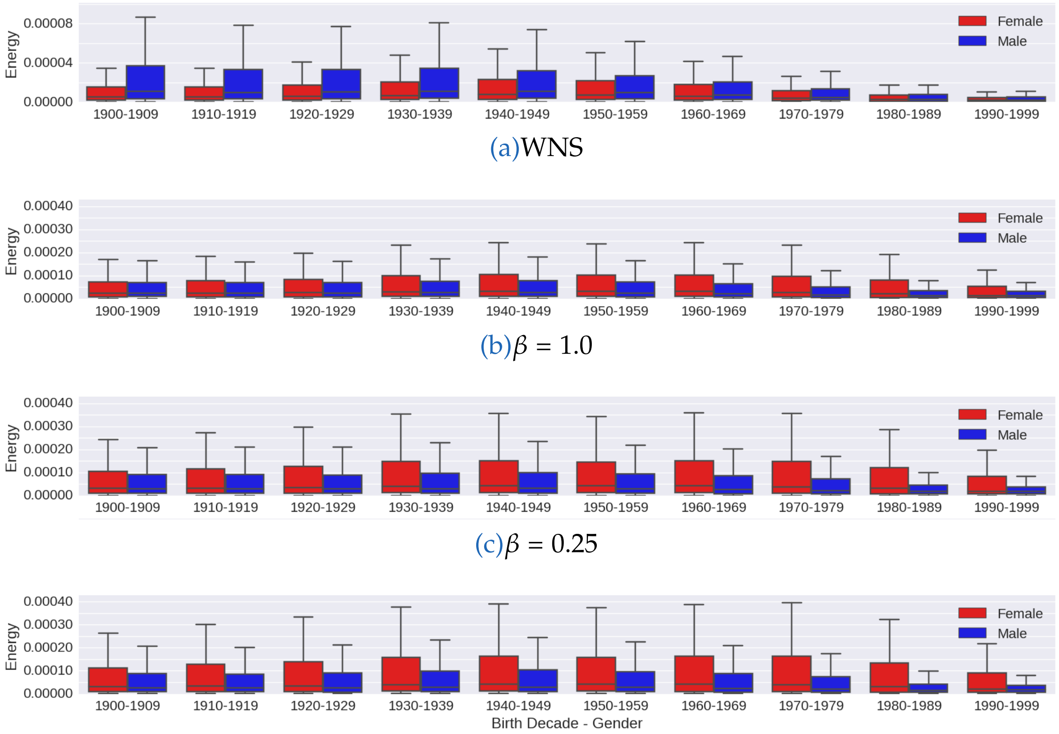
Investigating $\pi_{0.05}$, the 5th decade of the 20th century has the highest median values for females ($4.19 \cdot e^{-05}$) and males ($2.94 \cdot e^{-05}$). The following are examples of notable people born in this decade, who have the top energy values of $\pi_{0.05}$:

- **Jane Wild Hawking**: female, born in 1944
- **Margrethe II of Denmark**: female, born in 1940
- **Camilla, Duchess of Cornwall**: female, born in 1947
- **Charles, Prince of Wales**: male, born in 1948
- **George W. Bush**: male, born in 1946
- **Stephen Hawking**: male, born in 1942

This experiment also states that, on average, the base line prefers male articles to female ones if a grouping by birth decade is done. Compared to the base line, the stationary distributions calculated with different β-factors show a shift towards female articles. With increasing impact of the Wikipedia Clickstream data it is more likely to visit a Wikipedia article which is about a female than a male Wikipedia article.

6.2.3. Gender Tendency grouped by Birth Country

The results of this subsection are influenced by grouping the genders according to their birth countries. Here the base line, on average, prefers male articles too. All values for the three quartiles of the Wikipedia network structure stationary distribution for males for every birth country are higher than for females except for Austria. In π_{WNS} the median for female pages



(a)WNS

(b)β = 1.0

(c)β = 0.25

(d)β = 0.05

Figure 6.5.: **Energy values of genders for base line and different β-factors grouped by birth decade in the 20^{th} century:** Each row illustrates the boxplots of the stationary distributions grouped by birth decade calculated with different β-factors. (from top to bottom): π_{WNS}, $\pi_{1.0}$, $\pi_{0.25}$ and $\pi_{0.05}$.
The first row shows that on average male articles are preferred regardless of the decade of birth. The stationary distributions derived with different β-factors illustrates that regardless of the decade there is a shift of the gender tendency towards female articles. For better visibility the y-axis limits in 6.5a differ from the other three subplots.

is $1.08 \cdot e^{-05}$ whereas the lower male value is $8.55 \cdot e^{-06}$. The stationary distribution calculated with β-factor 1.0 shows that in most cases the quartile values are higher for females than for males. With increasing influence of the Wikipedia Clickstream data, almost all female quartile values are higher than their corresponding male quartile values. For example, $\pi_{0.05}$ shows that for every birth country every female quartile except for New Zealand is higher than the male ones. The female median for New Zealand is $2.09 \cdot e^{-05}$ and this is lower than the male value $2.11 \cdot e^{-05}$. In contrast, the female Q3 value $8.48 \cdot e^{-05}$ is higher than the corresponding male one $5.75 \cdot e^{-05}$.

India has the highest median value ($\pi_{0.05}$) for females of all countries ($8.46 \cdot e^{-05}$) and Germany for males ($3.40 \cdot e^{-05}$). The following are examples of notable people born in these countries, who have the top energy values of the corresponding countries of birth:

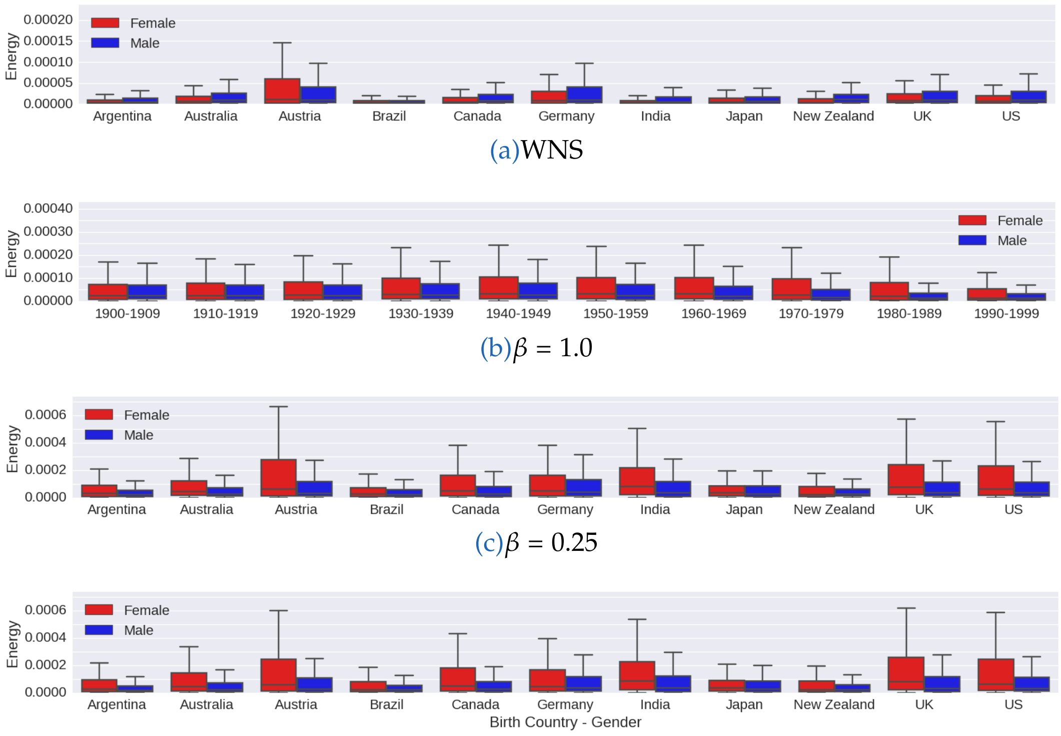
- **Indira Gandhi**: female, born in India
- **Joanna Lumley**: female, born in India
- **Padma Lakshmi**: female, born in India
- **Albert Einstein**: male, born in Germany
- **Martin Luther**: male, born in Germany
- **Pope Benedict XVI**: male, born in Germany

The results are listed in Table A.2 and illustrated in Figure 6.6.

This experiment highlights that also a grouping per birth country on average prefers male articles with the base line except Austria and that there is a shift towards females with decreasing β-factor. The higher the influence of the Wikipedia Clickstream dataset is, the more likely it is for the random surfer to visit a female Wikipedia article.

6.2.4. Gender Factor

The gender factor describes the gain or loss of energy of Wikipedia articles between π_{WNS} and $\pi_{0.05}$. On average, female values have a higher factor. The median for male articles is 3.14 whereas female Wikipedia articles have a median of 6.91. The quartile values are listed in Table 6.3 and the factor distribution and the corresponding boxplot is plotted in Figure 6.7.



(a)WNS

(b)$\beta = 1.0$

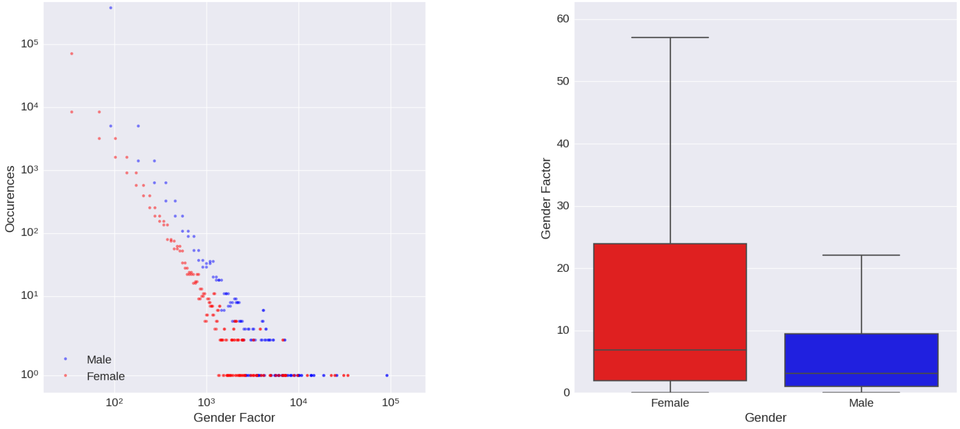
(c)$\beta = 0.25$

(d)$\beta = 0.05$

Figure 6.6.: **Energy values of genders for base line and different β-factors grouped by birth country:** Each row illustrates the boxplots of the stationary distributions grouped by birth country for different β-factors. (from top to bottom): π_{WNS}, $\pi_{1.0}$, $\pi_{0.25}$ and $\pi_{0.05}$. The list of countries includes the top two (amount of born people) countries per continent plus Austria.
The first row shows that on average male articles are preferred regardless of the country of birth except Austria. Here the base line on average favours female articles. The stationary distributions calculated with different β-factors illustrate that regardless of the country of birth there is a shift of the gender tendency towards female articles. For better visibility the y-axis limits in 6.6a differ from the other three subplots.

	Quartile 1	**Quartile 2 or Median**	**Quartile 3**
Females	1.91	6.91	23.97
Males	1.04	3.14	9.47

Table 6.3.: **Quartiles for genders of the factor distribution of $\frac{\pi_{0.05}}{\pi_{WNS}}$:** This table shows the quartiles for the calculated gender factors. The results illustrate that on average female articles profit more from the Wikipedia Clickstream data than male articles. For example, the median for female articles is 6.91 and for male articles 3.14.



(a)Gender factor distribution

(b)Boxplot of the gender factors

Figure 6.7.: **Factor distribution of $\frac{\pi_{0.05}}{\pi_{WNS}}$**: 6.7a illustrates the distribution of the factor values. This distribution shows that male articles have the highest factors but also the lowest factors.
6.7b shows the boxplot of the factor distribution. Here it can be seen that on average female articles have a significantly higher change factor than male articles.

On average, female articles have a significantly higher factor than males have. This indicates that Wikipedia articles that are about females profit more from the user click data than articles about males do. Similar to the previous results, the increasing influence of the Wikipedia Clickstream data leads to a shift from a male preferred Wikipedia network structure to a female favoured user navigation.

The distribution of the gender factors also follows a power law. Many of the articles have a lower factor and only a few have a higher factor. This skewness of the distribution can also be seen in the boxplot. The median is located at the lower half of the box and the boxes also start near zero.

6.2.5. Top Genders

Figures 6.8, 6.9 and 6.10 illustrate females and males of the top 50 energy values of each gender of the base line and the stationary distributions calculated with β-factors 1.0 and 0.05. One can see that the energy values of females are lower than the corresponding male values for the base line. Also the female energy values decrease faster than male ones.

In contrast, the user click influenced stationary distribution shows that the energy value differences between the top people are decreasing. In addition, the highest energy value changes from a male article to a female article compared to the base line. In addition to that female values of the stationary distributions calculated with different β-factors do not decrease as fast as female energy values of the base line.

The top genders of the stationary distribution of the base line illustrate the centrality of Wikipedia articles about people. As seen in Figure 6.8, all top 50 male Wikipedia articles have a higher energy value than the top 50 Wikipedia articles about women. Within the top 50 men there are *leaders* like actual and former American presidents, dictators or prime ministers, *inventors*, *philosophers* or *religious people* whereas in the top 50 females there are mainly *royals*, *politicians* or *scientists*. What these people, females and males, have in common is, that they are almost all well known persons and that publicity of a person leads to a high energy value within the Wikipedia network structure.

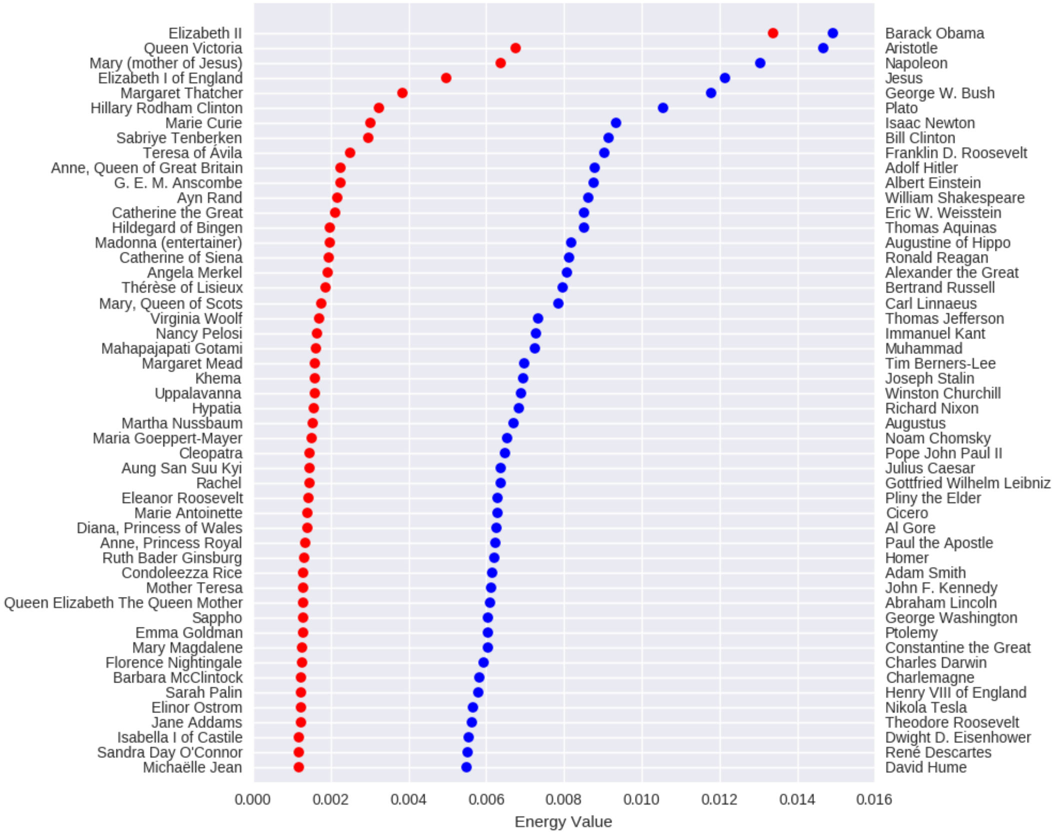


Figure 6.8.: **Top** 50 **females and males of** π_{WNS}: This figure shows that male articles have a higher energy value and thus are preferred in the Wikipedia network structure. Furthermore the energy values for female articles decrease faster than for male articles.

As described before, there are some clusters of persons with a certain occupation. For example, there are twelve presidents of the United States in this list. Additionally, nine other politicians like Angela Merkel, Winston Churchill or Sarah Palin, are also among the top 100 persons of the Wikipedia network structure. Religious people are also prevalent within the top people. Three out of the five major religious groups (Buddhism, Christianity, Islam) are represented by different persons. For example, Jesus, Mary Magdalene, Catherine of Siena (all Christianity), Muhammad (Islam), Mahapajapati Gotami and Khema (all Buddhism) are all people from different religions. However, Christianity has the most representatives within the top 100 persons of the Wikipedia network structure. Another big cluster consists of royal persons. Mostly represented by women (Elisabeth II, Diana, Princess of Wales, Anne, Queen of Great Britain), this person group also includes two male persons (Henry VIII of England and Charlemagne).

The top 100 persons of the base line also show that most of the people were born in North America or Great Britain. As the used Wikipedia network structure is based on the English language edition, this illustrates that the content of the information network (Wikipedia articles about notable persons who were born in an English speaking country) correlate with the language of users who navigated the information system (users of the English Wikipedia).

The top 50 females and males stationary distribution derived with β-factor 1.0 shows that with the influence of the Wikipedia clickstream dataset, the top position changes from a Wikipedia article about a male person (Barrack Obama) to a female Wikipedia article (Elizabeth II). Additionally, this top gender list also illustrates that certain types of occupation clusters exist. For female Wikipedia articles these clusters are *royal persons* and *politicians*, but also *artists* like singers and actresses are present. For male Wikipedia articles the clusters about persons' occupation is similar to the base line clusters: *Leaders*, like dictators and political leaders, *religious people* and *scientist*. This top gender list contains many royal people, especially the top 50 male Wikipedia articles include more of them compared to the base line. Nearly half of the Wikipedia articles about female persons (24 of 50) are about royal persons whereas the base line top 50 female Wikipedia articles include 20 % (10 of 50) royal persons. Also the male top 50 list has a lot of Wikipedia

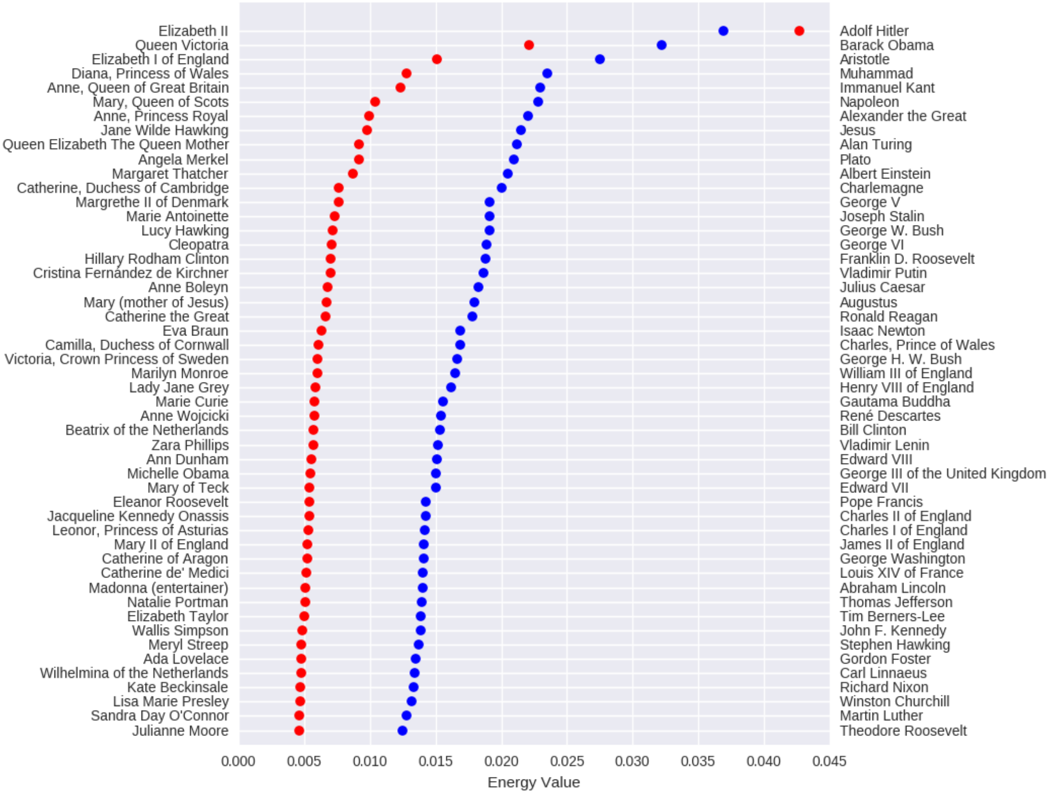


Figure 6.9.: **Top** 50 **females and males of** $\pi_{1.0}$: This figure shows that with β-factor 1.0 the top energy value changes from a male article to a female article (compared to the base line). Furthermore the distances between the genders also decrease and the energy values for male articles also decrease faster than the energy values for male articles of the base line.

articles about royal persons. 26 % (13 of 50) compared to just 4 % (2 of 50) of Wikipedia articles that are about royal males in the base line.

The second big cluster which both top 50 genders have in common, is the *leader* cluster, like political leaders, dictators and empresses or emperors (royal persons like Elisabeth II, who is also a figurative head of state are excluded from this cluster). Six out of 50 female persons are leaders and 21 out of 50 male persons are leaders. For instance, Margaret Thatcher, Cleopatra, Hillary Rodham Clinton, Napoleon, Vladimir Lenin and Winston Churchill. Altogether, 27 % of the top 100 persons are in the leader cluster of the stationary distribution calculated with β-factor 1.0 compared to 21 persons of the base line.

The amount of females if the religious cluster changed from eleven female Wikipedia articles from the base line to only one Wikipedia article about female persons from $\pi_{1.00}$. The sole female left is Mary, mother of Jesus, whereas the base line for instance also includes females like Mother Teresa, Catherine of Siena and Khema. The amount of persons of the male religious cluster changes from six ($\pi_{1.00}$) to five (π_{WNS}). This cluster of the stationary distribution derived with β-factor 1.0 includes the following three central figures of three major religions: Jesus (Christianity), Muhammad (Islam) and Gautama Buddha (Buddhism).

Another cluster that was mentioned before is the artist cluster. The base line included three female artists (Virginia Woolf, Madonna, Ayn Rand) and the amount increased by five for the stationary distribution calculated with β-factor 1.0. Contrary to that, the top 50 male list does not include any male artists for $\pi_{1.00}$ and only one (William Shakespeare) in the base line.

A comparison between the stationary distribution derived with β-factor 0.05 to the base line also shows that with increasing influence of the Wikipedia Clickstream data, the top position changes from a Wikipedia article about a male person to a Wikipedia article about a female person. Occupation clusters are also present in the top 50 females and males list of $\pi_{0.05}$. The top 50 female list has two major clusters (*royal* and *artist*) and one minor cluster (*leader*) whereas the top 50 males are divided into two major clusters (*royal* and *leader*) and three minor clusters (*scientist*, *philosopher* and *religious people*).

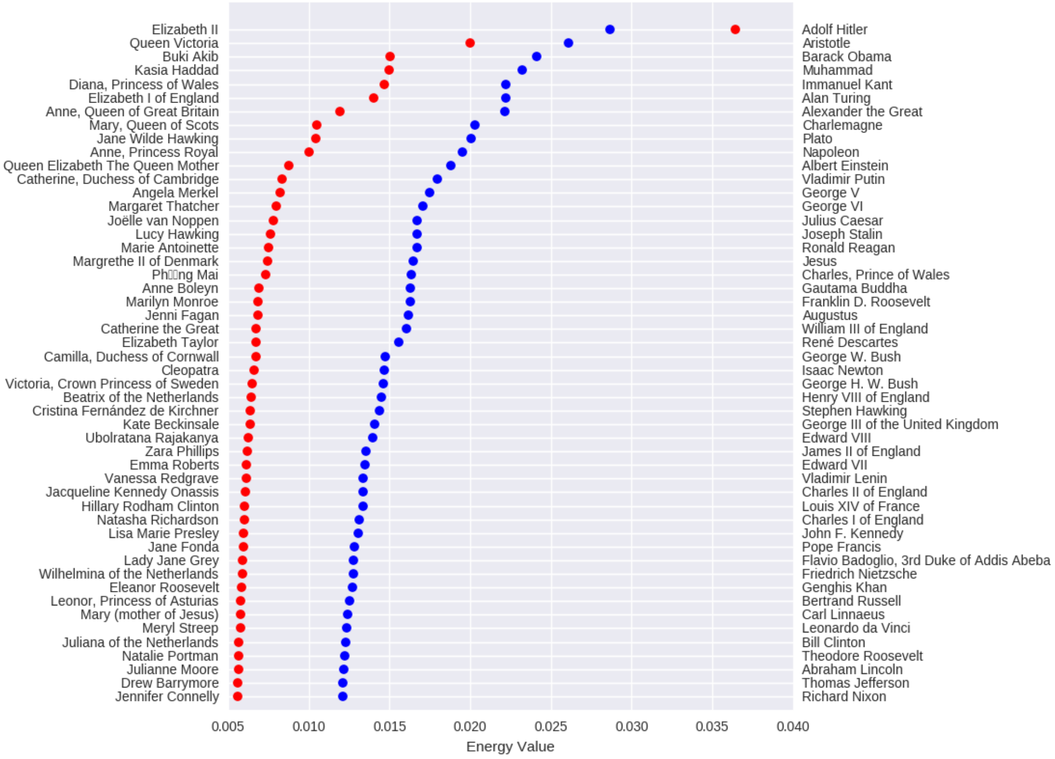


Figure 6.10.: **Top** 50 **females and males of** $\pi_{0.05}$: This figure shows that with β-factor 0.05 the top energy value changes from a male article to a female article (compared to the base line) and also the distance between the top male and female increases (compared to $\pi_{1.00}$). Furthermore the distances between the genders also decrease and the energy values for male articles also decrease faster than the energy values for male articles of the base line and $\pi_{0.05}$.

As mentioned before, both top gender lists contain many royal persons. 22 out of 50 Wikipedia articles about females and 14 out of 50 male Wikipedia articles are about royals. Altogether more than a third (36 %) of the Wikipedia articles about persons of the top 100 persons are about royal persons. Compared to the base line this cluster triples from 12 persons to 36 persons and has nearly the same size (37 persons) as the royal cluster of the stationary distribution calculated with β-factor 0.05.

The leader cluster, which is present in both top 50 gender list of $\pi_{0.05}$, consists of 6 female Wikipedia articles and 20 Wikipedia articles about males. Compared to the top gender list of the stationary distribution calculated with β-factor 1.0 the amount of female leaders stays the same and the amount of Wikipedia articles about male leaders decreases by one from 21 to 20.

The top 50 female list of $\pi_{0.05}$ also shows that, compared to the base line and the stationary distribution derived with β-factor 1.0, the amount of Wikipedia articles about artists increases. 17 women (14 actresses, two singers and one writer) are among the top 50 females of $\pi_{0.05}$. Compared to the base line this amount more than tripled from three to 14 and compared to $\pi_{1.0}$ the amount nearly doubled from eight to 14.

All of the top 50 females and males lists show that clusters with certain occupations of people exist. The base line illustrates that Wikipedia articles of female royals about the Wikipedia network structure have a high energy value and that the stationary distributions calculated with different β-factors also have such female royal clusters for the top 50 female persons. In contrast, the top 50 male persons have a cluster of male leaders which is present in the base line as well as in the stationary distributions derived with β-factors 1.0 and 0.05. The results of the stationary distributions also show that with increasing influence of the Wikipedia Clickstream data, two clusters have significantly grown. The artists cluster for the top female persons increased from three women to 17 females. One reason for these high energy values for female artists can be, than the Wikipedia Clickstream dataset is from February 2015 and in this month the Academy Awards (Oscars) were awarded and the top 50 females list of $\pi_{0.05}$ includes one award winner (Julianne Moore) and one award nominee (Meryl Streep). The second cluster that has grown is the royal cluster for male persons. The

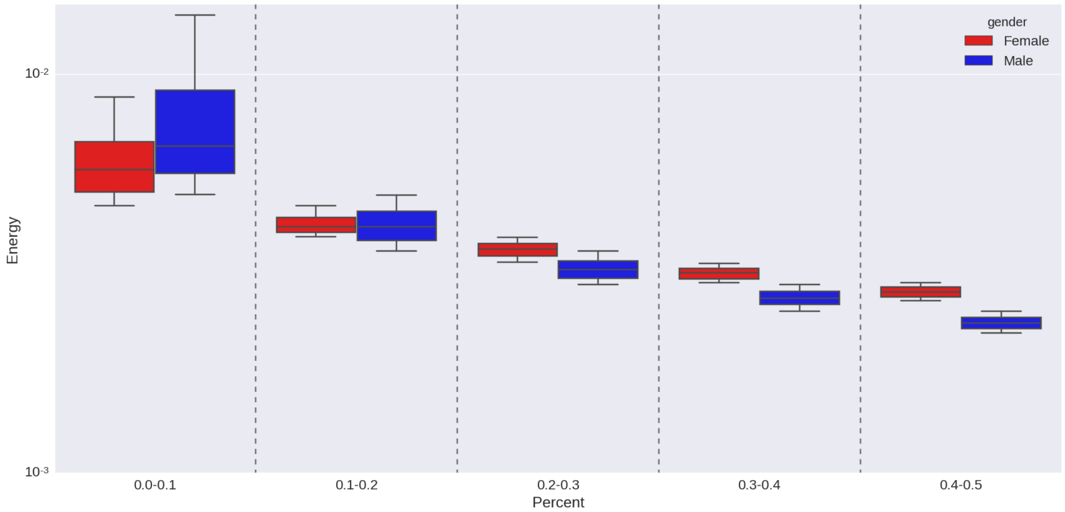


Figure 6.11.: **Top** 5 **per mill females and males of** $\pi_{0.05}$: This plot shows the top 5 per mill of the energy values of $\pi_{0.05}$. Within the first per mill, the energy values are on average higher for male articles. From the second per mill, the gender tendency shifts towards female preferred Wikipedia articles.

base line just included two male royals whereas the top 50 males list of the stationary distribution calculated with β-factor 0.05 consists of 14 Wikipedia articles about male royals.

The evolution of the top 50 females and males lists illustrate that nearly a third (32 %) of the Wikipedia articles about females are included in the base line and in the stationary distributions calculated with β-factors 1.0 and 0.05 and more than half (54 %) of the Wikipedia articles about males are present in π_{WNS}, $\pi_{1.0}$ and $\pi_{0.05}$. These top 50 females and males lists show that Wikipedia articles about persons who have a high centrality in the Wikipedia network structure also have high energy values in the stationary distributions that are influence by the Wikipedia Clickstream data.

Figure 6.11 shows that the energy values for the top per mill is higher for male articles than for females for $\pi_{0.05}$. From the second per mill the shift towards female preferred Wikipedia articles can be seen. The top gender factors of Wikipedia articles about males are significantly higher for the first per mill than corresponding values for Wikipedia articles about female persons. This fact is also apparent in the top 50 females and males list

Gender / Rank	Female	Male
1	34, 585.45	91, 650.85
2	31, 142.70	26, 119.27
3	24, 867.15	18, 850.53
4	22, 762.80	14, 747.87
5	12, 478.63	13, 877.28
6	9, 803.55	13, 664.49
7	8, 936.94	10, 854.18
8	7, 305.71	10, 273.82
9	7, 274.35	10, 033.81
10	7, 115.04	9, 954.01

Table 6.4.: **Top 10 factors for genders of the factor distribution of $\frac{\pi_{0.05}}{\pi_{WNS}}$:** This table shows the top ten change factors for genders. It illustrates that from the top notable people men benefit more from the Wikipedia Clickstream data than women. From the top 20 persons (ten female and ten male), seven males have a higher gender factor and only three females have a higher gender factor.

of $\pi_{0.05}$. In contrast, Wikipedia articles about females have, on average, a significantly higher energy value than men for $\pi_{0.05}$, which is illustrated in Figure 6.4. This shift from male preferred Wikipedia articles for the top energy values towards an average higher energy value for Wikipedia articles about females can be seen in the boxplots from the top per mill and top second per mill.

Figure 6.12 illustrates the top genders of the factor distribution $\frac{\pi_{0.05}}{\pi_{WNS}}$. The factor values for the top 10 females and males are listed in Table 6.4. These values show that concerning the top values of the factor distribution, males have a higher factor than females except for positions 2, 3 and 4. This list illustrates who profits most from the Wikipedia Clickstream dataset which reflects the user navigated Wikipedia articles. Within this top gender factor list males profit more from the user clicks than Wikipedia articles about women and also the clusters of persons with certain occupations differ. One cluster includes business related persons (*businesswomen* or *businessmen* and *entrepreneurs*) and is substantially bigger for Wikipedia

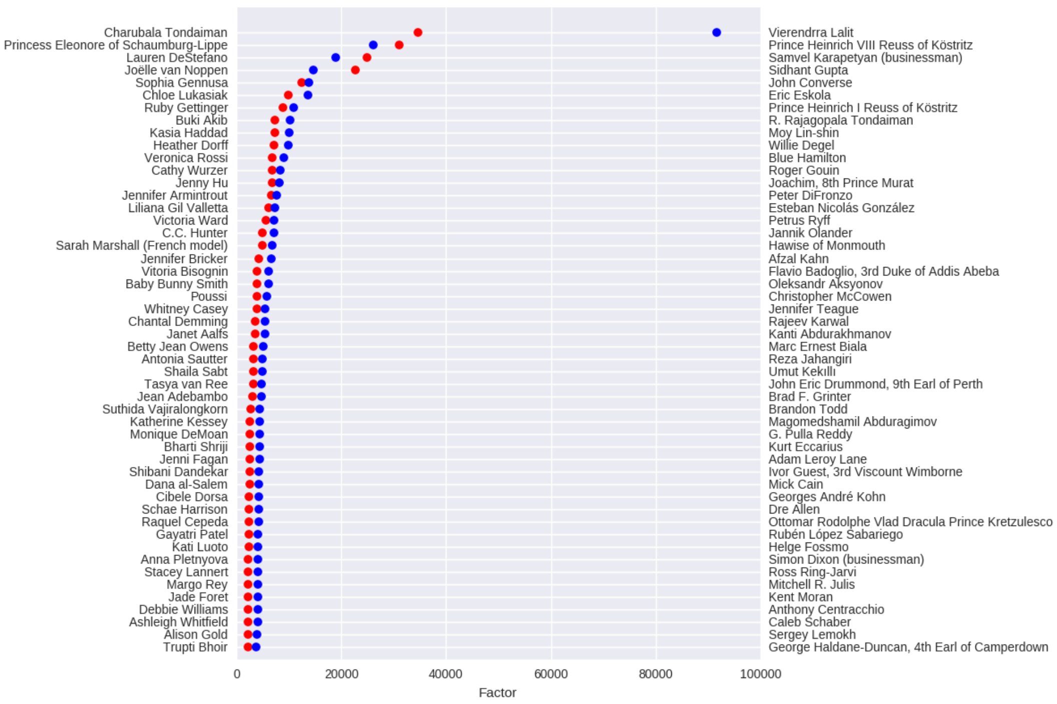


Figure 6.12.: **Top** 50 **females and males of** $\frac{\pi_{0.05}}{\pi_{WNS}}$**:** This figure illustrates that male articles have a higher gender factor concerning the top energy values. Only for positions 2, 3 and 4 female articles have a higher gender factor.

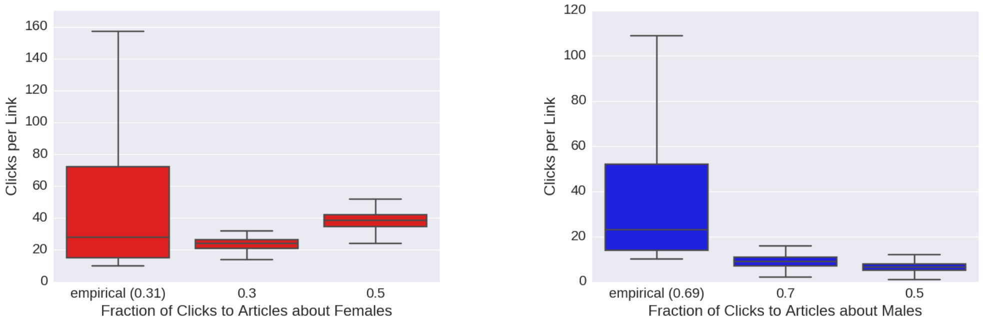
articles about males (eleven persons) than for females (one person). Also other occupation clusters like *royals*, *criminals* (five males), *sport persons* (five males) and *army related persons* (four males) have significantly more male persons than female persons. The previous four clusters only contain one female person per cluster for the top 50 gender factor list. In contrast, the top 50 female factor list includes 34 persons whose occupation is *artists* related (*actresses*, *singers*, *artists*, *writers* and *designer*) and another seven women who are *journalists* or *tv hosts* and *models*. Compared to the top 50 male factor list the corresponding clusters consist of ten people (four *actors*, two *directors*, two *journalists* and two *singers*). Since the gender factor describes the gain or loss of energy values of Wikipedia articles between the stationary distribution calculated with β-factor 0.05 and the base line, this top 50 gender factor list illustrates that the user navigated Wikipedia articles about persons favour artist related female persons and Wikipedia articles about men whose occupation is *businessmen* (and *entrepreneur*), *sport persons*, *criminals* and *army related persons*.

6.3. Results Summary

All results show that the base line prefers on average male Wikipedia articles. With increasing influence of the Wikipedia Clickstream data, the findings illustrate that there is a gender tendency in the user clicks with regard to female articles. For the random surfer it is also more likely to visit a female article when the user clicks navigation data set is used. The comparisons of the stationary distributions also show that articles about persons are generally more frequently visited than other articles, due to the increase of energy value of the articles about persons. The top 50 females and males of the base line and the stationary distributions calculated with β-factors 1.0 and 0.05 show that in these top lists Wikipedia articles about male persons have a higher energy value than Wikipedia articles about women. However, on average female persons have a higher energy value for the stationary distributions influenced by the user navigated Wikipedia articles. This shift from a male trended stationary distribution for the top energy values towards an average higher female energy value can be clearly observed within the first five per mill of the top energy values of $\pi_{0.05}$. The results also indicate that the top 50 female and male Wikipedia articles show clusters of persons with certain occupations and that there are different clusters for the genders. Also the top gender factors illustrate that the Wikipedia articles about persons show different occupations for females and males.

As the results show and as mentioned before, it is more likely for the random surfer to visit a Wikipedia article about a female person than a Wikipedia article about a male person. One reason for this probabilistic advantage of Wikipedia articles about women may be that the average click count of links leading to a Wikipedia article about females is higher than the average click count for links pointing to a Wikipedia article about males and that these higher link counts are distributed to fewer links pointing to articles about females. Altogether $3,659,843$ links point to a Wikipedia article about females and $20,802,490$ links point to an article about men. The users clicked on a link leading to an article about females $85,304,084$ times and on a link to Wikipedia articles about men $193,170,524$ times. So the average click count per link leading to Wikipedia articles about females is 23.31 and the corresponding value for Wikipedia articles about males is 9.29. The average click count values show that users on average click on links leading to articles about females more often than on links leading to Wikipedia articles about males. This behaviour might be because users have a fixed idea of what link to click next. Probably users decide beforehand if they want to visit a Wikipedia article about a female or a male and thus the users display a gender bias. To investigate in this hypothesis two splits (biased and unbiased) of the amount of user clicks each with two parts are made. The first part of the split is uniformly distributed to links leading to a Wikipedia article about females and the second part of the split is uniformly distributed to links leading to a Wikipedia article about males. The unbiased split has two equal parts of user clicks (50 % for female and 50 % for male articles) and parts of the biased split have nearly the same relation (30 % for female and 70 % for male articles) as the empirical data (actual amount of user clicks). The distributions of those two splits are compared to the distribution of the user clicks (without links that were never clicked) and the results are illustrated in Figure 6.13.

The distributions of the user clicks show that the median values for the biased click distributions are closer to the median values of the empirical user clicks than to the corresponding values of the unbiased click distri-



(a)Click distribution of Articles about Females

(b)Click distribution of Articles about Males

Figure 6.13.: This figure shows the three click distributions of links leading to Wikipedia articles about persons. Figure 6.13a illustrates the distributions of links pointing to articles about females and Figure 6.13b of links leading to Wikipedia articles about males. The first (left) boxplot of each subfigure illustrates the empirical clicks, the second (middle) boxplot of each subfigure shows the biased click distribution (30 % clicks to links pointing to an article about females and 70 % clicks to links pointing to an article about males) and the third (right) boxplot of each subfigure illustrates the unbiased click distribution.

butions. This fact might indicate that there is a gender bias located by the users which leads to a higher probability to visit a Wikipedia article about females for the random surfer. To further investigate this hypothesis it would be necessary to calculate stationary distributions with the unbiased and biased click distributions as link weights and compare the results to the corresponding stationary distribution represented in this thesis. This further process is not included in this thesis and is up to future work.

7. Conclusion & Outlook

This final chapter summarizes this thesis. Section 7.1 concludes the work and the resulting findings. In Section 7.2 limitations of the thesis are listed and an overview for future work is provided.

7.1. Conclusion

This thesis investigated user navigation behaviour of the online encyclopedia Wikipedia, one of the most frequently used information networks worldwide. Therefore a user click dataset (Wikipedia Clickstream) based on the user navigation sessions of Wikipedia articles was analyzed in-depth. The stationary distribution, which is a distribution of probability values that a random surfer visits a certain vertex (article) of a network in the limit of infinitely many steps, served as the basis for the analysis.

Focusing on Wikipedia articles about persons, the results showed that although the Wikipedia network structure favours articles about men, it is more likely for a random surfer to visit an article about a woman. Moreover, this shift from a male preferred network to a female favoured user navigation, was also visible when the analyzed Wikipedia articles about persons were categorized and grouped by year of birth or country of birth. Furthermore, this thesis outlined that even though the top articles about persons who benefited most from the user clicks were articles about males, on average articles about females profited more from the user navigation. All these results indicated a female gender tendency in the user navigation behaviour of Wikipedia. Finally, the results showed that Wikipedia articles about people are more likely to be visited by users than the Wikipedia network structure would indicate.

In conclusion, these results are a relevant finding for researches in the fields of user navigation and social studies as they provide a new insight into existing tendencies and trends in user navigation behaviour.

7.2. Limitations and Future Work

Since this thesis solely concentrates on the Wikipedia network structure and Wikipedia Clickstream dataset, it is up to future work to analyze user navigation behaviour for other information networks or social networks.

The dataset used for user navigation in this thesis was limited to the desktop version of the English Wikipedia of February 2015. For further research it would be worth considering datasets of different months as well as datasets from different Wikipedia language editions. As mobile usage of the World Wide Web is increasing year by year, it would be useful to investigate user navigation data created by mobile devices.

This thesis concentrated on analyzing user navigation tendencies and trends regarding Wikipedia articles about persons. It is up to future work to extend the spectrum of Wikipedia articles and also to use other categories beyond gender, year of birth and country of birth.

The Wikipedia Clickstream dataset is consolidated data of user clicks with no session information of the navigation process. It would be worth examining the navigation behaviour for different types of sessions (length of session, point in time the session was carried out) and for different kinds of users (sex, age, location).

Appendix

Appendix A.

Quartile Tables

	1900-1909	**1910-1919**	**1920-1929**	**1930-1939**	**1940-1949**	**1950-1959**	**1960-1969**	**1970-1979**	**1980-1989**	**1990-1999**
Females (π_{WNS})	$2.17 \cdot e^{-06}$ $5.27 \cdot e^{-06}$ $1.53 \cdot e^{-05}$	$1.97 \cdot e^{-06}$ $5.11 \cdot e^{-06}$ $1.51 \cdot e^{-05}$	$2.10 \cdot e^{-06}$ $5.58 \cdot e^{-06}$ $1.76 \cdot e^{-05}$	$2.48 \cdot e^{-06}$ $6.56 \cdot e^{-06}$ $2.07 \cdot e^{-05}$	$2.60 \cdot e^{-06}$ $7.48 \cdot e^{-06}$ $2.32 \cdot e^{-05}$	$2.55 \cdot e^{-06}$ $7.12 \cdot e^{-06}$ $2.18 \cdot e^{-05}$	$2.01 \cdot e^{-06}$ $5.72 \cdot e^{-06}$ $1.77 \cdot e^{-05}$	$1.48 \cdot e^{-06}$ $4.02 \cdot e^{-06}$ $1.15 \cdot e^{-05}$	$1.00 \cdot e^{-06}$ $2.72 \cdot e^{-06}$ $7.40 \cdot e^{-06}$	$7.08 \cdot e^{-07}$ $1.90 \cdot e^{-06}$ $4.58 \cdot e^{-06}$
Males (π_{WNS})	$3.57 \cdot e^{-06}$ $1.07 \cdot e^{-05}$ $3.69 \cdot e^{-05}$	$3.27 \cdot e^{-06}$ $9.91 \cdot e^{-06}$ $3.33 \cdot e^{-05}$	$3.29 \cdot e^{-06}$ $1.01 \cdot e^{-05}$ $3.29 \cdot e^{-05}$	$3.82 \cdot e^{-06}$ $1.10 \cdot e^{-05}$ $3.46 \cdot e^{-05}$	$3.63 \cdot e^{-06}$ $1.06 \cdot e^{-05}$ $3.18 \cdot e^{-05}$	$3.23 \cdot e^{-06}$ $9.43 \cdot e^{-06}$ $2.68 \cdot e^{-05}$	$2.59 \cdot e^{-06}$ $7.10 \cdot e^{-06}$ $2.02 \cdot e^{-05}$	$1.73 \cdot e^{-06}$ $4.73 \cdot e^{-06}$ $1.35 \cdot e^{-05}$	$1.08 \cdot e^{-06}$ $2.80 \cdot e^{-06}$ $7.59 \cdot e^{-06}$	$9.42 \cdot e^{-07}$ $2.20 \cdot e^{-06}$ $5.00 \cdot e^{-06}$
Females ($\pi_{1.0}$)	$8.25 \cdot e^{-06}$ $2.36 \cdot e^{-05}$ $7.26 \cdot e^{-05}$	$8.50 \cdot e^{-06}$ $2.39 \cdot e^{-05}$ $7.82 \cdot e^{-05}$	$8.64 \cdot e^{-06}$ $2.58 \cdot e^{-05}$ $8.47 \cdot e^{-05}$	$9.94 \cdot e^{-06}$ $2.92 \cdot e^{-05}$ $9.85 \cdot e^{-05}$	$1.12 \cdot e^{-05}$ $3.34 \cdot e^{-05}$ $1.04 \cdot e^{-04}$	$1.07 \cdot e^{-05}$ $3.22 \cdot e^{-05}$ $1.02 \cdot e^{-04}$	$9.79 \cdot e^{-06}$ $3.18 \cdot e^{-05}$ $1.02 \cdot e^{-04}$	$8.38 \cdot e^{-06}$ $2.75 \cdot e^{-05}$ $9.76 \cdot e^{-05}$	$6.65 \cdot e^{-06}$ $2.25 \cdot e^{-05}$ $8.09 \cdot e^{-05}$	$4.53 \cdot e^{-06}$ $1.38 \cdot e^{-05}$ $5.30 \cdot e^{-05}$
Males ($\pi_{1.0}$)	$9.38 \cdot e^{-06}$ $2.50 \cdot e^{-05}$ $7.10 \cdot e^{-05}$	$9.02 \cdot e^{-06}$ $2.43 \cdot e^{-05}$ $6.93 \cdot e^{-05}$	$8.71 \cdot e^{-06}$ $2.39 \cdot e^{-05}$ $7.00 \cdot e^{-05}$	$9.43 \cdot e^{-06}$ $2.56 \cdot e^{-05}$ $7.46 \cdot e^{-05}$	$9.67 \cdot e^{-06}$ $2.67 \cdot e^{-05}$ $7.78 \cdot e^{-05}$	$9.08 \cdot e^{-06}$ $2.51 \cdot e^{-05}$ $7.16 \cdot e^{-05}$	$7.74 \cdot e^{-06}$ $2.18 \cdot e^{-05}$ $6.48 \cdot e^{-05}$	$5.92 \cdot e^{-06}$ $1.69 \cdot e^{-05}$ $5.25 \cdot e^{-05}$	$5.35 \cdot e^{-06}$ $1.35 \cdot e^{-05}$ $3.47 \cdot e^{-05}$	$6.00 \cdot e^{-06}$ $1.44 \cdot e^{-05}$ $3.11 \cdot e^{-05}$
Females ($\pi_{0.25}$)	$1.07 \cdot e^{-05}$ $3.14 \cdot e^{-05}$ $1.04 \cdot e^{-04}$	$1.04 \cdot e^{-05}$ $3.28 \cdot e^{-05}$ $1.15 \cdot e^{-04}$	$1.06 \cdot e^{-05}$ $3.43 \cdot e^{-05}$ $1.25 \cdot e^{-04}$	$1.24 \cdot e^{-05}$ $3.94 \cdot e^{-05}$ $1.49 \cdot e^{-04}$	$1.36 \cdot e^{-05}$ $4.32 \cdot e^{-05}$ $1.51 \cdot e^{-04}$	$1.32 \cdot e^{-05}$ $4.21 \cdot e^{-05}$ $1.45 \cdot e^{-04}$	$1.18 \cdot e^{-05}$ $4.17 \cdot e^{-05}$ $1.50 \cdot e^{-04}$	$1.05 \cdot e^{-05}$ $3.84 \cdot e^{-05}$ $1.48 \cdot e^{-04}$	$8.63 \cdot e^{-06}$ $3.18 \cdot e^{-05}$ $1.20 \cdot e^{-04}$	$6.13 \cdot e^{-06}$ $1.94 \cdot e^{-05}$ $8.23 \cdot e^{-05}$
Males ($\pi_{0.25}$)	$1.01 \cdot e^{-05}$ $2.95 \cdot e^{-05}$ $8.97 \cdot e^{-05}$	$1.01 \cdot e^{-05}$ $2.89 \cdot e^{-05}$ $8.98 \cdot e^{-05}$	$9.68 \cdot e^{-06}$ $2.80 \cdot e^{-05}$ $8.96 \cdot e^{-05}$	$1.05 \cdot e^{-05}$ $3.01 \cdot e^{-05}$ $9.76 \cdot e^{-05}$	$1.08 \cdot e^{-05}$ $3.21 \cdot e^{-05}$ $1.00 \cdot e^{-04}$	$1.02 \cdot e^{-05}$ $2.97 \cdot e^{-05}$ $9.28 \cdot e^{-05}$	$8.73 \cdot e^{-06}$ $2.62 \cdot e^{-05}$ $8.58 \cdot e^{-05}$	$6.79 \cdot e^{-06}$ $2.07 \cdot e^{-05}$ $7.18 \cdot e^{-05}$	$6.82 \cdot e^{-06}$ $1.71 \cdot e^{-05}$ $4.41 \cdot e^{-05}$	$7.74 \cdot e^{-06}$ $1.79 \cdot e^{-05}$ $3.82 \cdot e^{-05}$
Females ($\pi_{0.05}$)	$9.21 \cdot e^{-06}$ $3.10 \cdot e^{-05}$ $1.11 \cdot e^{-04}$	$9.16 \cdot e^{-06}$ $3.33 \cdot e^{-05}$ $1.27 \cdot e^{-04}$	$9.41 \cdot e^{-06}$ $3.45 \cdot e^{-05}$ $1.39 \cdot e^{-04}$	$1.10 \cdot e^{-05}$ $3.90 \cdot e^{-05}$ $1.58 \cdot e^{-04}$	$1.21 \cdot e^{-05}$ $4.19 \cdot e^{-05}$ $1.63 \cdot e^{-04}$	$1.19 \cdot e^{-05}$ $4.14 \cdot e^{-05}$ $1.57 \cdot e^{-04}$	$1.08 \cdot e^{-05}$ $4.10 \cdot e^{-05}$ $1.62 \cdot e^{-04}$	$9.63 \cdot e^{-06}$ $3.84 \cdot e^{-05}$ $1.63 \cdot e^{-04}$	$7.91 \cdot e^{-06}$ $3.23 \cdot e^{-05}$ $1.33 \cdot e^{-04}$	$5.67 \cdot e^{-06}$ $1.99 \cdot e^{-05}$ $9.07 \cdot e^{-05}$
Males ($\pi_{0.05}$)	$9.03 \cdot e^{-06}$ $2.69 \cdot e^{-05}$ $8.80 \cdot e^{-05}$	$8.49 \cdot e^{-06}$ $2.62 \cdot e^{-05}$ $8.53 \cdot e^{-05}$	$8.09 \cdot e^{-06}$ $2.54 \cdot e^{-05}$ $9.03 \cdot e^{-05}$	$8.73 \cdot e^{-06}$ $2.75 \cdot e^{-05}$ $9.88 \cdot e^{-05}$	$9.00 \cdot e^{-06}$ $2.94 \cdot e^{-05}$ $1.03 \cdot e^{-04}$	$8.46 \cdot e^{-06}$ $2.75 \cdot e^{-05}$ $9.47 \cdot e^{-05}$	$7.29 \cdot e^{-06}$ $2.40 \cdot e^{-05}$ $8.78 \cdot e^{-05}$	$5.83 \cdot e^{-06}$ $1.92 \cdot e^{-05}$ $7.34 \cdot e^{-05}$	$6.26 \cdot e^{-06}$ $1.59 \cdot e^{-05}$ $4.30 \cdot e^{-05}$	$7.17 \cdot e^{-06}$ $1.63 \cdot e^{-05}$ $3.64 \cdot e^{-05}$

Table A.1.: **Quartiles for genders grouped by birth decade of the 20th century of the following stationary distributions: π_{WNS}, $\pi_{1.0}$, $\pi_{0.25}$ and $\pi_{0.05}$.** Each cell contains the corresponding quartiles where the first line in the cell is for Q1, the second line for Q2 or median and the third line for Q3.

	Argentina	Australia	Austria	Brazil	Canada	Germany	India	Japan	New Zealand	United Kingdom	United States
Females (π_{WNS})	$1.11 \cdot e^{-06}$ $3.58 \cdot e^{-06}$ $9.67 \cdot e^{-06}$	$1.97 \cdot e^{-06}$ $5.69 \cdot e^{-06}$ $1.84 \cdot e^{-05}$	$2.38 \cdot e^{-06}$ $1.08 \cdot e^{-05}$ $6.04 \cdot e^{-05}$	$8.30 \cdot e^{-07}$ $2.57 \cdot e^{-06}$ $8.41 \cdot e^{-06}$	$1.88 \cdot e^{-06}$ $5.11 \cdot e^{-06}$ $1.51 \cdot e^{-05}$	$2.25 \cdot e^{-06}$ $7.13 \cdot e^{-06}$ $2.95 \cdot e^{-05}$	$8.47 \cdot e^{-07}$ $2.57 \cdot e^{-06}$ $8.42 \cdot e^{-06}$	$1.57 \cdot e^{-06}$ $4.52 \cdot e^{-06}$ $1.40 \cdot e^{-05}$	$1.50 \cdot e^{-06}$ $3.54 \cdot e^{-06}$ $1.26 \cdot e^{-05}$	$2.75 \cdot e^{-06}$ $7.61 \cdot e^{-06}$ $2.37 \cdot e^{-05}$	$2.07 \cdot e^{-06}$ $5.81 \cdot e^{-06}$ $1.95 \cdot e^{-05}$
Males (π_{WNS})	$1.51 \cdot e^{-06}$ $4.63 \cdot e^{-06}$ $1.37 \cdot e^{-05}$	$2.90 \cdot e^{-06}$ $9.01 \cdot e^{-06}$ $2.53 \cdot e^{-05}$	$2.60 \cdot e^{-06}$ $8.55 \cdot e^{-06}$ $4.06 \cdot e^{-05}$	$9.72 \cdot e^{-07}$ $2.62 \cdot e^{-06}$ $7.96 \cdot e^{-06}$	$2.60 \cdot e^{-06}$ $7.04 \cdot e^{-06}$ $2.21 \cdot e^{-05}$	$2.90 \cdot e^{-06}$ $9.93 \cdot e^{-06}$ $4.07 \cdot e^{-05}$	$1.52 \cdot e^{-06}$ $4.81 \cdot e^{-06}$ $1.67 \cdot e^{-05}$	$2.12 \cdot e^{-06}$ $5.69 \cdot e^{-06}$ $1.60 \cdot e^{-05}$	$3.56 \cdot e^{-06}$ $1.05 \cdot e^{-05}$ $2.27 \cdot e^{-05}$	$2.89 \cdot e^{-06}$ $8.60 \cdot e^{-06}$ $2.98 \cdot e^{-05}$	$2.88 \cdot e^{-06}$ $9.53 \cdot e^{-06}$ $3.02 \cdot e^{-05}$
Females ($\pi_{1.0}$)	$7.28 \cdot e^{-06}$ $2.05 \cdot e^{-05}$ $6.15 \cdot e^{-05}$	$1.04 \cdot e^{-05}$ $3.22 \cdot e^{-05}$ $8.31 \cdot e^{-05}$	$1.17 \cdot e^{-05}$ $4.70 \cdot e^{-05}$ $2.21 \cdot e^{-04}$	$5.52 \cdot e^{-06}$ $1.82 \cdot e^{-05}$ $5.41 \cdot e^{-05}$	$1.02 \cdot e^{-05}$ $3.37 \cdot e^{-05}$ $1.16 \cdot e^{-04}$	$1.18 \cdot e^{-05}$ $3.74 \cdot e^{-05}$ $1.17 \cdot e^{-04}$	$1.45 \cdot e^{-05}$ $4.85 \cdot e^{-05}$ $1.32 \cdot e^{-04}$	$9.10 \cdot e^{-06}$ $2.55 \cdot e^{-05}$ $6.44 \cdot e^{-05}$	$6.23 \cdot e^{-06}$ $1.71 \cdot e^{-05}$ $5.81 \cdot e^{-05}$	$1.58 \cdot e^{-05}$ $5.26 \cdot e^{-05}$ $1.59 \cdot e^{-04}$	$1.31 \cdot e^{-05}$ $4.39 \cdot e^{-05}$ $1.58 \cdot e^{-04}$
Males ($\pi_{1.0}$)	$7.07 \cdot e^{-06}$ $1.72 \cdot e^{-05}$ $4.32 \cdot e^{-05}$	$9.07 \cdot e^{-06}$ $2.28 \cdot e^{-05}$ $5.50 \cdot e^{-05}$	$9.57 \cdot e^{-06}$ $2.56 \cdot e^{-05}$ $9.27 \cdot e^{-05}$	$6.13 \cdot e^{-06}$ $1.85 \cdot e^{-06}$ $4.48 \cdot e^{-05}$	$8.05 \cdot e^{-06}$ $2.16 \cdot e^{-05}$ $6.20 \cdot e^{-05}$	$1.10 \cdot e^{-05}$ $3.15 \cdot e^{-05}$ $1.16 \cdot e^{-04}$	$7.67 \cdot e^{-06}$ $2.48 \cdot e^{-05}$ $8.32 \cdot e^{-05}$	$8.06 \cdot e^{-06}$ $2.07 \cdot e^{-05}$ $6.27 \cdot e^{-05}$	$9.90 \cdot e^{-06}$ $2.24 \cdot e^{-05}$ $4.88 \cdot e^{-05}$	$9.86 \cdot e^{-06}$ $2.78 \cdot e^{-05}$ $8.75 \cdot e^{-05}$	$1.02 \cdot e^{-05}$ $2.89 \cdot e^{-05}$ $8.67 \cdot e^{-05}$
Females ($\pi_{0.25}$)	$9.21 \cdot e^{-06}$ $2.81 \cdot e^{-05}$ $8.93 \cdot e^{-05}$	$1.44 \cdot e^{-05}$ $4.45 \cdot e^{-05}$ $1.23 \cdot e^{-04}$	$1.32 \cdot e^{-05}$ $6.23 \cdot e^{-05}$ $2.76 \cdot e^{-04}$	$7.12 \cdot e^{-06}$ $2.50 \cdot e^{-05}$ $7.24 \cdot e^{-05}$	$1.34 \cdot e^{-05}$ $4.68 \cdot e^{-05}$ $1.61 \cdot e^{-04}$	$1.38 \cdot e^{-05}$ $4.76 \cdot e^{-05}$ $1.63 \cdot e^{-04}$	$2.08 \cdot e^{-05}$ $7.92 \cdot e^{-05}$ $2.15 \cdot e^{-04}$	$1.22 \cdot e^{-05}$ $3.52 \cdot e^{-05}$ $8.45 \cdot e^{-05}$	$7.09 \cdot e^{-06}$ $2.15 \cdot e^{-05}$ $8.03 \cdot e^{-05}$	$2.13 \cdot e^{-05}$ $7.69 \cdot e^{-05}$ $2.41 \cdot e^{-04}$	$1.71 \cdot e^{-05}$ $6.06 \cdot e^{-05}$ $2.31 \cdot e^{-04}$
Males ($\pi_{0.25}$)	$8.12 \cdot e^{-06}$ $2.01 \cdot e^{-05}$ $5.31 \cdot e^{-05}$	$1.05 \cdot e^{-05}$ $2.59 \cdot e^{-05}$ $7.08 \cdot e^{-05}$	$1.10 \cdot e^{-05}$ $3.04 \cdot e^{-05}$ $1.16 \cdot e^{-04}$	$7.56 \cdot e^{-06}$ $2.30 \cdot e^{-05}$ $5.68 \cdot e^{-05}$	$9.44 \cdot e^{-06}$ $2.63 \cdot e^{-05}$ $8.08 \cdot e^{-05}$	$1.28 \cdot e^{-05}$ $3.73 \cdot e^{-05}$ $1.32 \cdot e^{-04}$	$9.19 \cdot e^{-06}$ $3.42 \cdot e^{-05}$ $1.18 \cdot e^{-04}$	$9.68 \cdot e^{-06}$ $2.73 \cdot e^{-05}$ $8.32 \cdot e^{-05}$	$1.10 \cdot e^{-05}$ $2.43 \cdot e^{-05}$ $6.01 \cdot e^{-05}$	$1.15 \cdot e^{-05}$ $3.34 \cdot e^{-05}$ $1.14 \cdot e^{-04}$	$1.19 \cdot e^{-05}$ $3.35 \cdot e^{-05}$ $1.11 \cdot e^{-04}$
Females ($\pi_{0.05}$)	$8.47 \cdot e^{-06}$ $2.76 \cdot e^{-05}$ $9.34 \cdot e^{-05}$	$1.31 \cdot e^{-05}$ $4.37 \cdot e^{-05}$ $1.42 \cdot e^{-04}$	$1.06 \cdot e^{-05}$ $5.70 \cdot e^{-05}$ $2.46 \cdot e^{-04}$	$7.26 \cdot e^{-06}$ $2.34 \cdot e^{-05}$ $8.14 \cdot e^{-05}$	$1.24 \cdot e^{-05}$ $4.82 \cdot e^{-05}$ $1.80 \cdot e^{-04}$	$1.08 \cdot e^{-05}$ $4.50 \cdot e^{-05}$ $1.65 \cdot e^{-04}$	$2.12 \cdot e^{-05}$ $8.46 \cdot e^{-05}$ $2.28 \cdot e^{-04}$	$1.16 \cdot e^{-05}$ $3.37 \cdot e^{-05}$ $9.03 \cdot e^{-05}$	$6.43 \cdot e^{-06}$ $2.09 \cdot e^{-05}$ $8.48 \cdot e^{-05}$	$2.02 \cdot e^{-05}$ $8.18 \cdot e^{-05}$ $2.58 \cdot e^{-04}$	$1.57 \cdot e^{-05}$ $6.16 \cdot e^{-05}$ $2.43 \cdot e^{-04}$
Males ($\pi_{0.05}$)	$6.83 \cdot e^{-06}$ $1.80 \cdot e^{-05}$ $5.07 \cdot e^{-05}$	$8.84 \cdot e^{-06}$ $2.31 \cdot e^{-05}$ $7.17 \cdot e^{-05}$	$9.52 \cdot e^{-06}$ $2.65 \cdot e^{-05}$ $1.05 \cdot e^{-04}$	$6.92 \cdot e^{-06}$ $2.13 \cdot e^{-05}$ $5.51 \cdot e^{-05}$	$8.30 \cdot e^{-06}$ $2.47 \cdot e^{-05}$ $8.03 \cdot e^{-05}$	$1.10 \cdot e^{-05}$ $3.40 \cdot e^{-05}$ $1.16 \cdot e^{-04}$	$8.15 \cdot e^{-06}$ $3.32 \cdot e^{-05}$ $1.22 \cdot e^{-04}$	$8.68 \cdot e^{-06}$ $2.63 \cdot e^{-05}$ $8.54 \cdot e^{-05}$	$9.66 \cdot e^{-06}$ $2.11 \cdot e^{-05}$ $5.75 \cdot e^{-05}$	$9.88 \cdot e^{-06}$ $3.10 \cdot e^{-05}$ $1.16 \cdot e^{-04}$	$1.02 \cdot e^{-05}$ $3.05 \cdot e^{-05}$ $1.12 \cdot e^{-04}$

Table A.2.: **Quartiles for genders grouped by birth country of the following stationary distributions: π_{WNS}, $\pi_{1.0}$, $\pi_{0.25}$ and $\pi_{0.05}$.** Each cell contains the corresponding quartiles where the first line in the cell is for Q1, the second line for Q2 or median and the third line for Q3. The list of countries includes the top two (amount of born people) countries per continent plus Austria.

Bibliography

Acker, Joan (2006). “Inequality regimes gender, class, and race in organizations.” In: *Gender & society* 20.4, pp. 441–464.

Annandale, Ellen and Kate Hunt (2000). *Gender inequalities in health*. Ballmoor, Buckingham, UK: Open University Press Buckingham. ISBN: 1-33520-364-7.

Aragón, Pablo et al. (2012). “Biographical social networks on Wikipedia: a cross-cultural study of links that made history.” In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. ACM, p. 19.

Artazcoz, Lucia, Carme Borrell, and Joan Benach (2001). “Gender inequalities in health among workers: the relation with family demands.” In: *Journal of Epidemiology and Community Health* 55.9, pp. 639–647.

Barreto, Manuela and Naomi Ellemers (2005). “The burden of benevolent sexism: How it contributes to the maintenance of gender inequalities.” In: *European journal of social psychology* 35.5, pp. 633–642.

Beneria, Lourdes and Gita Sen (1982). “Class and gender inequalities and women’s role in economic development: Theoretical and practical implications.” In: *Feminist Studies* 8.1, pp. 157–176.

Boguna, Marian, Dmitri Krioukov, and Kimberly C Claffy (2009). “Navigability of complex networks.” In: *Nature Physics* 5.1, pp. 74–80.

Brin, Sergey and Lawrence Page (2012). “Reprint of: The anatomy of a large-scale hypertextual web search engine.” In: *Computer networks* 56.18, pp. 3825–3833.

Buchmann, Claudia, Thomas A DiPrete, and Anne McDaniel (2008). “Gender inequalities in education.” In: *Annu. Rev. Sociol* 34, pp. 319–337.

Colclough, Christopher, Pauline Rose, and Mercy Tembon (2000). “Gender inequalities in primary schooling: The roles of poverty and adverse cultural practice.” In: *International Journal of Educational Development* 20.1, pp. 5–27.

Collier, Benjamin and Julia Bear (2012). "Conflict, criticism, or confidence: an empirical examination of the gender gap in wikipedia contributions." In: *Proceedings of the ACM 2012 conference on computer supported cooperative work*. ACM, pp. 383–392.

Cortes, Corinna and Vladimir Vapnik (1995). "Support-Vector Networks." In: *Machine Learning* 20.3, pp. 273–297.

Forsythe, Nancy, Roberto Patricio Korzeniewicz, and Valerie Durrant (2000). "Gender inequalities and economic growth: A longitudinal evaluation." In: *Economic Development and Cultural Change* 48.3, pp. 573–617.

Geigl, Florian et al. (2015). "Random Surfers on a Web Encyclopedia." In: *CoRR* abs/1507.04489. URL: http://arxiv.org/abs/1507.04489.

Ghosh, Rumi and Kristina Lerman (2012). "Rethinking centrality: the role of dynamical processes in social network analysis." In: *arXiv preprint arXiv:1209.4616*.

Helic, D. (2012). "Analyzing user click paths in a Wikipedia navigation game." In: *MIPRO, 2012 Proceedings of the 35th International Convention*, pp. 374–379.

Helic, Denis et al. (2013). "Models of human navigation in information networks based on decentralized search." In: *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. ACM, pp. 89–98.

Hill, Benjamin Mako and Aaron Shaw (2013). "The Wikipedia gender gap revisited: Characterizing survey response bias with propensity score estimation." In: *PloS one* 8.6, e65782.

Lam, Shyong Tony K et al. (2011). "WP: clubhouse?: an exploration of Wikipedia's gender imbalance." In: *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*. ACM, pp. 1–10.

Milgram, Stanley (1967). "The small world problem." In: *Psychology today* 2.1, pp. 60–67.

Okojie, Christiana EE (1994). "Gender inequalities of health in the Third World." In: *Social science & medicine* 39.9, pp. 1237–1247.

Peixoto, Tiago P. (2014). "The graph-tool python library." In: *figshare*. DOI: 10.6084/m9.figshare.1164194. URL: http://figshare.com/articles/graph_tool/1164194.

Porter, M. F. (1997). *Readings in Information Retrieval*. Ed. by Karen Sparck Jones and Peter Willett. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. Chap. An Algorithm for Suffix Stripping, pp. 313–316. ISBN: 1-55860-454-5.

Ratkiewicz, J., A. Flammini, and F. Menczer (2010). "Traffic in Social Media I: Paths Through Information Networks." In: *Social Computing (SocialCom), 2010 IEEE Second International Conference on*, pp. 452–458. DOI: 10.1109/SocialCom.2010.72.

Reagle, Joseph and Lauren Rhue (2011). "Gender bias in Wikipedia and Britannica." In: *International Journal of Communication* 5, p. 21.

Selfe, Cynthia L and Paul R Meyer (1991). "Testing claims for on-line conferences." In: *Written communication* 8.2, pp. 163–192.

Wagner, Claudia et al. (2015). "It's a man's wikipedia? assessing gender inequality in an online encyclopedia." In: *arXiv preprint arXiv:1501.06307*.

West, Robert and Jure Leskovec (2012). "Automatic Versus Human Navigation in Information Networks." In: *ICWSM*.

Wikipedia (2016). *Criticism of Wikipedia*. URL: https://en.wikipedia.org/wiki/Criticism_of_Wikipedia (visited on 03/14/2016).

Wulczyn, Ellery and Dario Taraborelli (2015). "Wikipedia Clickstream." In: URL: http://dx.doi.org/10.6084/m9.figshare.1305770.v12.