Max Planck Institute
for Evolutionary Anthropology

UNIVERSITÄT LEIPZIG

Master Thesis

Finding and Analyzing Social Networks in unstructured web log data using probabilistic topic modeling

by

Patrick Jähnichen

Natural Language Processing Group
University of Leipzig, Germany

in cooperation with
Max-Planck-Institute for Evolutionary Anthropology
Leipzig, Germany

patrick_jaehnichen@informatik.uni-leipzig.de

Supervisors

Prof. Dr. Gerhard Heyer, Natural Language Processing Group, University of Leipzig
Dr. Colin Bannard, Department of Comparative Psychology, Max-Planck-Institute for Evolutionary Anthropology, Leipzig

Acknowledgement

I want to thank Prof. Dr. Gerhard Heyer and the Natural Language Processing Group for giving me the opportunity to write this thesis in a very helpful and pleasant working environment. I address very special thanks to Dr. Colin Bannard for acting as a supervisor who was always at hand with a useful piece of advice and for pointing me into this research direction in the first place. The work with him at the Max-Planck-Institute for Evolutionary Anthropology also always was a pleasure. I also want to express my gratitude towards my wife Julia, who supported me over the whole time of thesis work and without whom, this would not have been possible. Thanks to Mathias Seifert, who is responsible that there are considerably less typos and incomprehensible passages in this text and to Frank Binder who was a most welcome dialog partner.

Abstract

Web logs and other platforms used to organize a social life online have achieved an enormous success over the last few years. Opposed to applications directly designed for building up and visualizing social networks, web logs are comprised of mostly unstructured text data, that comes with some meta data, such as the author of the text, its publication date, the URL it is available under and the web log platform it originates from. Some basics on web logs and a description of such data is given in chapter 1. A way to extract networks between authors using the meta data mentioned is discussed and applied in chapter 2. The required theoretical background on graph theory is covered in this chapter and it is shown that the networks exhibit the Small World Phenomenon. The main question posed in this theses is discussed in chapters 3 and 4, which is, if these networks may be inferred not by the available meta data, but by pure natural language analysis of the text content, allowing inference of these networks without any meta data at hand. For this, different techniques are used, namely a simplistic frequentist model based on the "bag-of-words" assumption and so called Topic models making use of Bayesian probability theory. The Topic models used are called Latent Dirichlet Allocation and, expanding this model, the Author-Topic model. All these techniques and their foundations are thoroughly described and applied to the available data. After this, the possibility of predicting the distance between two authors of web log texts in a social network by comparing term frequency vectors(bag-of-words) or probability distributions produced by the Topic models in terms of different metrics. After comparing these different techniques, a new model, also building on Latent Dirichlet Allocation, is introduced in the last chapter, together with possible ways to improve prediction of social networks based on content analysis.

Contents

Chapter 1

Introduction, Preliminary ideas and expectations

1.1 Motivation

Over the last years, there has been an apparent trend towards Web 2.0 tools in the world wide web. Web 2.0 applications center on user generated content and present it to (potentially) all other users on the internet. This paradigm is a fundamental change in the traditional content generator-user interaction. Web content used to be generated by a well-known source(a company, a service provider or a private person) and presented to the users of the internet. This has changed, such that now, 'well-known' sources, which are a variety of open source and commercial organizations, provide tools to make use of the described paradigm by allowing users to generate content and at the same time providing a web-portal(or something similar), in which this content may be presented by the creators to other users. These platforms have reached a very high popularity and are heavily used, the video platform Youtube[1], the social network Facebook[2] and the weblog platform Wordpress[3] just being among the most famous and intensely used ones. As this thesis focusses on the latter type of Web 2.0 applications, a first question has to be answered: *What exactly is a weblog?*

Weblogs

A weblog (or blog as it is often abbreviated) is *not* a tool or application as such. It is a website with separate articles, displayed in reverse timely order(i.e. newest on top). Weblog portals or platforms equip users with the ability to run such websites without having the necessary technical skills to do so by themselves. This opportunity has caused a massive use of blogs by all types of people and for all kinds of topics. In fact, the use

[1] http://www.youtube.com
[2] http://www.facebook.com
[3] http://www.wordpress.com

of blogs has increased so rapidly, that Technorati[4] recently determined the total number of new blog articles to an astounding value of ten new articles per second [5]. All blog articles (or blog posts) have an author(the so called "blogger") and as blogs can address all topics from technical reviews over book recommendations to diary entries of private nature, it appears to be likely that some blog posts link to others(not necessarily in the same domain) by hyperlinks.

This fact raises some interesting questions: *How dense are blog posts interlinked? What information about authors can be revealed? Are these structures in any way comparable to known characteristics of social networks? And are connections correlated with similarity in the authors' posts?*

It is the aim of this thesis paper to show that blog post interlinkage projected on a graph with authors as nodes, follows the same principles as social networks. Furthermore, it will be shown that using even simple measures of document similarity, a high correlation to both connectedness and distance in the network can be found. Additionally, more complex probabilistic methods(Latent Dirichlet Allocation, Author-Topic-Model) are studied, demonstrated and discussed, on which an overall new approach is based to infer social networks directly from blog post content instead of monitoring Uniform Resource Locators(URLs) in the content. This is done, because links between blog posts are by no means complete, do *not* necessarily mean that two authors are connected via a social bond and might give an insufficient view of the bloggers' social network.

1.2 Weblogs and Social Networks

One of the main questions, this thesis deals with, is the question *when* and *how* social networks arise in weblog communities and how this can be discovered.

1.2.1 Author connectivity

Blog posts, as any other textual content on the internet, consist of plain text and hyperlinks. Additionally, they all have a unique URL, under which the blog post can be reached in the world wide web. By extracting the hyperlinks in the content and comparing them to the URLs of all other blog posts, matching URLs are considered to create a link between them. Although such links are always directed, two blog posts are considered to be connected, if one post links to the other or vice versa. This means, for a connection, a symmetric hyperlink is *not* needed (this is done to keep the complexity of graphs to a useful level, see Chapter 2). As every blog post has a distinct and unique author[6], it is assumed that, if two blog posts are connected, their authors may be considered connected as well in terms of a social network.

[4] http://www.technorati.com

[5] http://www.marketingcharts.com/interactive/blogging-hits-mainstream-integral-to-media-ecosystem-6256/technorati-state-of-blogosphere-size-2008jpg/

[6] blog posts that fail to have this prerequisite will be ignored

1.2.2 The data

The data ([9]) used throughout for computations described in this thesis, originates from a data set provided by spinn3r.com[7] for ICWSM (International Conference for Weblogs and Social Media) 2009[8]. Spinn3r is a company that focuses on indexing blog websites and providing an interface to access them. The entirety of blog websites is also called the *blogosphere*. For the ICWSM, Spinn3r provided a two months crawl of weblogs, consisting of 127GB of uncompressed data. Blog websites were rated during the crawling process according to *tailrank* and organized into different tiergroups[9].

Structure of the data

Only the first tiergroup of the data is used for further computations, which results in about 36GB of raw data or 5.3 million English blog posts used for data analysis. On one hand, as tailrank favors blog posts that were cited by others, the expectation to find a useful author-connection-graph is high, on the other hand, this helps to handle the sheer mass of data by using just 30% of the data. The data is provided in XML files, in which for each blog post a variety of meta data, such as timestamp, language, source or author are stored. This is used for dividing the data set into nine time slices each consisting of the blog posts of one week and restricting it to blog posts of English language only. This is done because the majority of blog posts available are in English language and the second largest language subset consists of blog posts with unknown language (see Fig.1.1).

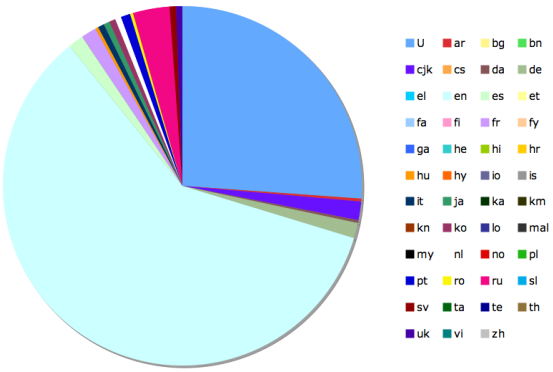


Figure 1.1: Language distribution

[7] http://www.spinn3r.com

[8] http://www.icwsm.org/2009/

[9] Tailrank computes a list of "top stories", consisting of blog entries that have been cited very often. Older items obtain a lower score than newer ones, same as less popular gain lower score than more popular entries(popularity again in terms of being cited by other blog entries). Entries high on the list tend to cause a significant activity in the blogosphere for about a day. The lower the tiergroup, the more successful are its contained blogs in placing blog entries onto the list of "top stories"

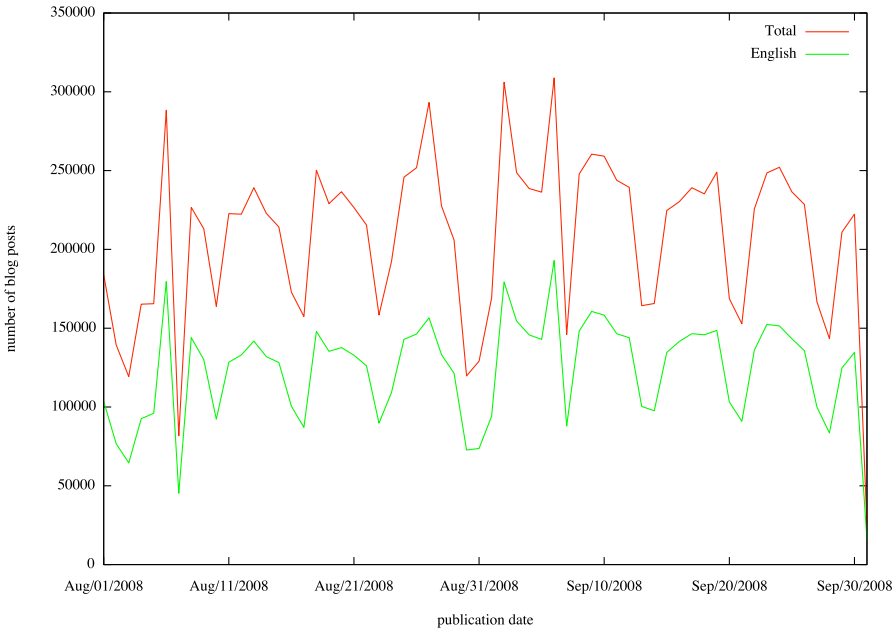


Figure 1.2: Publication date distribution

Furthermore, it can be stated that the emission of new blog posts is *not* constant over time. As seen in Fig.1.2, there are several bursts in blog post publication at certain dates. Significant peaks are observed on August, 6th, 13th, 18th, 27th, September 2nd and 6th, as well as on September 9th, 19th and 24th. Some of those may be explained with known events such as the Olympic Games in Beijing and the Tibetan Unrest just before the Games, the U.S. presidential election campaigns of the Democrat and Republican Party or (the later peaks) with the start of the worldwide economical crisis.

Making the data usable

The amount of data forces any analysis software to come up with a solution that is able to handle corpora that can not be held in memory for the whole analysis process. To overcome this challenge, an implementation based on relational databases has been chosen. For each document[10] a reference to its author, its URL and its source(i.e. the name of the blog) is stored. Beyond, the publication date, the title and the language of the document are written to the database. The actual content of the document is cleaned before storing it, by removing and storing all hyperlinks that occur in it(contained_links) and removing all HTML-Tags, punctuation and stop words. After that, the content is converted into a stream of integer values to save disk space and speed up performance. The actual strings are being held in a vocabulary from which for each integer the string value can be obtained. This makes up the following work flow for each blog post:

[10]From now on, blog post and document is used interchangeably.

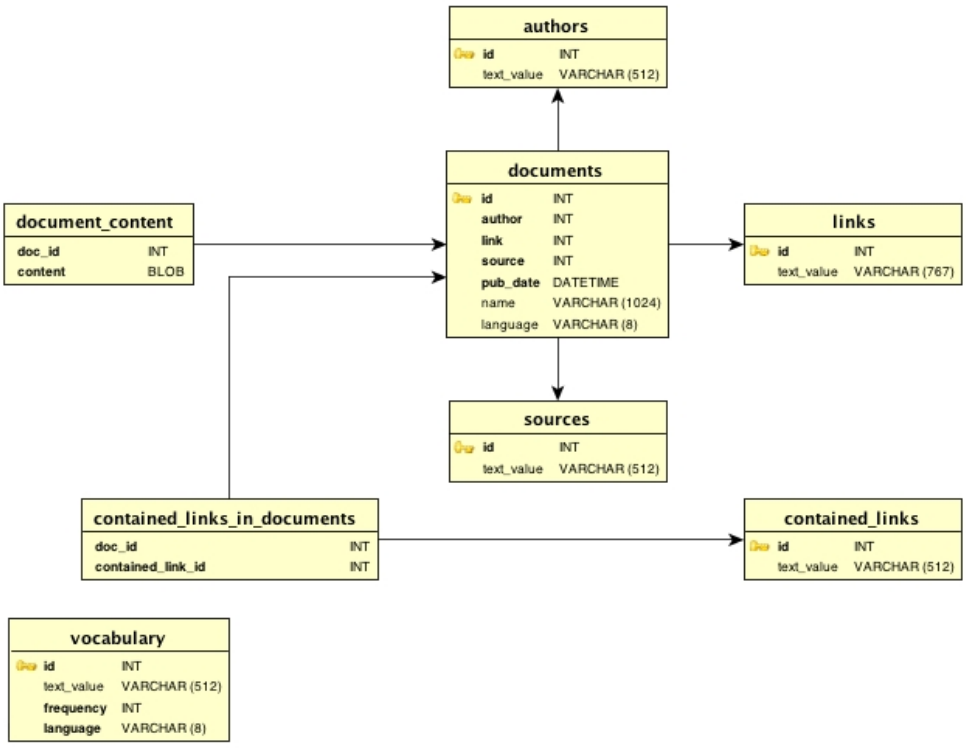


Figure 1.3: ER model of the blogdata database

1. Extraction of meta data, such as author, language, link (this blog post is originally available at), title, content, publication data and source(in most cases the name of the blog portal provider).

2. "Translate" author, link and source to integer ids, each representing a unique table entry in the relational database(allowing them to be taken from the database if already existent).

3. Extract all hyperlinks in the blog post's content, store them and delete them from it.

4. Delete all other html code from the content.

5. Delete stop words.

6. For all words in the content, add them to the vocabulary table of the database, resulting in an integer id for each unique word, then store the content as an integer array.

Chapter 2

Social Networks and Graph Theory

According to [49], social networks are comprised of "social entities" and their "relationships with each other". In such a network, "the social environment can be expressed as patterns or regularities in relationships among interacting units"[1]. The following remarks are intended to give a brief introduction into the theory of social network research. To do so, it is essential to start with some preliminary explanations of graph theory and the introduction of used mathematical symbols. After that, the experiments carried out in the 1960s by Stanley Milgram are described and discussed, as they laid out the basis of modern social network analysis. The first complete mathematical model in graph theory, able to describe the structures found by Milgram has been introduced by Duncan J. Watts and Steve Strogatz and will be discussed and applied to the available data set at the end of this chapter.

2.1 Some basics on graph theory

Graph theory is the study of graphs, whereas graphs are, at the lowest level of abstraction, simple data structures comprised of a set of points that are connected arbitrarily by a set of lines. The following definition of a graph is taken from [52]:

> A *graph* G consists of a nonempty set of elements, called *vertices*, and a list of unordered pairs of these elements, called *edges*. The set of vertices of the graph G is called the *vertex set* of G, denoted by $V(G)$, and the list of edges is called the *edge list* of G, denoted by $E(G)$. If v and w are vertices of G, then an edge of the form vw is said to *join* or *connect* v and w.

The number of vertices in $V(G)$ is called the *order* of the graph (further on termed n), the number of edges in $E(G)$ is called the *size* (M) and the number of edges connecting a vertex v to other vertices is called the *degree* of v, k_v. In practice, graphs can be

[1] [49], p. 3

used to model all kinds of networks. Here, vertices represent network elements (such as people, computers, cities) and edges represent relationships between these network elements (such as friendship bonds, ethernet connections, railroad connections).

2.1.1 History

The theory of graphs is based on the problem of the *Seven Bridges of Königsberg* which was firstly formulated by Leonhard Euler in [14]. Königsberg[2] is located at both sides of the river Pregel and includes two large islands, connected to the rest of the city through seven bridges (Fig.2.1). The idea of the problem is to find a walk through the city on

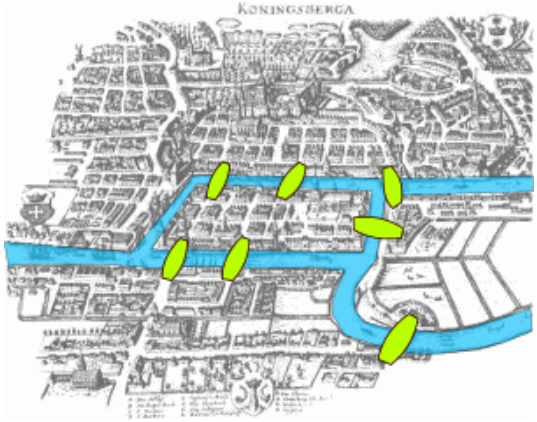


Figure 2.1: The seven bridges of Königsberg [51]

which each bridge is crossed once and only once. Euler found out, that the key problem to be solved is, to find the sequence of bridges crossed. This enabled him to reshape the problem into a mathematical structure of a graph. Here the different land masses represent the vertices of the graph and the bridges are edges between the vertices (Fig. 2.2). Euler pointed out that, during any walk, whenever one enters a land mass through a bridge, one has to leave it again over another bridge (since every bridge shall be crossed only once). This means that, if every bridge is crossed only once, the number of bridges touching a land mass has to be *even* (except for the land masses one defines to be start or finish of the walk). Since all four land masses are touched by an *odd* number of bridges, this leads to a contradiction. In other words, Euler proved that such a walk(later called a *Eulerian walk* in his honor) exists, if and only if (a) the graph is connected and (b) there exist exactly two or zero vertices with odd degree[3]. By stating and solving this problem, Euler laid down the principles of modern graph theory.

[2]Königsberg and the area around it have been conquered by the soviet Red Army during World War II and integrated into the Soviet Union and its name has been changed to Kaliningrad. The area still is an exclave to Russia nowadays.

[3]In fact, Euler stated and showed that this is a necessary condition for a Eularian walk. That it is also sufficient was only stated by him and later shown by Carl Hierholzer in [26]

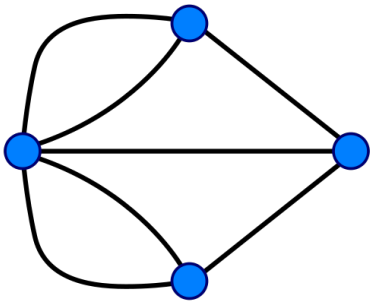


Figure 2.2: The seven bridges of Königsberg as a graph problem [51]

2.1.2 Properties of graphs

Due to the ability of representing arbitrary networks, it is clear, that vertices and edges can possess a virtually infinite number of characteristics. In contrast, graph theory deals only with the number of vertices in the graph and the characteristics of their relationships with each other, hence the characteristics of the edge set. The following restricting characterization of graphs is taken from [50] and is applied to the graphs considered in this thesis:

1. *Undirected.* Edges exhibit no inherent direction, implying that any relationship represented is symmetric.

2. *Unweighted.* Edges are not assigned any a priori strengths. Hence any importance that specific edges may later assume, derives solely from their relationship with other edges.

3. *Simple.* Multiple edges between the same pair of vertices or edges connecting a vertex to itself are forbidden.

4. *Sparse.* For an undirected graph, the maximal size (M) of $E(G) = \binom{n}{2} = \frac{n(n-1)}{2}$, corresponding to a "fully connected" or *complete* graph. Sparseness implies $M \ll \frac{n(n-1)}{2}$.

5. *Connected.* Any vertex can be reached from any other vertex by traversing a path consisting of only a finite number of edges.

Of course, these theoretical restrictions prevent graphs from modeling a variety of real world networks, as some connections between network components often are more important than others, and lots of real networks are not completely connected. But, in spite of simplifying the graph to a minimum, this model offers a minimal amount of structure about which meaningful statements can be made. Another property, a graph might exhibit, is the presence of *shortcut edges*. These are edges that connect vertices that would be widely separated otherwise. Their number is quantified by a model-independent paramter ϕ.

2.1.3 Graph metrics

By now, graphs always appeared as drawings and failed to be real mathematical structures. To overcome this flaw, the concepts of *adjacency matrix* and *adjacency list* are introduced. An *adjacency matrix* $\mathbf{M}(G)$ is a $n \times n$ matrix(with n being the order of the graph) in which $M_{i,j}$ is either 1 (if vertex i is connected to vertex j) or 0 (if otherwise). This matrix tends to be very sparse what makes it easy to store it quite efficiently in computer memory. The *adjacency list* is a simple list of all vertices of the graph with all vertices to which they are connected next to them. In other words, in a graph adjacency list, the adjacency lists of all vertices are stored. The size of vertex v's adjacency list is equal to its degree k_v. Given the graph adjacency list, it is easy to compute the average degree k of the graph, which is done by $\frac{1}{n}\sum_v k_v$. This also means that, for undirected graphs, k describes the relationship between the number of vertices n and the number of edges M, in other words, the relationship between *order* and *size* of the graph, by $M = \frac{n \cdot k}{2}$.

Another important property of graphs considered in this thesis, is the *characteristic path length* $L(G)$, that is the typical distance $d(i, j)$ between any vertex and every other vertex. In this context, distance does *not* refer to the well known "distance" concept known from metric euclidean spaces. In fact, the distance $d(i, j)$ refers to the number of edges that have to be crossed to reach vertex j from vertex i, in other words the *shortest path length* between vertices i and j. A formal definition of $L(G)$ is given in [50]:

> The *characteristic path length* (L) of a graph is the *median* of the *means* of the *shortest path lengths* connecting each vertex $v \in V(G)$ to all other vertices. That is, calculate $d(v, j) \forall j \in V(G)$ and find $\bar{d}_v$ for each v. Then define L as the median of $\{\bar{d}_v\}$.

Considering a social network, the degree of a vertex v can be interpreted as the number of people a person representing v knows[4]. To get hold of this in a graph theoretical context, the notion of *neighborhood* $\Gamma(v)$, also Γ_v, of a vertex v is introduced. The neighborhood is defined to be the subgraph consisting of all vertices adjacent to v (not including v itself). Now it also is interesting, how probable it is, that the people v knows, do also know each other. This is measured by the *clustering coefficient* γ_v, which is defined to be the fraction of existing edges and possible edges in v's neighborhood or formally,

$$\gamma_v = \frac{|E(\Gamma_v)|}{\binom{k_v}{2}}. \tag{2.1}$$

The clustering coefficient of a graph G, γ, is defined to be the average over all vertices' clustering coefficients,

$$\gamma = \frac{1}{n}\sum_v \gamma_v. \tag{2.2}$$

[4]Of course, knowledge is a wide and interpretable term here, providing flexibility in the problem in question.

One of the first authors drawing a bow between the distance of individuals in a social network, their degree and their social status was Harary in [24] where he defined the social status of a person as follows:

> *The status s(A)* of a person A in an organization Θ is the number of his immediate subordinates plus twice the number of their immediate subordinates (who are not immediate subordinates of A) plus three times the number of their immediate subordinates (not already included), etc.

Though Harary does not cover networks as such and restricts himself to trees (which are a special type of graphs[5]), he draws a direct conclusion of the social status of a person (A, vertex v), the degree of v, k_v and the vertices with shortest path length of 1, 2, etc. ($\forall j \in V(G) : d(v, j) = 1, 2, ...$) and their degrees.

2.2 The small-world phenomenon or the six degrees of separation

The concept of short paths from an individual to another unknown individual in a society is quite an old one. It often occurs when people meet and after some initial conversation find out about a common friend. A very likely exclamation in such a case might be: "Well, it's a small world!", giving this phenomenon its name: the *small-world phenomenon*. The first author to mention acquaintances between absolute strangers over just a few links was Figyes Karinthy, a Hungarian poet, in his short story "Chains" (Láncszemek)[6]. He describes three anecdotes, the first two of which are about connecting the author to well known persons, he never met before, e.g. Nobel Prize winner Zelma Lagerlöf. By thinking of just a few logic steps[7], the narrator arrives at very short paths connecting him to famous persons. Karinthy even takes the possibility into account that such links do not exist at all[8].

Over twenty years later, in the early 1950s, mathematician Manfred Kochen and political scientist Ithiel de Sola Pool wrote an article called "Contacts and Influences" in which they picked up the concept of finding acquaintances in what they called "contact nets", but

> [...] before we can decide what to collect [empirical information] we need to to think through the logical model of how a human contact net works. [9]

[5] A tree is a graph without any cycles, hence the cluster coefficient is zero.

[6] [29]

[7] e.g. Lagerlöf accepted the Nobel Prize by Swedish King Gustav and can be said to be acquainted with him and King Gustav being an enthusiastic tennis player who, besides others, already has played against a man called Kehrling with whom the author is acquainted with

[8] In the third example he states that it was impossible to find such a chain between Julius Caesar and an Aztec or Maya priest in less than 300, let alone six links, by simply taking into account, that Romans were not even aware of the existence of a (South-)American continent.

[9] [44], p. 10

They conducted first empirical experiments, trying to find the diversity of contacts for different groups in society and also the frequency of such contacts. Furthermore, they introduced graphs to

> [...]describe with precision the structure of human acquaintance networks, and the mechanisms by which social and political contacts can be established[...][10]

They concluded that

> [...]despite the effects of structure, the modal number of intermediaries in the minimum chain between pairs of Americans chosen at random is 2.[...][11]

Furthermore, they stated that

> [...]it is practically certain that any two individuals can contact one another by means of at least two intermediaries. In a structured population it is less likely, but still seems probable. And perhaps for the whole worlds population probably only one more bridging individual should be needed.[12]

But de Sola Pool and Kochen also were aware, that their analysis of these problems raised a lot of questions that could not be answered, so they held back publication of this article for over 20 years until a new journal ("Social Networks") was published. Here finally, they presented their thoughts, findings and unanswered questions, one of which is the question about the degree of separation in real social networks, to the research community. But still, in the 20 years before its publication, the draft circulated among researchers and also reached young psychologist Stanley Milgram, who visited the University of Paris, where Kochen and de Sola Pool were working at that moment. He took the challenge of unanswered questions and, back at Harvard University, he developed a series of experiments, enabling him to find chains in the U.S. American population. The experiment started by choosing people in Omaha, Nebraska and Wichita at random[13] to be the starting points and endpoint of a correspondence, respectively. A package and a letter describing the goal of the study and giving information about the target person in Boston was sent to the persons defined as starting points. They were asked if they personally[14] knew the target person and, in this case, to forward the package directly to that person. Otherwise, they were instructed, they should think of a person(friend, relative etc.) they knew personally that is more likely to know the target person and to forward the package to that person. Additionally, they were to send a postcard to

[10] [44], p. 29
[11] [44], p. 42
[12] ibid.
[13] Milgram chose these cities due to a large distance both socially and geographically, although, by using commercial address lists, the sampling process was not random at all, since companies collecting addresses for reselling them have an interest to primarily collect addresses of socially high leveled individuals, which in turn affects the experiments (due to [30])
[14] Throughout the whole experiment, to know someone personally means to address him/her by first name.

Harvard which allowed the researchers to follow the chain as the package went on. Furthermore, they were able to identify the breaking point of a chain, in case the package never arrived at the target person.

In [37], where Milgram published the outcome of his studies and in [48], where he together with Jeffrey Travers examined his findings critically, it turned out, that a huge problem of the studies was, that people often refused to take part and did not forward their package. In one of the experiments, 232 of 296 packages never reached the target person which is equivalent to over 78% loss. Nevertheless, the average chain length of packages actually reaching the target person was around 5.5 or 6 links and the authors concluded, that every American might be connected to every other American by a chain of only six intermediaries. Also, the authors found out, that often the package reaches a close geographic proximity relatively quickly, but it takes some time to find the correct contact to actually reach the target person. The friends or acquaintances through which packages reached the target person tend to be very few in their number, finishing a high number of chains respectively.

In concluding an average chain length of 6 links between all Americans, Milgram supported the acceptance of the ideas firstly expressed by Karinthy, although there are some serious critical points in the interpretation of his data. The main critique concerns the lack of completed chains which is only 30% at most. If one considers every person in the chain to be equally probable to refuse participation, longer chains are much less probable than shorter ones, leading to an underestimation of chain length[15]. Additionally, as indicated above, by using commercial address lists, participants tended to belong to a social group of certain level which means that starting points of a correspondence cannot be said to be randomly chosen from whole society and are also more likely to be able to complete a chain. Another critique is, that although showing that there exists a degree of separation of a quite small dimension, the actual network model able to reproduce real world networks, stayed hidden and continued to be until in the mid 1990s, when Duncan J. Watts and Steven H. Strogatz proposed a formal mathematical graph model for small world networks, as described in the next section.

Milgram, Kochen and Karinthy are often incorrectly credited as creators of the phrase "six degrees of separation", yet it most probably originates from a play of this name by John Guare[16].

2.3 Social Networks by Watts/Strogatz

2.3.1 Preconditions

Watts, as he describes in [50], was studying the synchronization of biological oscillators in a population of crickets. While trying to attract attention, male insects synchronized their chirps by listening to other crickets and adjusting their "responses" accordingly. The interesting question was now:

[15] [30]
[16] [23]

> [...]Who was listening to whom? One could always assume that, within a given tree, every cricket was listening to every other. On the other hand, maybe they paid attention to only a single nearby competitor. Or maybe it was some complicated combination in between. In any case, did it even matter how they were 'connected', or would the population as a whole do roughly the same thing regardless?[...][17]

Watts bases his studies on the work of Ray Solomonoff and Anatol Rapoport, who in 1951 introduced the notion of "random-biased nets". They model disease spreading in populations of different levels of structure in [45], p. 1:

> Consider an aggregate of points, from each of which issues some number of outwardly directed lines (axones). Each axone terminates upon some point of the aggregate, and the probability that an axone from one point terminates on another point is the same for every pair of points in the aggregate. The resulting conguration constitutes a *random net*.

With this theory comes the idea of propagation in a randomly connected network, in that all nodes are to have the same number of links to other nodes. These findings led to a variety of theories building on them, and more realistic views on the problem (as listed in [50], p. 13):

1. *Structural biases*, specifically *homophily*, meaning the tendency to connect with people "like" yourself, *symmetry* of edges, leading to undirected instead of directed edges, and *triad closure*, meaning the probability that your acquaintances are also acquainted with each other.[18]

2. *Social differentiation* of a population into heterogenous subgroups.[19]

Later, Granovetter introduced a distinction between the connections in links between nodes by the attribute of strength, which he defines in [20], p. 1362:

> Consider, now, any two arbitrarily selected individuals - call them A and B - and the set, $S = C, D, E, ...$, of all persons with ties to either *or* both of them. The hypothesis which enables us to relate dyadic ties to larger structures is: the stronger the tie between A and B, the larger the proportion of individuals in S to whom they will *both* be tied, that is, connected by a weak or strong tie. This overlap in their friendship circles is predicted to be least when their tie is absent, most when it is strong, and intermediate when it is weak.

Consider persons A, B, C, representing the vertices v, w, x in a graph G. Then, A and B are connected via a strong tie, if and only if $vw \in E(G)$ is true, that is, there exists a direct connection between A and B. Conversely, a weak tie would be a connection between B and C that is not directly but indirectly existent (see Fig. 2.3).

[17] [50], p. xiii
[18] [17]
[19] [43]

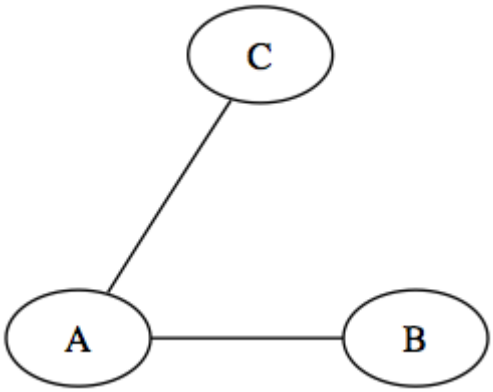


Figure 2.3: Weak tie in a social network

By this definition, it also turns out, that, if A is strongly tied to *both* B and C, there must exist a connection between B and C, be it a weak or a strong one. If there is a strong connection between B and C, then $wx \in E(G)$ must be true, otherwise B and C are connected indirectly via A. The (impossible) absence of this connection, Granovetter called the *forbidden triad*. He also found out, that weak ties are often much more important for the length of the chain between two arbitrarily chosen vertices in the graph, since strong ties primarily exist inside clique-like subsets of persons in the net and therefore do not contribute to short paths between persons in these cliques and others outside them. Weak ties instead, form the "bridges" between the cliques and thus predominantly contribute to shorter paths between arbitrarily chosen persons. This point of view allows it, to model a global view of a network, while keeping information about its local structure. As shown in [50], p. 18, social network analysis takes different attributes into account:

1. The statistical analysis of pathways through networks with varying degrees of local structure.

2. The qualitative description of the structure of networks in terms of local (e.g., clustering) and nonlocal (e.g., weak ties) features.

3. The renormalisation of networks, viewed as meta-networks of highly clustered or equivalent subnetworks.

4. The embedding of networks into (hopefully low-dimensional) spaces where the coordinates are readily interpretable and relationships between members can be visualized more easily .

2.3.2 From order to randomness

As Watts proceeds, it becomes clear, that one of the main problems in finding a theoretical model for social networks lies in the fact, that they exist in an intermediate range between what he calls "order" and "randomness". That is to say, there exist theoretical models for either completely ordered networks (e.g., d-dimensional lattices, see Fig. 2.4) or completely random ones (e.g., Rapoport's random-biased nets) and social networks must be located at any point between them. [50], p. 24, defines a d-dimensional lattice to be

> [...] a labelled, unweighted, undirected, simple graph that is similar to a Euclidean cubic lattice of dimension d in that any vertex v is joined to its lattice neighbours, u_i and w_i, as specified by
>
> $$u_i = \left[\left(v - i^{d'}\right) + n\right] (\mathrm{mod} n),$$
>
> $$w_i = \left(v + i^{d'}\right) (\mathrm{mod} n),$$
>
> where $1 \leq i \leq \frac{k}{2}, 1 \leq d' \leq d$, and it is generally assumed that $k \geq 2d$.

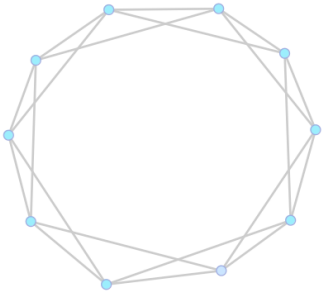


Figure 2.4: 1-dimensional lattice with k=4

There are two different types of random graphs, the first of which, $G(n, M)$, is a graph with n vertices and M randomly chosen edges(M often depends on n) and is mostly abbreviated G_M (see Fig. 2.5). The second type of graph, $G(n, p)$, consists of n vertices and each of the $\binom{n}{2}$ edges between them exists with probability p(Fig. 2.6) and mostly is abbreviated G_p. As the properties of graphs are usually examined for $n \to \infty$, G_M and G_p are practically interchangeable as $M \simeq pN$[20].

What is apparent, is, that both extremes of the range have the common characteristic of showing the same structure in both local and global level, in other words they "look" the same everywhere. Given the two extremes of completely ordered and completely random networks and what is known about structure and behavior of those networks, Watts poses the following question (in [50], p. 24):

[20] [50], p. 35

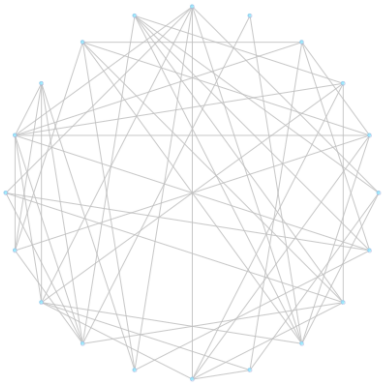


Figure 2.5: Random graph with n=20, M=55

Figure 2.6: Random graph with n=20, p=0.1

> [...]*Does the Small-World Phenomenon arise at some point in the transition from order to disorder, and if so, what is responsible for it?* In other words, *What is the most general set of characteristics that guarantee that a system will exhibit the small-world phenomenon, and can those characteristics be specified in a manner independent of the model used to create the network?*

To answer this question, it is necessary to recall the definitions of both *characteristic path length* and *clustering coefficient* (eq. 2.1, 2.2). A large benefit of using lattices as "ordered" graphs is, that in lattices (in this case the subclass of 1-lattices), both graph properties can be computed directly by

$$L = \frac{n\,(n+k-2)}{2k\,(n-1)}$$

and

$$\gamma = \frac{3}{4}\frac{(k-2)}{(k-1)}$$

and characterize the graph sufficiently. From these, it is clear, that $L\,(G)$ scales linearly with n, the number of vertices in the graph, and inversely so with k, the average degree of the graph. The clustering coefficient, γ, on the other side, is independent of n and also of k (for large k). This holds for higher dimensional lattices as well.
In random graphs, a finite value for $L\,(G)$ only exists if the graph is connected. As shown in [13], this is true for "almost any" random graph with $M > \frac{n}{2\ln(n)}$ edges (this is the same as setting $k \gtrsim \ln(n)$).

To recapitulate, Watts examines the properties of "ordered" and "random" graphs (namely characteristic path length and clustering coefficient) and how they behave when transforming a graph from ordered to random (or vice versa). To gain more insight into this, he starts with a model that he calls the "connected-cavemen world". It is constructed out of several clusters, each consisting of complete graph[21](see Fig. 2.7). In each cluster, all vertices share the same degree, k, and thus consist of $n_{local} = k + 1$ vertices[22]. By altering the clusters as shown in Fig. 2.8, and using the recurved edge to connect the clusters, the connected-cavemen graph results. The recurved edges serve as short cuts, as they connect edges that were not connected at all before, hence they reduce the path lengths between vertices of different clusters from ∞ to some finite value. Watts argues, that this resulting graph is the most highly clustered one, that also meets the requirements of connectedness and sparseness[23]. The notion of shortcuts is quite important, because they act as a separating property of edges for different definitions of the degree (taken from [50], p. 105):

> The *effective local degree* k_{local} is the average number of edges per vertex that have a range $r = 2$. That is, *local edges* are part of at least one triad,

[21]In a complete graph, all vertices are adjacent to all others, hence the clustering coefficient is 1
[22]In Fig. 2.7, k=4 and thus n=5
[23]This assumption is not proven, though if the graph is not the most highly clustered one, the graph that is can only be more clustered by an amount of $\mathcal{O}\left(\frac{1}{k^2}\right)$, which is negligible.

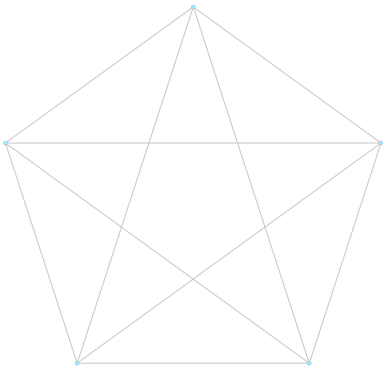


Figure 2.7: A cluster in the connected cavemen graph consisting of 5 vertices

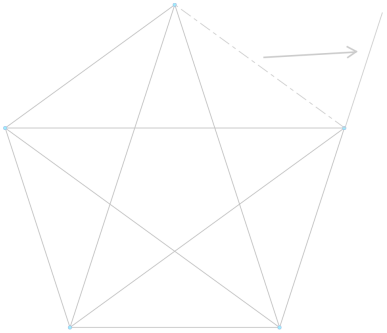


Figure 2.8: Modification of clusters to allow connections

so k_{local} captures the number of edges belonging to a typical vertex that are *not* shortcuts.

The *effective clustering degree* $k_{cluster}$ is the average number of vertices $u_{j\neq i} \in \Gamma(v)$ that each $u_i \in \Gamma(v)$ is connected to. That is, $k_{cluster}$ quantifies how many of v's *other* neighbours, each of v's neighbours is connected to.

Using these new definitions of degrees, it is also possible to redefine the clustering coefficient by

$$\gamma = \frac{k_{local}\left(k_{cluster}-1\right)}{k\left(k-1\right)} \tag{2.3}$$

and for the connected-cavemen world, it can be shown that by averaging over all vertices,

$$\begin{aligned} k_{local_{cc}} &= \frac{1}{k+1}\left[\left(k-2\right)k+\left(k-1\right)+k+\left(k-1\right)\right] && (2.4)\\ &= \frac{k^2+k-2}{k+1} && (2.5)\\ &= k-\frac{2}{k+1} && (2.6) \end{aligned}$$

and

$$k_{cluster_{cc}} = k-\frac{4}{k}+\mathcal{O}\left(\frac{1}{k}\right)^3. \tag{2.7}$$

The same distinction(*local* inside the cluster and *global* in the graph) is also valid for the property of characteristic path lengths. Since most pairs of vertices in the same cluster are adjacent to one another in the connected-cavemen graph (which means that $d(i,j)=1$), and because most pairs consist of vertices of different clusters (due to $n \gg k$), the characteristic path length is dominated by the shortest path length between clusters. This means, there exist two different distances measuring the distance between two vertices of the same cluster (d_{local}) and between two vertices of different clusters (d_{global}). As shown in [50], p.107, for the connected-cavemen graph it is straight forward to show that only two out of $\frac{(k+1)k}{2}$ pairs have $d(i,j)=2$ and the rest has $d(i,j)=1$. Therefore

$$\begin{aligned} d_{local} = L_{local} &= \frac{2}{\left(k+1\right)k}\left[\left(\frac{\left(k+1\right)k}{2}-2\right)\cdot 1+2\cdot 2\right] && (2.8)\\ &= 1+\frac{4}{\left(k+1\right)k} && (2.9) \end{aligned}$$

which in case $k \gg 1$ simplifies to $d_{local} \approx 1$. Looking at each cluster as an individual vertex implies that d_{global} depends on both L_{local} and the *global length scale* L_{global}. L_{global} is given by the characteristic length of a ring with $n_{global} = \frac{n}{(k+1)}$ and $k_{global} = 2$, such that

$$L_{ring} = \frac{n\left(n+k-2\right)}{2k\left(n-1\right)}$$

and thus

$$L_{global} = \frac{\left(\frac{n}{k+1}\right)^2}{4\left(\frac{n}{k+1} - 1\right)}. \tag{2.10}$$

By looking on the connected-cavemen graph, it is apparent that the path between vertex v in one cluster to vertex w in another one consists of three different steps:

1. The number of steps to get out of v's "home cluster" (L_{local})
2. The number of steps to reach w's cluster. This includes two steps per cluster(there are $L_{global} - 1$ of them), one global edge to reach the cluster and one local edge to traverse it.
3. The number of steps to get into the target cluster containing w ($L_{local} + 1$)[24].

Summing this leads to

$$d_{global} = L_{local} + 2\left(L_{global} - 1\right) + \left(1 + L_{local}\right) \tag{2.11}$$

$$= \frac{8}{k\left(k+1\right)} + \frac{\left(\frac{n}{k+1}\right)^2}{2\left(\frac{n}{k+1} - 1\right)} + 1 \tag{2.12}$$

and for $n \gg k \gg 1$ simplifies to

$$d_{global} \approx \frac{n}{2\left(k+1\right)}. \tag{2.13}$$

Given the fact that there are

$$\frac{\left(k+1\right)k}{2} \cdot \frac{n}{k+1} = \frac{n \cdot k}{2} = N_{local}$$

pairs of vertices in a cluster and

$$\frac{n}{2\left(k+1\right)}\left[\left(\frac{n}{k+1} - 1\right) \cdot \left(k+1\right)^2\right] = \frac{n\left(n-k-1\right)}{2} = N_{global}$$

pairs in different clusters, summing to

$$N = \frac{n\left(n-1\right)}{2},$$

the average distance between all pairs of vertices is given by

$$L_{cc} = \frac{1}{N}\left[N_{local} \cdot d_{local} + N_{global} \cdot d_{global}\right] \tag{2.14}$$

$$\approx \frac{2}{n\left(n-1\right)}\left[\frac{n \cdot k}{2} \cdot 1 + \frac{n\left(n-k-1\right)}{2} \cdot \frac{n}{2\left(k+1\right)}\right] \tag{2.15}$$

$$= \frac{k}{n-1} + \frac{n\left(n-k-1\right)}{2\left(k+1\right)\left(n-1\right)} \tag{2.16}$$

$$\approx \frac{n}{2\left(k+1\right)} \qquad \left(n \gg k \gg 1\right). \tag{2.17}$$

[24]The extra edge is needed to reach the cluster.

To approximate random graphs, Watts uses Moore graphs as described in [8], p. 252. A Moore graph is a *perfectly expanding graph* in that every vertex is connected to exactly k other vertices, none of which is adjacent to one another. On a local scale, a Moore graph looks like Fig. 2.9. To compute the characteristic path length in such a graph,

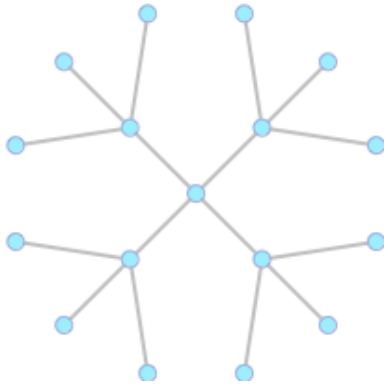


Figure 2.9: A Moore graph at local scale

consider the following: at distance d from a vertex v, $k\,(k-1)^{d-1}$ other vertices can be reached and when reaching the most distant parts of the graph, D steps away from v,

$$S = \sum_{d=1}^{D-1} k\,(k-1)^{d-1}$$

vertices are already included. The $(n - S - 1)$ vertices that are not included are at distance D from v by definition. Hence the sum of distances from v to other vertices is

$$\sum_i L_{v,i} = \sum_{d=1}^{D-1} d \cdot k\,(k-1)^{d-1} + (n - S - 1) \cdot D.$$

Seeing that L_v is the same for all v and averaging over $n-1$ vertices(omitting v) leads to

$$L_M = \frac{1}{n-1}\left[\sum_{d=1}^{D-1} d \cdot k\,(k-1)^{d-1} + (n - S - 1) \cdot D\right]$$

D now being the *diameter* of the graph. Further, Watts shows that for $k > 2$, L_M can be expressed as

$$L_M = D - \frac{k\,(k-1)^D}{(n-1)\,(k-2)^2} + \frac{k\,(D\,(k-2)+1)}{(n-1)\,(k-2)^2} \tag{2.18}$$

where D can be approximated by

$$D = \left\lfloor \frac{\ln\left[\frac{(k-2)}{k}\,(n-1)+1\right]}{\ln\,(k-1)} + 1 \right\rfloor.$$

For $k = 2$, it follows that

$$L_M = \frac{n^2}{4\,(n-1)}.$$

Although a Moore graph is *not* a random graph, connections between v's neighbors result in a redundancy of only $\mathcal{O}\left(\frac{k}{n}\right)$ for every vertex independently, thus L will vary only slightly as a result of negligible clustering.

By definition of a perfectly expanding graph, a Moore graph does not contain any triads and thus

$$\gamma_M = 0.$$

Again, this result has to be altered slightly, as in random graph triads may occur at a random chance. As Watts describes, in a k-regular random graph, the clustering coefficient is expected to be

$$\gamma_{random} = \frac{k-1}{n} \approx \frac{k}{n}$$

which becomes insignificantly small for $n \gg k$. Now a *small-world graph* can be defined exclusively in terms of n and k as Watts does in [50], p. 114:

> A *small-world graph* is a graph with n vertices and average degree k that exhibits $L \approx L_{random}\,(n,k)$, but $\gamma \gg \gamma_{random} \approx \frac{k}{n}$.

The analysis of Moore graphs and the connected-cavemen graph allows an analytical formulation of the transition between ordered and random graphs which is predominantly propelled by the introduction of shortcuts. Again, consider the connected-cavemen graph in that different clusters are treated as vertices that have their own local length scaling. By randomly replacing edges that connect vertices in the same cluster by edges that connect vertices in different clusters, these edges are pushed from a local to a global length scale. This causes L_{local} to increase and simultaneously L_{global} to decrease. As stated above, the overall characteristic path length is dominated by d_{global} and Watts even goes a step further in approximating $L \approx d_{global}$. By definition of the connected-cavemen graph, edges are assumed to connect vertices in the same cluster in just one step. By removing edges from the clusters, this obviously does not hold any longer. Using equation 2.11 and assuming that $L \approx d_{global}$ it follows that

$$\begin{aligned} L &\approx L_{local} + (1 + L_{local})\,(L_{global} - 1) + 1 + L_{local} && (2.19)\\ &= L_{local} + L_{global}\,(1 + L_{local}) && (2.20) \end{aligned}$$

Assuming that short cuts are introduced at a rate linearly dependent on ϕ[25], it follows for n and k:

$$\begin{aligned} k_{local} &= (1-\phi)\left(k - \frac{2}{k+2}\right) \\ n_{local} &= k+1 \\ k_{global} &= 2(1-\phi) + k(k+1)\phi \\ n_{global} &= \frac{n}{k+1} \end{aligned}$$

Using these, the characteristic path length can be expressed as a function of ϕ:

$$L(\phi) = L_M(n_{local}, k_{local}(\phi)) + L_M(n_{global}, k_{global}(\phi)) \times [L_M(n_{local}, k_{local}(\phi)) + 1] \tag{2.21}$$

$$= L_M\left(k+1, (1-\phi)\left(k - \frac{2}{k+2}\right)\right) + L_M\left(\frac{n}{k+1}, 2(1-\phi) + k(k+1)\phi\right) \times \left[L_M\left(k+1, (1-\phi)\left(k - \frac{2}{k+1}\right)\right) + 1\right] \tag{2.22}$$

where L_M is given by equation 2.18. The case of the connected-cavemen graph can be constructed by setting $\phi = 0$ which leads equation 2.22 to the expression L_{cc}(equations 2.14 through 2.17) in the limit $\phi = 0$. For this expression, several assumptions[26] are made, that are likely not to hold in a real world scenario, especially for large ϕ. As ϕ increases, obviously, both local and global length scale will converge until they are the same. After converging to a shared threshold, the approximation gets invalid and L_{local} will increase considerably. An important consequence of equation 2.22 is,

> [...]that when $k_{global} = 2$(that is, $\phi = 0$), $L \propto \frac{n}{k}$, but that for $k_{global} > 2$(that is, $\phi > 0$), even infinitesimally, $L \propto \frac{\ln(n)}{\ln(k)}$[...][27]

Additionally, it is also possible to make an analytical approximation for the cluster coefficient in terms of ϕ, according to ϕ's interpolation between order and randomness. By increasing ϕ, both $k_{cluster}$ and k_{local} linearly decrease. Inserting into equation 2.2 yields

$$\begin{aligned} \gamma &= \frac{(1-\phi)\,k_{local}\left[(1-\phi)\,k_{cluster} - 1\right]}{k(k-1)} \\ &= \frac{(1-\phi)\left(k - \frac{2}{k+1}\right)\left[(1-\phi)\left(k - \frac{4}{k}\right) - 1\right]}{k(k-1)} \end{aligned}$$

[25] Recall that ϕ is a parameter defining the number of short cut edges in the graph
[26] [50], p. 117
[27] ibid.

Neglecting $\mathcal{O}\left(\frac{1}{k}\right)^3$ terms results in

$$\gamma = 1 - 2\phi + \phi^2 - \left(\phi - \phi^2\right)\frac{1}{k} + \left(11\phi - 5\phi^2 - 6\right)\frac{1}{k^2} + \mathcal{O}\left(\frac{1}{k}\right)^3. \tag{2.23}$$

Equations 2.22 and 2.23 show that, while having a vast impact on $L(\phi)$, shortcuts effect $\gamma(\phi)$ only linearly.

2.4 Social networks in the corpus

Having covered the theoretical foundations, the social network that exists within the corpus may be described. To do this, the corpus has been partitioned into nine segments, comprising the blog posts of one week each(exact dates are shown in table 2.1).

segment	dates
1	Aug 01 to Aug 08
2	Aug 08 to Aug 15
3	Aug 15 to Aug 22
4	Aug 22 to Aug 29
5	Aug 29 to Sep 05
6	Sep 05 to Sep 12
7	Sep 12 to Sep 19
8	Sep 19 to Sep 26
9	Sep 26 to Oct 01

Table 2.1: Segments and dates

All documents have been stored together with the URLs contained in their text content and the URL under which they were originally available(see chapter 1). Now, if a text-contained URL matches the URL of another document, this means that the author of the first(taken to be author A) has linked in one of his blog posts to another document, of which the author is known(taken to be author B). As the overall goal is, to predict social networks by analyzing text content similarity, consider that, if A links to another document in one of his/her posts, it is highly likely that the other document's content has "something to do" with the content of his/her own blog post. Additionally, A has to know the blog post, he/she has linked to. This is only possible, if (a) A knows B and regularly reads B's blog posts or (b) another author(author C) that A is acquainted[28] to, is also acquainted to B, giving A the possibility to come across B's post by reading C's posts and following the link there.
The second possibility might also be extended to a chain of arbitrary length, although, as explained earlier, the longer this chain, the lesser the probability of its existence.

[28]being acquainted or to know each other is used interchangeably to represent the fact that an author links to a document of another author in one of his/her blog posts

Therefore, for each document in a segment, the following steps have been applied to an initially empty graph G:

1. Determine author and hyperlinks contained in the text content of the document.
2. Compare the hyperlinks to a list of links of other documents(each document is uniquely described by such a link).
3. If a text-contained link in a document matches the unique link of another document and given that the matched document also belongs to this time segment:
   (a) add both documents' authors(A and B) to $V(G)$, such that $V(G) = V(G) \cup \{A\} \Leftrightarrow A \notin V(G)$ and $V(G) = V(G) \cup \{B\} \Leftrightarrow B \notin V(G)$.
   (b) add an edge (A, B) to $E(G)$, such that $E(G) = E(G) \cup \{(A, B)\} \Leftrightarrow (A, B) \notin E(G) \wedge (B, A) \notin E(G)$

The maximum number of nodes in such a social network is equal to the number of people that authored a blog post in the according segment of data, but the actual order of the network is much lower, since only a fraction of authors are connected through hyperlinks(columns labeled "%" in table 2.2). Also, the size of the graph is much lower than that of a fully connected graph of the same size(columns labeled "%$\cdot 10^{-4}$" in the same table). To fulfill all graph properties described in 2.1.2., only the largest connected component of the resulting graph is considered.

	maximal values		found values				largest connected component			
week	order	size	order	%	size	%$\cdot 10^{-4}$	order	%	size	%$\cdot 10^{-4}$
1	87831	$3.9 \cdot 10^9$	5932	6.75	6611	1.69	3830	4.36	5368	1.37
2	104440	$5.45 \cdot 10^9$	7999	7.65	9344	1.71	5390	5.16	7785	1.42
3	102027	$5.2 \cdot 10^9$	7827	7.67	9023	1.73	5129	5.03	7371	1.41
4	101315	$5.13 \cdot 10^9$	8093	7.98	9288	1.81	5361	5.29	7684	1.49
5	99786	$4.97 \cdot 10^9$	9092	9.1	11117	2.23	6383	6.4	9554	1.92
6	109155	$5.95 \cdot 10^9$	8917	8.17	10665	1.79	6041	5.53	8945	1.5
7	107841	$5.81 \cdot 10^9$	8550	7.9	10363	1.78	5851	5.43	8632	1.48
8	112153	$6.28 \cdot 10^9$	9244	8.24	12213	1.94	5965	5.32	8896	1.42
9	82846	$3.43 \cdot 10^9$	6705	8.1	7698	2.24	4080	4.92	5533	1.61

Table 2.2: Comparison of maximal possible, found and largest connected component of found networks in the data

Considering table 2.2, the property of sparsity is also fulfilled by the networks found in the data. In a next step, the average path lengths and clustering coefficients of the networks are computed. Furthermore, for each network, a random network of equal order and roughly the same size has been generated. For these random networks, average path lengths and clustering coefficients have been computed as well and the results can be inspected in table 2.3. Comparing the found networks[29] with random networks, it

	network graph			random graph		
week	L	γ	D	L	γ	D
1	6.3	0.092	17	7.8	$3.2 \cdot 10^{-4}$	26
2	6.2	0.11	7.9	18	$4.7 \cdot 10^{-4}$	30
3	6.15	0.099	21	7.99	$5.9 \cdot 10^{-4}$	30
4	6.15	0.115	22	8.1	$5.3 \cdot 10^{-4}$	30
5	5.35	0.113	18	7.9	$3.1 \cdot 10^{-4}$	23
6	5.6	0.107	18	7.94	$3.2 \cdot 10^{-4}$	23
7	5.84	0.099	20	7.94	$3.5 \cdot 10^{-4}$	26
8	5.76	0.098	20	7.86	$3.2 \cdot 10^{-4}$	21
9	6.29	0.104	19	8.02	$3.2 \cdot 10^{-4}$	25

Table 2.3: Average path lengths, cluster coefficients and diameters of networks extracted from the data set and corresponding random networks.

becomes clear, that all networks extracted from the data set exhibit the small world phenomenon, as they have similar characteristic path lengths as the corresponding random networks, but differ greatly in their cluster coefficients(the cluster coefficient of the networks found is up to 360 times bigger than in random networks) and additionally fulfill all demanded network properties. Therefore, the networks found in the data set are taken to be social networks of authors. As segment 5's data set comprises the most[30] authors, it will be used for all upcoming computations. A visualization of this network, centered around the node with maximum degree, is shown in Fig. 2.10. In it, it can clearly be seen that most nodes are at a distance of 5 or 6 hops away from each other and that the diameter of this network is just as given in table 2.3.

[29]The "found networks" are actually the largest connected components of the found networks, to insure connectedness.

[30]in percentage to the actual number of authors in the segment and the actual number of edges between them

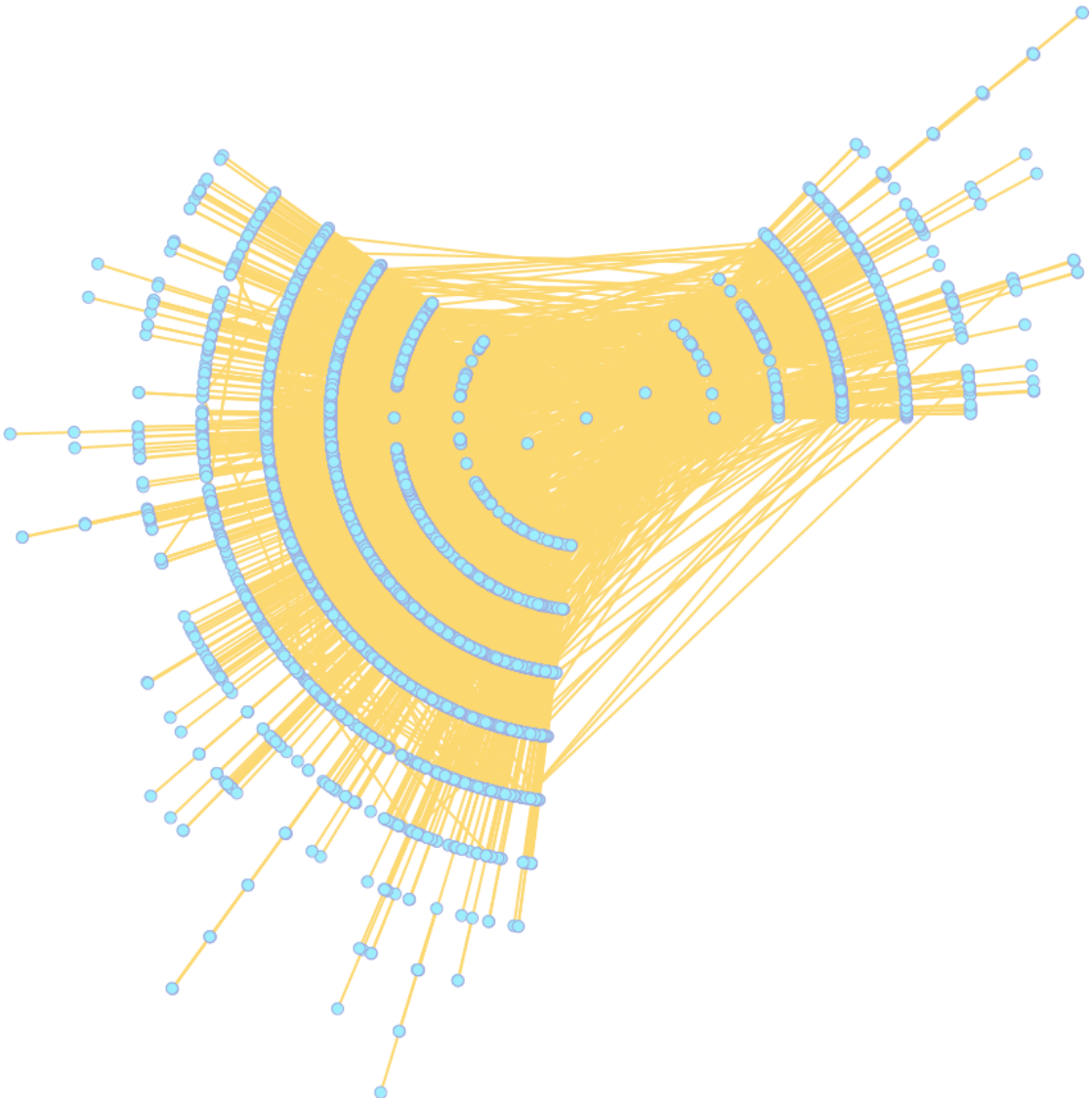


Figure 2.10: A visualization of an extracted social network.

Chapter 3

Frequency probability and a simple application

This chapter gives an overview over the field of statistics and introduces inferential statistics and frequency probability. After discussing these topics, a simple frequentist model is applied to the data set available, which is called the bag-of-words model.

3.1 Statistics - some introductory remarks

Statistics is a mathematical science concerned with the collection, analysis, interpretation and evaluation of data. There are two different main tasks, statistics is able to fulfill:

Statistics may be used to summarize, analyze or explain a collection of data. This is helpful for interpreting experiments, finding trends in the data and so forth. This is called *descriptive statistics*.

Finding randomness and uncertainty in observing certain patterns in the data and drawing conclusions about the process that generated the data is also possible using statistical methods. This is called *inferential statistics*.

Descriptive statistics is primarily used to quantitatively summarize or characterize a held out data set. Rather than drawing inferential statements about the real-life subject the data is thought to represent, as inferential statistics does, descriptive statistics may find rules or even laws in the data, but does not claim them to be true in any other data set or situation. Because the analysis of blog data tries to do just that(inferring the membership in author communities of future documents based solely on content and authorship information of available ones), the second manifestation of statistics is being focused on in this thesis.

3.2 Inferential statistics

As mentioned before, inferential statistics tries to infer knowledge about future events(to a definite extent of uncertainty) from observed events of the same kind(e.g. weather forecasts). Two, sometimes referred to as rivaling, different main concepts in inferential statistics exist, that are (a) frequency probability and (b) Bayesian probability.

3.2.1 Frequency probability

The concept of frequency probability is based on the idea to restrict oneself to the actually measurable. This means that, in contrast to Bayesian probability, hypotheses can *not* have probabilities in frequency probability, as they are no random variables. So, frequency probability is restricted solely on random events that *always* do or do not occur. This also implies, that data can only be measured and interpreted in terms of what is actually observable. Consider a bogus coin, for which it has been shown in a number of trials, that head occurs twice as often as tail. A frequentist model will be perfectly able to predict the possibility of getting a tail when the coin is thrown, but it is not capable to infer from that, with what probability the guy you are playing with tries to cheat on you. Rather it will only be able to accept or reject this *hypothesis*. Frequency probability deals with the relative frequency of an event, given the whole of all events and claims that, in the long run, this is equal to the probability of the event. Consider an event e that appears n_e times in n trials of a random experiment. Then the relative frequency $n_{relative}$ and so an approximate of the probability of event e is given by

$$\begin{aligned} n_{relative} &= \frac{n_e}{n} \\ &\approx p\left(e\right). \end{aligned}$$

and, taking the long run into account, it follows

$$p\left(e\right) = \lim_{n\to\infty} \frac{n_e}{n}. \tag{3.1}$$

In the example above, let e be the event of tossing tail, then, as $n \to \infty$, $p\left(e\right) = \frac{1}{3}$. If event e always shows the same relative frequency, no matter what happens before, it can be said to be *independent* of other events. If this is not the case, it exhibits a *conditional probability* conditioned on the event(s) that occurred before. Consider two different bags, one of green, the other of blue color. In the green bag, there are two red stones and four black stones, in the blue bag however, there are three red stones and one black stone (Fig. 3.1). The experiment is conducted as follows: randomly choose one of the two bags and from that bag, randomly take a stone. After noting what color that stone has, it is put back into its bag. This may be repeated for a large number of times, after which the outcome is analyzed. Suppose that, looking at the results of the experiment, it turns out, that the green bag is chosen 30% and the blue one 70% of the time. Furthermore, for each stone in the chosen bag, it is equally likely to pick

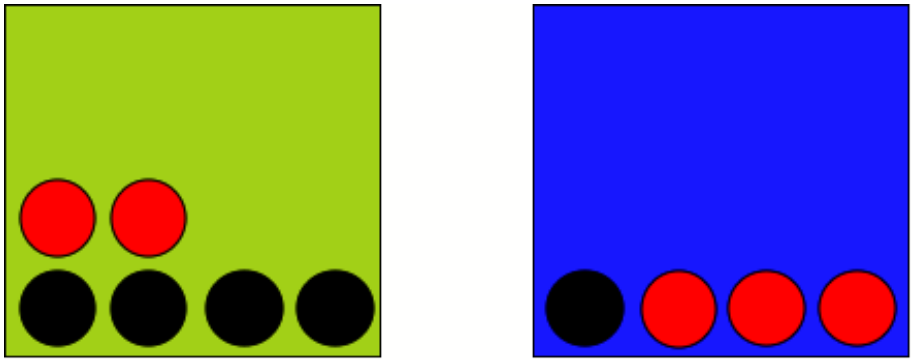


Figure 3.1: Red and black stones in two bags

it afterwards. In this example, the color of the bag is a *random variable* (noted by B), which can take two possible values, namely green (g) or blue (b). The color of the stone picked is another random variable denoted by S and can also take two values, that of red (r) and black (bl). Applying equation 3.1 leads to the probabilities of either choosing the green or the blue bag, which is denoted by $p(B = g) = \frac{3}{10}$ and $p(B = b) = \frac{7}{10}$. As seen here, the values of probabilities always lay in the interval of $[0, 1]$ and if they are *mutually exclusive*, that is, if always exactly one of the events takes place, they must sum up to one. Two questions concerning the problem of conditional probabilities might be: *What is the overall probability that a red stone is picked in the experiment?* or *If a black stone is picked, what is the probability, that it has been picked out of the green bag?* To answer these questions, consider a more general approach[1]. Let X and Y be two random variables of an experiment where X can take values x_i and Y can take values y_j with $i = 1, \ldots, M$ and $j = 1, \ldots, L$ respectively. Consider N iterations of the experiment and let the number of times where $X = x_i$ *and* $Y = y_j$ be n_{ij} (to visualize this example, consider the occurrences of every individual outcome possible to be recorded as in Figure 3.2). Now let the number of times where $X = x_i$ *independent* of Y be c_i and similarly the number of times where $Y = y_j$ be r_j.
Then, the probability that X takes value x_i and Y takes value y_j is called the *joint* probability and is given by

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N} \tag{3.2}$$

Note that according to equation 3.1, the limit $N \to \infty$ is considered here.
The probability that X takes value x_i independently of Y is given by

$$p(X = x_i) = \frac{c_i}{N} \tag{3.3}$$

Since c_i is the number of all occurrences in column i, it is also valid to assert that $c_i = \sum_j n_{ij}$. Using equations 3.2 and 3.3 it is already possible to derive the *sum rule* of

[1] Figure 3.2 and the general example explaining the product and sum rule of probability theory is taken from [5], pp. 13–16

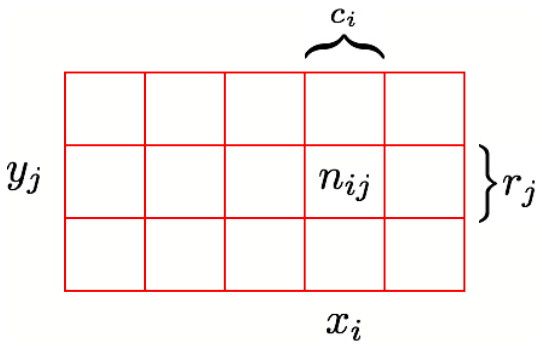


Figure 3.2: Probability matrix for product and sum rule example

probability which is given by

$$p\left(X = x_i\right) = \sum_{j=1}^{L} p\left(X = x_i, Y = y_j\right) \tag{3.4}$$

The left hand side of this equation, $p\left(X = x_i\right)$ is also called the *marginal* probability due to the process of marginalizing, that is summing out, the other variables(Y in this case).

Considering only those occurrences where $X = x_i$, the probability for those occurrences for which $Y = y_j$ holds additionally, is called the *conditional* probability of $Y = y_j$ given $X = x_i$ and is defined by

$$p\left(Y = y_j | X = x_i\right) = \frac{n_{ij}}{c_i} \tag{3.5}$$

From equations 3.2, 3.3 and 3.5, it follows that

$$\begin{aligned} p\left(X = x_i, Y = y_j\right) &= \frac{n_{ij}}{N} = \frac{n_{ij}}{c_i} \cdot \frac{c_i}{N} \\ &= p\left(Y = y_j | X = x_i\right) p\left(X = x_i\right) \end{aligned} \tag{3.6}$$

which is called the *product rule* of probability. By now, a strict distinction between a random variable(B in the bag example) and the values it can take(in this case the values g and b), has been made. Although this is helpful in avoiding misunderstandings, it leads to a rather complicated notation, so that from now on $p\left(B\right)$ will specify the *distribution* over the random variable and $p\left(g\right)$ the distribution evaluated for the case that B takes the value g. Using this notation, the *rules of probability* in terms of random variables take the form

$$\textbf{sum rule} \qquad p\left(X\right) = \sum_{Y} p\left(X, Y\right) \tag{3.7}$$

$$\textbf{product rule} \qquad p\left(X, Y\right) = p\left(Y | X\right) p\left(X\right) \tag{3.8}$$

From the product rule and using the property of symmetry in $p(X,Y) = p(Y,X)$ it immediately follows that

$$p(Y|X) = \frac{p(X|Y)\,p(Y)}{p(X)} \tag{3.9}$$

what is called *Bayes' theorem* and will play a central role in further discussions. Its left hand side would be the *conditional probability* of Y given X, whereas the first term of the right hand side's nominator shall be called the *likelihood function* or simply *likelihood* of X given Y[2]. The denominator can be expressed by

$$p(X) = \sum_{Y} p(X|Y)\,p(Y) \tag{3.10}$$

by using both sum and product rule and again the property of symmetry. Note that Bayes' theorem is a central theorem in probability theory, especially in the field of conditional probabilities. But basically, it has nothing to do with the concept of Bayesian probability which is discussed in the next chapter, although, of course, its statements do hold there as well. Coming back to the example using bags and stones, consider $N \to \infty$. It is already known that,

$$p(g) = \frac{3}{10}$$

and

$$p(b) = \frac{7}{10}.$$

If now, the green bag is chosen, the probability of picking a red stone is equivalent to the fraction of red stones in the green bag, namely $\frac{1}{3}$, and thus $p(r|g) = \frac{1}{3}$. In fact, all four possible conditional probabilities are given by

$$p(r|g) = \frac{1}{3} \tag{3.11}$$

$$p(bl|g) = \frac{2}{3} \tag{3.12}$$

$$p(r|b) = \frac{3}{4} \tag{3.13}$$

$$p(bl|b) = \frac{1}{4} \tag{3.14}$$

Note that again, these probabilities sum up to one as in

$$p(r|g) + p(bl|g) = 1$$

and

$$p(r|b) + p(bl|b) = 1.$$

Using the sum and product rules enables to answer the questions posed above:

[2]In other words, if the probability defines the outcome given a set of parameters, than the likelihood allows to find the parameters given a specified outcome

- *What is the overall probability that a red stone is picked in the experiment?*

$$\begin{aligned} p(r) &= p(r|g)\,p(g) + p(r|b)\,p(b) \\ &= \frac{1}{3}\cdot\frac{3}{10} + \frac{3}{4}\cdot\frac{7}{10} = \frac{5}{8} \end{aligned}$$

- *If a black stone is picked, what is the probability, that it has been picked out of the green bag?* Note that by finding $p(r) = \frac{5}{8}$, $p(bl) = \frac{3}{8}$ by definition of mutual exclusivity.

$$\begin{aligned} p(g|bl) &= \frac{p(bl|g)\,p(g)}{p(bl)} \\ &= \frac{2}{3}\cdot\frac{3}{10}\cdot\frac{8}{3} = \frac{8}{15} \end{aligned}$$

The second question allows some more interpretation of Bayes' theorem. The only information available, if it is not known, which stone is taken from the bag, is the color of the bag. Since this is known *beforehand*, $p(B)$ is called the *prior probability*. After having learnt, which stone was picked from the bag, the conditional probability of $p(B|S)$ can be computed via Bayes' theorem (eq. 3.9) and is called the *posterior probability*, since it is known after information about the stone is at hand. What is quite peculiar in the above example is, that given only the prior probabilities, three out of ten times, the green bag is chosen. Hence the prior probability is $\frac{3}{10}$. But given the information that a black stone is picked, the posterior probability of the green bag raises to $\frac{8}{15}$, which is considerably higher. Of course, this reflects the common intuition that, given that the ratio of black stones is much higher in the green bag, it is also more likely to have chosen the green bag, if it is known before, that a black stone has been picked. Furthermore, it shall be stated, that, if the joint probability of two events factorizes directly into their respective individual probabilities, that is

$$p(X,Y) = p(X)\,p(Y)\,,$$

the events are said to be *independent* of each other. This also implies that

$$p(Y|X) = p(Y)$$

in this case.

Frequency probability deals with probabilities in terms of frequencies of events in a random and repeatable experiment. It is also referred to as the *classical* or *frequentist* interpretation of probability. But the content of texts can not be derived or imitated by random and repeatable processes. The content exists, and there is no way of arriving at the very same content by conducting a random experiment, which means that there have to be other ways to analyze the inner structure of words and concepts using probability theory. This is called Bayesian probability and is discussed in the next chapter.

3.3 A naive Bag-of-words approach

The usage of a naive bag-of-words approach is the simplest method used to analyze the data and relies on the "bag-of-words" assumption[3]. Here, for each author, the words used in the author's documents are counted and stored in a frequency vector. After that, the vector is divided by a scalar equal to the size of the vocabulary. Although this vector can not be interpreted as a probability distribution[4], information radius and skew divergence distance measures are applied to compute the distance between two vectors, hence authors. As the bag-of-words model is based on frequency vectors, it is also straight forward to use the cosine distance measure for computing the similarity between two frequency vectors, as is often done in natural language processing or information retrieval tasks[5]. As expected, the relation between cosine distance of two authors' word frequency vectors and their corresponding distance in the network graph is considerable(see Fig. 3.3), although, for low cosine distances(i.e. up to a value of 0.7) the path distance is at the level of the average path length of the network. Only at higher values, the path distance between two authors falls clearly. This might be explained with the fact, that cosine distance completely discards terms that only appear in one of the authors' term frequency vectors. That means, at high cosine distance values, the vectors have more terms in common than at lower values and additionally a higher frequency of common terms. This also fits the expectation that authors, that share common interests have a lower network path distance than authors with less common interests.

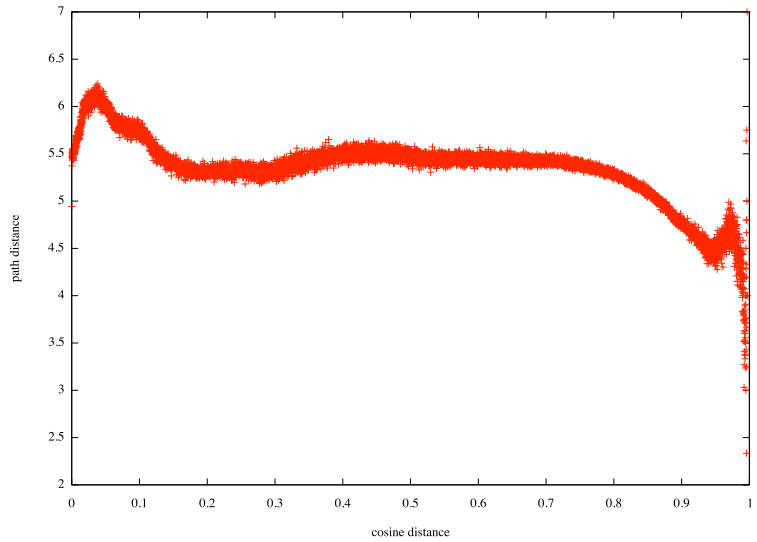


Figure 3.3: Cosine distance against network path length - Bag-of-words model

[3]this is the assumption that, for computing the overall probability of a document, the order of the document's words can be neglected

[4]as this is a normalized frequency vector, its components do not necessarily sum up to one

[5]For a description of the distance measures used, see Appendix B.

When measuring the distance between authors' frequency vectors using information radius(Fig. 3.4), no correlation between its value and the distance in the network graph can be found. Again this has been expected, as frequency vectors are not comparable to probability distributions and therefore can not be dealt with this kind of measure. In fact, the path length between two authors falls, as the information radius between their word frequency vector rises(which means, that they are less similar).

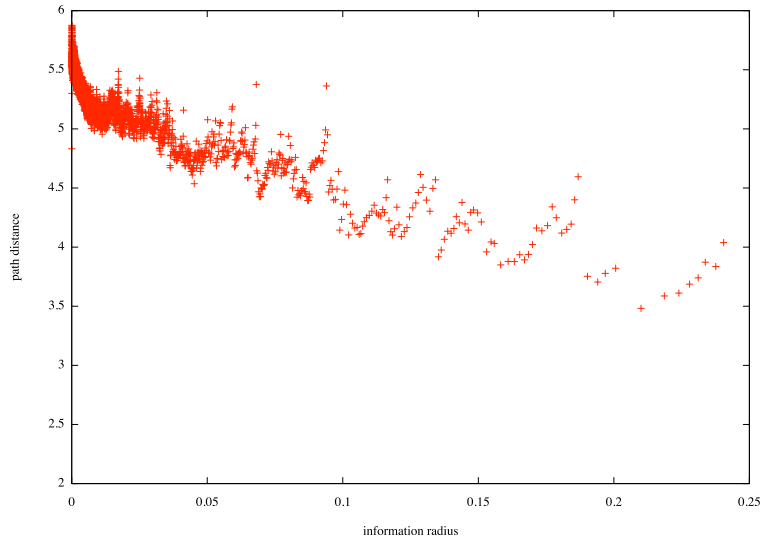
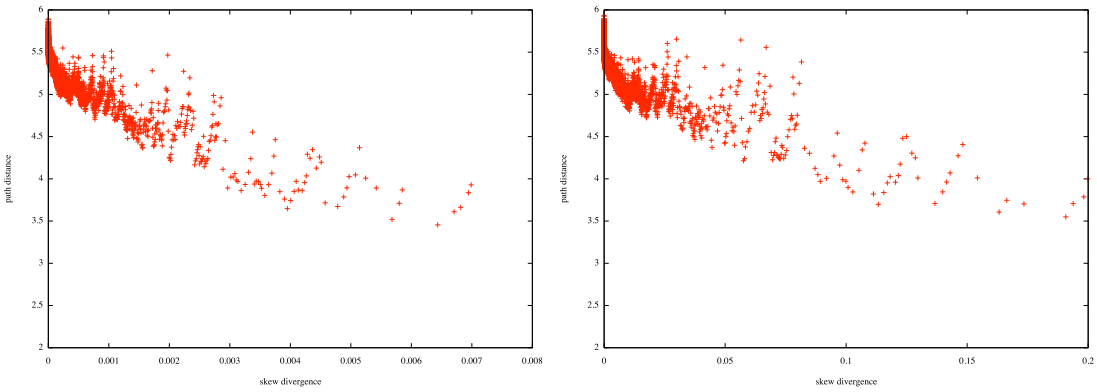


Figure 3.4: Information radius against network path length - Bag-of-words model

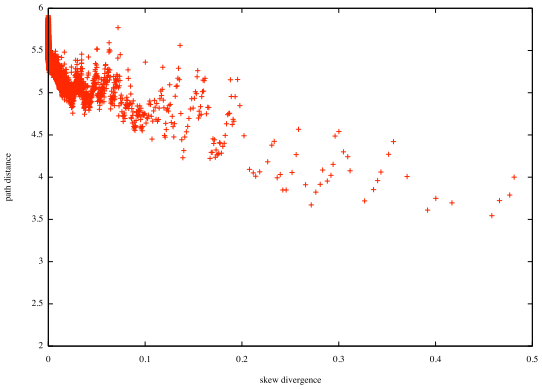
Using skew divergence for the same task shows similar results, although the developing of values seems to be more smooth. As skew divergence behaves alike(see Fig. 3.5) for all different values of α, skew divergence for $\alpha = 0.99$ will be considered from now on[6]. The fact, that for both information radius and skew divergence, the path length falls, as

[6]skew divergence with lower values of α result in lower values of skew divergence and a very tight alignment of values, so for easier interpretation a high value of α is used

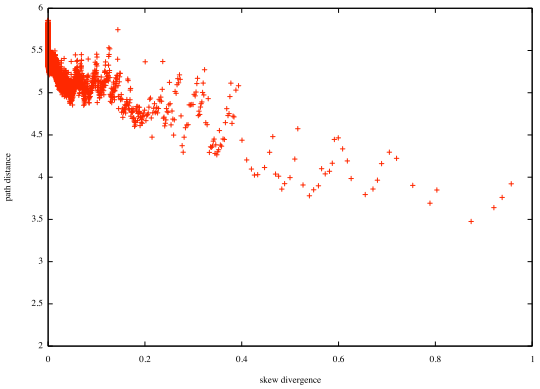


(a) $\alpha = 0.01$

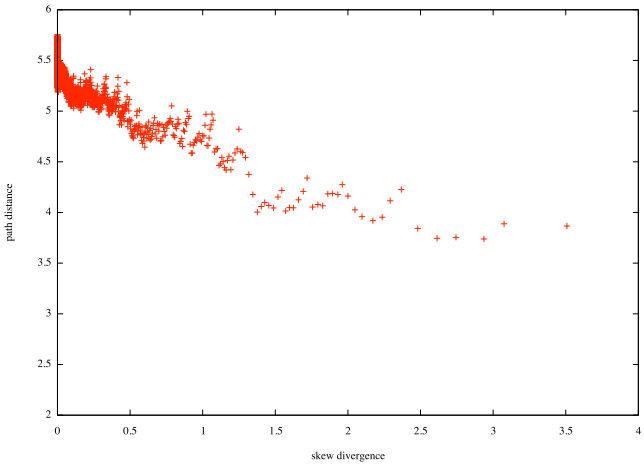
(b) $\alpha = 0.25$



(c) $\alpha = 0.5$



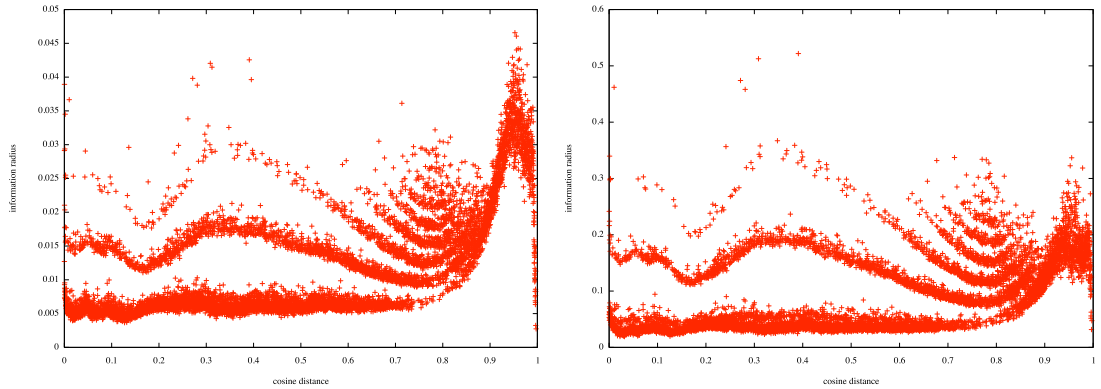
(d) $\alpha = 0.75$



(e) $\alpha = 0.99$

Figure 3.5: Skew divergences against network path length - Bag-of-words model

the distance between vectors rises, might be explained by looking at the term frequency vectors. The vectors potentially have a very high count of zero frequency entries, that, using the cosine measure are omitted, but, using skew divergence do contribute. This means, that for low values of the distance measure(where the rapid decrease in path length may be observed), it "penalizes" an author pair, where one of them uses some very rare words, whereas it strongly rewards overlap in the term frequency vectors. Hence, using Kulback-Leibler based distance measures in a term frequency setting produces exactly the opposite of what has been expected. Looking at Fig. 3.6, this can also be shown empirically, as for high values of cosine distance(hence a high similarity), the KL-divergence based distance measures rise considerably, indicating less similarity. Additionally, for the symmetric distance measure, this effect seems to be even stronger.



(a) Information radius against cosine distance **(b)** Skew divergence against cosine distance

Figure 3.6: KL-divergence based distance measures against cosine distance - Bag-of-words model

Chapter 4

Bayesian probability and Machine Learning

Having covered the basics of frequency probability, another forming of inferential statistics will be covered in this chapter. Bayesian probability tries to measure the state of knowledge and, in contrast to frequency probability, hypotheses are treated as random variables that can take on a probability. Additionally, the technique of latent variable models, on which all of the more complex models base, is introduced. Special emphasis is lain onto Latent Dirichlet Allocation (LDA), also known as the Topic model. Furthermore, a machine learning approach to the generally intractable "solution" of this model, the Gibbs sampler, is introduced and explained. Further topic models in discussion are based on LDA and expand this model by adding variables and certain semantics useful for the analysis of online correspondences.

4.1 Bayesian probability

The term of Bayesian probability goes back to Thomas Bayes, who, in 1764, published what is nowadays known as Bayes' theorem. Although Bayes laid down the principle of this field of probability theory in [3], it was not until Pierre-Simon Laplace, who introduced a more general version of this theorem and used it to address problems like medical statistics, reliability estimates and even jurisprudence[1]. According to [28], in Bayesian probability, the concept of probability is construed as "a measure of a state of knowledge", whereas the classical view interprets it as a frequency or property of a system. Again, Bayesian probability decomposes into two different points of view, namely the *objectivist view* and the *subjectivist view*. The subjectivist view treats the state of knowledge to be some kind of "personal belief"[2], whereas the objectivist view refers to strong logical reasoning, given explicitly defined, unambiguous data to be the state of knowledge. In [28], p. 44f, Edwin James(who heavily relies on logical reasoning)

[1] [46]
[2] [15]

describes that

> [...]any probability assignment is necessarily 'subjective' in the sense that it describes only a state of knowledge, and not anything that could be measured in a physical experiment.

But he further points out, that, if starting with a predefined set of logical suppositions(here called "desiderata"),

> Anyone who has the same information, but comes to a different conclusion [...], is necessarily violating one of those desiderata.

If this is the case, he claims,

> [...] it appears to us that a rational person, should he discover that he was violating one of them, would wish to revise his thinking[...]

Concluding, it is the aim of the desiderata

> [...] to make these probability assignments completely 'objective' in the sense that they are independent of the personality of the user.

This reasoning goes back to Richard Cox, who in [11] showed that, if rational numbers are used to represent different degrees of belief, then a simple set of axioms coding these beliefs(in Jaynes' case the "desiderata") is sufficient to define a distinct set of rules for manipulating the degrees of belief which exactly correspond to the sum and product rule of probability theory. As these rational numbers behave consistent to the rules of probability, they are called *Bayesian probabilities.*
Considering the notion of verifying(or determining the degree of belief) that a certain hypothesis about some data is correct, a parameter vector $\mathbf{w}$ is introduced, representing the parameters of the hypothesis H. Additionally, the data to be analyzed is denoted $\mathcal{D}$. Using equation 3.9, it now is obtained that

$$p\left(H|\mathcal{D}\right) = \frac{p\left(\mathcal{D}|H\right)p\left(H\right)}{p\left(\mathcal{D}\right)},$$

thus

$$p\left(\mathbf{w}|\mathcal{D}\right) = \frac{p\left(\mathcal{D}|\mathbf{w}\right)p\left(\mathbf{w}\right)}{p\left(\mathcal{D}\right)} \tag{4.1}$$

where $p\left(\mathcal{D}\right)$ is the prior probability of getting $\mathcal{D}$ under all possible hypotheses. Given a set[3] of hypotheses H_i, this can be rewritten as

$$p\left(\mathcal{D}\right) = \sum_i p\left(\mathcal{D}, H_i\right) = \sum_i p\left(\mathcal{D}|H_i\right)p\left(H_i\right) \tag{4.2}$$

[3]The set must be exhaustive and its content mutual exclusive, that is, it must comprise all possible hypotheses and exactly one of them can be correct.

that is, the posterior probability for a given hypothesis given the data is proportional to the likelihood of the data as a function of the parameters $\mathbf{w}$ under this hypothesis times the prior probability of the hypothesis, hence

$$posterior \propto likelihood \times prior. \tag{4.3}$$

By comprising all possible hypotheses, equation 4.2 serves as a normalization constant, ensuring that the left hand side of equation 4.1 is a valid probability and integrates to one. In fact, according to [5], p.22, integrating both sides of equation 4.1 with respect to $\mathbf{w}$ leads to an expression of the denominator in Bayes' theorem in terms of the likelihood function and prior distribution:

$$p(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})\, p(\mathbf{w})\, \mathrm{d}\mathbf{w} \tag{4.4}$$

The uncertainty in the parameters $\mathbf{w}$, that the hypothesis they describe is actually correct for the observed data $\mathcal{D}$, is expressed via a probability distribution over $\mathbf{w}$ (see next section). Although the basics of Bayesian probability are known since the 18th century, it is only in the last decade, due to more and more computation speed and lesser memory limitations, that practical applications of these methods gained more importance. The full procedure of solving problems in terms of Bayesian probability involves(as seen in equation 4.2 and 4.4) the marginalization(summation or integration) over the whole parameter space, which tends to grow faster, the more complex a problem gets. Only after the advent of machine learning techniques, such as the Markov chain Monte Carlo algorithms(see section 4.3) allowed the application of Bayesian probability to a wide range of problems.

4.2 Related work

4.2.1 Latent variable models

Latent variables in a probabilistic model are variables that cannot be observed directly, but can be inferred from other variables, the *observable variables*, that are directly measurable. In [4], the method of adding latent variables to a set of observable variables is considered a "powerful approach to probabilistic modeling". Its key concept is to define a joint probability distribution over both observed and latent variables and then to determine the corresponding distribution over the observed variables by using Bayes' theorem and marginalization. Models that try to explain observed variables by finding a distribution over (theoretical) latent variables that influence the distribution over observable variables are called *Latent variable models*. Considering the field in which these models are used here, the latent variables may also be called *hypothetical variables* as they are not physically existent and not measurable, but represent abstract concepts that are only defined within a corresponding model.

Mathematical notation and terminology

For the discussion of the following techniques, the introduction of a common mathematical notation and terminology is essential. Throughout the following explanations, these terms will be used:

- The single most basic unit of data is a *word*. It is defined to be part of a vocabulary denoted by an index out of $\{1, \ldots, V\}$. A word token representing the vth word of the vocabulary shall be denoted by w^v.
- A *document* is a sequence of N words and is represented by a vector $\mathbf{w} = (w_1, w_2, \ldots, w_N)$.
- A *corpus* is a collection of M documents represented by $D = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M\}$.

4.2.2 Generative models

As mentioned above, latent variable models statistically infer unobservable variables by analyzing available data. Considering corpus data, this means inferring the latent structure from which the documents in the corpus were generated by analyzing the available data, i.e. the set of words. In [22], Griffiths et. al. call this the extraction of the gist. To do this, any form of semantical representation might be used, the one thing that needs to be specified though, is the *process* in which a set of word is generated by this representation of their gist. Such a probabilistic process is also called a *generative model*. According to [22], it combines the strengths of two more traditional techniques, which are structured representation and statistical learning. Considering all techniques for corpus and natural language analysis as models to reduce the dimensionality of the data, some advantages of generative models can be pointed out. Other methods for dimensionality reduction are *tf-idf*[4], which reduces documents of arbitrary length to fixed-length lists of numbers, *LSI*[5], which already is able to capture some basic linguistic aspects such as synonymy and polysemy and *pLSI*[6]. The latter is also called the *aspect model* and is a true probabilistic but *not* a generative model. Although in pLSI, documents are described by a distribution over topics, there is no generative process for generating this distribution. From this, serious problems arise:

- the number of parameters grows linearly with the size of the data set, leading to problems with overfitting
- documents that lay outside the training set, i.e. unseen documents, can not be assigned with a probability distribution, destroying the predictability property of such a model

[4] [42]
[5] [12]
[6] [27]

4.3 A Markov chain Monte Carlo method and its application

Since an exact inference is intractable in most probabilistic models, it is essential to find some form of approximation for it. A common way to do this, is to use approximate inference methods that are based on numerical sampling, also known as Monte Carlo methods. The aim of Monte Carlo methods is to find the expectation of a function $f(z)$ with respect to a probability distribution $p(z)$ where z can either be a discrete or continuous variable. The evaluation of the expectation is done by

$$\mathbb{E}(f) = \int f(z)\, p(z)\, dz$$

which is too complex to evaluate exactly by analytical means. To overcome this problem, with a Monte Carlo method, a set of samples z_l where $l = 1, \ldots, L$, drawn independently from $p(z)$ is obtained. Here the expectation is given as an approximation by the finite sum

$$\hat{f} = \frac{1}{L} \sum_{l=1}^{L} f(z_l).$$

When z_l is drawn from $p(z)$, it can be said that

$$\mathbb{E}\left(\hat{f}\right) = \mathbb{E}(f).$$

Also, it is notable, that the overall accuracy does not depend on the dimensionality of z, meaning that even a small number of samples is sufficient to reach a high accuracy. The problem is though, that subsequent samples are highly correlated and to approximate the expectation, independently drawn samples are needed. Additionally, using simple sampling techniques leads to problems and severe limitations in spaces of high dimensionality[7]. To overcome these limitations, the Markov chain Monte Carlo(MCMC) technique is used, which allows sampling from a large class of probability distributions and scales well with dimensionality in the sample space. In particular, Gibbs sampling is used as a special type of MCMC. The motivation for using Markov chains together with the Monte Carlo method lies in the specific properties of Markov chains. Consider a Markov chain of first order and a set of random variables $z^{(1)}, \ldots, z^{(M)}$ such that for $m \in \{1, \ldots, M-1\}$

$$p\left(z^{(m+1)}|z^{(1)}, \ldots, z^{(m)}\right) = p\left(z^{(m+1)}|z^{(m)}\right)$$

holds true. To specify the Markov chain, it is necessary to give a probability distribution for the initial variable $p\left(z^{(0)}\right)$ together with the conditional probabilities for subsequent variables, also called *transition probabilities*

$$T_m\left(z^{(m)}|z^{(m+1)}\right) \equiv p\left(z^{(m+1)}|z^{(m)}\right).$$

[7] [5], pp. 526–536

A Markov chain will be stationary, that is invariant, if each step of the Markov chain leaves this distribution invariant. This means, that a Markov chain with $T(z', z)$ and $p^\star(z)$ will be stationary, if

$$p^\star(z) = \sum_{z'} T\left(z', z\right) p^\star\left(z'\right).$$

Also, it is required that, as $m \to \infty$, $p\left(z^{(m)}\right)$ converges to the required distribution $p^\star(z)$. The property of a Markov chain to reach a stationary condition is called *ergodicity* and the invariant distribution is the *equilibrium distribution* of the chain.
Gibbs sampling is a simple MCMC method used for approximation of the equilibrium distribution of a Markov chain, hence the distribution from which the samples are drawn. It originates from mathematical physics and was introduced into the field of inferential statistics by Geman and Geman in [19]. Consider the distribution to sample from to be $p(\mathbf{z}) = p(z_1, \ldots, z_M)$ and a Markov chain with chosen initial state. Each step of the Gibbs sampling process replaces the value of one variable by a value drawn from the distribution of that variable conditioned on the values of all other variables. In other words, value z_i is replaced by a value drawn from $p\left(z_i|z_{\setminus i}\right)$ where z_i is the ith component of $\mathbf{z}$ and $z_{\setminus i}$ is $z_1, \ldots, z_M$ with z_i omitted. This process is repeated by continuous cycling through all variables. To show that this actually finds the equilibrium distribution, it is necessary to show that $p(\mathbf{z})$ is an invariant of each step of the Gibbs sampling process and so of the Markov chain. This is true, because by sampling from $p\left(z_i|z_{\setminus i}\right)$, the marginal distribution $p\left(z_{\setminus i}\right)$ is clearly invariant, as $z_{\setminus i}$ keeps unchanged. Also each step samples from the correct conditional distribution $p\left(z_i|z_{\setminus i}\right)$ by definition. Because conditional and marginal distribution together define the joint posterior distribution, the joint distribution is itself invariant. Also, to show that the samples are from the correct distribution, it has to be ergodic. This can be shown by proving that the conditional probability is nowhere zero, that is from any point in z space, any other point can be reached in a finite number of steps(including one update of each component variable per step). To complete the Gibbs sampling algorithm, the distribution of initial states must be specified. Of course, samples depend on that initial state but will become independent after some time(the so called *burn-in period*). Also successive samples are correlated, which makes it necessary to subsample the sequence in order to obtain nearly independent samples.

4.4 Latent Dirichlet Allocation and the Topic model

4.4.1 The model

Latent Dirichlet Allocation(LDA) is a "generative probabilistic model for collections of discrete data such as text corpora" first described by Blei, Ng and Jordan in [7]. The predecessor methods described in the previous chapter are based on the "bag-of-words" assumption. Interpreting that assumption in probability theory, it is an assumption of

exchangeability of the words in a document[8]. Additionally, this assumption concerns the order of documents in a corpus as well, so that a document in a corpus can also be treated as an exchangeable random variable. Keeping this in mind and seeing that de Finetti states in [15] that any collection of exchangeable random variables follows a mixture distribution, in general an infinite one, it is straight forward to treat this as the probability distribution over (a)latent topic variables in documents and (b)words in topics. Blei et. al. emphasize the importance that

> [...]an assumption of exchangeability is not equivalent to an assumption that the random variables are independent and identically distributed. Rather, the exchangeability essentially can be interpreted as meaning "*conditionally* independent and identically distributed,"[...][9]

the latter being conditioned with respect to the underlying latent parameter of the probability distribution. This means that the joint distribution of the random variables is quite simple and factored while marginally over the latent parameter very complex joint distributions may occur. Making use of this, Blei et. al. define a three-level hierarchical generative model, in which

> [...] each item of a collection is modeled as a finite mixture over an underlying set of topics. Each topic is, in turn, modeled as an infinite mixture over an underlying set of topic probabilities.[10]

The use of natural language includes "[...]retrieval of concepts from memory in response to an ongoing stream of information."[11] In [6], Griffiths, Steyvers and Tenenbaum describe a way of analyzing large natural language corpora and the extraction of meaning out of the documents which they comprise for which they use LDA. Their idea is, to treat topics as the gist of a set of words which in turn can be represented as a probability distribution over topics. This can be used to explicitly represent a document. The content of the topics is described by the words to which they assign a high probability. Using LDA, their *topic model* learns a set of topics

> [...]automatically from a collection of documents, as a computational analogue of how human learners might form semantic representations through their linguistic experience[12]

The key concept of LDA is to infer the underlying latent topic structure of documents by describing them as random mixtures over the latent topic variables and topics by distributions over words. The following generative process for the documents in a corpus D is described in [7]:

1. Choose $N \sim Poisson\,(\xi)$,

[8] [1]
[9] [7], p. 995
[10] [7]
[11] [22]
[12] [6], p. 212

2. Choose $\theta \sim Dir\,(\alpha)$,
3. Choose $\phi \sim Dir\,(\beta)$,
4. For each of the n words w_n:
   (a) Choose a topic $z_n \sim Multinomial\,(\theta)$.
   (b) Choose a word w_n from $p\,(w_n|\phi_{z_n})$, a multinomial probability conditioned on the topic z_n.

This has also been depicted in the paper using the plate notation, shown in Figure 4.1.

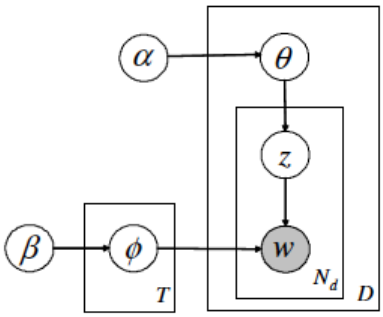


Figure 4.1: The LDA model in plate notation

Some simplifying assumptions are made in this model:

1. The dimensionality k and j of the Dirichlet distributions are assumed known and fixed (k being the dimensionality of the latent topic variable z).
2. The assumption of document length following a Poisson distribution is ignored, because it is not critical to any computations and N is independent of other data generating variables(which are θ, ϕ and $\mathbf{z}$).

As θ is a k-dimensional Dirichlet random variable, it can take values in the $(k-1)$-simplex[13] and has the probability density

$$p\,(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k}\alpha_i\right)}{\prod_{i=1}^{k}\Gamma\,(\alpha_i)}\theta_1^{\alpha_1-1}\cdots\theta_k^{\alpha_k-1}, \tag{4.5}$$

over this simplex. Similarly, ϕ's probability density is given by

$$p\,(\phi|\beta) = \frac{\Gamma\left(\sum_{i=1}^{j}\beta_i\right)}{\prod_{i=1}^{j}\Gamma\,(\beta_i)}\phi_1^{\beta_1-1}\cdots\phi_j^{\beta_j-1}. \tag{4.6}$$

[13] a k-dimensional vector lies in the $(k-1)$-simplex if all components are positive(including zero) and they sum up to one

The parameters α and β are a k-vector and j-vector respectively, with components $\alpha_i > 0$ and $\beta_i > 0$ and $\Gamma(\cdot)$ being the Gamma function. Incorporating the parameters α and β, the topic mixture θ's and the word mixture ϕ's joint probability distributions, a set of N topic assignments $\mathbf{z}$ and N words $\mathbf{w}$ (one assignment for each word) is given by

$$p(\mathbf{w}, \mathbf{z}, \theta, \phi | \alpha, \beta) = p(\theta|\alpha) \prod_{j=1}^{T} p(\phi_j|\beta) \prod_{n=1}^{N} p(z_n|\theta)\, p(w_n|\phi_{z_n}). \tag{4.7}$$

Here $p(z_n|\theta)$ is θ_i for a unique i such that $z_n^i = 1$. By integrating over θ and ϕ and summing over z, the marginal distribution of a single document can be derived:

$$p(\mathbf{w}|\alpha, \beta) = \int_{\theta} p(\theta|\alpha) \left(\prod_{n=1}^{N} \sum_{z_n} p(z_n|\theta) \right) d\theta \int_{\phi} \prod_{j=1}^{T} p(\phi_j|\beta) \left(\prod_{n=1}^{N} \sum_{z_n} p(w_n|\phi_{z_n}) \right) d\phi \tag{4.8}$$

By taking the product of the documents' marginal probabilities, the overall probability of a corpus is obtained by

$$p(D|\alpha, \beta) = \prod_{d=1}^{D} \int_{\theta_d} p(\theta_d|\alpha) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}|\theta_d) \right) d\theta_d \int_{\phi} \prod_{j=1}^{T} p(\phi_j|\beta) \left(\prod_{d=1}^{D} \prod_{n=1}^{N_d} \sum_{z_{dn}} p(w_{dn}|\phi_{z_{dn}}) \right) d\phi \tag{4.9}$$

Note that, as LDA is a three level probabilistic model, the parameters α and β are considered corpus wide variables that are sampled only once for a whole corpus, θ_d are document level parameters and are sampled once per document accordingly whereas z_{dn} and w_{dn} are word level variables that are sampled once for each word in each document in the corpus. Recalling the assumption of exchangeability, it may be stated that random variables $\{z_1, \ldots, z_N\}$ are exchangeable, if their joint distribution is invariant to exchanging their position in the set, in other words, replacing their indices by a permutation of the integers from 1 to N:

$$p(z_1, \ldots, z_N) = p\left(z_{\pi(1)}, \ldots, z_{\pi(N)}\right)$$

An infinite sequence of random variables is said to be *infinitely exchangeable* if all finite subsequences are exchangeable. As de Finetti's theorem states, that the joint distribution of an infinitely exchangeable sequence of random variables is the same as if a random parameter were taken from a probability distribution and, conditioned on this parameter, the random variables were independent and identically distributed. Using this, Blei et. al. point out, that by assuming that words are generated by topics and that the topics are infinitely exchangeable in a document, the probability of a sequence of words and topics must be

$$p(\mathbf{w}, \mathbf{z}) = \int_{\theta} p(\theta) \left(\prod_{n=1}^{N} p(z_n|\theta) \right) d\theta \int_{\phi} \prod_{j=1}^{T} p(\phi_j) \left(\prod_{n=1}^{N} p(w_n|\phi_{z_n}) \right) d\phi,$$

θ and ϕ being the random parameters of a multinomial over topics[14] and a multinomial over words respectively. The actual LDA distribution on documents, i.e. the distribution over topics in a document, is obtained by marginalizing out the topic variables in equation 4.7. Additionally, θ and ϕ follow a Dirichlet distribution.

4.4.2 Sampling

Resorting to the ideas of Blei et. al., Griffiths and Steyvers came up with a simple and efficient algorithm in [21] to approximate the LDA distribution using Gibbs sampling. Consider a document to contain different topics. Then a simple way to describe the contribution of these topics is, to view topics as a probability distributions over words and documents as a probabilistic mixture over these topics. Taking the number of topics to be fixed to T, the probability of a word is given by

$$p(w_i) = \sum_{j=1}^{T} p(w_i|z_i = j)\, p(z_i = j) \tag{4.10}$$

Here $p(w|z)$ is high for words that are important to "their" respective topics and $p(z)$ is "[...]the prevalence of those topics within a document."[15] Looking at the whole corpus and considering that the T different topics are represented by a vocabulary of V (unique) words, $p(w|z)$ can be represented as a set of T multinomial distributions ϕ over V words and is given by $p(w|z = j) = \phi_w^{(j)}$. $p(z)$ is then a set of D multinomial distributions θ over T topics and the topic probabilities for a word in document d is given by $p(z = j) = \theta_j^{(d)}$. Furthermore, they do not treat ϕ and θ as parameters to be estimated, instead they focus on the posterior distribution over the assignments of words to topics which is given by $p(\mathbf{z}|\mathbf{w})$. ϕ and θ are then approximated by analyzing the posterior distribution. They also add a Dirichlet prior on ϕ so that the probability model becomes:

$$\begin{aligned} w_i|z_i, \phi^{(z_i)} &\sim Discrete\left(\phi^{(z_i)}\right) \\ \phi &\sim Dirichlet(\beta) \\ z_i|\theta^{(d_i)} &\sim Discrete\left(\theta^{(d_i)}\right) \\ \theta &\sim Dirichlet(\alpha) \end{aligned}$$

In their model, α and β are specified as hyperparameters, defining the priors on θ and ϕ. Also, both are considered to be symmetric Dirichlet priors, i.e. both are not vectors as proposed in [7], but have a single value. They are conjugates to the multinomial distributions θ and ϕ, making it possible to compute the joint distribution $p(\mathbf{w}, \mathbf{z})$ by integrating out both multinomials. As

$$p(\mathbf{w}, \mathbf{z}) = p(\mathbf{w}|\mathbf{z})\, p(\mathbf{z})$$

[14] [7], p. 998
[15] [21], p. 5228

and ϕ and θ appear only in the first and second term respectively, separate integration is possible. Doing this yields

$$p(\mathbf{w}|\mathbf{z}) = \left(\frac{\Gamma(V\beta)}{\Gamma(\beta)^V}\right)^T \prod_{j=1}^{T} \frac{\prod_w \Gamma\left(n_j^{(w)} + \beta\right)}{\Gamma\left(n_j^{(\cdot)} + V\beta\right)} \tag{4.11}$$

for the first term where $n_j^{(w)}$ is the number times, word w has been assigned to topic j and $n_j^{(\cdot)}$ is the number of all words that have been assigned to topic j,and

$$p(\mathbf{z}) = \left(\frac{\Gamma(T\alpha)}{\Gamma(\alpha)^T}\right)^D \prod_{d=1}^{D} \frac{\prod_j \Gamma\left(n_j^{(d)} + \alpha\right)}{\Gamma\left(n_\cdot^{(d)} + T\alpha\right)} \tag{4.12}$$

where $n_j^{(d)}$ is the number of words from document d that have been assigned to topic j and $n_\cdot^{(d)}$ is the number of words in document d that have been assigned to any topic (and so effectively the length of document d). The overall goal is still to evaluate

$$p(\mathbf{z}|\mathbf{w}) = \frac{p(\mathbf{w},\mathbf{z})}{\sum_z p(\mathbf{w},\mathbf{z})},$$

the probability of a set of topics, given a set of words, i.e. a document. Because the sum in the denominator of this equation can not be split into factors and involves T^n terms, this is generally intractable. This is, where the application of Markov chain Monte Carlo and Gibbs sampling comes into place. Each state of the Markov chain is an assignment of values to the sampled variable $\mathbf{z}$. Recall that, in Gibbs sampling, the next state of the chain is reached by sampling all variables from their distribution conditioned on all other variables and the available data. This full conditional $p\left(z_i|\mathbf{z}_{\setminus i}, \mathbf{w}\right)$ is given by equations 4.11 and 4.12 and after simplifications[16] yields

$$p\left(z_i = j|\mathbf{z}_{\setminus i}, \mathbf{w}\right) \propto \frac{n_{\setminus i,j}^{(w_i)} + \beta}{n_{\setminus i,j}^{(\cdot)} + V\beta} \frac{n_{\setminus i,j}^{(d_i)} + \alpha}{n_{\setminus i}^{(d_i)} + T\alpha} \tag{4.13}$$

where $n_{\setminus i}^{(\cdot)}$ means the count excluding the current assignment of z_i. The first term of this equation corresponds to the probability of word w_i in topic j and the second term is the probability of topic j in document d_i. As the counts in equation 4.13 are the only information needed to compute the full conditional distribution, this can be done very efficiently by storing a sparse $V \times T$ matrix for counting the number of times a word has been assigned to a topic and another sparse $T \times D$ matrix for storing the number of times a topic occurred in a document. Applying the Gibbs sampling as described in section 4.3, i.e. finding an initial state of the chain, subsequently sampling from the data for a number of iterations (*burn-in period*) and then taking samples from the chain

[16] see Appendix C.1

at a suitable lag to ensure a low dependency, gives a set of samples from the posterior distribution $p(\mathbf{z}|\mathbf{w})$. With each of these samples, ϕ and θ can be estimated from the value of $\mathbf{z}$ by

$$\hat{\theta}_j^{(d)} = \frac{n_j^{(d)} + \alpha}{n_\cdot^{(d)} + T\alpha} \tag{4.14}$$

$$\hat{\phi}_j^{(w)} = \frac{n_j^{(w)} + \beta}{n_j^{(\cdot)} + V\beta}. \tag{4.15}$$

Griffiths et. al. also state, that these additionally "[...]correspond to the predictive distributions over new words w and new topics z conditioned on $\mathbf{w}$ and $\mathbf{z}$."[17]

4.5 The Author-Topic model

4.5.1 The model

The Author-Topic model is a probabilistic model extending LDA that can be used to include authorship information into the topic model proposed by LDA. In it, authors take the place of documents in the LDA model, in that each author(instead of each document) is assigned a multinomial distribution over topics and each topic is assigned a multinomial distribution over words. This results in the model being capable of examining both the content of documents as well as the interests of authors. Consider a group of authors $\mathbf{a}_d$ that decide to write a document d. For each word in a document, an author is chosen uniformly from $\mathbf{a}_d$. Now, a topic is chosen from a multinomial distribution over topics characteristic to that author and then the word is generated from the chosen topic.

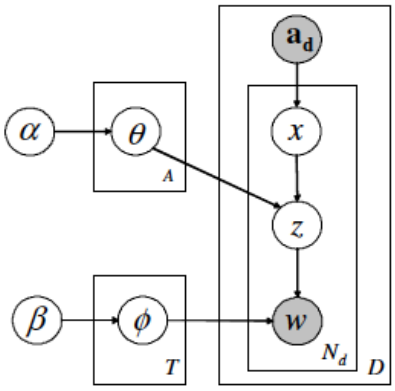


Figure 4.2: The Author-Topic model in plate notation

[17] [21], p. 5230

The plate notation of this model has been published in [40] and is shown in Figure 4.2. Here, x represents the author responsible for a specific word and is drawn from $\mathbf{a}_d$. Each author is associated with a distribution θ over topics, chosen from a symmetric Dirichlet(α) prior. The mixture weights corresponding to the chosen author are then used to select a topic z and a word is generated according to a distribution ϕ over words corresponding to this topic, drawn from a symmetric Dirichlet(β) prior. The complete generative process, due to [39], is:

1. for each author $a = 1, \ldots, A$ choose $\theta_a \sim Dir\,(\alpha)$
   for each topic $t = 1, \ldots, T$ choose $\phi_t \sim Dir\,(\beta)$,
2. for each document $d = 1, \ldots, D$:
   (a) for each word w_i indexed by $i = 1, \ldots, N_d$
      i. conditioned on $\mathbf{a}_d$ choose author $x_i \sim Uniform\,(\mathbf{a}_d)$,
      ii. conditioned on x_i choose topic $z_i \sim Discrete\,(\theta_{x_i})$,
      iii. conditioned on z_i choose word $w_i \sim Discrete\,(\phi_{z_i})$.

By estimating ϕ and θ, information about the typical topics, an author writes about, may be obtained and, additionally, a representation of the content of each document in terms of these topics. Formally, the overall probability of a corpus is given by

$$p\,(\mathcal{D}|\theta, \phi, \mathcal{A}) = \prod_{d=1}^{D} p\,(\mathbf{w}_d|\theta, \phi, \mathbf{a}_d) \tag{4.16}$$

with $\mathcal{D}$ being the set of all documents and $\mathcal{A}$ being the set of all authors appearing in the corpus. The probability of a single document can then be obtained by

$$\begin{aligned}
p\,(\mathbf{w}|\theta, \phi, \mathcal{A}) &= \prod_{n=1}^{N} p\,(w_n|\theta, \phi, \mathbf{a}_d) \\
&= \prod_{n=1}^{N} \sum_{a=1}^{A} \sum_{j=1}^{T} p\,(w_n, z_n = j, x_n = a|\theta, \phi, \mathbf{a}_d) \\
&= \prod_{n=1}^{N} \sum_{a=1}^{A} \sum_{j=1}^{T} p\,(w_n|z_n = j, \phi_j)\, p\,(z_n = j|x_n = a, \theta_a)\, p\,(x_n = a|\mathbf{a}_d) \\
&= \prod_{n=1}^{N} \frac{1}{A_d} \sum_{a \in \mathbf{a}_d} \sum_{j=1}^{T} p\,(w_n|z_n = j, \phi_j)\, p\,(z_n = j|x_n = a, \theta_a)
\end{aligned}$$

with A_d being the number of authors for a given document(i.e. $A_d = |\mathbf{a}_d|$). Recalling the notation used in the previous section, this leads to

$$p\,(\mathbf{w}|\theta, \phi, \mathcal{A}) = \prod_{n=1}^{N} \frac{1}{A_d} \sum_{x_n} \sum_{z_n} p\,(w_n|\phi_{z_n})\, p\,(z_n|\theta_{x_n})\,. \tag{4.17}$$

Integration over θ and ϕ and incorporating their joint probability distributions yields

$$p\left(\mathbf{w}|\alpha,\beta,\mathcal{A}\right)=\iint\prod_{a=1}^{A}p\left(\theta_a|\alpha\right)\prod_{j=1}^{T}p\left(\phi_j|\beta\right)\prod_{n=1}^{N}\frac{1}{A_d}\sum_{x_n}\sum_{z_n}p\left(w_n|\phi_{z_n}\right)p\left(z_n|\theta_{x_n}\right)d\theta d\phi \tag{4.18}$$

as the probability of a single document. Restructuring the equation and multiplying over all document leads to

$$\begin{aligned}p\left(\mathcal{D}|\alpha,\beta,\mathcal{A}\right)=&\int_{\theta}\prod_{a=1}^{A}p\left(\theta_a|\alpha\right)\prod_{d=1}^{D}\prod_{n=1}^{N_d}\sum_{x_{dn}}\sum_{z_{dn}}p\left(z_{dn}|\theta_{x_{dn}}\right)d\theta\\&\int_{\phi}\prod_{j=1}^{T}p\left(\phi_j|\beta\right)\prod_{d=1}^{D}\prod_{n=1}^{N_d}\sum_{z_{dn}}\sum_{x_{dn}}p\left(w_{dn}|\phi_{z_{dn}}\right)d\phi\end{aligned}$$

Under consideration of exchangeability, the joint probability distribution of $\mathbf{w},\mathbf{z}$ and $\mathbf{x}$ for a single document is then given by

$$p\left(\mathbf{w},\mathbf{z},\mathbf{x}|\alpha,\beta,\mathcal{A}\right)=\iint\prod_{a=1}^{A}p\left(\theta_a|\alpha\right)\prod_{j=1}^{T}p\left(\phi_j|\beta\right)\prod_{n=1}^{N}\frac{1}{A_d}p\left(w_n|\phi_{z_n}\right)p\left(z_n|\theta_{x_n}\right)d\theta d\phi \tag{4.19}$$

where θ consists of A topic distributions(one for each author) and ϕ consists of T word distributions(one for each topic). Again, θ and ϕ follow a Dirichlet distribution.

4.5.2 Sampling

As the Author-Topic model is based on Latent Dirichlet Allocation, the sampling process described in [39] also expands the Gibbs sampling process of LDA. In the Author-Topic model, there are two sets of latent variables, z and x, and each (z_i,x_i) pair is drawn as a block, conditioned on all other variables through

$$p\left(z_i=j,x_i=k|w_i=m,\mathbf{z}_{\backslash i},\mathbf{x}_{\backslash i},\mathbf{w}_{\backslash i},\mathbf{a}_d\right)\propto\frac{C_{mj}^{WT}+\beta}{\sum_{m'}C_{m'j}^{WT}+V\beta}\frac{C_{kj}^{AT}+\alpha}{\sum_{j'}C_{kj'}^{AT}+T\alpha} \tag{4.20}$$

Here, $z_i=j$ and $x_i=k$ represent assignments of the ith word in a document to topic j and author k respectively. $w_i=m$ represents that the ith word in a document is the mth word in the vocabulary and $\mathbf{z}_{\backslash i}$ and $\mathbf{x}_{\backslash i}$ represent the all topic and author assignment excluding the current one. C_{mj}^{WT} and C_{kj}^{AT} are the number of times a word m has been assigned to topic j and the number of times a an author k is assigned to topic j respectively. Equation 4.20 is the conditional probability derived by marginalizing out the random variables θ and ϕ, that, in the Author-Topic model represent the probability of a word given a topic(θ) and the probability of a topic given an author(ϕ). That means, that the markov chain constructed, is designed to converge to this posterior

distribution over $\mathbf{x}$ and $\mathbf{z}$ conditioned on the training data $\mathcal{D}$, α and β. Again, tracking only small amounts of information of a corpus is sufficient. For using the algorithm, only a $V \times T$(word by topic) and a $A \times T$(author by topic) matrix are needed, both of which can be stored using a sparse matrix storage format. The algorithm is started by randomly assigning words to topics and authors(the initial configuration of the markov chain) and in each subsequent iteration of it, equation 4.20 is used to reassign topics and authors for each individual word in the corpus. Similarly to the sampling process in LDA, the chain is run for certain number of iterations(burn in) to let the chain reach its equilibrium distribution. After that, samples are taken from the posterior distribution and the random variables θ and ϕ can be estimated from them by

$$\phi_{mj} = \frac{C_{mj}^{WT} + \beta}{\sum_{m'} C_{m'j}^{WT} + V\beta} \tag{4.21}$$

$$\theta_{kj} = \frac{C_{kj}^{AT} + \alpha}{\sum_{j'} C_{kj'}^{AT} + T\alpha} \tag{4.22}$$

4.6 The Author-Recipient-Topic model

The Author-Recipient-Topic(ART) model has been introduced in [35] as a Bayesian network, able to model the distinct features of social networks in email messages. This model also builds on LDA and, furthermore on the Author-Topic model, although the authors argue that both preceding models are not able to capture social network connectivity. To overcome this problem, they added

> [...] the key attribute that the distribution over topics is conditioned distinctly on both the sender and recipient - steering the discovery of topics according to the relationships between people.[18]

Also, they note that

> (t)here has not [...] been significant work by researchers with backgrounds in statistical natural language processing, nor analysis that captures the richness of the *language contents* [...] and other high dimensional specifics of the interactions between people.[19]

Although tightly knit groups of users may be identified by looking at the email messages they trade amongst each other, these information are not enough to discover all roles they might fulfill. The cause of this is, that in a group of users, email messages are sent to each other in a "roughly symmetric fashion"[20] and therefore, the users appear to fill similar roles. Now by looking at the content of these messages as well, a more precise distinction of the users' roles becomes possible. Imagine a manager that communicates

[18] [36]
[19] ibid.
[20] ibid.

with the team by email. Clearly, it is possible to assign the manager and the team to a group that is closely connected. Analyzing the content of their messages now allows the distinction between team and manager and so gives a more realistic picture of this real life situation.

4.6.1 The model

The ART model treats the words in a message to be generated given the author of the message and a set of recipients. It is similar to the Author-Topic model, but conditions the topic distribution of individual messages on the author and individual recipients, instead of just authors as in the Author-Topic model. This incorporates the social structure of the authors and recipients of mail messages into the process of topic discovery. In this model, each topic is a multinomial probability distribution over words, whereas each author-recipient pair is assigned a multinomial distribution over topics. The benefit of the model(and in fact also of the Author-Topic model) is, that people can be compared by comparing the preconditioned topic distributions by means of probability distribution similarity and also can be clustered into groups by this means. ART is modeled due to the special structure of email messages opposed to other text documents. A message has one unique author and (in general) a set of recipients. The authors describe it to be able to

> [...]capture(s) topics and the directed social network of senders and recipients by conditioning the multinomial distribution over topics distinctly on both the author and one recipient of a message.[21]

The model is also designed as a Bayesian network and models both the text content of the messages sent in a social network as well as the network itself, retaining the direction of ties in it. The plate notation[22] of this network is given in Figure 4.3
The model's generative process is given by:

1. for each author-recipient pair (a,x) with $a = 1, \ldots, A$ and $x = 1, \ldots, A$ choose $\theta_{ax} \sim Dir(\alpha)$
2. for each topic $t = 1, \ldots, T$ choose $\phi_t \sim Dir(\beta)$,
3. for each message d
   (a) observe the author a_d and the set of recipients $\mathbf{r}_d$
   (b) for each word w in d
      i. choose recipient $x \sim Uniform(\mathbf{r}_d)$
      ii. choose topic $z \sim \theta_{a_d x}$
      iii. choose word $w \sim \phi_z$

[21] ibid.
[22] [35]

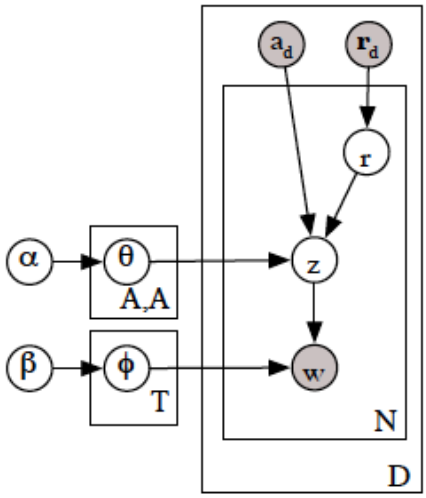


Figure 4.3: The Author-Recipient-Topic model in plate notation

The joint probability distribution of θ and ϕ is given by

$$p\left(\theta,\phi,\mathbf{x},\mathbf{z},\mathbf{w}|\alpha,\beta,\mathbf{a},\mathbf{r}\right)=\prod_{i=1}^{A}\prod_{j=1}^{A}p\left(\theta_{ij}|\alpha\right)\prod_{t=1}^{T}p\left(\phi_{t}|\beta\right)\prod_{d=1}^{D}\prod_{i=1}^{N_d}\left(p\left(x_{di}|\mathbf{r}_{d}\right)p\left(z_{di}|\theta_{a_d x_{di}}\right)p\left(w_{di}|\phi_{z_{di}}\right)\right)$$

and by marginalizing out θ and ϕ by integration and summing over $\mathbf{z}$ and $\mathbf{x}$, the marginal distribution of a corpus is obtained to be

$$p\left(\mathbf{w}|\alpha,\beta,\mathbf{a},\mathbf{r}\right)$$
$$=\iint\prod_{i=1}^{A}\prod_{j=1}^{A}p\left(\theta_{ij}|\alpha\right)\prod_{t=1}^{T}p\left(\phi_{t}|\beta\right)\prod_{d=1}^{D}\prod_{i=1}^{N_d}\sum_{x_{di}=1}^{A}\left(p\left(x_{di}|\mathbf{r}_{d}\right)\sum_{z_{di}=1}^{T}\left(p\left(z_{di}|\theta_{a_d x_{di}}\right)p\left(w_{di}|\phi_{z_{di}}\right)\right)\right)d\theta d\phi$$

4.6.2 Sampling

In [36], the authors describe a Gibbs sampling procedure for the ART model. To avoid repetition, only a brief description of the sampling process is given here. The distribution to be estimated is

$$p\left(x_{di},z_{di}|\mathbf{x}_{\backslash di},\mathbf{z}_{\backslash di},\alpha,\beta,\mathbf{a},\mathbf{r}\right)\propto\frac{\alpha_{z_{di}}+n_{a_d x_{di} z_{di}}-1}{\sum_{t=1}^{T}\left(\alpha_t+n_{a_d x_{di} t}\right)-1}\frac{\beta_{w_{di}}+m_{z_{di} w_{di}}-1}{\sum_{v=1}^{V}\left(\beta_v+m_{z_{di} v}\right)-1}$$

or, when α and β are treated as scalar values instead of vectors(similar to both models before), given by

$$p\left(x_{di},z_{di}|\mathbf{x}_{\backslash di},\mathbf{z}_{\backslash di},\alpha,\beta,\mathbf{a},\mathbf{r}\right)\propto\frac{\alpha+n_{a_d x_{di} z_{di}}-1}{\sum_{t=1}^{T}n_{a_d x_{di} t}+T\alpha-1}\frac{\beta+m_{z_{di} w_{di}}-1}{\sum_{v=1}^{V}m_{z_{di} v}+V\beta-1}$$

where n_{ijt} represents the number of tokens assigned to topic t and the author-recipient pair (i, j) and m_{tv} represents the number of times, word v has been assigned top topic t. Collecting samples over a training set allows to compute posterior estimates of θ and ϕ via

$$\hat{\theta}_{ijz} = \frac{\alpha + n_{ijz}}{\sum_{t=1}^{T} n_{ijz} + T\alpha}$$
$$\hat{\phi}_{tw} = \frac{\beta + m_{tw}}{\sum_{v=1}^{V} m_{tv} + V\beta}$$

4.7 Applying Latent Dirichlet Allocation

By applying the LDA model, a topic probability distribution is determined for each document. To generate a topic probability distribution for authors, all documents of an author are determined and their probability distributions are averaged. Finally, the similarity between the so generated topic probability distributions of authors is computed and compared to the actual path length in the social network. In contrast to the bag-of-words model, cosine distance shows only a minor correlation to network path length using LDA(Fig. 4.4). For values up to a cosine distance of 0.7 the network path length lies around the average path length of the author network, whereas at higher values, the network path length drops. Recall, that cosine distance normally omits zero valued components in vectors, but in this case topic probability distributions are considered that have a value unequal zero for all components, hence for every topic, there is a defined probability. This implies, that only for authors with overlapping high probability topics, higher cosine distance values are produced. Additionally, KL-divergence based distance measures are lower for these authors, resulting in lower distance values for higher cosine distances(Fig. 4.5).
Using topic models, the KL-based distances stay indifferent for low values of cosine distance and then decrease as cosine distance increases, whereas cosine distance still gives a similar, although not as distinct as with using bag-of-words, and suitable approximation between similarity and path length (Fig. 4.4).
On the contrary, KL-divergence based distance measures, that have been explicitly proposed for comparing probability distributions, naturally perform better. Also, as a sign for better estimation of path lengths, both information radius(Fig. 4.6) and skew divergence(Fig. 4.7) no longer show shorter path lengths for higher distance values as in the bag-of-words model, but for low values even show a growing path distance as similarity decreases, i.e. the distance value increases, and stabilize at around average path length.

Skew divergence seems to perform slightly better in this task, as the path distance interval is narrower than with using information radius.

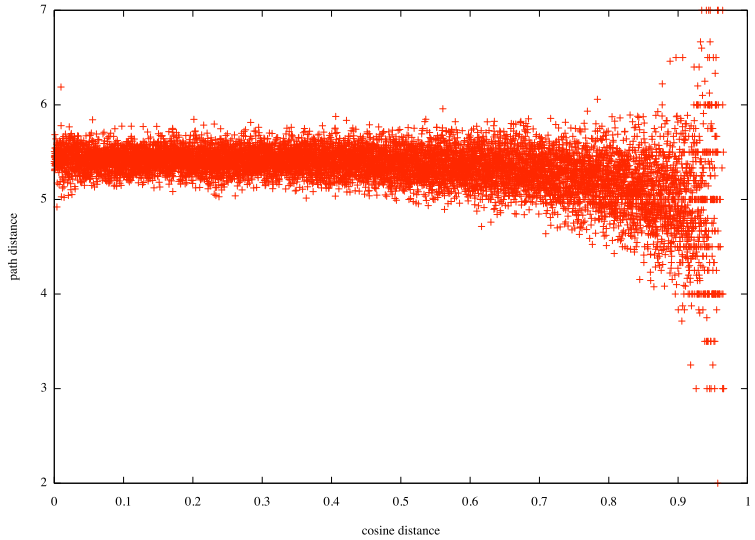
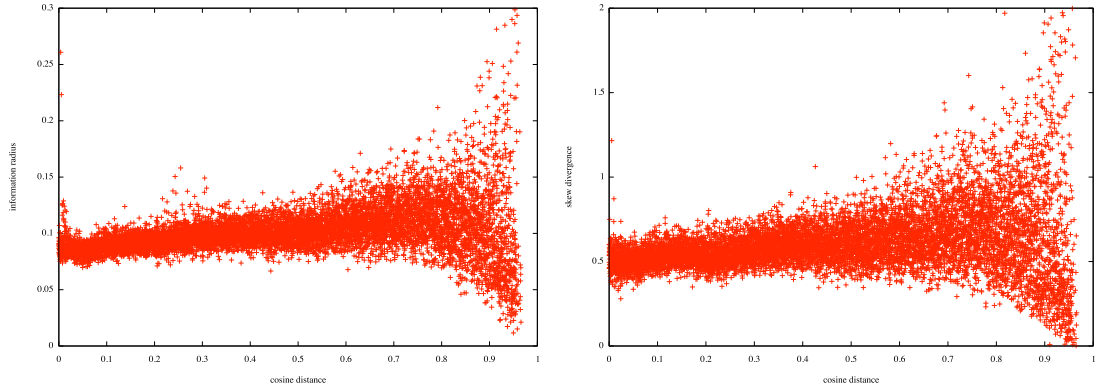


Figure 4.4: Cosine distance against network path length - LDA model



(a) Information radius against cosine distance **(b)** Skew divergence against cosine distance

Figure 4.5: KL-divergence based distance measures against cosine distance - LDA model

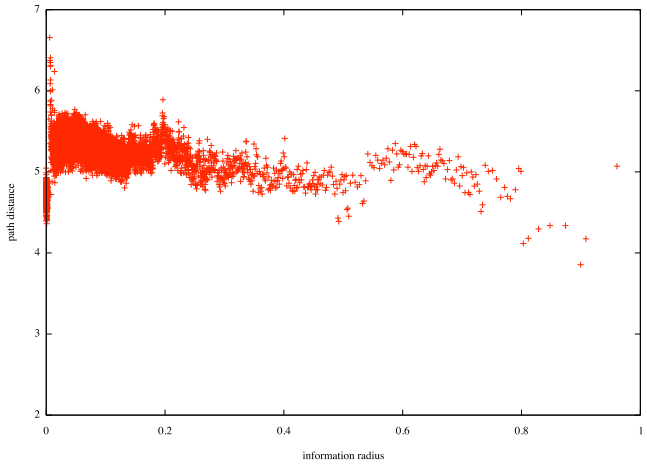


Figure 4.6: Information radius against network path length - LDA model

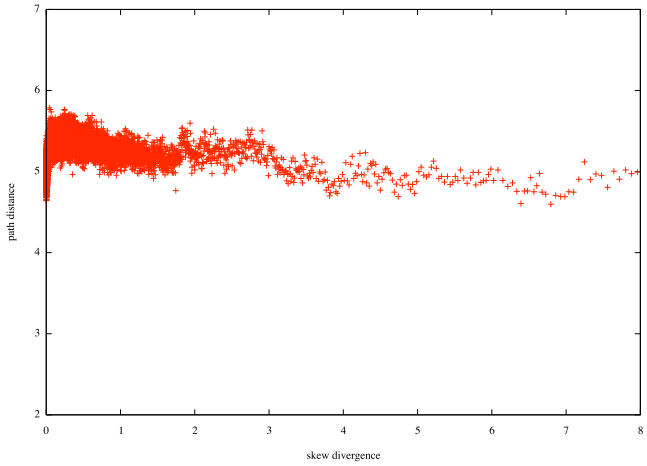


Figure 4.7: Skew divergence against network path length - LDA model

4.8 Applying the Author-Topic model

The benefit of the AT model is, that for each author a topic probability distribution is directly given by the model. Hence, it is straight forward to compute similarity between different authors and compare it to path lengths in the social network. Doing so, again shows that cosine distance is not suitable for facing this task, results using it(Fig. 4.8) strongly resemble those using LDA(Fig. 4.4), although in the AT model, path distance continuously decreases with increasing cosine distance and also keeps within a narrower interval.

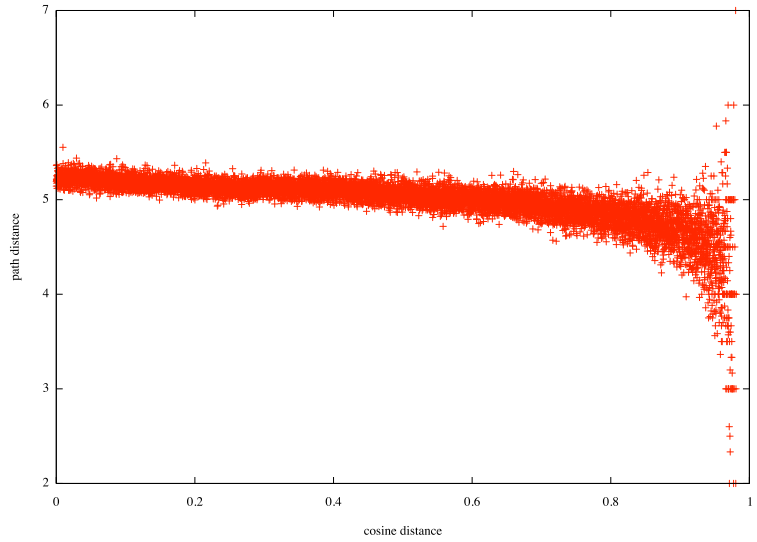
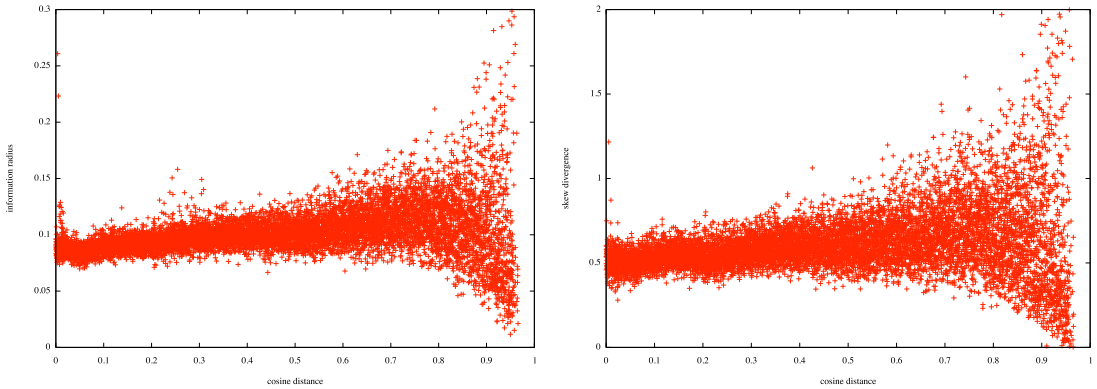


Figure 4.8: Skew divergence against network path length - AT model

Looking at the correlation between cosine distance and both KL-divergence based distance measures(Fig. 4.9), shows that both divergence based measures stay constant for low cosine distances and decrease for higher ones, which resembles results from the other topic model used. But using the AT model, divergence based measures center around cosine distances more tightly.
Considering the correlation between KL-divergence based measures and the path distance in the network, it can be seen, that for low divergences, path distance increases and stagnates at around average path length. The fact, that with increasing similarity measure values(and hence with less similarity), path distances in the social network grow, is shown even better than in the LDA model.
In the AT model, as opposed to LDA, the information radius distance measure backs expectations to a higher degree and thus performs better.



(a) Information radius against cosine distance **(b)** Skew divergence against cosine distance

Figure 4.9: KL-divergence based distance measures against cosine distance - AT model

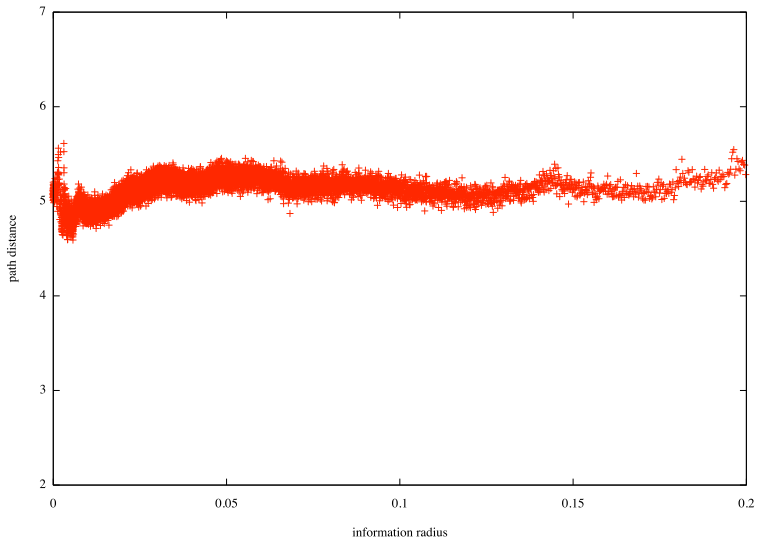


Figure 4.10: Information radius against network path length - AT model

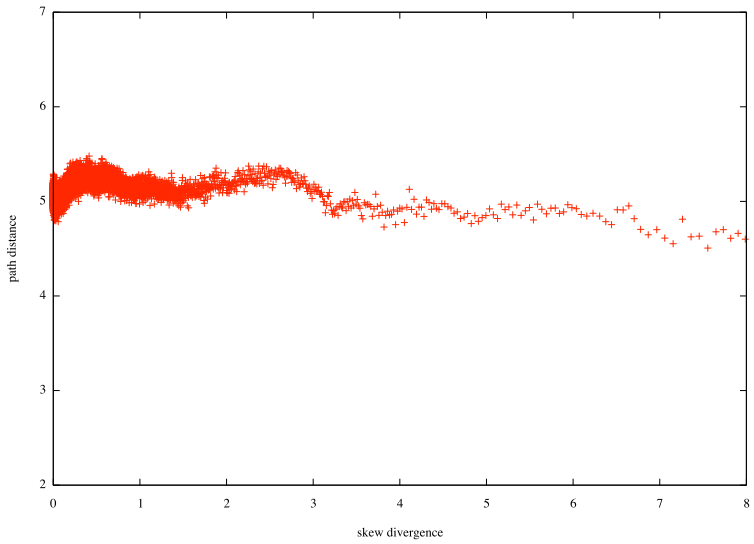


Figure 4.11: Skew divergence against network path length - AT model

Chapter 5

Conclusion

In this thesis, different techniques for natural language analysis have been described and studied. Furthermore, an introduction into the theory of social networks has been given and the possibility of inferring social networks by analyzing text content has been examined. Results encourage that this is the case when using probabilistic topic modeling to determine author specific probability distributions over topics. The models under examination were a simple frequentist bag-of-words model, Latent Dirichlet Allocation and the Author-Topic model. The least two have shown to produce an approximation of the social networks extracted as described in chapter 2.

5.1 Comparison of used techniques

For a better interpretability, results from the three models are compared. For a better overview, bezier fitted curves are used to show value distribution. As already signified before, cosine distance as a measure of similarity and thus for path length looses its applicability with more complex models. For both topic models, cosine distance does not behave as clearly and distinct related to path distance as it does in the bag-of-words model(Fig. 5.1). Interestingly, the path distance stabilizes at a higher value for the LDA model than for the AT model, which might be caused by the process of simply averaging over all documents of an author instead of directly using author topic probabilities provided by the AT model.
Also the correlation between cosine distance and KL-divergence based distances change with the models used.
KL-divergence based distance measures obviously are no option when dealing with term frequency models like bag-of-words. For topic models on the other side, both measures show an expected correlation between their values and path distances, where a lower author similarity(and thus fewer shared interests) result in a higher distance in the social network. In comparison, the skew distance measure exhibits a smoother behavior, especially when inspecting the AT model. Here the slope of the correlation graph is lower, which is also more plausible as there is no expected threshold of skew divergence at which the path distance ”jumps” from a low to a high level.

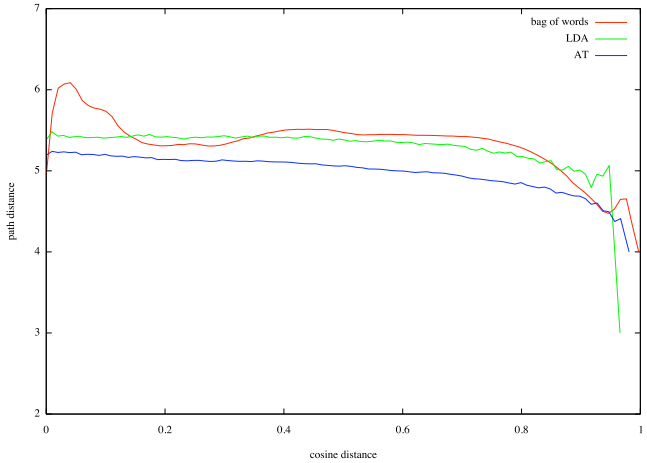
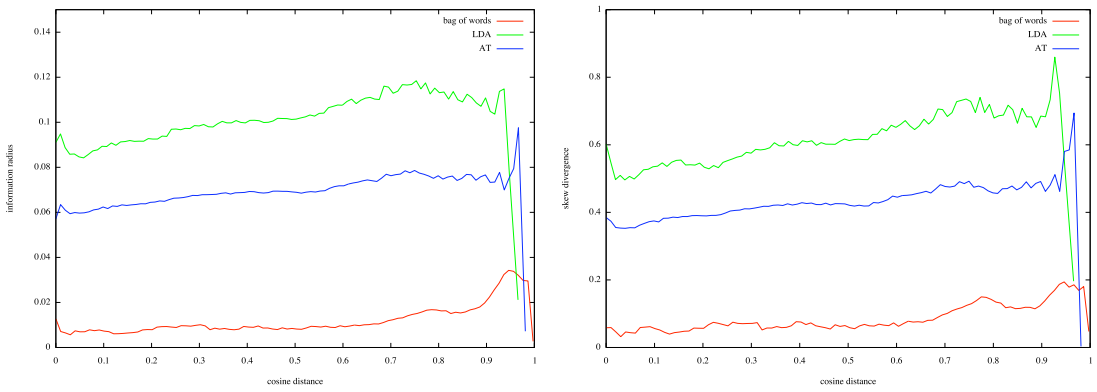


Figure 5.1: Cosine distance against network path length - all models



(a) Information radius against cosine distance **(b)** Skew divergence against cosine distance

Figure 5.2: KL-divergence based distance measures against cosine distance - all models

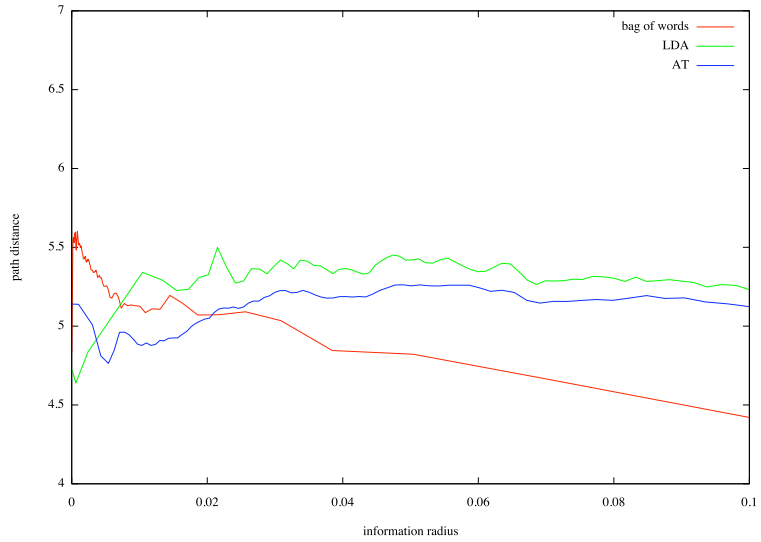


Figure 5.3: Information radius against network path length - all models

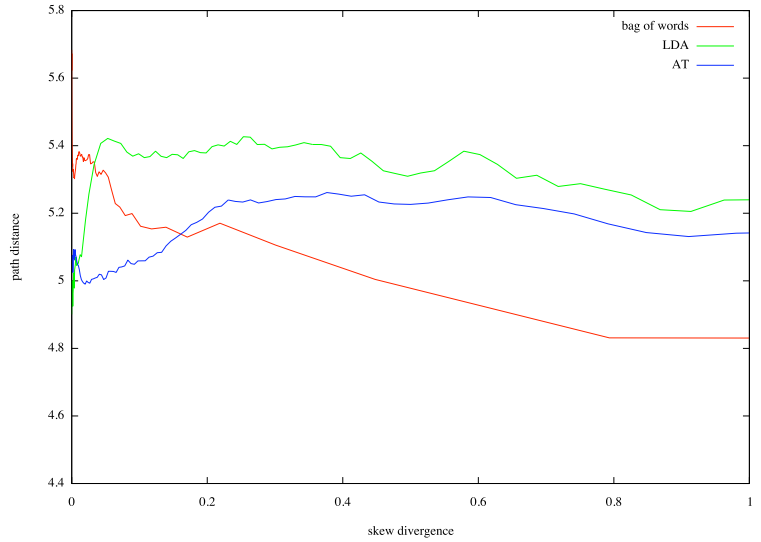


Figure 5.4: Skew diversion against network path length - all models

5.2 A proposal for the Community-Author-Topic model

As could be seen, the correlation between authors' topic probability distributions(or some sort of approximation to it as in the bag-of-words model and in LDA) and path lengths in a social network formed by these authors grows[1] with the complexity of the used models. Furthermore, using topic models also gives the possibility to classify previously unknown documents in terms of topics and author interests(in AT model). The knowledge of the structure of social networks, namely the local clustering of cliques, proposes a more complex three level topic model comparable to the Author-Recipient-Topic model, which integrates the following generative process:

1. for each author a with $a = 1, \ldots, A$, choose $\eta_a \sim Dir(\alpha)$
2. for each community c with $c = 1, \ldots, C$, choose $\theta_c \sim Dir(\beta)$
3. for each topic t with $t = 1, \ldots, T$, choose $\phi_t \sim Dir(\gamma)$
4. for each document d with $d = 1, \ldots, D$
   (a) observe author a_d of document d
   (b) choose community $x_d \sim \eta_{a_d}$
   (c) for each word w with $w = 1, \ldots, N_d$
      i. choose topic $z_{dn} \sim \theta_{x_d}$
      ii. choose word $w_{dn} \sim \phi_{z_{dn}}$

A graphical representation of this model, which is called the Community-Author-Topic model is given in plate notation in Fig. 5.5.

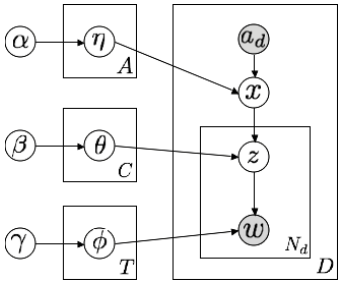


Figure 5.5: Community-Author-Topic model in plate notation

Following plate notation, the joint posterior distribution of η, θ and ϕ for a document is given by

$$p(\mathbf{w}, \mathbf{c}, \mathbf{z}, \eta, \theta, \phi | \alpha, \beta, \gamma, a_d) = \prod_{a=1}^{A} p(\eta_a|\alpha) \prod_{c=1}^{C} p(\theta_c|\beta) \prod_{t=1}^{T} p(\phi_t|\gamma) \, p(x|\eta_{a_d}) \prod_{n=1}^{N} p(z_n|\theta_x) \, p(w_n|\phi_{z_n}) \tag{5.1}$$

[1] i.e. less similar authors show a greater network path distance than more similar authors

Integrating over η, θ and ϕ and taking all D documents into account yields

$$p(\mathbf{w}, \mathbf{c}, \mathbf{z}|\alpha, \beta, \gamma, a_d) =$$
$$\iiint \prod_{a=1}^{A} p(\eta_a|\alpha) \prod_{c=1}^{C} p(\theta_c|\beta) \prod_{t=1}^{T} p(\phi_t|\gamma) \prod_{d=1}^{D} p(x_d|\eta_{a_d}) \prod_{n=1}^{N_d} p(z_{dn}|\theta_{x_n}) p(w_{dn}|\phi_{z_{dn}}) \, d\eta d\theta d\phi \tag{5.2}$$

As shown in Appendix C.2, estimates for η, θ and ϕ are given by

$$\hat{\eta}_{a,c} = \frac{n_c^a + \alpha}{n_{(\cdot)}^a + C\alpha} \tag{5.3}$$

$$\hat{\theta}_{c,j} = \frac{n_c^j + \beta}{n_c^{(\cdot)} + T\beta} \tag{5.4}$$

$$\hat{\phi}_{j,v} = \frac{n_v^j + \gamma}{n_{(\cdot)}^j + V\gamma} \tag{5.5}$$

where n_c^a is the number of times community c has been assigned to author a, n_c^j is the number of times, topic j has been assigned to community c and n_v^j is the number of times, word v has been assigned to topic j.

5.3 Further work

Their has been no chance so far, to implement and test the Community-Author-Topic model, so a first step in further work would be to do so. Results from this should show, whether this approach is useful for (more or less) exactly inferring social network by pure content analysis. This would enable to provide a service of author proposal by means of short paths in an underlying social network instead of content-motivated decisions only. Another advancement applicable to all described probabilistic model is described in [47]. It is possible to infer the number of topics(in the LDA, AT and ART models) and the number of communities in the CAT model by using hierarchical dirichlet processes. In [47] a technique called the stick breaking procedure is applied to arrive at a suitable value for these variables. In all examined cases, these variables have been treated as known and fixed. As Teh et. al. state,

> [...]the number of mixture components is unknown a priori and is to be inferred from the data.[...][2]

Another possibility to arrive at suitable values for this number is, to use what is called the "Chinese Restaurant process"[3]. To further test the ability to correctly predict social networks, the data available could be partitioned into a training set and test data. This

[2] ibid.
[3] ibid.

would allow to estimate relations between authors of a week, based on the data of all other weeks and then test how well the model predicts the actual existent network in the data.

Bibliography

[1] D. Aldous. *Exchangeability and related topics.* In: *École d'été de probabilités de Saint-Flour, XIII.* Springer, 1985.

[2] S. Ali and S. Silvey. *A general class of coefficients of divergence of one distribution from another.* Journal of the Royal Statistical Society. Series B (Methodological), (1966).

[3] T. Bayes and M. Price. *An Essay towards Solving a Problem in the Doctrine of Chances.* Philosophical Transactions (1683-1775), (1763).

[4] C. M. Bishop. *Latent variable models.* In: M. I. Jordan (Ed.), *Learning in Graphical Models.* MIT Press, 1999, pp. 371–403.

[5] C. M. Bishop. *Pattern Recognition and Machine Learning.* Springer, 2006.

[6] D. M. Blei, M. I. Jordan and A. Y. Ng. *Hierarchical Bayesian Models for Applications in Information Retrieval.* In: J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith and M. West (Eds.), *Bayesian statistics 7, Proceedings of the seventh Valencia International Meeting*, vol. 7. 2003, pp. 25–43.

[7] D. M. Blei, A. Ng and M. Jordan. *Latent dirichlet allocation.* The Journal of Machine Learning Research, (2003).

[8] B. Bollobás. *Random graphs.* Cambridge University Press, 2001, 2nd ed.

[9] K. Burton, A. Java and I. Soboroff. *The ICWSM 2009 Spinn3r Dataset.* In Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009), San Jose, CA, 2009.

[10] B. Carpenter, G. Fox, S. Ko and S. Lim. *mpiJava 1.2: API Specification.* Rice University, September, (1999).

[11] R. T. Cox. *Probability, Frequency and Reasonable expectation.* American Journal of Physics, vol. 14 (1946)(1):pp. 1–13.

[12] S. Deerwester, S. Dumais, T. Landauer, G. Furnas and R. Harshman. *Indexing by latent semantic analysis.* Journal of the American Society for Information Science, vol. 41 (1990)(6):pp. 391–407.

[13] P. Erdös and A. Renyi. *On random graphs.* Publicationes Mathematicae Debrecen, vol. 6 (1959):pp. 290–297.

[14] L. Euler. *Solutio problematis ad geometriam situs pertinentis.* Commentarii academiae scientiarum imperialis Petropolitanae, vol. 8 (1741):pp. 128–140.

[15] B. de Finetti, A. Machí and A. Smith. *Theory of probability: a critical introductory treatment.* Wiley, 1974, 2nd ed.

[16] M. Flynn. *Some computer organizations and their effectiveness.* IEEE Transactions on Computers, (1972).

[17] C. C. Foster, A. Rapoport and C. J. Orwant. *A study of a large sociogram II. Elimination of free parameters.* Behavioral science, vol. 8 (1963):pp. 56–65.

[18] I. Foster. *Designing and Building Parallel Programs: Concepts and Tools for Parallel Software Engineering.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1995.

[19] S. Geman and D. Geman. *Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images.* IEEE Transactions on Pattern Analysis and Machine Intelligence, (1984)(6):pp. 721—741.

[20] M. S. Granovetter. *The Strength of Weak Ties.* American Journal of Sociology, vol. 78 (1973)(6):p. 1360.

[21] T. Griffiths and M. Steyvers. *Finding scientific topics.* Proceedings of the National Academy of Sciences, (2004).

[22] T. Griffiths, M. Steyvers and J. Tenenbaum. *Topics in semantic representation.* Psychological review, vol. 114 (2007)(2):pp. 211–244.

[23] J. Guare. *Six Degrees of Separation: A play.* New York: Vintage, 1990.

[24] F. Harary. *Status and Contrastatus.* Sociometry, vol. 22 (1959)(1):pp. 23–43.

[25] M. Hearst. *TextTiling: Segmenting text into multi-paragraph subtopic passages.* Computational linguistics, vol. 23 (1997)(1):pp. 33–64.

[26] C. Hierholzer and C. Wiener. *Ueber die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren.* Mathematische Annalen, vol. VI (1873):pp. 30–32.

[27] T. Hofmann. *Probabilistic latent semantic indexing.* Proceedings of the Twenty-Secong Annual International SIGIR Conference, (1999).

[28] E. T. Jaynes and G. L. Bretthorst. *Probability Theory: The Logic of Science.* Cambridge University Press, 2003.

[29] F. Karinthy. *Chains*, 1929.

[30] J. S. Kleinfeld. *The Small World Problem.* Society, vol. 39 (2002)(2):pp. 61–66.

[31] S. Kullback and R. Leibler. *On Information and Sufficiency.* The Annals of Mathematical Statistics, vol. 22 (1951)(1):pp. 79–86.

[32] L. Lee. *On the effectiveness of the skew divergence for statistical language analysis.* Artificial Intelligence and Statistics, (2001):pp. 65–72.

[33] J. Lin. *Divergence measures based on the Shannon entropy.* IEEE Transactions on Information theory, vol. 37 (1991)(1):pp. 145–151.

[34] C. D. Manning and H. Schütze. *Foundations of statistical natural language processing.* MIT Press, 1999.

[35] A. McCallum, A. Corrada-Emmanuel and X. Wang. *Topic and role discovery in social networks.* Proceedings of IJCAI, (2005).

[36] A. McCallum, X. Wang and A. Corrada-Emmanuel. *Topic and role discovery in social networks with experiments on Enron and academic email.* Journal of Artificial Intelligence Research, (2007)(29).

[37] S. Milgram. *The Small World Problem.* Psychology today, vol. 1 (1967)(1).

[38] D. Newman, P. Smyth and M. Steyvers. *Scalable Parallel Topic Models.*

[39] M. Rosen-Zvi, T. Griffiths and M. Steyvers. *Learning Author Topic Models from Text Corpora.* The Journal of Machine Learning Research, (2005).

[40] M. Rosen-Zvi, T. Griffiths, M. Steyvers and P. Smyth. *The author-topic model for authors and documents.* Proceedings of the 20th conference on Uncertainty in Artificial Intelligence, (2004).

[41] G. Salton and C. Buckley. *Term-weighting approaches in automatic text retrieval.* Information processing & management, vol. 24 (1988)(5):pp. 513–523.

[42] G. Salton and M. J. McGill. *Introduction to modern information retrieval.* McGraw-Hill, 1983.

[43] J. Skvoretz and T. J. Fararo. *Connectivity and the Small World Problem.* In: M. Kochen (Ed.), *The Small World*, chap. 15. Ablex, 1989, pp. 296–326.

[44] I. de Sola Pool and M. Kochen. *Contacts and Influence.* Social Networks, vol. 1 (1978/79):pp. 5–51.

[45] R. Solomonoff and A. Rapoport. *Connectivity of random nets.* Bulletin of Mathematical Biology, vol. 13 (1951)(2):pp. 107–117.

[46] S. M. Stigler. *The History of Statistics.* Belknap Press/Harvard University Press, 1986.

[47] Y. Teh, M. Jordan, M. Beal and D. Blei. *Hierarchical dirichlet processes.* Journal of the American Statistical Association, (2006).

[48] J. Travers and S. Milgram. *An Experimental Study of the Small World Problem.* Sociometry, vol. 32 (1969)(4):pp. 425–443.

[49] S. Wasserman and K. Faust. *Social Network Analysis: Methods and applications.* Cambridge University Press, 1994.

[50] D. J. Watts. *Small worlds: The Dynamics of Networks between Order and Randomness.* Princeton University Press, 1999.

[51] Wikipedia. *Seven Bridges of Königsberg.* Online, accessed 9 July 2009. URL `http://en.wikipedia.org/w/index.php?title=Seven_Bridges_of_K%C3%B6nigsberg&oldid=302172204`.

[52] R. J. Wilson and J. J. Watkins. *Graphs: An Introductory Approach : a First Course in Discrete Mathematics.* Wiley, 1990.

List of Figures

Appendix A

A framework for distributed problem solving

Although using only a subset of the whole data set, it is still a challenge to deal with this amount of data. For the purpose of importing blog posts into a relational database and analyzing the data afterwards, a framework for distributed computing has been developed and used. The software uses the Message Passing Interface(MPI) and the master worker paradigm to distribute and recollect computation units and results, respectively. MPI is the most prominent technology used in high-performance computing today. It is not a programming language, rather it is a library that enables users of a specific programming language to use message passing by calling the library's subprograms. Currently there exist implementations in C that can be called from C, C++ and Fortran programs as well as in Python. Additionally, Bryan Carpenter introduced a JAVA JNI binding to the C library of MPI in [10], that enables JAVA users to also benefit from the possibilities of large computer clusters.

A.1 The Message Passing Interface

MPI is a protocol used for message-based communication on distributed memory systems. That is, it provides the ability to exchange messages between processes on different machines, working on the same problem in parallel and using shared (hard disk) memory. Thus parallelism is achieved by passing messages, which is no more than explicitly transmitting data from one process to another. MPI supports both point-to-point, as well as collective communication. During collective communication, messages can be sent from one source to many receivers or received by only one receiver from many sources. In point-to-point communication there always exists exactly one source and receiver, respectively. Following Flynn's taxonomy[1], a multiple instruction multiple data stream architecture has been implemented, more specifically the Single Process, Multiple Data technique. Using this, the problem to solve is divided into smaller subproblems that can

[1] [16]

be processed by a single machine each, resulting in shorter processing time.

A.2 Implementation

As mentioned, the implementation of the software uses the Master-Worker paradigm. Here, a supervising process, the master, takes care of splitting up the main task into smaller subtasks and distributing the subtasks to all other available processes, the workers. Furthermore, it recollects the results from the workers and is responsible for combining them to the original task's result. In case of data import, the master receives the directory of the xml files to import, scans this directory and all subdirectories for any xml files and generates a list of files to import. After that, the master process checks how many worker processes are available and splits up the file list accordingly. Each worker process is given one part of the list of files that is to be imported into the relational database. The worker then independently process the work flow chain described in chapter 1.2.2. During this task, no communication between the worker processes is needed, as the relational database takes care of any synchronization and data parallelism issues. This kind of parallelized task is also known as an embarrassingly parallel workload[2].

Parallelization of topic models has first been described in [38]. To implement this, again the master-worker paradigm is used. The master process splits up the corpus into smaller sets of documents and sends them to the worker processes. Additionally, the whole vocabulary used in the model is determined, as well as all authors whose documents appear in the corpus, and are sent via broadcast to all processes by the master. Next, each worker process initiates the markov chain and starts it. After each iteration of the chain, the different count matrices are sent from the workers back to the master, are combined and then redistributed back to the workers. Other algorithms, such as computing the current log-likelihood of the model, given the data, are also applied in this step. In the next iteration, all workers then start with updated count matrices, keeping differences between a topic model run on a single processor and that run on a computer cluster negligibly low.

[2] [18], section 1.4.4

Appendix B

Comparing author similarity to path distance

The goal is now, to inspect, if authors, "judged" solely on the text content they publish and the resulting similarities between them, can be linked to the distances they take in the empirically found social networks. To do this, a suitable way to compute these similarities has to be found. In literature, several ways of defining a similarity are to be found, namely *cosine distance*, *Kullback-Leibler divergence*, *Jenson-Shannon divergence* and *skew divergence*(among many others, see [34], 8.5.1 for more). The similarity in all of these forms is computed between two vectors of the same length. Additionally, all but the cosine distance measure are distance measures used for comparing probability distributions. Hence two vectors $\vec{i}$ and $\vec{j}$, both having dimension n are considered in all measures.

The cosine distance is often used in information retrieval tasks[1], where the vectors $\vec{i}$ and $\vec{j}$ mostly represent term frequency vectors. It is equivalent to the cosine of the angle between $\vec{i}$ and $\vec{j}$ in $\mathbb{R}^n$, hence the name and is defined[2] as

$$cos\left(\vec{i},\vec{j}\right) = \frac{\vec{i}\cdot\vec{j}}{|\vec{i}||\vec{j}|} = \frac{\sum_{m=1}^{n} i_m j_m}{\sqrt{\sum_{m=1}^{n} i_m}\sqrt{\sum_{m=1}^{n} j_m}}.$$

The Kullback-Leibler(KL) divergence has been introduced in [31] by Salomon Kullback and Richard Leibler and is also known as information divergence, information gain or relative entropy. It is a non-symmetric measure of the difference of two probability distributions and is defined as

$$D_{KL}\left(P||Q\right) = \sum_{i} P\left(i\right) log\left(\frac{P\left(i\right)}{Q\left(i\right)}\right)$$

with P and Q being the two probability distributions to examine. In most applications, P refers to the distribution of the test data, whereas Q typically represents a theoretic

[1] see [25], section 4.4. or [41] as examples

[2] [34], p. 300

measure or model or an approximation of P. The fact that the KL divergence is non-symmetric, hence

$$D_{KL}(P||Q) \neq D_{KL}(Q||P)$$

implies that the KL divergence is not a true metric, rather it belongs to broader class of so called f-divergences[3]. Also note, that the KL divergence is undefined per definitionem, if $\exists i : P(i) \neq 0 \wedge Q(i) = 0$.

Basing on the KL divergence, Jenson-Shannon(JS) divergence is a smoothed and symmetric version of the former proposed in [33] and is also known as *information radius*. Its definition[4] is

$$JSD(P,Q) = \frac{1}{2}(D_{KL}(P||M) + D_{KL}(Q||M))$$

with

$$M = \frac{1}{2}(P+Q)$$

Smoothed and symmetric means, that JS divergence always takes on a defined value and $JSD(P||Q) = JSD(Q||P)$.

Another smoothed but non-symmetric version of the KL divergence is the skew divergence proposed by Lilian Lee in [32]. It smoothes one of the distributions by mixing it with the other, by a degree determined by a parameter α:

$$s_\alpha(P,Q) = D_{KL}(Q||\alpha P + (1-\alpha)Q)$$

Although Lee argues that the skew divergence $s_{\alpha=0.99}(P,Q)$ yields best results in a problem regarding co-ocurrence probabilities, different measures have been used and analyzed.

[3] [2]
[4] [34], p. 304

Appendix C

Sampling in Topic models

C.1 LDA model

Consider the plate model of Latent Dirichlet Allocation shown in chapter 3 (Fig. 3.3). Following the plate notation, the overall probability of a model is given by

$$p\left(\mathbf{w}, \mathbf{z}, \theta, \phi, \alpha, \beta\right) = \prod_{j=1}^{T} p\left(\phi_j|\beta\right) \prod_{d=1}^{D} p\left(\theta_d|\alpha\right) \prod_{n=1}^{N_d} p\left(z_{dn}|\theta_d\right) p\left(w_{dn}|\phi_{z_{dn}}\right). \tag{C.1}$$

Integrating over both θ and ϕ yields

$$p\left(\mathbf{w}, \mathbf{z}, \alpha, \beta\right) = \int_{\theta}\int_{\phi} p\left(\mathbf{w}, \mathbf{z}, \theta, \phi, \alpha, \beta\right) d\theta d\phi \tag{C.2}$$

$$= \int_{\theta} \prod_{d=1}^{D} p\left(\theta_d|\alpha\right) \prod_{n=1}^{N_d} p\left(z_{dn}|\theta_d\right) d\theta \int_{\phi} \prod_{j=1}^{T} p\left(\phi_j|\beta\right) \prod_{d=1}^{D}\prod_{n=1}^{N_d} p\left(w_{dn}|\phi_{z_{dn}}\right) d\phi \tag{C.3}$$

As all different θ are independent of each other, and the same holds true for all ϕ, all θ and ϕ can be examined separately. In case of θ, this gives

$$\int_{\theta} \prod_{d=1}^{D} p\left(\theta_d|\alpha\right) \prod_{n=1}^{N_d} p\left(z_{dn}|\theta_d\right) d\theta = \prod_{d=1}^{D} \int_{\theta_d} p\left(\theta_d|\alpha\right) \prod_{n=1}^{N_d} p\left(z_{dn}|\theta_d\right) d\theta_d \tag{C.4}$$

and focussing on only one θ_d and replacing probability by a true distribution expression yields

$$\int_{\theta_d} p\left(\theta_d|\alpha\right) \prod_{n=1}^{N_d} p\left(z_{dn}|\theta_d\right) d\theta_d = \int_{\theta_d} \frac{\Gamma\left(\sum_{j=1}^{T}\alpha_j\right)}{\prod_{j=1}^{T}\Gamma\left(\alpha_j\right)} \prod_{j=1}^{T} \theta_{dj}^{\alpha_j - 1} \prod_{n=1}^{N_d} p\left(z_{dn}|\theta_d\right) d\theta_d. \tag{C.5}$$

Now let n_{jr}^{i} be a 3-dimensional variable, representing the number of tokens in the j-th document of type r(i.e. the r-th word in the vocabulary) that are assigned to topic i. If

any dimension is unbounded, use $(\cdot)$, i.e. $n^i_{j(\cdot)}$ would be the number of all tokens in the j-th document that have been assigned to topic i. Following this assumption, the last term of the right hand side of Eq. B.5 can be replaced by

$$\prod_{n=1}^{N_d} p\left(z_{dn}|\theta_d\right) = \prod_{j=1}^{T} \theta_{dj}^{n^j_{d(\cdot)}}$$

which leads the right hand side of B.5 to

$$\int_{\theta_d} \frac{\Gamma\left(\sum_{j=1}^T \alpha_j\right)}{\prod_{j=1}^T \Gamma\left(\alpha_j\right)} \prod_{j=1}^{T} \theta_{dj}^{\alpha_j - 1} \prod_{j=1}^{T} \theta_{dj}^{n^j_{d(\cdot)}} d\theta_d = \int_{\theta_d} \frac{\Gamma\left(\sum_{j=1}^T \alpha_j\right)}{\prod_{j=1}^T \Gamma\left(\alpha_j\right)} \prod_{j=1}^{T} \theta_{dj}^{n^j_{d(\cdot)} + \alpha_j - 1}. \tag{C.6}$$

Now consider the definition of the Dirichlet distribution

$$\begin{aligned} f\left(x_1, \ldots, x_{k-1}; \alpha_1, \ldots, \alpha_k\right) &= \frac{1}{B\left(\alpha\right)} \prod_{i=1}^{k} x_i^{\alpha_i - 1} \\ &= \frac{\Gamma\left(\sum_{i=1}^k \alpha_i\right)}{\prod_{i=1}^k \Gamma\left(\alpha_i\right)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}. \end{aligned}$$

This yields the property, that

$$\int_{\theta_d} \frac{\Gamma\left(\sum_{j=1}^T n^j_{d(\cdot)} + \alpha_j\right)}{\prod_{j=1}^T \Gamma\left(n^j_{d(\cdot)} + \alpha_j\right)} \prod_{j=1}^{T} \theta_{dj}^{n^j_{d(\cdot)} + \alpha_j - 1} d\theta_d = 1 \tag{C.7}$$

and finally to

$$\int_{\theta_d} p\left(\theta_d|\alpha\right) \prod_{n=1}^{N_d} p\left(z_{dn}|\theta_d\right) d\theta_d \tag{C.8}$$

$$= \int_{\theta_d} \frac{\Gamma\left(\sum_{j=1}^T \alpha_j\right)}{\prod_{j=1}^T \Gamma\left(\alpha_j\right)} \prod_{j=1}^{T} \theta_{dj}^{n^j_{d(\cdot)} + \alpha_j - 1} \tag{C.9}$$

$$= \frac{\Gamma\left(\sum_{j=1}^T \alpha_j\right)}{\prod_{j=1}^T \Gamma\left(\alpha_j\right)} \frac{\prod_{j=1}^T \Gamma\left(n^j_{d(\cdot)} + \alpha_j\right)}{\Gamma\left(\sum_{j=1}^T n^j_{d(\cdot)} + \alpha_j\right)} \underbrace{\int_{\theta_d} \frac{\Gamma\left(\sum_{j=1}^T n^j_{d(\cdot)} + \alpha_j\right)}{\prod_{j=1}^T \Gamma\left(n^j_{d(\cdot)} + \alpha_j\right)} \prod_{j=1}^{T} \theta_{dj}^{n^j_{d(\cdot)} + \alpha_j - 1} d\theta_d}_{=1} \tag{C.10}$$

$$= \frac{\Gamma\left(\sum_{j=1}^T \alpha_j\right)}{\prod_{j=1}^T \Gamma\left(\alpha_j\right)} \frac{\prod_{j=1}^T \Gamma\left(n^j_{d(\cdot)} + \alpha_j\right)}{\Gamma\left(\sum_{j=1}^T n^j_{d(\cdot)} + \alpha_j\right)} \tag{C.11}$$

The same holds true for all ϕ. Again consider only one ϕ_j out of

$$\int_{\phi} \prod_{j=1}^{T} p\left(\phi_j|\beta\right) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p\left(w_{dn}|\phi_{z_{dn}}\right) d\phi \tag{C.12}$$

$$= \prod_{j=1}^{T} \int_{\phi_j} p\left(\phi_j|\beta\right) \prod_{d=1}^{D} \prod_{n=1}^{N_d} p\left(w_{dn}|\phi_{z_{dn}}\right) d\phi_j \tag{C.13}$$

in which the true distribution expression is inserted to give

$$\int_{\phi_j} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \prod_{v=1}^{V} \phi_{jv}^{\beta_v - 1} \prod_{v=1}^{V} \phi_{jv}^{n^j_{(\cdot)v}} d\phi_j \tag{C.14}$$

$$= \int_{\phi_j} \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \prod_{v=1}^{V} \phi_{jv}^{n^j_{(\cdot)v} + \beta_v - 1} d\phi_j \tag{C.15}$$

$$= \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V} \Gamma\left(n^j_{(\cdot)v} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n^j_{(\cdot)v} + \beta_v\right)} \underbrace{\int_{\phi_j} \frac{\Gamma\left(\sum_{v=1}^{V} n^j_{(\cdot)v} + \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(n^j_{(\cdot)v} + \beta_v\right)} \prod_{v=1}^{V} \phi_{jv}^{n^j_{(\cdot)v} + \beta_v - 1} d\phi_j}_{=1} \tag{C.16}$$

$$= \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V} \Gamma\left(n^j_{(\cdot)v} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n^j_{(\cdot)v} + \beta_v\right)}. \tag{C.17}$$

Comparing Eq. B.11 and B.17 to Eq. 3.28 and 3.29 shows that an estimation of θ and ϕ is given by

$$\frac{n^{(d)}_j + \alpha}{n^{(d)}_{\cdot} + T\alpha} \propto \frac{\Gamma\left(\sum_{j=1}^{T} \alpha_j\right)}{\prod_{j=1}^{T} \Gamma\left(\alpha_j\right)} \frac{\prod_{j=1}^{T} \Gamma\left(n^j_{d(\cdot)} + \alpha_j\right)}{\Gamma\left(\sum_{j=1}^{T} n^j_{d(\cdot)} + \alpha_j\right)} \tag{C.18}$$

$$\frac{n^{(w)}_j + \beta}{n^{(\cdot)}_j + V\beta} \propto \frac{\Gamma\left(\sum_{v=1}^{V} \beta_v\right)}{\prod_{v=1}^{V} \Gamma\left(\beta_v\right)} \frac{\prod_{v=1}^{V} \Gamma\left(n^j_{(\cdot)v} + \beta_v\right)}{\Gamma\left(\sum_{v=1}^{V} n^j_{(\cdot)v} + \beta_v\right)} \tag{C.19}$$

C.2 Community-Author-Topic model

Following the plate notation of the Community-Author-Topic model, the overall probability of the model is given by

$$p\left(\mathbf{w},\mathbf{x},\mathbf{z},\eta,\theta,\phi|\alpha,\beta,\gamma\right) = \prod_{a=1}^{A} p\left(\eta_a|\alpha\right) \prod_{c=1}^{C} p\left(\theta_c|\beta\right) \prod_{j=1}^{T} p\left(\phi_j|\gamma\right) p\left(x|\eta_{a_d}\right) \prod_{n=1}^{N} p\left(z_n|\theta_x\right) p\left(w_n|\phi_{z_n}\right) \tag{C.20}$$

Integrating over η, θ and ϕ yields

$$\begin{aligned} &p\left(\mathbf{w},\mathbf{x},\mathbf{z}|\alpha,\beta,\gamma\right) \\ &= \int_{\eta}\int_{\theta}\int_{\phi} \prod_{a=1}^{A} p\left(\eta_a|\alpha\right) \prod_{c=1}^{C} p\left(\theta_c|\beta\right) \prod_{j=1}^{T} p\left(\phi_j|\gamma\right) p\left(x|\eta_{a_d}\right) \prod_{n=1}^{N} p\left(z_n|\theta_x\right) p\left(w_n|\phi_{z_n}\right) d\eta d\theta d\phi \end{aligned} \tag{C.21}$$

Again, all different η, θ and ϕ may be treated as independent of each other and are examined separately. In case of η this leads to

$$\int_{\eta} \prod_{a=1}^{A} p\left(\eta_a|\alpha\right) \prod_{d=1}^{D} p\left(x_d|\eta_{a_d}\right) d\eta \qquad = \prod_{a=1}^{A} \int_{\eta_a} p\left(\eta_a|\alpha\right) \prod_{d=1}^{D} p\left(x_d|\eta_{a_d}\right) d\eta_a \tag{C.22}$$

Focussing on only one η_a and replacing probability by a true distribution expression yields

$$\int_{\eta_a} \frac{\Gamma\left(\sum_{c=1}^{C} \alpha_c\right)}{\prod_{c=1}^{C} \Gamma\left(\alpha_c\right)} \prod_{c=1}^{C} \eta a_d, c^{\alpha_c - 1} \prod_{d=1}^{D} p\left(x_d|\eta_{a_d}\right) d\eta_a \tag{C.23}$$

where

$$\prod_{d=1}^{D} p\left(x_d|\eta_{a_d}\right) = \prod_{c=1}^{C} \eta_{a_d,c}^{n_c^{a_d}}$$

with $n_c^{a_d}$ being the number of times, author a_d has been assigned to community c. This leads to

$$\int_{\eta_a} \frac{\Gamma\left(\sum_{c=1}^{C} \alpha_c\right)}{\prod_{c=1}^{C} \Gamma\left(\alpha_c\right)} \prod_{c=1}^{C} \eta a_d, c^{\alpha_c - 1} \prod_{c=1}^{C} \eta_{a_d,c}^{n_c^{a_d}} d\eta_a \tag{C.24}$$

$$= \int_{\eta_a} \frac{\Gamma\left(\sum_{c=1}^{C} \alpha_c\right)}{\prod_{c=1}^{C} \Gamma\left(\alpha_c\right)} \prod_{c=1}^{C} \eta a_d, c^{n_c^{a_d} + \alpha_c - 1} d\eta_a \tag{C.25}$$

$$= \frac{\Gamma\left(\sum_{c=1}^{C} \alpha_c\right)}{\prod_{c=1}^{C} \Gamma\left(\alpha_c\right)} \frac{\prod_{c=1}^{C} \Gamma\left(n_c^{a_d} + \alpha_c\right)}{\Gamma\left(\sum_{c=1}^{C} n_c^{a_d} + \alpha_c\right)} \underbrace{\int_{\eta_a} \frac{\Gamma\left(\sum_{c=1}^{C} n_c^{a_d} + \alpha_c\right)}{\prod_{c=1}^{C} \Gamma\left(n_c^{a_d} + \alpha_c\right)} \prod_{c=1}^{C} \eta a_d, c^{n_c^{a_d} + \alpha_c - 1} d\eta_a}_{=1} \tag{C.26}$$

$$= \frac{\Gamma\left(\sum_{c=1}^{C} \alpha_c\right)}{\prod_{c=1}^{C} \Gamma\left(\alpha_c\right)} \frac{\prod_{c=1}^{C} \Gamma\left(n_c^{a_d} + \alpha_c\right)}{\Gamma\left(\sum_{c=1}^{C} n_c^{a_d} + \alpha_c\right)} \tag{C.27}$$

Therefore, an estimation of η is given by

$$\eta_{a_d,c} \propto \frac{n_c^{a_d} + \alpha}{n_{(\cdot)}^{a_d} + C\alpha}. \tag{C.28}$$

Note that η is a document variable, sampled once for a document, whereas θ and ϕ are word level variables, sampled once for each word in each document. Considering θ,

$$\int_{\theta} \prod_{c=1}^{C} p\left(\theta_c|\beta\right) \prod_{n=1}^{N_d} p\left(z_n|\theta_c\right) d\theta$$

$$= \prod_{c=1}^{C} \int_{\theta_c} p\left(\theta_c|\beta\right) \prod_{n=1}^{N_d} p\left(z_n|\theta_c\right) d\theta_c \tag{C.29}$$

is given. Focussing on only one θ_c and inserting the true distribution gives

$$\int_{\theta_c} \frac{\Gamma\left(\sum_{j=1}^{T} \beta_j\right)}{\prod_{j=1}^{T} \Gamma\left(\beta_t\right)} \prod_{j=1}^{T} \theta_{c,j}^{\beta_j - 1} \prod_{n=1}^{N} p\left(z_n|\theta_c\right) d\theta_c \tag{C.30}$$

Again,

$$\prod_{n=1}^{N} p\left(z_n|\theta_c\right) = \prod_{j=1}^{T} \theta_{c,j}^{n_c^j}$$

with n_c^j being the number of times, a token is assigned to topic j and community c. This leads to

$$\int_{\theta_c} \frac{\Gamma\left(\sum_{j=1}^T \beta_j\right)}{\prod_{j=1}^T \Gamma\left(\beta_t\right)} \prod_{j=1}^T \theta_{c,j}^{n_c^j+\beta_j-1} d\theta_c \tag{C.31}$$

$$= \frac{\Gamma\left(\sum_{j=1}^T \beta_j\right)}{\prod_{j=1}^T \Gamma\left(\beta_t\right)} \frac{\prod_{j=1}^T \Gamma\left(n_c^j+\beta_j\right)}{\Gamma\left(\sum_{j=1}^T n_c^j+\beta_j\right)} \underbrace{\int_{\theta_c} \frac{\Gamma\left(\sum_{j=1}^T n_c^j+\beta_j\right)}{\prod_{j=1}^T \Gamma\left(n_c^j+\beta_j\right)} \prod_{j=1}^T \theta_{c,j}^{n_c^j+\beta_j-1} d\theta_c}_{=1} \tag{C.32}$$

$$= \frac{\Gamma\left(\sum_{j=1}^T \beta_j\right)}{\prod_{j=1}^T \Gamma\left(\beta_t\right)} \frac{\prod_{j=1}^T \Gamma\left(n_c^j+\beta_j\right)}{\Gamma\left(\sum_{j=1}^T n_c^j+\beta_j\right)} \tag{C.33}$$

Thus, an estimation of θ is given by

$$\theta_{c,j} \propto \frac{n_c^j+\beta}{n_c^{(\cdot)}+T\beta}.$$

Inspection of ϕ gives

$$\prod_{j=1}^T \int_{\phi_j} p\left(\phi_j|\gamma\right) \prod_{n=1}^N p\left(w_n|\phi_{z_n}\right) d\phi_j \tag{C.34}$$

Looking at only one ϕ_j and given that

$$\prod_{n=1}^N p\left(w_n|\phi_{z_n}\right) = \prod_{v=1}^V \phi_{j,v}^{n_v^j}$$

with n_v^j being the number of times, the v-th word in the vocabulary has been assigned to topic j, this leads to

$$\int_{\phi_j} \frac{\Gamma\left(\sum_{v=1}^V \gamma_v\right)}{\prod_{v=1}^V \Gamma\left(\gamma_v\right)} \prod_{v=1}^V \phi_{j,v}^{n_v^j+\gamma_v-1} \tag{C.35}$$

$$= \frac{\Gamma\left(\sum_{v=1}^V \gamma_v\right)}{\prod_{v=1}^V \Gamma\left(\gamma_v\right)} \frac{\prod_{v=1}^V \Gamma\left(n_v^j+\gamma_v\right)}{\Gamma\left(\sum_{v=1}^V n_v^j+\gamma_v\right)} \underbrace{\int_{\phi_j} \frac{\Gamma\left(\sum_{v=1}^V n_v^j+\gamma_v\right)}{\prod_{v=1}^V \Gamma\left(n_v^j+\gamma_v\right)} \prod_{v=1}^V \phi_{j,v}^{n_v^j+\gamma_v-1} d\phi_j}_{=1} \tag{C.36}$$

$$= \frac{\Gamma\left(\sum_{v=1}^V \gamma_v\right)}{\prod_{v=1}^V \Gamma\left(\gamma_v\right)} \frac{\prod_{v=1}^V \Gamma\left(n_v^j+\gamma_v\right)}{\Gamma\left(\sum_{v=1}^V n_v^j+\gamma_v\right)} \tag{C.37}$$

which gives an estimate for ϕ of

$$\phi_{j,v} \propto \frac{n_v^j+\gamma}{n_{(\cdot)}^j+V\gamma}.$$