DISSERTATION

submitted

to the

Combined Faculty for the Natural Sciences and Mathematics

of

Heidelberg University, Germany

for the degree of

Doctor of Natural Sciences

Put forward by

M.A. Hui Li

Born in: Nanjing, Jiangsu, China

Oral examination:.....

Social Network Extraction and Exploration of Historic Correspondences

Advisor: Prof. Dr. Michael Gertz

Abstract

Historic correspondences, in the form of letters, provide a scenario in which historic figures and events are reflected and thus play a ubiquitous role in the study of history. Confronted with the digitization of thousands of historic letters and motivated by the potentially valuable insights into history and intuitive quantitative relations between historic persons, researchers have recently focused on the network analysis of historic correspondences. However, most related research constructs the correspondence networks only based on the sender-recipient relation with the objective of visualization. Very few of them have proceeded beyond the above stage to exploit the detailed modeling of correspondence networks, let alone to develop novel concepts and algorithms derived from network analysis or formal approaches to the data uncertainty issue in historic correspondence.

In the context of this dissertation, we develop a comprehensive correspondence network model, which integrates the personal, temporal, geographical, and topic information extracted from letter metadata and letter content into a hypergraph structure. Based on our correspondence network model, we analyze three types of person-person relations (sender-recipient, co-sender, and co-recipient) and two types of person-topic relations (author-topic and sender-recipient-topic) statically and dynamically. We develop multiple measurements, such as local and global reciprocity for quantifying reciprocal behavior in weighted networks, and the topic participation score for quantifying interests or the focus of individuals or real-life communities. We investigate the rising and the fading trends of topics in order to find correlations among persons, topics, and historic events. Furthermore, we develop a novel probabilistic framework for refinement of uncertain person names, geographical location names, and temporal expressions in the metadata of historic letters.

We conduct extensive experiments using letter collections to validate and evaluate the proposed models and measurements in this dissertation. A thorough discussion of experimental results shows the effectiveness, applicability and advantages of our developed models and approaches.

Zusammenfassung

Historische Korrespondenzen in Briefform stellen ein Szenario dar, in dem historische Figuren und Ereignisse dokumentiert werden und spielen eine zentrale Rolle in den Geschichtsstudien. Aufgrund der Digitalisierung von tausenden historischen Briefen und motiviert durch potenzielle Einblicke in die Geschichte und die quantitativen Beziehungen zwischen historischen Personen, haben sich Forscher in letzter Zeit auf die Netzwerkanalyse historischer Korrespondenzen konzentriert. Die meisten dieser Forschungen konstruieren Korrespondenznetzwerke jedoch nur auf Grundlage von Sender-Empfänger-Beziehungen mit dem Ziel der Visualisierung. Nur sehr wenige Arbeiten darüber hinaus und nutzen die detaillierte Modellierung von Korrespondenznetzwerken aus, um neue Konzepte und Algorithmen aus der Netzwerkanalyse zu verwenden oder Ansätzen zum Datenunsicherheitsproblem in historischer Korrespondenzen zu entwickeln.

Im Kontext dieser Dissertation wird ein umfassendes Korrespondenznetzwerkmodell vorgestellt, das persönliche, zeitliche, geografische und thematische Informationen aus Briefmetadaten und Briefinhalten in eine Hypergraphstruktur integriert. Basierend auf dieser Grundlage werden drei Typen von Personen-Personen-Beziehungen (Sender-Empfänger, Mitsender, und Mitempfänger) und zwei Typen von Person-Themen-Beziehungen (Autor-Thema und Sender-Empfänger-Thema) analysiert. Verschiedene Messverfahren werden entwickelt, wie lokale und globale Reziprozität zur Quantifizierung der reziproken Verhaltens in gewichteten Netzwerken, sowie ein Topic Participation Score zur Quantifizierung von Interessen oder Foki von Einzelpersonen bzw. realen Communities. Zunehmende und abnehmende Trends in Topics dienen dazu, die Korrelationen zwischen Personen, Themen und historischen Ereignissen zu identifizieren. Außerdem wird ein neuartiges probabilistisches Rahmenwerk zur Verfeinerung von unsicheren Personennamen, geografischen Ortsnamen und zeitlichen Ausdrücken in Brief-Metadaten entwickelt.

Anhand von umfangreichen Experimenten mit großen historischen Briefsammlungen werden die vorgeschlagenen Modelle und Messungen in dieser Dissertation validiert und evaluiert. Eine detaillierte Diskussion der Ergebnisse der Experimente dient dazu, die Effektivität, die Anwendbarkeit, und die Vorteile der entwickelten Modelle und Ansätze aufzuzeigen.

Acknowledgements

This dissertation could not be considered completed if I do not acknowledge the assistance, patience, and support of many individuals and institutions. First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Dr. Michael Gertz, of the Faculty of Mathematics and Computer Science at Heidelberg University, for giving me the honor to come to his group and inspiring me with the prime research overview that leads to the possibility of the research of this dissertation. Not only that, I would like to thank him for imparting me his valuable advice, profound knowledge and tremendous insights, as well as his firm support and enduring patience to my research progress and academic writing over the years. Before I came to the Database Research System Group in 2013, I was a casual student who had trivial knowledge about my late research. But that did not deter my supervisor from being understanding, responsible and compassionate in supervising me with my research.

During the past four years, I have met many problems and difficulties, however, the support and guidance of my supervisor helped me stay committed to research and forging through the difficult times to the final part of the dissertation. I sincerely appreciate that he took time out of his busy schedule to go through the draft of my dissertation and met me after only a few days with comments and suggestions on almost every page. My supervisor sets a great example of excellence as a researcher to me and I sincerely thank him for his kindness and tolerance to my study from the bottom of my heart.

I would like to express my gratitude to my second supervisor, Prof. Dr. Stefan Riezler, for his warm-hearted support in my research, the insightful comments I have received, as well as the fruitful weekly colloquiums and delicious cakes he has provided. Besides my supervisors, I am also grateful to the rest of the committee members PD Dr. Wolfgang Merkle and Prof. Dr. Gerhard Reinelt for the genuine comments I will receive in the defense.

I would like to express my gratitude to the Heidelberg Graduate School of Mathematical and Computational Methods for the Sciences (HGS MathComp) for the financial support so that this dissertation has become possible. I am also grateful to Dr. Michael Winckler at HGS MathComp, for giving me consistent encouragement and useful advice upon my doctoral research. I would also like to express my gratitude to all the following institutions that provide data sources for my research that made my dissertation possible: Melanchthon Forschungsstelle at Heidelberg University, Department of Theology at Heidelberg University, Darwin Correspondence Project at Cambridge University, Wallace Correspondence Project, and Mark Twain Project.

During my PhD study, I had a wonderful time to work at our Database Research System Group, in particular thanks to a circle of friendly and warm-hearted people. Thank you Dr. Jannik Strötgen for asking me to collaborate with him on Heideltime and helping me publish my first paper in this group. Thank you Dr. Christian Sengstock for helping me solve a lot of questions related to data mining and social network analysis. Thank you Thomas Bögel, you are a really smart person and an ideal office mate. Thank you Andreas, for your valuable encouragement and thoughtful discussions from time to time. Many thanks to HiWis such as Leonard Henger and Niels Bernlöhr, who helped me a lot with software installation and server configuration. Thank you, Lutz Büch, you are a really humorous and optimistic person, who always brings happiness to everyone. Thank you Kai Chen, for imparting you excellent programming skills and experiences to me.

My thanks also go to Dr. Tran Van Canh, Dr. Van Quoc Anh, Dr. Ayser Armiti, Dr. Hamed Abdelhaq, Dr. Florian Flatow, Julian Zell, Ruobing Shen, Xiaoyu Chen, Zhen Dong, Diego Costa, Catherine Proux-Wieland, and everyone else in our lunch group for your friendship and support over the years. I am very grateful to Mrs. Natalia Ulrich, Mrs. Dorothea Heukäufer, and Mrs. Anke Sopka, for assisting me in many different ways and handling the paperwork sincerely. I would additionally like to express my gratitude to Elke Pürzer, Johannes Huber, Anna and Rudi for giving me working opportunity and teaching me a lot in other areas besides research. I wish furthermore to thank Jonathan Griffiths as freelance English-language proofreader for having corrected and vastly improved both the language and the style of this dissertation. He was also especially patient and prompt in his work and contact.

I would like to express my gratitude to my family and friends for caring and trusting me all the time. Thank you so much, my mum and dad, for you persistent love, support and understanding. Thank you, my best friend Li Qi, for always being there to listen when I need an ear and comfort me with her words and charm. Thank you, Prof. Dr. Georg Wolschin and Dr. Sabine Wolschin, my friends and landlords, for providing me fantastic accommodation and heartful care all these years. My thanks also go to Xiwen Lin, Weite Zhang, Ruihan Zhang, Zhiqi Lu, Xiaoyu Ye, Chen Chen and everyone else for your friendship and support over the years. A special thank is dedicated to my love Ming, who cooks tasty food and kicks me on my backside to push me forward whenever I need one. Thank you all with all my heart and soul.

CONTENTS

If you want to change the world, pick up your pen and write.

— Martin Luther

CHAPTER 1

INTRODUCTION

With the advance of technology and the growth of digital culture in the past decade, researchers have come to pay increasing attention to historic texts, namely books, journals, correspondences, newspapers, maps, manuscripts, photographs and so on, which are being digitized and thus made more available as forms of machine-readable texts, photographic images, maps, audio files, among others [1]. Technologies such as infrared scans and optical character recognition are employed in the digitization process in order to enrich the online historic materials or to make new discoveries. Historic correspondences, in the form of letters, spread knowledge and materials in national and international communications between friends, colleagues, family members, collaborators, to name but a few [2]. Scholarly letters, in the past, constituted one of the most direct and important means of interactions between scholars at home and abroad [3]. These letters are invaluable since they not only provide information about the history of languages and the dynamic genres over time, but they also contribute to the reconstruction of relationships and exchanges of information in the past [4].

The general trend of transcribing manuscripts in the digitization of historic correspondences is presently inclined towards the use of Extensible Markup Language (XML) [5], Text Encoding Initiative (TEI) Guidelines [6] or PDF formats. The expected digitized outputs in many ongoing correspondence projects [3, 7, 8] are scanned images of the original handwritten letter as well as typewritten transcriptions. Figure 1.1 shows a digitized letter as a scanned image of the original writing in the letter collection of Charles Robert Darwin [9]. Facilitated by digitization,

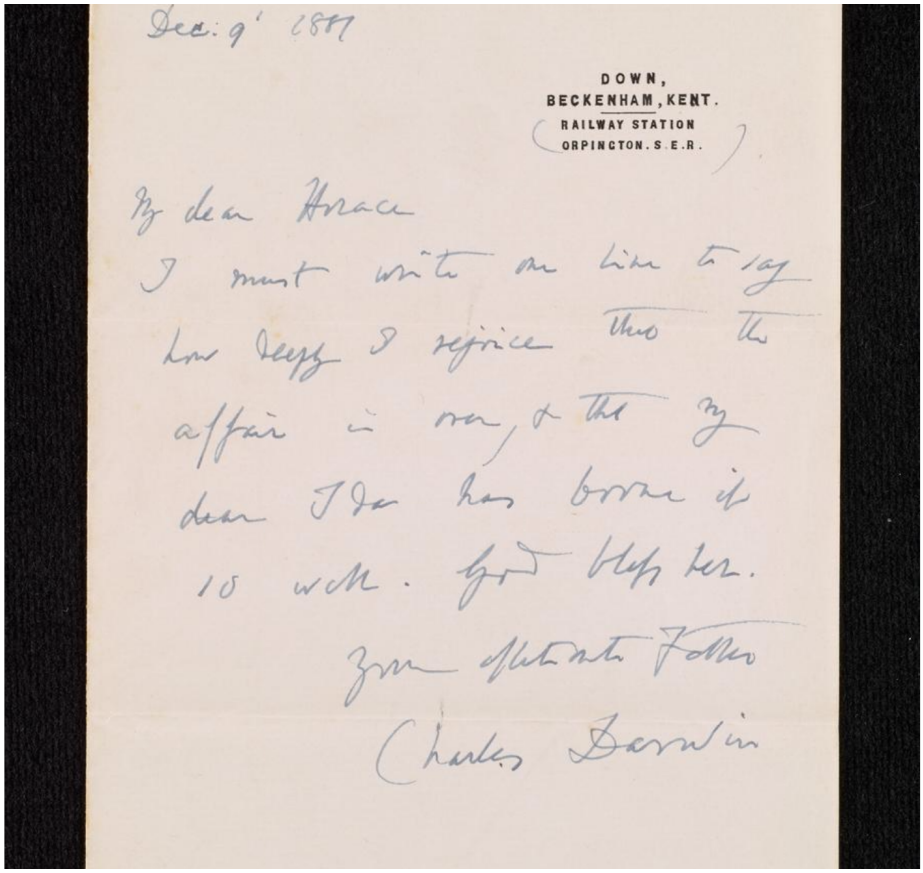


Figure 1.1: An image of the original letter written by Charles Darwin and sent to his son Horace Darwin about the birth of his grandson on December 9th, 1881 in Darwin's letter collection. https://www.darwinproject.ac.uk/letter/?docId=letters/DCP-LETT-13541.xml;query=darwin;brand=default [Last accessed: September 30, 2017].

emerging academic interests in historic correspondences have recently given rise to several fields, for instance, historic linguistics [10], sociolinguistics [11], literary studies [12], history studies [13], and computer science [14], which explore these epistolary collections from different perspectives. This allows researchers from different disciplines to collaborate and address questions that could not be answered before, but which are of central importance for understanding the spread of epistolary information in specific historic periods.

Most transcripts comprise two parts: letter metadata and letter contents. The recorded metadata of most letters consist of the names of correspondents (i.e., sender(s) and recipient(s)), the date on which a certain letter was written, the origin or the location from which the letter was sent and the destination or the location to which it was sent. The availability of such structured metadata

reveals details about the correspondents and gives us a glimpse into individuals' movements over their lives, and a description of their social networks over time. The preserved and digitized contents of most letters are unstructured texts, in which the linguistic and rhetoric features constitute part of the social, academic, economic, and political culture marking up the past. The availability of such texts reveals further and new details about the interests and expectations of correspondents, and shows us certain trends of diverse topics across their lives.

However, collecting the scattered letter collections from different institutions and libraries is a non-trivial task, which makes the full appreciation of historic times based on the large-scale correspondence network analysis very difficult. Furthermore, most studies only focus on the network visualization of sender-recipient relationships, with the result that research upon the strength and interconnectedness of these networks is still at its very beginning. In addition, the transcription of historic correspondences and the combination of letter collections from various libraries and institutions always cause vague, uncertain or even conflicting information [142], but very few of the previous studies have explicitly addressed this problem.

Different from the previous research, we make particular efforts to answer the following questions:

- In a way different from the popular visualization of correspondence networks, can we model the correspondences in such a way that we can analyze the knowledge circulation and recognize interesting evolving patterns? How can we integrate the personal, temporal, geographical and content information into the network? How can we analyze different types of relationships such as person-person relations or person-topic relations that are embedded in the network model?

- For a deeper analysis of correspondence networks, it would be fruitful to review the personal interactions regarding different topics and to take both static and dynamic views into account. How can we extract the topics from letters and link them to specific historic persons? How did these topics change over time?

- Due to the time and the quality of the handwriting, letter collections are rarely complete and always contain uncertain entities such as ambiguous

person names, missing locations or incomplete dates. Although historians have drawn attention to the normalization of historic texts, a fully complete and accurate transcription is time-consuming and has never been realized. Therefore, how can we deal with these uncertain entities that exist in various letter collections? Can we refine these uncertain entities by taking the advantage of correspondence networks and letter contents instead of using manual annotations alone?

1.1 Motivations and Challenges

In this section, we discuss the motivations and major challenges for the further tasks addressed in this dissertation. These motivations and challenges will serve as the basis and the objectives for the empirical study in this dissertation.

- **Modeling Correspondence Networks**. The typical goals of previous studies were restricted to the visualization and observations of networks and patterns, while fewer studies focused on the formal modeling and network measurements of historic correspondences. For the purposes of simplification and vivid visualization, most previous studies separated the letter content from the letter metadata in the graph construction. They therefore overlook the interesting relations between correspondents and word patterns embedded in correspondences. In this dissertation, we aim to create a correspondence network that closely resembles today's social networks based on the metadata and the content of historic letters. It is our hypothesis that such a correspondence network provides a more comprehensive view of the academic circles, individual movements, and individual influences across the lives of the letter-writers and letter-communicators, and the correlation between dynamic interactions and historic events, than what would be perceivable through intuitive graph visualization alone.

- **Measuring Relationships**. The most frequently studied relationship in the previous research is "who wrote to whom", or in other words, the sender-recipient relation. However, there are a lot more latent relations that can be extracted from historic correspondence networks, e.g., the relation between

co-senders, the relation between co-recipients, and the relation between correspondents and topics. These latent relationships not only shed light on social interactions and information exchanges in the past, but also reveal the interpersonal relationships among people and topics that are mentioned in any particular letter. However, most studies only touch the surface of these relationships instead of analyzing them in depth. In this dissertation, we aim to generalize and develop measurements in combination with statistical techniques and network structures in order to observe different types of relations that are embedded in correspondence networks. It is our hypothesis that these measurements provide a more quantitatively accurate description of the period of time, the latent relationships between different entities, and the trends in topics that are embedded in letters than what would be perceivable through the analysis of the sender-recipient relation alone.

- **Data Uncertainty in Historic Correspondences**. There has been a notable dearth of research on data uncertainty in the digital humanities and social network analysis, even though historic correspondences inevitably contain more than a few fragmentary and uncertain data such as anonymous letter writers or missing dates. Without refinement of unknown or imprecise data, tasks such as literature study or information retrieval can only be carried out on a coarse-grained level. Therefore, in this dissertation, we aim to develop a probabilistic framework in combination with topic modeling techniques, network structures and the co-occurrences of entities in the letter metadata. It is our hypothesis that our probabilistic approach contributes to a more precise and effective refinement of the uncertain entities in the letter metadata than what would be achieved through letter content analysis alone.

1.2 Major Contributions

By addressing the tasks and challenges above, the contributions of this dissertation in the context of social network analysis are as follows.

- We propose a comprehensive ***correspondence network model*** that integrates the personal, temporal, geographical and content information of historic letters within a hypergraph representation. This model enables

us to analyze ***multiple relations***, i.e., three types of person-person relations (sender-recipient, co-sender, and co-recipient relations) and two types of person-topic relations (individual and sender-recipient pair topics) from both static and dynamic points of view.

- We develop ***local reciprocity*** and ***global reciprocity*** measurements in order to quantify the reciprocal relations in weighted networks. We develop measurements to analyze the ***topic participation*** of a node or pairs of nodes in the networks. Based on the topic participation score, we examine the ***trends*** of different topics in the historic correspondences and investigate the ***rising*** and the ***fading*** trends of topics accordingly.

- We develop a novel ***probabilistic framework*** in order to refine uncertain entities (e.g., ambiguous person names, incomplete locations, and missing dates) that exist in the correspondence metadata. We leverage the co-occurrences of entities in the metadata, network structures, and topic modeling techniques, to refine missing or ambiguous entities effectively.

- We conduct extensive experiments in order to evaluate the effectiveness of the proposed models and corresponding measurements using various letter collections. The results obtained are further discussed as well in order to show the applicability and advantages of the models and approaches introduced.

In addition, our contributions in this dissertation in the context of digital humanities are as follows.

- Our correspondence network model and the corresponding measurements are not limited to a specific size of letter collection, but can be applied to correspondence collections at any size.

- Our proposed measurements such as topic participation scores can be applied to the area of ***digital humanities*** and ***information retrieval*** for tasks such as individual topic retrieval [3] and representative topic/letter recommendation [15].

- Our probabilistic framework that is developed in order to refine the uncertain entities in the letter metadata can also be applied in the area of digital humanities such as ***stylometry*** [16, 17] to predict the potential author(s) of anonymous historic letters.

- We employ different approaches and conduct comparative experiments in order to evaluate the impact of pre-processing on the task of topic extraction from letter collections. This can be a valuable ***reference and guide*** for researchers doing similar projects in future.

1.3 Outline of the Dissertation

The rest of the dissertation is organized as follows.

Chapter 2. In this Chapter, we present the background and related work relevant to our study in this dissertation. We first introduce a brief overview of social networks and a selection of corresponding examples. After this we discuss three research fields that are addressed in this dissertation, namely fundamental concepts and methods in network science, historic correspondence research, and data uncertainty in digital humanities. The material in this chapter forms the basis for the further studies in Chapters 3 and 4.

Chapter 3. In this Chapter, we concentrate on the metadata of historic letters and develop a correspondence network model with three derived graphs regarding different relations between correspondents. Specific network measures are generalized or developed for different graphs, respectively. Furthermore, in order to observe the evolution of a correspondence network and the contact patterns of individuals over time, we use not only the representation of contact sequences since most letter repositories do not have the duration of a letter, but also the concept of graphlets in the representation in order to treat the network as a time-series of static graphs. We then apply our models and measurements to empirical datasets, in order not only to obtain an overview of the correspondence networks in the given historic periods, but also to gain significant insights into individual correspondence activity and to note interesting (fluctuating and stable) patterns of networks over time.

Chapter 4. In this Chapter, we concentrate on the content of historic letters and extend our correspondence model to provide a stage on which personal, temporal and geographical information in the metadata and the contents of letters are interconnected. We combine topic modeling techniques and network structures to extract and explore the correspondent-specific (author-only or sender-recipient pair) topics effectively. Moreover, we explicitly investigate the trends of topics over time in order to correlate significant fluctuations in topics with the social or life events of historic persons. In addition, we propose a probabilistic framework in combination with topic distributions, network structures and entity co-occurrences in order to refine the uncertain entities in the letter metadata. We then apply our models and measurements to empirical datasets, in order not only to explore relations between correspondents and topics dynamically, but also to evaluate the effectiveness of our data uncertainty approach.

Chapter 5. This final Chapter presents a summary of the dissertation and makes a range of suggestions for future studies.

If you young fellows were wise, the devil couldn't do anything to you, but since you aren't wise, you need us who are old.

— Martin Luther

CHAPTER 2

BACKGROUND AND BASIC CONCEPTS

In this chapter, we will present the background of this dissertation and introduce a range of related basic concepts that are most relevant to our study. We begin in Section 2.1 with a brief introduction to social networks and a selection of corresponding examples. Section 2.2 is devoted to the first research field that is addressed in this dissertation, namely graph principles. A number of representations and measurements are discussed in this section. In Section 2.3, a selection of community structures and community detection techniques will be introduced. The temporal study in social networks will be discussed in Section 2.4. In Section 2.5, we will briefly present the second research field that is addressed in this dissertation, namely historic correspondence research. In this section, we reviewed the related study on correspondence networks in the area of digital humanities. In Section 2.6 we will introduce the basic concept of topic modeling and its corresponding application in digital humanities and email research. In Section 2.7 we will present the third research field that is addressed in this dissertation, namely data uncertainty in digital humanities. In this section, we will discuss the related work on named entity disambiguation and text similarity. The material in this chapter forms the basis for the further studies in the rest of this dissertation.

2.1 Social Networks

We begin, in this section, with an introduction of social networks and a brief description of two important examples of social networks. The term *network* is

used at various levels of formality. The Oxford English Dictionary [18] defines the word *network* in its most general form: "any netlike or complex system or collection of interrelated things, as topographical features, lines of transportation, or telecommunications route". In a simple form, *network* is defined as a collection of points joined together in pairs by line [19]. *Social network* is one typical type of networks and is defined as "a system of social interactions and relationships; a group of people who are socially connected to one another" [20]. We are surrounded by *social networks* in our daily life and some of the well-known examples nowadays are online human communication networks such as *Facebook*[1]and *Twitter*[2]. Although they are not the first social networks created, they are probably the examples most well-known to most people.

Facebook is one of the world's largest social networks [21]. Founded in 2004, *Facebook* expanded from a campus networking service to a worldwide social networking platform. Until 2015, *Facebook* had over 1.59 billion monthly active users [22]. There are many features in *Facebook* that socially connect among people, some of which are listed as follows.

- **Friends**. A *Facebook* user connect to other users by sending a friend request. And once two individuals become *friends*, the relation can be categorized with labels such as "friends", "family members" and "acquaintances".

- **Profile**. Each user has a personal profile that includes basic information such as name, the declared partner relation with others and a *timeline* for posting messages, images, events, etc., with friends together.

- **Reaction Button**. *Facebook* provides a thumb-up symbol called *like* button, which is available not only for *Facebook* users, but also for websites outside *Facebook*. It represents a positive feedback to postings, pages, advertisements, and so on.

Twitter is one of the world's most popular social networking service [23]. Founded in 2006, it rapidly gained millions of worldwide users. Until 2015, *Twitter* had over 332 million active users [24]. *Twitter* also provides many features to users and some of the features are listed as follows.

0 www.facebook.com [Last accessed: April 2, 2017].
1 www.twitter.com [Last accessed: April 2, 2017].

- **Followers**. A *Twitter* user can post messages up to 140 characters long. These short messages are called *tweets*. Users who subscribe to and receive other users' tweets are called *followers*. Such a relationship on *Twitter* is not reciprocal [30].

- **Contextual Links**. A *Twitter* user can use @ sign with a username following to indicate the *tweet* is mentioned or replied to other users.

- **Trending Topics**. *Hashtags* are short, descriptive words or phrases identified by the # sign preceded. Users can keep track of a certain topic by searching for specific *hashtags* and repost the related messages using *retweet* button.

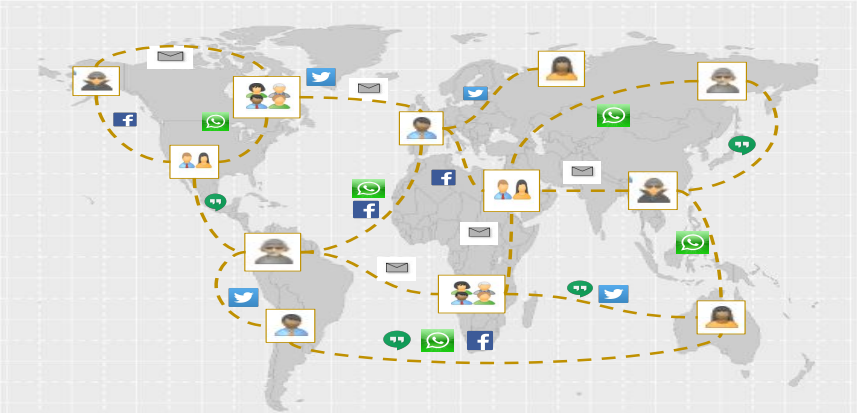


Figure 2.1: A simple example of a social network: people around the world (mini-figures) and their mutual interactions (dashed lines). The icons on each dashed line illustrate the mode of communication that different people use (e.g., Facebook, Twitter, Google Chat, and so on).

Users and relationships embedded in these networks can be represented using graph structures conveniently. In the following section, we will introduce how to represent a social network as a graph and introduce the basic graph measures.

2.2 Graph Principles

In this section we introduce the basic theoretical mechanisms that originate from graph theory in order to describe and analyze networks. In this section, we only

focus on the concepts in graph theory that are most associated with the research of social networks. We first present the most noteworthy definitions and representations of graphs along with the mathematical concepts that are to be applied in Section 2.2.1. Then we introduce statistical measures that are necessary and important for our research in quantifying graph structure in Section 2.2.2.

2.2.1 Graph Definitions and Representations

The Oxford English Dictionary [25] defines the word *Graph* as "a kind of symbolic diagram, in which a system of connections is expressed by spots or circles, some pairs of which are colligated by one or more line". In other words, a *graph* is a mathematical system representation consisting of a set of objects where some pairs of the objects are connected by some types of links. Objects are called vertices or nodes, and links are also called edges. *Networks* are frequently used interchangeably with the term *graph* and are most commonly represented in various graph forms. Formally, a graph is defined as follows.

Definition 2.1. Graph. A graph is represented as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{V}$ denotes a set of nodes or vertices and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ denotes a set of edges or links. The notions $|\mathcal{V}|$ and $|\mathcal{E}|$ denote the number of vertices and edges, respectively.

Definition 2.2. Subgraph. A graph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ is a *subgraph* of $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ if $\mathcal{V}' \subseteq \mathcal{V}$ and $\mathcal{E}' \subseteq \mathcal{E}$.

Graphs can be classified into *undirected* and *directed* graphs in terms of the directions of edges. In an *undirected* graph, there is no ordering between the vertices while defining an edge. In other words, if (u, v) is an edge of the graph, so is (v, u). Comparatively, if an edge between the two vertices has an ordering in the graph, it means that the edge from v to u $\langle v, u \rangle$ is not the same as the edge from u to v $\langle u, v \rangle$ and this graph is called a *directed graph*.

Graphs can also be classified into *unweighted* and *weighted* graphs in terms of values associated with edges. In a *weighted* graph, weights, strengths or values are associated with edges. For instance, on an airline map nodes represent airports and edges represent the routes among them. The weight associated with each

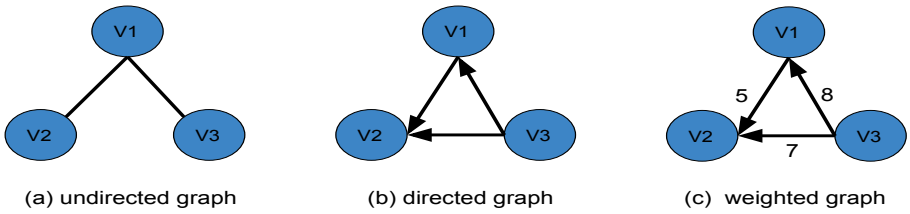


Figure 2.2: (a), (b) and (c) are separate examples of three types of graphs: undirected graph, directed graph and weighted graph.

edge represents the distance between each two airports.

Besides, graphs can be classified into *simple* graph and *multigraph* in terms of the number of edges between any pair of nodes. In a *simple* graph, there exists only one edge between any two nodes. A *multigraph* is a graph in which two or more edges (*multiedges*) between two nodes are allowed. The generalizations of graphs admitting an edge joining more than two nodes at a time are called *hypergraphs* [47].

Definition 2.3. Hypergraph. An undirected hypergraph is represented as $H = (V, E)$ where V denotes a set of nodes and $E \subseteq V \times V$ denotes a set of edges. Each edge consists of a subset of nodes and each edge joins more than two nodes. Such an edge is called a hyperedge.

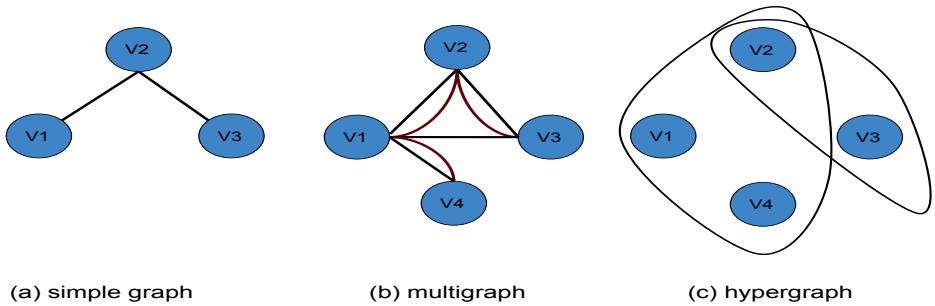


Figure 2.3: (a), (b) and (c) are separate examples of three types of graphs: simple graph, multigraph and hypergraph.

The definition of hypergraph above will be extended for the directed multi-edge situation in Chapter 3 as the basis of our correspondence network model.

2.2.2 Adjacency Matrix

Graphs are frequently represented in the form of *adjacency matrices* (also known as *Sociomatrices*). i and j in a graph are *adjacent* when i and j are connected via an edge and these two nodes are called the *endpoints* of this edge. We define an adjacency matrix of a graph $\mathcal{G}$ with n nodes as follows.

Definition 2.4. Adjacency Matrix. Given a graph $\mathcal{G}$ with n nodes, its adjacency matrix A is represented as a $n \times n$ matrix, in which each of its elements represents the existence and possibly the weight of each edge.

In other words, the adjacency matrix $\mathbf{A}$ of a (unweighted) graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is the matrix with elements A_{ij} such that

$$A_{ij} = \begin{cases} 1 & \text{if (i,j)} \in \mathcal{E} \\ 0 & otherwise. \end{cases}$$

The entry A_{ij} in the adjacency matrix A is non-zero when there is an edge from node i to node j in the graph, otherwise 0. Sometimes an entry A_{ij} in the adjacency matrix for a weighted graph is equal to the corresponding weight for the edge (i, j), and in this case, the matrix is also called *weight matrix*.

2.2.3 Paths

A path in a graph is a sequence of nodes, such that every consecutive pair of nodes in the sequence is joined by an edge [19]. In this section, we discuss several path-related concepts, i.e., the loop, the path, the shortest path, the density, the neighbor, to name but a few. These concepts can be useful to describe the graph structure.

Loop. A *loop* (also called a self-loop) is an edge that starts and ends on the same node. A *walk* is a sequence of adjacent nodes in which edges and nodes can be visited more than once. A walk is open if it starts and ends at two different nodes, and closed if it starts and ends at the same node. For example, an open walk from node i to node j is a sequence of nodes $\{v_1 = i, v_2, ..., v_k = j\}$, where the starting node of the walk is i and the ending node is j.

Path. A *path* is a walk in which each node is visited once. In an *undirected* network, edges can be traversed in both directions, but in a *directed* network each edge in a path must be traversed in the correct direction for that edge [19]. Given two nodes v_1 and v_k, a directed path between two nodes is defined as follows.

Definition 2.5. Path. A directed path from v_1 to v_k is represented as a sequence of nodes $p(i,j)$ starting from node i and ending with node j: $p(i,j) := \langle v_1, ..., v_k \rangle \quad s.t., \langle v_x, v_{x+1} \rangle \in \mathcal{E}, 1 \leq x \leq k-1, v_1 = i \wedge v_k = j$, such that the directed edge $\langle v_x, v_{x+1} \rangle$ is traversed.

Given that there might be more than one path from i to j, we denote as $\mathcal{P}(i,j)$ the set of all the paths from i to j. In unweighted networks, the length of a path is calculated as the number of edges traversed in the path; the path length for a weighted network will be introduced in Chapter 3.

Connectedness. A node i is *connected* to another node j if there exists a path from i to j. Two edges are *incident* in an undirected graph when they share one endpoint. In a directed graph, two edges are incident if the ending of one is the beginning of the other [26]. A graph is *connected* if there is a path between any pairs of nodes in the graph, otherwise it is *disconnected*. In a connected graph, all pairs of nodes are reachable; in a disconnected graph, the nodes can be divided into subsets in which there is no path between the nodes in different subsets [27]. A *component* of a graph is a maximal connected subgraph. The strongly connected component of a directed graph is represented as the largest set of nodes in the network in which there is a directed path between any pair of nodes.

Definition 2.6. Neighbors. Given a node $i \in V$, the set of neighbors $N(i)$ is represented as the set of nodes that are directly connected with i.

$$N(i) := \{j \in V, j \neq i \mid (i,j) \in \mathcal{E}\}. \tag{2.1}$$

Let $N_e(i)$ be the set of edges whose endpoints are both in the set of neighbors $N(i)$:

$$N_e(i) := \{(j,k) \in \mathcal{E} \mid j \in N(i), k \in N(i)\}. \tag{2.2}$$

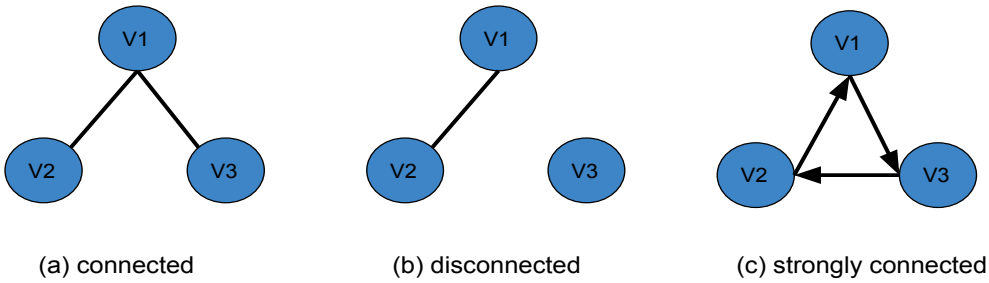


Figure 2.4: (a), (b) and (c) are separate examples of connectivities in graphs. (a) is a connected undirected graph and each node is connected to each other. (b) is a disconnected graph with two components. (c) is a directed graph with one strongly connected component.

The Shortest Path. The *shortest path* is a path that has the minimal path length between two nodes. The distance between two nodes is measured by the *shortest path length*, which is also called the geodesic distance and denoted as $d(i, j)$. The *diameter* of a graph is the maximal geodesic distance between any pair of nodes in the graph.

Density. The *density* of a graph is calculated as the fraction of the actual number of edges divided by the maximum possible number of edges in the graph. In an undirected graph $\mathcal{G}$ that has no self-loops and no multiple edges, the density of a graph is calculated as:

$$\rho(\mathcal{G}) := \frac{|\mathcal{E}|}{|\mathcal{V}|(|\mathcal{V}|-1)/2}\,, \tag{2.3}$$

where in this case, the maximum number of edges is $|\mathcal{V}|(|\mathcal{V}|-1)/2$, and the value of density lies in the range $[0, 1]$. $\rho(\mathcal{G}) = 0$ when there is no edge in the graph, whereas $\rho(\mathcal{G}) = 1$ when each node in the graph is connected to every other node. For a directed graph, the factor of $\frac{1}{2}$ in the Equation 2.3 is dropped.

2.2.4 Centrality Measures

One of the primary uses of graph theory in social network analysis is the identification of the most important nodes in a social network [27]. Centrality defines how important a node is within a given network. Several centrality measures are designed to quantify *importance* and thereby to facilitate the answering of centrality questions [28]. Here we focus primarily on the most common versions

of four classic types of centrality measures, namely *degree*, *betweenness*, *closeness*, and *eigenvector* centrality, respectively.

Degree. For a given node, *degree* centrality is measured by the number of connections that it has with other nodes in a graph. In an *undirected* graph, the degree centrality is measured by the number of edges that are linked to a given node v and denoted as $deg(v)$. In a *directed* graph, the degree can be divided into in-degree $deg^{in}(v)$ and out-degree $deg^{out}(v)$. $deg^{in}(v)$ is the number of edges that point towards node v, while $deg^{out}(v)$ is the number of edges that originate at node v and point towards other nodes. In social networks, in-degree measures the popularity (prestige) of a node, and out-degree measures the node's own preference (gregariousness) for connections [26]. The measure of degree centrality mentioned here will be extended to weighted networks in Chapter 3, when we come to deal with correspondence networks.

The degree distribution of all the nodes in a network is one of the most fundamental aspects of network properties [19]. Networks following *power-law* degree distribution are called *scale-free* networks [29]. The probability of nodes having degree k, for large values of k, follows $P(k) \sim k^{-\gamma}$, where γ is called scale-free exponent and usually ranges in $[2, 3]$ for real networks [26]. In such networks, most nodes have low degrees and only a few nodes have much higher degrees.

Betweenness. *Betweenness* centrality is a measurement of how important a given node is in connecting other nodes. There are two types of betweenness. One is *node* betweenness and the other is *edge* betweenness. *Node* betweenness measures how often a given node lies on the shortest path between two other nodes. Nodes of high betweenness appear more frequently in the shortest paths between other nodes in the graph, and therefore play critical roles in the network structure [30]. The most commonly used betweenness centrality for a node i, is calculated as:

$$B_c(i) := \sum_{i \neq u \neq v} \frac{|u, i, v|}{|u, v|}, \tag{2.4}$$

where $|u, v|$ denotes the number of shortest paths between nodes u and v, and $|u, i, v|$ denotes the number of the shortest paths between u and v that pass through i. Take Figure 2.5 as an example. Node v_3 lies on the shortest path between v_1

and v_4 or v_5 and between v_2 and v_4 or v_5, therefore the betweenness centrality of node v_3 is $B_c(v_3) := 2 \times (1/1 + 1/1 + 1/1 + 1/1) = 8$. As node v_1 is not on the shortest path between all other pairs of nodes, the betweenness centrality of v_1 is 0.

On the other hand, the *edge* betweenness measures the number of shortest paths between two nodes that pass through a certain edge. It is also widely used in community detection [31]. The betweenness centrality for an edge e is calculated as:

$$B_c(e) := \sum_{u \neq v} \frac{|u, e, v|}{|u, v|}, \tag{2.5}$$

where $|u, e, v|$ denotes the number of the shortest paths between nodes u and v that contain edge e. For example, in Figure 2.5, the edge between node v_1 and node v_2 lies on the shortest path between v_1 and other nodes in the graph, therefore the betweenness centrality of (v_1, v_2) is $B_c((v_1, v_2)) := 2 \times (1/1 + 1/1 + 1/1 + 1/1) = 8$.

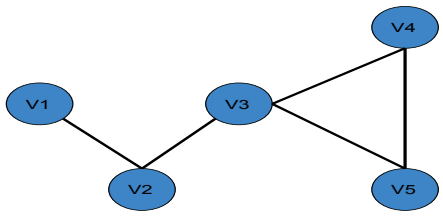


Figure 2.5: An example of a simple graph: an undirected and unweighted graph with five nodes and five edges.

Closeness Centrality. *Closeness* centrality focuses on how close a given node is to all the other nodes in the network [32]. The idea is that the more central a node is, the smaller the total geodesic distance to other nodes is, and the higher is the closeness score that the node gets. This is calculated as:

$$CL_c(i) := \frac{1}{\sum_{j \neq i} d(i, j)}, \tag{2.6}$$

where d denotes the geodesic distance between node i and node j. For instance, in Figure 2.5, the closeness centrality for node v_3 is $Cl_c(v_3) := 1/((2+1+1+1)/4) = 0.8$.

Eigenvector. *Eigenvector* centrality is a measurement of how central a node is with regard to the importance of its neighbors in a graph [19]. Many centrality measures, such as the Katz centrality [33] and the PageRank [34] centrality, are eigenvector-based. The eigenvector centrality is computed on the basis of the idea that the more central the neighbors of any given node are, the more central that the node itself is. In other words, the eigenvector score of a node should be proportional to the sum of the scores of its neighbors as:

$$C_{Eig}(i) := \beta \sum_{(i,j)\in E} C_{Eig}(j) = \beta \sum_{j\in N(i)} \mathcal{A}_{ij} C_{Eig}(j)\,, \tag{2.7}$$

where β is a fixed constant, and $\mathcal{A}_{ij}$ is the entry in the adjacency matrix corresponding to the edge between i and its neighbor j. By assuming that $C_{Eig} := (C_{Eig}(1), C_{Eig}(2), ..., C_{Eig}(n))^T$ is the centrality vector, we can rewrite the above equation as:

$$C_{Eig} := \beta A C_{Eig} \text{ or } A C_{Eig} = \lambda C_{Eig}\,, \tag{2.8}$$

where $\lambda = \frac{1}{\beta}$ is called an *eigenvalue* of adjacency matrix $\mathcal{A}$.

2.2.5 Reciprocity and Transitivity

Various kinds of relations exist between pairs of nodes in a network, of which one simple relation is "connected by an edge" [19]. If the "connected by an edge" relation is extended to "connected by edges in both directions", this relation can be called *reciprocal*. If the "connected by an edge" relation is extended to three nodes, for example, in the instance that node i is connected to node j and node j is connected to node k, this relation can be called *transitive*. In this section, we look at two measures, reciprocity and transitivity, to describe the behavior of mutual contacts of individuals in the real world.

Reciprocity. If there is at least one edge from node u to node v and at least one edge from node v to node u, these two nodes u and v are reciprocal to each other. A network with many mutual ties between nodes indicates a set of strong

social relationships. *Reciprocity* is calculated as the fraction of edges that are reciprocated and thus only appropriate for directed graphs [19]. It is measured as:

$$Reci(i) := \frac{1}{|\mathcal{E}|} \sum_{ij} A_{ij} A_{ji} \,, \tag{2.9}$$

where A_{ij} and A_{ji} are two entries of adjacency matrix A. The product of $A_{ij}A_{ji}$ is equal to 1 if and only if there are separate edges from i to j and from j to i, otherwise 0. However, this equation does not take the weight of reciprocal edges into consideration. We will discuss the weight situation in Chapter 3 and propose an approach to measuring reciprocity in weighted networks.

Clustering Coefficient. The *clustering coefficient*, which is also called *transitivity*, measures the degree to which nodes tend to form groups together [27]. It is defined on the basis of the number of triangles and triples in a graph. A *triple* is a path over a set of three nodes, whereas a *triangle* is a closed triple in which three nodes are all connected to each other. There are two types of clustering coefficient: the *global* clustering coefficient and the *local* clustering coefficient. The *global* clustering coefficient is an indicator of the clustering in the whole network. It is calculated as the fraction of the number of triangles (closed triplets) divided by the total number of triples:

$$CC(\mathcal{G}) := \frac{\text{number of triangles}}{\text{number of triples}}. \tag{2.10}$$

The *local* clustering coefficient $CC(i)$, for a node i, is the fraction of the number of edges that join pairs of neighbors of i over the total number of possible edges between the node's neighbors. $CC(i)$ measures the probability that a pair of nodes, which are both neighbors of i, are connected. It is calculated as:

$$CC(i) := \frac{2|N_e(i)|}{|N(i)||N(i)-1|} \,, \tag{2.11}$$

where $|N_e(i)|$ represents the number of edges connecting node $i's$ neighbors, and $|N(i)|$ represents the number of neighbors that i has.

2.3 Community Detection

In this section, we begin with a brief discussion of the approaches to defining communities in Section 2.3.1 and the representations of community structures including cliques and the star structure in Section 2.3.2. Then we describe a series of techniques that depend on graph clustering algorithms for discovering communities in Section 2.3.3.

2.3.1 Defining Communities

The definition of *community* is still the subject of debate and depends on the specific system and algorithm [35]. The general idea that most definitions of *community* share is that a greater number of edges connect nodes within a community than the number of edges connecting nodes of the community with other nodes in the graph. [30].

Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ and a subgraph $\mathcal{C} = (\mathcal{V}_c, \mathcal{E}_c)$ within $\mathcal{G}$, the *internal* degree $d^{int}(v)$ is measured as the number of edges connecting a node v in $\mathcal{C}$ to other nodes in $\mathcal{C}$, and the *external* degree $d^{ext}(v)$ is measured as the number of edges connecting $v \in \mathcal{V}_c$ to the rest of the graph. The internal degree $d^{int}(\mathcal{C})$ and external degree $d^{ext}(\mathcal{C})$ of subgraph $\mathcal{C}$ are the sum of the internal degrees and external degrees of all the nodes in $\mathcal{C}$, respectively. Based on these notations, the internal density $\rho^{int}(\mathcal{C})$ of $\mathcal{C}$ is measured as the fraction of the number of internal edges of $\mathcal{C}$ divided by the number of all possible internal edges using the following formula:

$$\begin{aligned} \rho^{int}(\mathcal{C}) &:= \frac{\text{number of internal edges in } \mathcal{C}}{\text{number of all possible edges in } \mathcal{C}} \\ &= \frac{\frac{d^{int}(\mathcal{C})}{2}}{\frac{|\mathcal{V}_c| \times (|\mathcal{V}_c|-1)}{2}} = \frac{d^{int}(\mathcal{C})}{|\mathcal{V}_c| \times (|\mathcal{V}_c|-1)}. \end{aligned} \tag{2.12}$$

Similarly, the external density $\rho^{ext}(\mathcal{C})$ is measured as the fraction of the number of edges from the nodes within $\mathcal{C}$ to the rest of the graph divided by the maximum possible number of this kind of edges using the following formula:

$$\rho^{ext}(\mathcal{C}) :== \frac{d^{ext}(\mathcal{C})}{|\mathcal{V}_c| \times (|\mathcal{V}| - |\mathcal{V}_c|)}. \tag{2.13}$$

Suppose that $\mathcal{C}$ is a community, then the internal density $\rho^{int}(\mathcal{C})$ is expected to be significantly larger than the external density $\rho^{ext}(\mathcal{C})$. The maximization of the difference between internal density and external density is the basis of most clustering algorithms of community detection in networks.

2.3.2 Community Structures

In real networks, the degree distribution always follows the power-law effect, i.e., most nodes have low degree and a few nodes have much larger degree. High concentrations of edges exist within specific groups of nodes, whereas low concentrations of edges exist between these groups [35]. This feature of real networks is called *community structure* [36], or *clustering*. Communities can thus also be called *clusters*.

The community structure, or structures, in a graph can be defined from either a *local* (subgraph) or a *global* (whole graph) perspective. The *local* definition focuses on the subgraph and regards the corresponding communities as maximal subgraphs, i.e., the nodes in a community are strongly connected to such an effect that no more nodes and edges can be added to the subgraph without losing the property of their *strong connectedness* [30]. On the other hand, the *global* definition focuses on the whole graph. A graph has a community structure if it is different from a random graph [35], or in other words, a group of nodes in the graph must have a significantly higher number of edges than the expected number of edges in a random graph.

A *partition* is a division of a network into groups, in which each node in the network belongs to one group. The two terms *graph partitioning* and *graph clustering* are both frequently used in network analysis. Generally speaking, both of these two techniques aim at the identification of groups of nodes with many internal and few external edges, but they are different with respect to whether the number and the size of the clusters in a graph is predefined or not. In contrast with graph clustering, graph partitioning usually implies that the number of partitions is fixed and the task is to partition the node set into blocks of almost equal size.

A *clique* is a maximal complete subgraph in which nodes are fully connected to each other [37]. The *triangle* is the simplest form of a clique and often appears in real networks. The *star* structure in a network consists of a central node and every other node in the network is connected to it. We define the star structure as:

Definition 2.7. Star Structure. Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a star structure is represented as a subgraph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}')$ of $\mathcal{G}$ in which there is a central node that is incident to all other nodes.

Star structures are very common in social networks, e.g., a community formed by Facebook users where a user has many friends [30]. Our correspondence networks are also typical star-structure networks and we will analyze them in Chapter 3.

2.3.3 Graph Clustering Approaches

Detecting community structures in a graph can generally be considered as a clustering problem [30]. A large number of clustering algorithms have been proposed and developed in fields such as computer science and sociology [38]. In this section, we discuss graph clustering approaches, which are divided into six categories and some of them will be useful in community detection for correspondence networks in Chapter 3.

Hierarchical Clustering. *Hierarchical clustering*, which was introduced by Johnson [39], is one of the earliest clustering methods. It functions by grouping nodes into a tree of cluster and is further classified into two categories, namely the *agglomerative* approach and the *divisive* approach.

- **The Agglomerative Approach to Hierarchical Clustering**. The *agglomerative* approach follows a *bottom-up* strategy. At the beginning, each node starts in its own cluster, and then the closest pairs of clusters are merged as one cluster. This merging process is repeated until all nodes have become clustered into one single cluster. There are three common ways to measure how similar two clusters are, namely the methods of *single linkage*, *complete linkage*, and *average linkage* [40].

- **The Divisive Approach to Hierarchical Clustering**. The *divisive* approach follows a reverse *top-down* strategy, which starts with all nodes in a single cluster and then splits them into smaller clusters recursively.

Partitional Clustering. Generally, *partitioning clustering* is the method whereby each data point is assigned to one cluster, given a number k of mutually exclusive clusters. One of the most commonly used partitional clustering algorithms is the k-means clustering [41]. This starts with k number of clusters set by the users and the centroid (center) of each cluster is initiated. Then each point in the dataset is assigned to the nearest cluster that has the closest distance from the node to the specific centroid. When all points have been assigned, the new centroids are re-calculated. These two steps are repeated until no more changes for centroids can be made. The k-means algorithm uses squared error (SE) as an objective function by which to minimise the total intra-cluster variance.

Spectral Clustering. *Spectral clustering* includes all methods and techniques that partition a set of objects into clusters by using the eigenvectors of matrices [35]. In particular, the objects could be either points in some metric space, or the nodes of a graph. This approach was first proposed by Donath and Hoffmann [42], who used the eigenvectors of the adjacency matrix for graph partitions. Spectral Clustering consists of the transformation of the initial set of objects into a set of points in space, whose coordinates are elements of eigenvectors: the set of points is then clustered via standard techniques. The change of representation induced by the eigenvectors makes the cluster properties of the initial dataset much more evident.

Density-based Clustering. *Density-based* clustering aims at finding clusters whose points appear within each cluster with a density that is considerably higher than it would be outside of the cluster [43]. DBSCAN is a well-known algorithm that follows this intuition. The key idea of DBSCAN is that for each point belonging to a cluster, the neighborhood of a given radius has to contain at least a minimum number of points, or in other words, the density in the neighborhood has to exceed some threshold.

Overlapping Communities. Most of the methods discussed in the previous paragraphs aim at detecting exclusive communities in a graph, i.e., each node is assigned to a single community. However, in real graphs, nodes are often shared

between communities, and the issue of detecting overlapping communities has become quite popular in the last few years [35]. The most popular technique for overlapping communities detection is the Clique Percolation Method (CPM) [44]. The concept of CPM presupposes that the internal edges of a community are likely to form cliques due to their high density. Palla *et al.* [44] proposed the term *k-clique* to indicate a complete graph with k number of nodes. Two k-cliques are adjacent if they share $k-1$ nodes. The union of adjacent k-cliques is called a *k-clique chain*. Two k-cliques are connected if they are part of a k-clique chain. Finally, a *k-clique community* is defined as the largest connected subgraph formed by the union of a k-clique and all k-cliques that are connected to it.

Quality Function. A *quality function* is a function to measure how good a graph partition is. It assigns a score to each partition and the partitions are ranked on the basis of this score. Some well-known instances of quality functions are as follows.

- **Performance**. The performance function counts the number of pairs of nodes that are correctly clustered into groups [45]. This function can be the number of pairs of nodes that are connected by an edge and are clustered into the same group, or the number of pairs of nodes that are not connected by an edge and are clustered into different communities. For a certain partition $\mathcal{P}_\mathcal{G}$ of a graph, the performance function is calculated using the following equation:

$$f(\mathcal{P}_\mathcal{G}) := \frac{\sum_{i,j,i<j}^{N_c} (|E(\mathcal{C}_i)| + |(u,v) \notin \mathcal{E} \mid u \in \mathcal{C}_i, v \in \mathcal{C}_j|)}{\mathcal{V} \times (\mathcal{V}-1)/2}, \tag{2.14}$$

  where $\mathcal{P}_\mathcal{G} = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_{N_c}\}$ is a partition of a given graph $\mathcal{G}$ and $|E(\mathcal{C}_i)|$ denotes the number of edges within a community $\mathcal{C}_i$. u and v correspond to two nodes in $\mathcal{G}$. The value of $f(\mathcal{P}_\mathcal{G})$ is within the range of $[0, 1]$. The higher the value is, the denser the intra-communities and the sparser the inter-communities are.

- **Modularity**. The *modularity* function, proposed by Newman and Girvan [31, 46], is a popular quality function based on the comparison between the

actual density of edges in a community and the expected density of edges in a *null model*, which is a random graph with the same expected degree sequence of the original graph [35]. Modularity is not only a basic concept that lies behind the global definition of a community, but also a quality function and the key ingredient of the most well-known graph clustering techniques. It can be written as,

$$Q(\mathcal{P}_{\mathcal{G}}) := \sum_{i=1}^{N_c} (\frac{|E(\mathcal{C}_i)|}{|\mathcal{E}|} - (\frac{D(\mathcal{C}_i)}{2\,|\mathcal{E}|})^2)\,, \tag{2.15}$$

where $E(\mathcal{C}_i)$ represents the number of total edges within the community $\mathcal{C}_i$, $N_{\mathcal{C}}$ denotes the number of communities in the graph $\mathcal{G}$, and $D(\mathcal{C}_i)$ denotes the sum of the degree of the nodes in $\mathcal{C}_i$. The first part of the equation above denotes the fraction of number of edges inside the community $\mathcal{C}_i$ divided by the number of total edges of graph $\mathcal{G}$, and the second part indicates the expected fraction of edges within a random graph (null model) that have the same degree for each node as $\mathcal{G}$. A subgraph is a community in which the corresponding contribution to modularity in sum is positive. The denser is the internal edges of the cluster in comparison with the expected number, the better the community is.

Louvain algorithm [48] is one of the modularity optimization algorithms that have gained popularity recently, since its accuracy is comparable to the accuracy of other algorithms but offer better scalability [49]. It consists of two steps which are repeated iteratively. First, given a weighted network of N nodes, initially each node is assigned to a different community. Then for each node i, the change in modularity is calculated if we position i into the community of each neighbor j of i. The change in modularity ΔQ is calculated using the following formula:

$$\Delta Q := \left[\frac{\sum_{in}+k_{i,in}}{2W} - \left(\frac{\sum_{tot}+k_i}{2W}\right)^2\right] - \left[\frac{\sum_{in}}{2W} - \left(\frac{\sum_{tot}}{2W}\right)^2 - \left(\frac{k_i}{2W}\right)^2\right], \tag{2.16}$$

where $\sum_{in}$ denotes the sum of the weights of all the edges inside the community $\mathcal{C}_i$; $\sum_{tot}$ denotes the sum of the weights of the edges of all the nodes

in $\mathcal{C}_i$; k_i denotes the sum of the weights of the edges incident to node i; $k_{i,in}$ denotes the sum of the weights of the edges between i and other nodes in $\mathcal{C}_i$ and W denotes the sum of the weights of all edges in the network.

After calculating ΔQ for all communities and node i is placed into the community that achieves the largest positive change in modularity. If no positive change is found, i stays in its original community. This process is applied to all nodes until no further improvement can be achieved [49], finishing Step I.

Second, a new network whose nodes are the communities from the completed first step is constructed. Edges between nodes of the same community are represented by weighted self-loops. The weight of the edge between two nodes in the new network is the sum of the weights of the edges between the nodes in the corresponding two communities in step I. Once Step II is finished, we repeat the process of Steps I and II until there are no more changes and maximum modularity is attained.

In summary, we introduce the necessary concepts of community structures and well-known community detection techniques in this section. These community detection algorithms are frequently used in social network study and Louvain algorithm will help us to explore the collaborations embedded in correspondence networks in Chapter 3.

2.4 Temporal Properties of Social Networks

In this section, we first introduce the basic concepts and representations of dynamics in social networks in Section 2.4.1. Then we discuss several focuses of temporal properties and important ideas concerning temporal patterns in Section 2.4.2, which are most relevant to our research. We will not describe any actual algorithms of measurements in this section, but the ideas form a foundation for the measurements that will be proposed in Chapter 3.

2.4.1 Temporal Representations in Social Networks

Holme [50] refers to the basic unit of relation in static networks as a "link" and the basic unit of "interaction" in temporal study as a "contact". Temporal study of networks capture the information with respect to the interactions of two nodes, the

time of the interactions, and the duration of the interactions. But being connected in static networks might not be true in temporal study when the temporal order of paths is considered. In static networks, for instance, if there is an edge from node i to node j and another edge from node j to node k, there is a path from i to j via k, and in this way node i and node k is connected. However, it might not be true for network with temporal information. If the interaction between j and k takes place before i and j, i cannot reach k. The events in temporal study represent the temporal sequence of interactions between nodes [51], and many systems have been modeled integrated with temporal information such as collaboration networks [52], transportation networks [53] and human interaction networks [54]. The temporal information contained in these networks is non-trivial if we want to trace the transmission of a disease, the dynamic spread of tweets, or the influences of individuals over time, to name but a few possible examples.

We divide the dynamics in social networks into two types: *evolving* networks and *temporal* networks. *Evolving networks* focus on the question whether the topological properties of social networks follow specific patterns over time [30]. They are considered as a type of *generative model* [56]. For instance, the Barabási-Albert (BA) model is a widely accepted evolving network model, which is used to generate scale-free networks using preferential attachment [57]. Preferential attachment is a process that a new node joins the network and typically has a higher probability of forming a link to the more connected nodes than the less connected nodes in the network. It is applied in the context of academic networks to explain that new researchers are more likely to collaborate with well-known colleagues [58]. In contrast to evolving networks, *temporal networks*, also called time-varying networks or dynamic networks, are characterized by the activation of distinct nodes and edges [56]. In other words, temporal networks focus on the fluctuation in nodes and edges at discrete points in time and specific topics such as the role of individuals in the spreading of diseases and the influences of individuals.

Systems with interactions and corresponding temporal information can be divided into two major types of representations [55]. The first is called *contact sequences*. When the duration of the interactions are negligible, the system can be represented as a set of C contacts (i, j, t) where i and j are two interacting nodes and t denotes the time of the interaction. In other words, the interactions are assumed to be instantaneous [56]. Alternatively, the contact sequences can also be represented

as a set of edges where each edge is a pair of nodes and is associated with a set of the time of the interaction. Typical systems include email, instant message, text messages, among others.

The second type of representation is an *interval graph* in which the durations of the interactions are non-negligible. The edges are active over a set of time intervals instead of being active at a single point in time, e.g., telephone calls [50]. In other words, a network can be represented as a time-ordered sequence of *snapshots*, each of which is an observation of a network within a given *time window* [59]. The size of each time window can either be a point in time, or a time interval. In each time window, a snapshot either is a static graph at a specific point in time, or it consists of aggregated static graphs constructed by combining all edges present within a predefined time interval [60]. In the case of a snapshot-based network, one could study each snapshot independently via the existing methods for static network analysis and then analyze the time-series of the results [61]. However, this strategy treats each network snapshot in isolation and valuable temporal information might be lost between snapshots.

2.4.2 Temporal Measurements in Social Networks

In this section, we first review temporal path proposed for characterizing temporal-topological structure. We also discuss some related methods proposed for characterizing temporal patterns in networks that are relevant to our study.

Temporal Path. The *temporal path* is given different names by different researchers [56]. It has also been called a *journey* [62] or a *time-respecting* path [63][64]. Existing path-related measures for static graphs are based on a network model where edges and nodes are aggregated into a single static graph [65]. However, in networks with temporal information, the paths that traverse nodes through a network are not static but rather change in time. Hence, the paths in networks with temporal information are usually defined as sequences of contacts following a certain temporal-order and connecting sets of nodes [55]. In most cases, the temporal paths are measured with duration, i.e., they begin and end at a certain point in time. A crucial concept in graphs with temporal information is *journey*, which corresponds to the temporal extension of the notion of path, and

forms the basis of most recently introduced temporal concepts [67]. Journeys can be regarded as paths over time from a source to a destination and therefore have both a topological and a temporal length [68].

There are several different names of the *shortest paths* in networks concerning related temporal information and objectives, e.g., the *fastest* path, the *earliest-arrival* path and the *foremost* path [66]. The shortest path is defined as the journey that has geodesic distance within given observation windows. The fastest path is defined as the journey that has the minimum traverse time. The earliest-arrival path is defined as the journey that arrives at the target node the earliest. The foremost path is defined as the journey that reaches a target node the latest. These measurements of temporal paths focus on capturing both topological and temporal features. Similar approaches are also proposed [62][64][69] with still other names. For instance, the *fastest* time within which one node can reach another is called information latency [64], temporal distance [51][62], the reachability time [63] or temporal proximity [69].

As an extension of static networks, temporal extensions of networks adopt and generalize many concepts such as centrality measures from static network analysis. They are built on the basis of the temporal paths and pay more attention to the topological structures of the network. However, since our correspondence network only contains a point in time for each edge instead of durations and most edges exist between the central node and others, the temporal centrality measures and temporal path are not our focus in the following chapters.

Inter-event/Inter-contact Time. Techniques such as the burstiness [63] and persistent patterns [70], are exploited in networks to explore interesting temporal patterns. In a time series of contacts or events, the *inter-event* or *inter-contact* time distribution is the frequency distribution of the time between two consecutive contacts or events [50]. In empirical datasets, the inter-event time distribution is usually heavy-tailed, or even scale-free [71, 72]. In other words, the inter-event time distribution for human communication dynamics is often *bursty* [63] and the bursty time series are usually characterized by their coefficient of variation, which is also called *burstiness* [73]. Some researchers measure the burstiness of individual nodes and have discovered the bursty structure when people send e-mails [74].

Persistent Patterns. Temporal patterns such as certain links and subgraphs that do not change so much as others over time have also attracted the attention of researchers. Clauset and Eagle [70] proposed a similarity measure, which they called *adjacency correlation*, to compare a node's connectivity at one snapshot to the next. Similarly, Valdano *et al.* [75] defined the *loyalty* of a node as a local measure of its tendency to maintain contacts with the same elements for a pair of two consecutive snapshots, and uncovered non-trivial correlations with the node's "epidemic risk". Neiger *et al.* [76] addressed the problem inversely by measuring how connected changing links behave in the network. They constructed a network in which nodes were fixed and links changed from one time step to another. They analyzed the concentration of changes in dynamic networks and found a very restricted set of nodes of the network that could be responsible for changes.

Human Correspondence Activity. Oliveira *et al.* [78] indicated that Darwin's and Einstein's late responses for resumed correspondences were not singularities or exceptions, but rather represented a fundamental pattern of human dynamics, i.e., famous people were no better at escaping than the majority. Malmgren *et al.* [79] indicated that like emails, the correspondence patterns of 16 writers, performers, politicians, and scientists were well described by the circadian cycle, task repetition and changing communication needs. They discovered that human correspondences could be accurately modeled as a cascading non-homogeneous Poisson process and this process could give rise to heavy-tailed statistics, but not to power-law statistics characterized by critical exponents.

Many other measurements of temporal patterns in social networks have been studied in addition to the ones described in this section. However, considering that we are dealing with specific correspondence networks that many temporal patterns cannot be feasibly measured, we will focus mainly on the patterns we mentioned above.

2.5 Historic Correspondence Research

We begin in this section with an introduction of historic correspondence research and related projects in the area of digital humanities, before turning to focus on the current study on correspondence networks. After this we categorize them into different types in terms of the sources and objectives.

The term *historic* (or historical) is defined as "famous or important in history, or potentially so" by the Oxford English Dictionary [80] and the term *correspondence* is defined as "communication by exchanging letters (or other forms)" [81]. The term *historic texts* is defined as texts written in languages which are different from languages currently in use for the purpose of natural language processing [14]. *Historic Correspondences*, in the broadest sense, refers not only to letters which were written in historic languages, but also to commissions, petitions, instructions and speeches written in the past [109]. These historic documents bring the historic figures in them to life and allow us to explore the individuals and the society in the past.

Projects	Major Language	Time Period	Features
EMLO[1] [7]	English	1550–1750	search platform linked data
CEEC[2][82]	English	1418–1680	Pos-tagging
CELL[3][8]	English	1500–1800	search platform visualization
CERA[4][83]	Latin	1520–1770	digitization
CKCC[5][3]	Dutch	17th century	visualization topic modeling search platform
Republic of Letters [84]	English	1400–1800	search platform visualization
Electronic Enlightenment [85]	English	17th–19th century	search platform
Frühneuzeitliche Ärztebriefe [86]	German	1500–1700	linked data search platform
Italian Literary[87]	Italian	16th–17th century	digitization
Literary World [88]	English	19th century	digitization
Vernetzte Korrespondenzen [89]	German	1939–1945	search platform

Table 2.1: A list of well-known epistolary centers working on historic correspondences. In this table we list the major languages of historic letters, the time periods of the letter collections, and the research features of these projects, respectively.

2.5.1 Correspondence Analysis in Digital Humanities

The digitization of historic texts makes the correspondence collections more accessible to both academics and the general public who are interested in the lives of the letter writers [90]. Digitized historic correspondences embody a rapidly growing field of research and involves collaborations of historians, computer scientists, and researchers in language, literature, social science, physics, and so on. Tables 2.1

1 Early Modern Letters Online.
2 Early English Correspondence.
3 Center for Editing Lives and Letters.
4 Corpus Epistolicum Recentioris Aevi.
5 Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic.

and 2.2 list current well-known epistolary centers and their corresponding correspondence projects. Most of the previous work in this area can be organized into three general categories in terms of their objectives: a) the digitization of historic correspondences, b) the search platform for historic correspondences, and c) the representation of different types of relationships embedded in historic correspondences. The digitization of letters includes scanning, optical character recognition (OCR), correction, manual annotation, and markup on letter contents [83, 87, 88], to name but a few common tasks in this field. Based on the digitized letters, researchers make efforts to build a search platform for correspondences based on the metadata of digitized letters (e.g., person names, place names, dates, to name but a few) [3, 8, 85]. Due to the spelling variations, existence of ambiguous entities, and the effect of language changes, content analysis such as Part-of-Speech (POS) tagging and semantic analysis of letters in previous studies [9, 82] depends heavily on different researchers with their specialist expertise and languages. This causes differences and divergences when integrating corpora from different sources. The most typical and commonly represented type of relationships embedded in historic letters is the sender-recipient relationship in network visualization [3, 8, 84].

2.5.2 Correspondence Network Study

The term *network* has frequently been used in historic studies to describe the dissemination of correspondences by scholars in the seventeenth and eighteenth centuries [91]. Correspondence networks can be classified into *individual* networks and *group* networks, according to the scale and the size of a letter collection. Correspondence networks can also be classified into *sender-recipient* networks, *co-citation* networks, and *author-topic* networks, according to the research focus of a specific type of relationships embedded in letter collections. Moreover, correspondence networks can also be classified into *static* networks, *dynamic* networks, *geographical* networks and *spatio-temporal* networks, according to the dimensions of networks involved in visualization. In this section, we look in detail at these different types or focuses of correspondence networks.

on the one hand, an *individual* correspondence network is a network built from one person's own letter collection. It is an *ego-centric* network, in which one specific person is at the center of the graph. This star-structure network includes letters between this person and his/her correspondents, but the miscellaneous

Projects	Major Language	Time Period	Features
Martin Opitz von Boberfeld [92]	German	1597–1639	search platform
Thomas Bodley [93]	English	1585–1597	search platform visualization
Bess of Hardwick's Letters [94]	English	1550–1608	search platform
Electronic Capito [95]	English	16th century	digitization
Carolus Clusius [96]	Latin	16th–17th century	digitization
Alfred Newton [97]	English	1672–1676	digitization
The Spenser Letters [98]	English	1580–1589	digitization
Darwin Correspondence [9]	English	1837–1883	sentiment words timeline linked data
Thomas Gray [99]	English	1716–1771	search platform
Sir Hans Sloane [100]	English	1680–1745	search platform
Françoise de Graffigny [101]	French	18th century	digitization
Hugo Grotius [102]	Dutch	1597–1645	digitization
The Cullen Project [104]	English	1710–1790	tagged letters
Constantijn Huygens [105]	Dutch	1608–1687	search platform
Ioannes Dantiscus [106]	Latin	18th century	digitization
Johann Valentin Andreae [107]	German	17th century	search platform
The Linnaean Correspondence [108]	Swedish, Latin	18th century	search platform
William of Orange [109]	Dutch	1549–1584	search platform
Oswald Myconius [110]	German	1488–1552	digitization
William Dugdale [111]	English	1635–1686	digitization
Philipp Jakob Spener [112]	German	1691–1705	digitization

Table 2.2: A list of major projects working on individual correspondences. In this table, we list the major languages of historic letters, the time period of the letter collections, and the research features of these projects, respectively.

letters shared between his correspondents are seldom included in the network. On the other hand, a *group* correspondence network is a network built from the letter collections of groups of people living during the same time period and having intersections. It is a *socio-centric* network and can be viewed as a combination of many individual correspondence networks. Most studies have focused on the correspondences of individual scholars [9, 93, 97]. They create machine-readable catalogues of correspondence with appropriate metadata for analyzing academic exchanges [2]. Currently several projects [3, 7, 84] are aiming to transcend the limits of individual correspondences by integrating resources from different projects. However, these networks highly depend on the preservation of the letters written by both well-known individuals and less-known individuals. Letters of women, provincials, non-Europeans, and artisans are less well-preserved than the letters of male scholars from the upper European society concerning participation in scientific academies [2]. Moreover, collecting the scattered corpora of letters from countless libraries, archives and private collections has been a non-trivial task. These problems make it difficult to acquire a full appreciation of historic times based on the large-scale exploration and analysis.

Sender-recipient relations represent the most intuitive relationship embedded in historic correspondences and are frequently represented in network visualization [3, 84]. This "who is writing to whom" relation is visualized in a directed graph with the people being nodes and the letters being edges. *Co-citation* networks represent the co-occurrences of person names mentioned in the same letter. This "who is mentioned together with whom in a letter" relation is visualized in an undirected graph with the individuals being nodes and their co-occurrences in the letters being edges [3]. The person who is mentioned most in the letters appears in the center of the graph. *Author-topic* networks represent the relations between topics embedded in letter contents and authors of the letters. This "which topic is mentioned by whom" relation is visualized in an undirected graph with multi-type nodes and edges [113]. A few projects deal with the relations between multiple senders and recipients [114], or the relations between correspondents and people mentioned in the letters [93]. However, most research efforts have given attention only to the network visualization of relationships, with the result that research into the strength and interconnectedness of these networks is still in its early stage. Few projects have moved beyond the above stage to exploit the detailed modeling of correspondence networks, let alone to develop new concepts and algorithms derived from modern social network analysis.

Figure 2.6: Visualization of correspondence network of academic letters in the 17th century Dutch Republic as shown on a geographical map. http://ckcc.huygens.knaw.nl/epistolarium/# [Last accessed: September 25, 2015].

Recently, more and more researchers have paid attention to the geographical and temporal visualization of correspondence networks. Moreton [115] visualized the geographical distribution of correspondences on a map with the locations of senders and recipients being nodes and their letter interactions being edges. Heuvel

[116] proposed to combine the spatial distributions with the individual correspondence network in order to represent the exchange of knowledge in Early Modern Europe better. However, he did not give a detailed description of how to combine these two aspects, either the modeling of spatial distributions or the individual correspondence network. Circulation of Knowledge and Learned Practices in the 17th century Dutch Republic (CKCC) [3] visualized the locations of senders and recipients as points on a geographical map. They used undirected lines to connect locations, and the width of each line indicated the number of letters between two individuals. Independent from network visualization, CKCC [3] and the Darwin Correspondence Project [9] have both exploited interactive timelines to explore letters over time. Moreover, mapping the Republic of Letters [84] integrated the network with the spatial and temporal dimensions together, in order to present correspondences geographically and dynamically. Each sender-recipient network was embedded in a geographical map and combined with a timeline stacked bar chart. However, none of these projects gave a detailed description of how they dealt with the uncertain spatial or temporal entities in their historic letters. No meta-analytical literature exists regarding how they extrapolated and visualized the uncertain data in meaningful ways.

2.6 Topic Modeling

One emerging issue that researchers in digital humanities now pay increasing attention to is the possible application of statistical language models, such as topic modeling, to interpret the meaning embedded in letter contents. In this section, we first introduce the concept of topic in texts in Section 2.6.1, before we move to provide an overview of one of the most classical topic models, LDA, in Section 2.6.2. Then in Section 2.6.3, we will discuss the application of topic modeling in different areas, such as the digital humanities and email analysis, which are most relevant to our study.

2.6.1 Concept of Topic in Texts

A *topic* is defined by the Oxford English Dictionary as a matter or subject dealt within a text, discourse, or conversation [117]. David Blei defines *topic* in the context of modeling as a probability distribution over a fixed vocabulary of terms [118]. In other words, a topic is a group of words that tend to occur together in

the same context, while the same word is allowed to appear in different contexts. For instance, a list of words "football, basketball, swimming, badminton" with corresponding word frequency is labeled as the topic *Sports*. Another list of words such as "badminton, Olympic, games, winner" also contains the word "badminton" and is labeled as "Olympics". These two word-lists reflect different patterns of word usage that include the word "badminton". And these two topics themselves are the recurring verbal patterns of co-occurrences.

2.6.2 Latent Dirichlet Allocation

Topic modeling is an approach that is used to discover the hidden topics that pervade a large and unstructured collection of documents automatically [120]. One of the most classical topic models is Latent Dirichlet Allocation (LDA), which was presented by David Blei [118] in 2003. This approach involves Bayesian statistics and optimization algorithms. The intuition that lies behind LDA considers a document as a mixture of topics and a topic as a mixture of terms. There are three following major assumptions of LDA.

1. The order of the words within an analyzed text is irrelevant.
2. The order of the documents from an analyzed corpus is irrelevant.
3. The number of topics is previously known.

Formally a topic is a multinomial distribution of words, and a document is associated with a multinomial distribution of topics. We can describe LDA more formally with the help of the following notations.

In Table 2.3, K denotes the number of topics in the collection, ϕ denotes a discrete probability distribution over a fixed vocabulary that represents the corresponding topics, θ denotes a document-specific distribution over the available topics, z denotes the topic index for word w, and α and β denote hyperparameters for the symmetric Dirichlet distributions from which the discrete distributions are drawn. ϕ_z corresponds to the multinomial distribution of terms in a topic z, and each θ_d corresponds to the multinomial distribution of topics in document d.

Variable	Dimension & Type	Description
K	Integer	# of topics
V	Integer	# of unique terms
D	Integer	# of documents
N	Integer	# of tokens
θ	$D \times K$ of probabilities	Topic distribution in documents
ϕ	$K \times V$ of probabilities	Word distribution in topics
α	$D \times K$ of α priors	Dirichlet prior for θ
β	$K \times V$ of β priors	Dirichlet prior for ϕ
w	N-Vector of word identity	Words in documents
z	N-Vector of topic assignment	Topic assignment of words

Table 2.3: Notations for LDA and # means "the number of".

If one knows ϕ_z and θ_d beforehand, then the probability that a word w in d belongs to topic z is calculated as follows:

$$P(w, z|\phi_z, \theta_d) := P(z|\theta_d)P(w|\phi_z). \tag{2.17}$$

However, ϕ_z and θ_d are always hidden and only the documents are observed. In other words, the topic structure, i.e., the distribution of terms in any topic and the topic proportion of any documents, are hidden. Therefore, the observed documents should be exploited to infer the hidden variables. Suppose ϕ_z and θ_d are generated by two respective distributions $P(\phi_z|\beta)$ and $P(\theta_d|\alpha)$. In this scenario, the joint probability of word w and topic z in document d is represented as follows:

$$p(w, z, \theta_d, \phi_z|\alpha, \beta) := p(\phi_z|\beta)p(\theta_d|\alpha)p(z|\theta_d)p(w|\phi_z). \tag{2.18}$$

Dirichlet Distribution. Dirichlet distribution is an exponential distribution over the simplex of positive vectors that sum to one [119]. LDA employs the Dirichlet distribution, which is calculated as follows:

$$Dir(\theta|\alpha) := \frac{\Gamma(\sum_{i=1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}, \tag{2.19}$$

where Γ denotes the Gamma function, which can be regarded as a real-value extension of the factorial function. A *symmetric Dirichlet* is a Dirichlet where each component of the parameter is equal to the same value. According to Table refTab:tab:ldanotation, LDA contains two Dirichlet random variables, θ denotes the distribution over topics and ϕ denotes the distribution over the vocabulary. The joint distribution of the hidden and the observed variables corresponds to the following generative process for LDA.

1. Sample the distributions of terms in topics $\phi := \{\phi_z \sim Dir_{|V|}(\beta)\}$, in which $Dir_{|V|}(\beta)\}$ denotes a V-dimensional Dirichlet with hyperparameter β.
2. for each document d
   (a) sample topic proportion $\theta_d \sim Dir_{|z|}(\alpha)$
   (b) for each word w in document d
      i. sample a topic index $z \sim Mult(\theta_d)$
      ii. sample term w in the selected topic z, i.e., $w \sim Mult(\theta_z)$.

The computational problem of inferring the hidden topic structure from the documents is specifically the problem of computing the posterior distribution, i.e., the conditional distribution of the hidden variables given the documents.

Posterior Inference for LDA. This refers to the conditional distribution of the topic structure in the model given the observed data. Using our notation, the posterior is calculated by the following equation,

$$P(\theta, \phi, z|w, \alpha, \beta) := \frac{p(\theta, \phi, z, w|\alpha, \beta)}{p(w|\alpha, \beta)}, \tag{2.20}$$

where the numerator denotes the joint distribution of all the random variables, which can be easily computed for any setting of the hidden variables. The denominator represents the marginal probability of the observations, which is the probability of seeing the observed corpus under any topic distribution. However, the factor $p(w|\alpha, \beta)$ is very difficult to compute. A central research goal of modern probabilistic topic modeling is to develop efficient methods for approximating it [120]. Topic modeling algorithms generally fall into two categories, namely *sampling-based* algorithms and *variational* algorithms.

Sampling-based algorithms attempt to collect samples from the posterior in order to approximate it with an empirical distribution. The most commonly used algorithm for topic modeling is the Gibbs sampling algorithm [121], which is a family of algorithms from the Markov Chain Monte Carlo (MCMC) framework. Gibbs sampling aims to construct a Markov chain, i.e., a sequence of random variables, in which each of the variables is dependent on the previous one, whose stationary distribution is the target posterior. In other words, after a number of iterations of stepping through the chain, sampling from the distribution should converge to become close to sampling from the desired posterior. Rather than approximating the posterior with samples, *variational* methods posit a parameterized family of distributions over the hidden structure, and then find the member of that family that is closest to the posterior [122, 123].

Estimating the number of topics. LDA allows us to model the topic distribution of a given collection of texts. However, it requires a given number of topics to estimate topic and word distributions. Many researchers have tried to estimate the number of desired topics automatically [124, 125, 126]. Even though various methods differ in their respective focuses, they all follow the same idea of computing similarities (or distances) between pairs of topics over a series of instances of the model with varying numbers of topics. The desired number of topics of a given collection is reached when the overall dissimilarity between topics reaches its maximum value. In the following paragraphs, we describe several algorithms that are used in selecting the appropriate number of topics in the text collections.

A relatively simple way to find the "right" number of topics without training data that has been proposed by Griffiths *et al.* [127], is looping through models with different numbers of topics, in order to find the number with the maximum log-likelihood. Arun *et al.* [124] used a Singular Value Decomposition (SVD) to represent the separability between the words contained in the vocabulary. If the singular values of the topic-word matrix are equal to the norm of the rows in the matrix, this means that the vocabulary is well separated among the topics. This method is evaluated with the Kullback-Liebler divergence (KLD) metric for each topic space. But this method is time-consuming and not rigorous.

Teh *et al.* [125] proposed the Hierarchical Dirichlet Process (HDP) approach to find the appropriate number of topics in LDA, by assuming that the groups of data have a predefined hierarchical structure. Each pre-defined group is associated with a Dirichlet Process (DP) whose basic measure is sampled from a higher-level DP. HDP replaces the finite topic mixture in LDA with a DP, and gives the different mixing proportions to each doc-specific DP. Teh *et al.* constructed both the LDA model and the HDP model for one corpus, and found that the posterior sample of the number of topics used by HDP is consistent with the best parameter k of the LDA model. Being different from HDP, Cao *et al.* [126] found the connection between LDA model performance and topic correlations, and adaptively guided the generation of the topics by the topic density statistics in the parameter estimation process. However, the selection of topics is not made in terms of their significance, and no criterion is given for topic ranking.

2.6.3 Topic Modeling Applications

The objective of topic modeling, i.e., to discover the latent topic information in large text collections, attracts the attention of researchers from different areas. In this section, we discuss the related studies in the areas of digital humanities and email analysis, since they are most associated with our correspondence network study.

In the area of digital humanities, topic modeling is widely used as a discovery tool to navigate large archives [128], and topics are often described as *discourses*, *topoi* or *rhetorical frames*. The first publication of topic modeling in historic studies was the analysis of an American newspaper between 1728–1800, consisting of $80,000$ texts in the form of articles and advertisements [127]. The analysts discovered that most identified topics were trivial or just noise. In 2006, David J. Newman and Sharon Block [129] used topic modeling on the 18th century Pennsylvania Gazette. They made a list of the most likely words in a topic and the label they assigned to that topic. In this list, some of the topics are obvious, but others are not so easy to understand if you do not know the context of the corpus.

Most topic modeling-related research in digital humanities can be ascribed one of two approaches: the synchronic or the diachronic study. The synchronic study

analyzes topics at a specific time. For instance, Meeks [130] employed the topic modeling tool MALLET on 50 texts which discuss digital humanities and visualized document-topic networks. Rather than creating new models, researchers create networks out of topic models that have been generated by data. Using network visualization, they can see how documents relate to one another, how documents relate to topics, how topics are related to each other, and how all of those are related to words. In contrast, the diachronic study focuses on temporal dynamics of topics, i.e., by charting the topics over time. For example, Nelson [131] analyzed changes in topics over time to explore social and political life in Richmond during the American Civil War. Blevins [132] used topic modeling on the diary of Martha Ballard to identify topic trends over 27 years. Mimno [133] used topic modeling on 24 classics journals spanning over a century to observe how topics in the journals changed over time and how the journals became more different or more similar over time.

Email, similar to historic correspondence, was previously widely used as a highly effective communication tool for exchanging information among people. Each email message also contains two components: a) the header and b) the body. The header usually consists of a set of entities, such as the sender/recipient ("From/To"), the title, and the date. The body consists of texts, sometimes with attachments, figures, URL links, and so on. Considering the similarity between emails and correspondences, we also look into the existing research on email network analysis. Freeman [134] defined an email network as one type of social network in which the people who send and/or receive emails are nodes and the email messages themselves are links. Quite a few social network techniques have been applied in this area and one of the major datasets is Enron Corporation's email corpus. This is a publicly available corporate email collection containing 150 users (mostly senior management of Enron) and $0.5M$ messages [135].

There are two main approaches for email network analysis. One is the social network approach on the header of an email. This approach focuses on the topological structure of email networks and the dynamics of emails between people. Rowe *et al.* [136] proposed a social network analysis algorithm to rank the social hierarchy of users based on the responsive time of users and centrality measures. All the statistics are normalized and combined to calculate an overall social score with which the users are ultimately ranked. Diesner *et al.* [137]

used the social network analysis techniques to extract properties of the Enron network and to detect the key players around the time of Enron's crisis. However, admittedly, there is a lot more information embedded in the content of email.

The other approach is to enrich the header-based network analysis with email content. Common topics between two people can be learned from the content of emails between them. Rosen-Zvi *et al.* [199] proposed a LDA-based model, i.e., the *Author-Topic Model*, which assumes that topics are formed on the basis of the mixtures of authors writing a paper. McCallum *et al.* [138] extended it to their author-recipient-topic (ART) model, which consists of fitting a multinomial distribution over topics and authors/recipients of a message simultaneously. The topics generated in this model are for different sender-recipient pairs. However, few of them have focused on unified measurements on author or individual topics and sender-recipient pairs.

2.7 Data Uncertainty in Digital Humanities

Recent years have seen the emergence of historic network research, in which the handling of ambiguous historic data become an inevitable problem for comprehensive text and network analysis. In this section, we first introduce the concept of data uncertainty in Section 2.7.1, before turning to discuss the methods that are used in named entity recognition and disambiguation in Section 2.7.2, since they are applied in natural language processing and information retrieval to deal with the data uncertainty. In Section 2.7.3, we give an overview of measures of text similarity and the application of topic models in text similarity, since this is associated with one of our main focuses in Chapter 4 and will be integrated into our data uncertainty approach.

2.7.1 Concept of Data Uncertainty

Longley *et al.* [139] define *uncertainty* in general within this context as the acknowledgment and consideration of imperfections in information. Pitrowsky [14] attributes the cause of the uncertainty in historic texts to the fact that digital historic texts are not originals but transcriptions. The transcriptions always cause uncertainty, modifications and errors. Plewe [140] divides this uncertainty in historic records into three types.

- *Unknown Uncertainty*: the uncertainty is so great that the encoder cannot determine where, when, or how something exists.
- *Imprecise Uncertainty*: the exact value is not known, but can be limited to one or more possibilities that hopefully includes the correct value.
- *Inaccurate Uncertainty*: a record is in error, even if it appears precise.

The issue of uncertainty is particularly acute in interpreting historic data from early time periods [141]. The combination of diverse historic data from a variety of sources confronts us with vague, uncertain, or even conflicting information [142]. Uncertain data, in the form of entities (e.g., person names, location names, or dates), are either ambiguous, approximate, or missing in the metadata of letters. This can result in not only data redundancy [143], but also inaccuracies in information retrieval and knowledge extraction.

Andert *et al.* [144] developed a platform for capturing metadata from historic correspondence. They described the uncertainties that existed in person, address, and date information, and recorded the corresponding degrees of uncertainties. However, they did not propose any approach to refine uncertain data. Few studies in digital humanities have explicitly addressed the problem of uncertain entities in the metadata of historic correspondences. There is a notable lack of literature in this field that is relevant to exploring how uncertain data should be extrapolated in meaningful ways. In this dissertation, we confine the uncertain data in the historical correspondences to missing or ambiguous person names, location names and dates in the letter metadata. We consider the refinement of data uncertainty in historical letters as the disambiguation of entities, and introduce the major studies of named entity disambiguation in the following section.

2.7.2 Named Entity Disambiguation

Entity in the Oxford English Dictionary is defined as "a thing with distinct and independent existence" [145]. It can be nominal such as "consciousness". In the Sixth Message Understanding Conference (MUC-6), a named entity, proposed by Grishman and Sundheim [146], involved person names, organization names, and geographic locations as well as time, currency, and percentage expressions. Similarly, in the information extraction task of the Seventh Message Understanding

Conference (MUC-7)[147], named entities were defined as proper names (person, organization, and location names) and quantities of interest (dates, times, percentages, and monetary amounts). Named entity recognition (NER) and named entity disambiguation (NED) are two important subtasks of information extraction.

The task of named entity recognition (NER) is the process that any phrases referring to an entity are identified in a given text [148]. In this process, each phrase is called a *mention* [149]. While early studies of NER were mostly based on handcrafted rules, most recent studies now use supervised machine learning techniques from a training corpus [150]. Although there are many NER taggers such as the Stanford NER tagger[1] for modern languages, few of them deal with historic languages, not to mention a certain historic language within a specific time period. Grover *et al.* [151] built a rule-based NER system for person and place names with handcrafted rules in digitized records of British parliamentary proceedings from the late seventeenth to the early nineteenth centuries. Borin *et al.* [152] also built a rule-based NER system for nine types of entities with handcrafted rules on 19th Century Swedish Literature. In the past decade, many projects have devoted themselves to the name identification of locations but few to the further disambiguation task [155].

Considering that we do not have experienced historians or linguists to write rules, and that we lack labeled or annotated texts of letters for training, we have chosen to focus only on the entity disambiguation task, i.e., the process of resolving the appropriate meaning of an entity mention in a certain context [153]. The task of named entity disambiguation (NED) consists of the disambiguation of mentions of entities and the mapping of them onto the entities in a given entity collection or knowledge base [148]. Most of the proposed NED algorithms assume that a knowledge base can provide explicit and useful information to help disambiguate a mention to the right entity [154]. Smith *et al.* [155] used a gazetteer for identifying and disambiguating geographical names in a historic digital library. However, most existing knowledge bases such as Wikidata [156] are created and maintained by multiple editors (volunteer contributors) instead of by experts. Without enough context coverage and the verification of knowledge by historians, these knowledge bases are not appropriate or sufficient for the disambiguation of named entities, such as historic person names or place names.

Furthermore, not everything is added to the knowledge bases and many person names do not appear in Wikipedia. For instance, Wikidata may wrongly disambiguate the mention "Kaspar Müller" to a joiner, whereas actually in our corpus, he was a German *Kanzler* at that time. Not only that, knowledge bases such as Wikidata[2] and DBPedia[3] cannot provide sufficient information concerning the time when a certain letter was sent, even though both the sender and the recipient of the letter can be queried correctly in the knowledge bases. It seems infeasible to find the real author or potential date range of letters with the help of knowledge bases. Therefore, we shift our attention to the research into entity disambiguation (in which not only named entities are included, but also entities such as email address, files, homepage, among others) combined with network structures or additional information.

Generally speaking, three types of probabilities are explored in the NED task [154].

1. **Entity Popularity**. This is based on the assumption that the number of incoming and outgoing relationships among entities — specifically, the number of edges (links) — is its *popularity* [157]. The *prominence* or *popularity* of entities can be seen as a probabilistic prior to mapping a mention onto an entity [158]. Usually the popularity of an entity is estimated as the knowledge base frequencies of certain mentions in hyperlink anchor texts which refer to specific entities. However, if we merely depend on this probability, it will disambiguate all appearances of a mention to a fixed entity, rather than disambiguating the contexts along with them [154].

2. **Context Similarity**. Bunescu and Pasca [159] calculated context similarity by comparing the textual context around a mention to the Wikipedia categories associated with each entity candidate. Li *et al.* [154] defined context similarity between the text of the mention and the page describing the referred entity in Wikipedia. This probability complements the entity popularity and is widely used in NER task.

[1] https://nlp.stanford.edu/software/CRF-NER.shtml [Last accessed: February 3, 2017].
[2] https://www.wikidata.org/wiki/Wikidata:Data_access [Last accessed: February 3, 2017].
[3] http://wiki.dbpedia.org/OnlineAccess [Last accessed: February 3, 2017].
[4] https://wordnet.princeton.edu/ [Last accessed: February 3, 2017].

3. **Entity Coherence**. Entity coherence refers to the real-world relatedness of different entities that function as candidate interpretations of different textual mentions in the document [160]. Entity coherence is not based on the context, so it is always the same, regardless of the query document. This probability takes the cross-reference links of knowledge base(s) into account, and the coherence between two entities is quantified as the number of incoming links that their knowledge base articles share [158].

Graph-based Entity Disambiguation. The last approach to entity disambiguation that we describe in this section is the graph-based approach. Recently, more and more researchers have made use of network structures in entity disambiguation [158, 161, 162], sometimes together with additional information [163], in order to exploit various kinds of relations between entities. Bhattacharya *et al.* [164] constructed a reference graph, i.e., a graph of some collections of references to entities, where nodes correspond to references and hyperedges correspond to the relations that are observed to hold between the references. They used unsupervised clustering approaches to cluster references which map onto the same entity, in order to disambiguate author names of research papers. Malin *et al.* [162] addressed entity disambiguation based on the network structure alone. He constructed a network with actors being nodes and their common movies being edges, which was derived from the Internet Movie Database. He proposed an alternative similarity metric based on random walks of the network and achieved a significant increase in disambiguation capability of person names compared to previous models.

Minkov *et al.* [161] formulated disambiguation of person names in emails as the task of retrieving the person most related to a particular name mention. They employed a graph-based approach with multi-type nodes and labeled directed edges. Each node corresponds to an entity, e.g., person, email-address, file, among others. And each edge corresponds to a binary relation between any two nodes, e.g., "sent from", "sent to", "alias", to name but a few. They defined the weight of an edge as the probability of moving from a node to another node using a lazy-walk process. Based on the experimental results on CSpace email corpus, they showed that the graph-based approach improves substantially over plausible baselines. Similarly, Hermansson *et al.* [163] only used a base graph on its own. In their graph, each node represents one identifier that may correspond to one or several underlying entities. Each node is labeled as ambiguous or unambiguous.

Two nodes are connected if the two corresponding identifiers are related in some way, and edge weight represents the strength of the relation. They characterized the similarity between two nodes based on their local neighborhood structure using graph kernels, and they solved the resulting classification task using support vector machine (SVM). Having carried out experiments on two datasets (Recorded Future News Data and Internet Movie Database), they showed that their method was significantly better in terms of speed and accuracy compared to a state-of-the-art method.

Levin *et al.* [165] presented a study of the impact of adding social network analysis to traditional methods when it comes to the name disambiguation problem in digital libraries. In their network, authors and papers were nodes, and there were two types of undirected edges: edges between authors, and edges between authors and papers. Based on the experiments using library datasets, they showed that the use of social network analysis significantly improved the quality of results. Shen *et al.* [166] proposed a probabilistic approach in order to link the named entities in web text with a heterogeneous information network. Their probabilistic model *SHINE* consists of two components: the entity popularity model and the entity object model. The entity popularity model captures the popularity of an entity, and the entity object model captures the distribution of objects of different types appearing in the context of an entity by using random walks. They performed experiments on the DBLP bibliographic network and showed the effectiveness and efficiency of their model compared to the vector similarity-based method (VSim) and the entity popularity-based method (POP). But in their study they did not provide a description of entity matching in terms of different types and possible granularities. In Chapter 4, based on their model, we will develop our probabilistic framework for the refinement of uncertain entities in the metadata of historic letters, and we will make a detailed analysis of entity matching in terms of different types, namely person names, location names, dates, and possible granularities accordingly.

2.7.3 Text Similarity

In this section, we introduce a selection of measures in text similarity, since we will choose the most appropriate measurement for the candidate selection in the

refinement of uncertain entities in Chapter 4. Extensive text similarity measures can be divided into three major categories: *string-based* methods, *corpus-based* methods, and *knowledge-based* methods [167]. *String-based* methods are derived from the idea of *lexical* similarity: two documents are lexically similar if they have similar token sequences. Corpus-based and knowledge-based methods are derived from the idea of *semantic* similarity: two documents are semantically similar if they have similar semantic contents.

String-based Document Similarity. String-based measures focus on string sequences and character composition [167]. String-based measures can be divided into two categories: character-based and token-based similarity measures.

- **Character-based Similarity Measures**. Character-based measures quantify similarity between two strings at the level of character transformations [168]. The representative character-based metrics are Levenshtein distance [169], Longest Common SubString distance (LCS) [170], Jaro-Winkler [171], to name but a few.

  Levenshtein distance. The Levenshtein distance [169] between two strings is calculated as the minimum number of edit operations required to transform one string to the other, where the allowed edit operations include insertion, deletion, and substitution of a single character, or a transposition of two adjacent characters.

  LCS. The Longest Common SubString distance (LCS) [170] between two strings is calculated as the length of the longest contiguous characters that exist in both strings.

  Jaro-Winkler distance. The Jaro distance [171] has been successfully applied to short string matching, especially names and addresses [172]. Given two strings s_1 and s_2, the Jaro distance is calculated as:

$$d_j(s_1, s_2) := \frac{1}{3}\left(\frac{m}{|s_1|} + \frac{m}{|s_2|} + \frac{m-t}{2m}\right), \tag{2.21}$$

  where $|s_1|$ and $|s_2|$ indicate the number of characters in s_1 and s_2, respectively. m denotes the number of matching characters in two strings, and t denotes

the number of transpositions of character matching. The Jaro–Winkler distance [173] is an extension of Jaro distance by incorporating an extra penalty for character mismatches in the first four characters. It is calculated as:

$$d_{jw}(s_1, s_2) := d_j(s_1, s_2) + lp(1 - d_j(s_1, s_2)) \,, \tag{2.22}$$

where l is the length of the common prefix shared by two strings with a maximum of four characters. The factor p is a penalty factor, Winkler [173] and Cohen *et al.* [172] chose $p = 0.1$.

- **Token-based Similarity Measures**. Token-based similarity measures first transform strings into sets of tokens, and then use the set-based similarity metrics to quantify their similarity [168]. The Jaccard similarity [174] and Cosine similarity are two well-known token-based metrics. Jaccard similarity is computed as the number of shared terms over the number of all unique terms in two strings [167]. The cosine similarity measures the correlation between two vectors. When two strings are represented as two term vectors, the cosine similarity is quantified as the cosine of the angle between these two vectors [175].

While successful to some certain degree, the string-based similarity methods cannot always identify the semantic similarity of texts [176]. Not all the texts with a similar meaning necessarily share many of the same words. For instance, two strings "Tom has an animal" and "Tom owns a dog" are obviously similar to each other semantically, but string-based similarity metrics might fail in identifying the semantic connection between these two strings.

Corpus-based Document Similarity. Corpus-based document similarity calculates the similarity between words according to information that is exclusively derived from large corpora [167]. Latent Semantic Analysis (LSA) [177] is a well-known method in corpus-based similarity. LSA [178] used Singular Value Decomposition (SVD) to find the semantic representations of words by analyzing the statistical relationships among words in a large corpus of texts. The underlying assumption here is that words which are close in meaning will occur in similar contexts of texts [179]. When LSA is used to compute sentence similarity, a

vector for each sentence is formed in the reduced-dimensional space and is then measured by the cosine of the angle between their corresponding row vectors [180]. The dimension size of the word by context matrix is limited and fixed to several hundred because of the computational limit of SVD [182]. LSA yields a vector space model that allows for a homogeneous representation (and hence comparison) of words, word sets, and texts [176]. Topic models which are evolved from earlier dimensionality reduction techniques can be considered as a probabilistic version of LSA [183], and indeed topic models outperforms it [129].

Knowledge-based Document Similarity. Knowledge-based document similarity is also a semantic similarity measure that quantifies the degree to which two words are semantically related, using information derived from semantic networks (e.g., WordNet) [167]. WordNet is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets)[4]. Synsets are interlinked by means of conceptual-semantic and lexical relations. Leacock and Chodorow [184], Wu and Palmer [185], Resnik [186], Lin [187], and Jiang and Conrath [188] are all well-known measures that work well on the WordNet hierarchy and have a relatively high computational efficiency.

2.7.4 Topic Models in Text Similarity

The topic distribution derived from a text collection can also be used to calculate the similarity of documents: two documents are similar to the extent that the same topics appear in those documents [189]. In other words, the similarity between two documents can be measured by the similarity between their corresponding topic distributions. There are many similarity functions for probability distributions [190]. The Kullback-Leibler divergence (KLD) has been popularly used in the data mining literature to measure the difference or divergence between two probability distributions over the same variable x [191]. Given two probability distributions p and q, it is calculated as follows:

$$KL(p,q) := \sum_{x \in X} p(x) ln \frac{p(x)}{q(x)}. \tag{2.23}$$

If we replace p and q with the topic distribution of two texts, we obtain the KL distance between two documents. However, the KL divergence has two major

problems [192]. KL is not suitable for case $q(x)$ is zero and KL is not symmetric. In many applications, it is adapted to become symmetric as follows:

$$KL'(p,q) := \frac{1}{2}\left[KL(p,q) + KL(q,p)\right]. \tag{2.24}$$

Another option is to use the symmetric Jensen-Shannon (JS) [193] divergence as follows:

$$JS(p,q) := \frac{1}{2}\left[KL(p,\frac{p+q}{2}) + KL(q,\frac{p+q}{2})\right]. \tag{2.25}$$

It measures similarity between p and q through the average of p and q, i.e., two distributions p and q will be similar if they are similar to their average $(p+q)/2$. Further to this, it is also possible to consider the topic distributions as vectors and use similarity functions such as Euclidean distance or cosine.

Topic Models in Authorship Attribution. Text similarity techniques have been used for authorship attribution [194]. The main idea that lies behind statistically or computationally-supported authorship attribution is that, by measuring some textual features, we can distinguish between texts written by different authors [195]. The origins of this field go back to the eighteenth century, when the English logician Augustus de Morgan suggested that authorship might be settled by comparing the word length of one text to another [196]. His hypothesis was investigated by Mendenhall [197], who quantified the writing style of Shakespeare, Bacon, and Marlowe based on word length. A study conducted by Mosteller and Wallace opened up the field to the exploration of new types of textual features and new modeling techniques [194]. They applied Bayesian statistical analysis of the frequencies of a small set of function words on the Federalist papers in order to uncover the distinctive authors [198]. In recent years, thanks to advances in areas such as natural language processing, machine learning, and information retrieval [195], this research field has been developed significantly.

Topic models have been used for authorship attribution and have yielded good results. Rosen-Zvi *et al.* [16] defined an author-topic model (AT) for single-authored texts and indicated that topic models could be used to represent the interests of authors. Mimno and McCallum [200] proposed a model DMR to deal with authorship attribution of multi-authored documents. Pearl and Stevyers [17]

used topic distributions as part of features for authorship verification and found that topic model helped them to achieve state-of-the-art verification accuracy. Seroussi *et al.* [201] proposed a disjoint author-document topic model (DADT) to include the modeling of documents, and achieved better results than the AT model. However, few of them have moved beyond the attribution of authors to the refinement of other uncertain entities such as ambiguous locations or missing dates in the texts.

2.8 Summary of the Chapter

In this chapter, we have placed the dissertation within its research background, in particular its background of social network analysis, historic correspondence study, and the specific issue of data uncertainty in digital humanities. We began with a brief introduction of social networks in Section 2.1 and then presented a discussion of some of the most important and fundamental graph principles for performing network analysis in Section 2.2. Furthermore, in Section 2.3, we introduced studies of community detection and paid especial attention to the graph clustering approaches such as hierarchical and partitional clustering. We discussed several frequently encountered temporal properties such as persistent patterns and inter-event time of social networks in Section 2.4. From Section 2.1 to Section 2.4 we presented our first research field in this dissertation. We looked at the examples, definitions, representations, and measurements of network studies, and examined how they function and what kind of questions can be addressed with them.

Furthermore, in Section 2.5, we presented a general overview of the second research field addressed in this dissertation, namely historic correspondence research, and we highlighted major projects or studies that have focused on interpretation and visualization of relations between historic scholars. In Section 2.6, we introduced the concept of topic modeling and discussed its applications in digital humanities and email study.

In Section 2.7 we presented the third research field in this dissertation, namely data uncertainty in digital humanities. We consider this issue as an entity disambiguation task, and reviewed a range of studies in the area of named entity

disambiguation and text similarity. We discussed a variety of approaches in named entity disambiguation, especially in the area of information retrieval, which is the basis of our further research in Chapter 4. Three major categories of text similarity measures were briefly covered in Section 2.7.3, and the application of topic modeling in text similarity was also described in Section 2.7.4.

In the next chapter, we will propose a correspondence network model and corresponding measurements that take full advantages of the letter metadata. We will apply our models and measurements to empirical datasets in order to obtain a detailed analysis of the patterns and relations embedded in correspondence networks.

Sail away from the safe harbor. Catch the trade winds in your sails. Explore. Dream. Discover.

— Mark Twain

CHAPTER 3

CORRESPONDENCE NETWORKS: MODELS AND ANALYSIS

3.1 Overview and Objectives

The concepts and methods of social network research have in recent years become increasingly applied in areas such as digital humanities. Social networks provide theories and techniques to represent and investigate the structure, content and dynamics of interactions between people, organizations, infrastructures [27], to name but a few. In the previous studies of correspondence networks, correspondents (sender(s) and recipient(s)) are represented as nodes and the letters sent among them are represented as edges. This sender-recipient relation is the most frequent relation covered in the previous research of correspondence network, and other relations, such as co-sender, co-recipient or person-topic relations, are seldom mentioned.

In this chapter, we focus on the metadata of historic correspondences and formalize different types of entities in the metadata, namely person names, location names, and dates. We propose an in-depth correspondence network model with three graphs derived from it. The three derived graphs measure the relations between senders and recipients and between multiple senders/recipients, respectively. Furthermore, in order to uncover interesting patterns and provide a deeper insight into the historic periods, we study the temporal aspects of correspondence networks and use specific measures to observe the correspondence networks from a temporal point of view.

The first objective in this chapter is to propose a ***comprehensive correspondence network model*** that takes full advantage of the letter metadata. Compared to the previous correspondence network research (cf. Section 2.5), we focus more on the formal modeling of a complete correspondence network, and our model thus integrates the personal, temporal and geographical information into a hypergraph structure. We derive three graphs from the correspondence network model and develop corresponding measures in order to interpret different person-person relations embedded in the historic correspondences. In the following chapter, we will extend the correspondence network model with the textual information to discover more about person-topic relations.

The second objective in this chapter is to explore the ***dynamic patterns*** embedded in the historic correspondences. We deal exclusively with letter metadata to observe the evolution of the correspondence network and the contact patterns of individuals over time. We introduce not only the representation of contact sequences, since most letter repositories do not have the duration of a letter, but also the concept of graphlets in the representation, in order to treat the network as a time-series of static graphs. We focus on the fluctuating and persistent patterns embedded in individual correspondence activity and the evolving network structures.

This chapter is organized as follows. After this introduction and outline (cf. Section 3.1), Section 3.2 illustrates the problems that we deal with in this chapter: modeling relationships and temporal analysis. Section 3.3 then goes on to describe the typical components of the metadata of letters and the related issue of data uncertainty. Section 3.4 presents our correspondence network models with three graphs derived from it. The temporal study of the correspondence network is discussed in Section 3.5. We use various datasets for experimental evaluations. The details of the dataset and the results of our experiments are presented in Section 3.6. We summarize this chapter in Section 3.7.

3.2 Problem Statements

The main issues to be addressed in this chapter are summarized by the following problem statements.

- **Modeling Relationships**. The first challenge is to model the *relationships between correspondents* from a computational modeling point of view. The most typical approach in the discipline of digital humanities is to describe the direct relationship "who wrote to whom" using graph visualization. Although this relation can be easily derived from letter metadata, this approach neglects other potential relations such as co-sender or co-recipient relation in the historic correspondences. Therefore, in this chapter, we propose a correspondence network model with three derived graphs in order to explore three types of relations, and specific metrics are developed in order to discover the latent relations between historic persons and interesting patterns embedded in the correspondence network.

- **Temporal Analysis**. The second challenge is to discover *temporal* patterns embedded in the correspondence network. In most letter repositories, each letter only records the date of writing, but not the date on which the letter was received. Moreover, since most letters exist between one particular individual and others in the individual correspondences, it is not feasible for us to implement measures such as the temporal paths and temporal centrality. Therefore, in this chapter, we use two types of representations, namely contact sequences and graphlet sequences, to describe the evolving correspondence network structures. For the contact sequences, we focus on the contact patterns of inter-contact and reciprocal time, in order to obtain significant insights of individual correspondence activity in the historic times. For the graphlet sequences, we generalize the refreshing rate and persistent rate of nodes and edges in order to obtain fluctuating and stable patterns of networks over time.

3.3 Building Blocks for Correspondence Networks

The metadata of historic correspondences consist of typical components of letters, such as senders, recipients, the date when the letter was written, and locations of the sender and the recipient (these locations are also named as origin and destination). These components are the building blocks for constructing a correspondence network structure. Although historians and linguists make efforts in annotating letters with clear and complete metadata information, some ambiguous or uncertain entities are still inevitable in historic correspondences. In this section, we first

categorize and describe different types of entities in the metadata of letters, and then we introduce our approach to formalize each of these entities, respectively.

3.3.1 Data Uncertainty in Historic Correspondences

Due to the fact that digitized historic letters are not originals but transcriptions based on different principles, the integration of diverse letter collections from various sources not only enriches our knowledge of historic texts, but also confronts us with vague, uncertain or even conflicting information. The attributes, which these letters share at a basic level, are a sender, a recipient, an origin (the location from which a letter was sent), a destination (the location from which a letter was received) and a date. These five elements not only bring together different letter corpora for comparison and collaboration, but also provide a macro analysis to the correspondents and the society within which they wrote letters. In the following section, we generally categorize the uncertain entities into two types in terms of degree of uncertainty.

1. **Unknown Entities**. This category refers to missing values for entities when no information concerning the entity is available either in the metadata or in the content of a letter. For instance, a letter contains no information with respect to the date of writing and the name of the sender(s).

2. **Imprecise Entities**. This category refers to incomplete or ambiguous mentions for different types of entities in the metadata. For instance, only the first name of the recipient is recorded but his or her last name is unknown.

In the following part, we introduce the three types of entities in the metadata of historic letters, i.e., dates, person names, and location names.

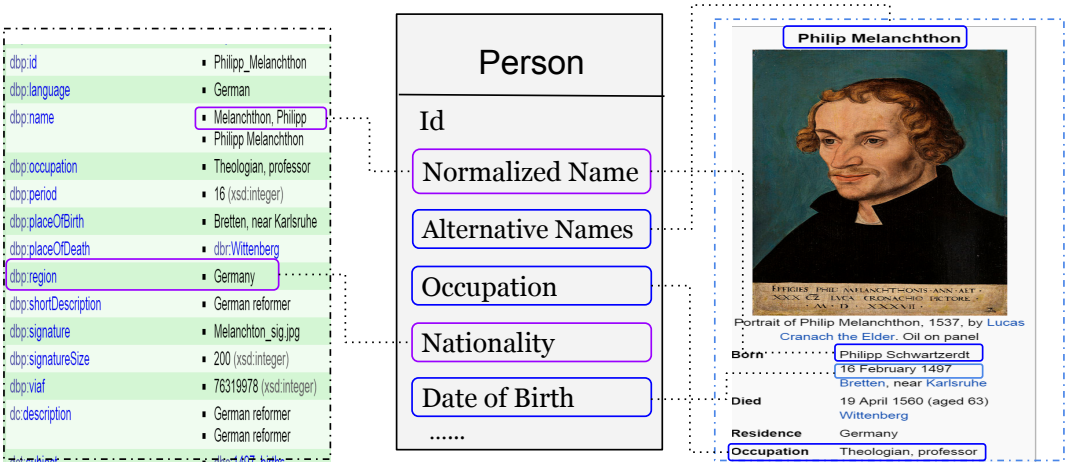
- **Dates of Writing**. Most correspondence repositories record the exact dates of writing and letters are organized in a chronological order. However, sometimes only partial or approximate dates are known, e.g., “July ?? 1650” or “1650 or 1651”. This might due to the imprecise recognition of handwritten date on the age-worn envelope. In order to capture dates and store them in the database in a compatible way, many correspondence projects not only

record the original textual form of the date, but also normalize the date information to a standardized format [144] (e.g., day/month/year or year-month-day). It can be a single day, a date range or a non-empty sorted set containing both single days and date ranges. For instance, if the date of writing is "July 1650", it is recognized as a period from the first day of July to the last day of July [115]. Some projects keep the year information separately in records as another field [84] for the timeline visualization.

- **Person Names**. The spelling variations of person names not only vary from historic languages to modern languages, but also among different languages. For instance, the Polish reformer Jan Laski is named as "Johannes Laski" in Early New High German (ENHG) and "Johannes a Lasco" in modern German. Moreover, two individuals might share one identical name. For example, the German scholar "Joachim Camerarius" also named his son "Joachim Camerarius". Besides, the abbreviations of person names are also common in letter collections, e.g., "R. F. Cooke", in Charles Darwin's letter collection[1], was actually Robert Francis Cooke, Darwin's publisher. Without additional information, we cannot directly identify the exact person by his or her name alone. Many correspondence projects follow the TEI markup guidelines [6] to organize information relating to person names, e.g., first name, last name, occupation, birthdate, deathdate, and so on [115]. However, not all person information is available within the metadata and the content of the letter, and additional information from other knowledge repositories is needed to be captured.

- **Locations**. In a similar way to person names, spelling variations also exist for location names, i.e., the origin and the destination of a letter. Some of the names of recorded locations, due to the changes of spelling and administrative boundaries, have been forgotten or are no longer in use. For instance, Bardejov is a small town in Slovakia, but it was named as "Bartfeld" in Early New High German. Moreover, locations in historic correspondences have different levels of granularity. Although most locations in letter repository are accurate to the city or town or mailing address level of granularity, there are a few locations that only have the name of a country or a state. For

[1] https://www.darwinproject.ac.uk/letter/?docId=nameregs/nameregs_1053.xml [Last accessed: January 11, 2017].



(a) DBPedia Page (b) Data Model (c) Wikipedia Infobox

Figure 3.1: (*a*) and (*c*) are DBPedia and Wikipedia page of Philipp Melanchthon (1497–1560), respectively. (*b*) is a display of the simplified data model for a correspondent in a corpus of historic correspondence.

Source: (*a*) http://de.dbpedia.org/page/Philipp_Melanchthon
(*c*) https://en.wikipedia.org/wiki/Philip_Melanchthon

instance, in the place names of "Whitehall Court, London" and "Germany", these two expressions refer to a place in a city and a country, respectively. Many correspondence projects organize the location information [115], e.g., street name, village/town/city, region, country and GIS coordinates in the database. Considering the differences and connections between historic place names and modern place names, additional gazetteers are exploited to capture supplementary information to the recorded location names.

3.3.2 Entity Formalization and Standardization

In this section, we focus on formalizing the three types of entities described above, i.e., person names, location names and dates of writing. With the help of existing tools and knowledge bases, we make an effort to associate the person names with additional information, standardize the location names with geographic coordinates, and normalize the temporal expressions of dates. These are the basic building blocks for constructing a correspondence network structure.

- **Dates of Writing**. We denote the set of all possible dates as T and assume a timeline, with days being the finest level of granularity. A date $t \in T$ in

the correspondence is comprised of a textual form of the temporal expression recorded in the metadata of a certain letter, a normalized value and the degree of certainty. The original input could be a single date, a date range, or even an empty set. The normalized value is the normalized semantic representation of a date. The degree of certainty is an annotation concerning the precision of a temporal expression, since a letter might only contain part of the date. For the sake of simplicity, we use only three granularities: day, month, and year. If a date is complete, such as "26.07.1514", the normalized value is "1514-07-26" (yyyy-mm-dd) according to the ISO-timeML annotation standard [202] and the certainty is "day", whereas if the input is "1655", the normalized value is a range [1655-01-01, 1655-12-31] and the certainty is "year". If the input is an expression such as "1615 or 1616", the normalized value is "[1615-01-01, 1615-12-31]∪[1616-01-01, 1616-12-31] and the certainty is "year". We represent the missing dates as "NA" in the original input and as "0" in the normalized value and the degree of certainty.

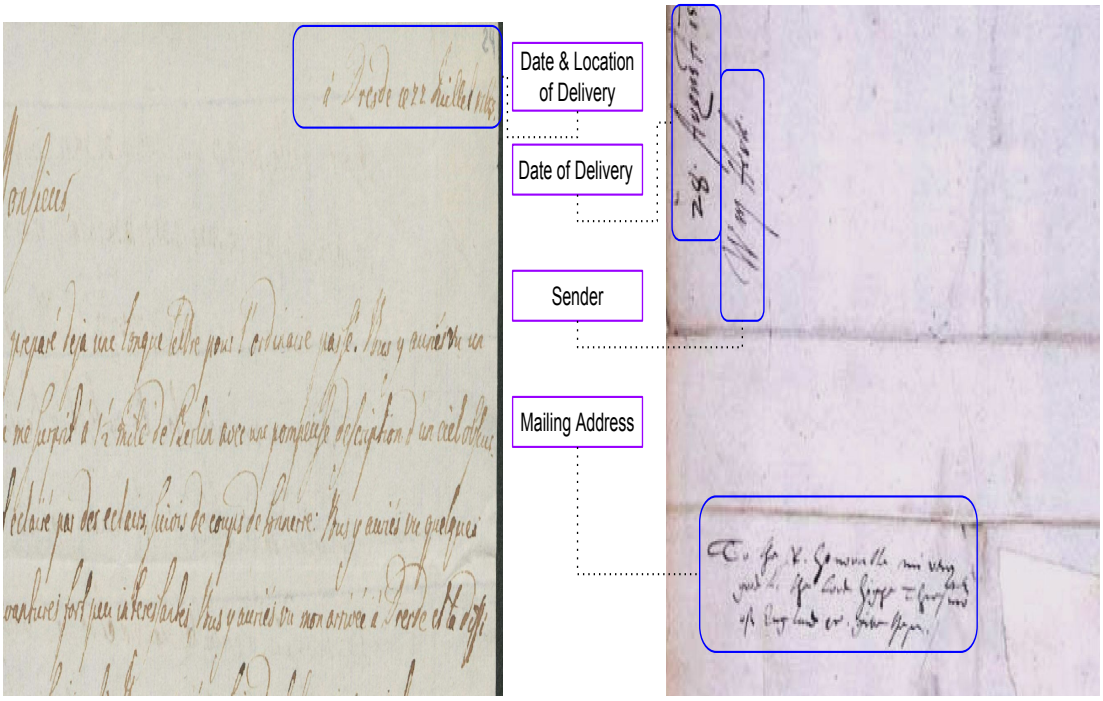
- **Person Names**. Let P denote a set of individuals. We assume that senders and recipients are human beings, although there might be letters that have a more abstract sender/recipient, e.g., a letter is sent to an organization by a university. The model of a person in a correspondence is briefly illustrated in Figure 3.1. The model of each person $p \in P$ consists of an id, a normalized name, a set of alternative names, and other descriptive attributes, e.g., occupation, nationality, date of birth, among others. Knowledge bases such as Wikidata [156] are employed in the extraction of profile information of historic persons, and historians assist us to ensure the accuracy of the extracted profile information. Knowledge bases such as Wikidata provide access to query the person names directly for corresponding information. The corresponding query is a set with the restriction of time period to reduce the name duplication. Besides this, the date of writing is also taken into account as a filtering condition, i.e., for a letter, the date of writing should be after the writer's date of birth and before the date of death of him/her. We represent the missing names as NA in all fields. Although we do not deal with organizations in this data model, the mechanism described above can be easily extended to include organizations as well.

- **Locations**. Locations are described as a city, a town or a detailed mailing address in the metadata of historic correspondences. We define L as the set of all locations and use a geographic hierarchy, with city/town denoting the finest level of granularity: country, state and city/town. Although many more geographic granularities exist (e.g., address or suburb), for the sake of simplicity we only record these two geographic granularities. A location *loc* consists of a textual form of the geographical expression recorded in the metadata of a certain letter, a normalized description, and the geo-coordinates of the location. The original location is the place where the letter was sent or received. Although the geographical expression of a location is referred to not only by specific geographic coordinates, but also by its region, for the sake of simplicity we keep the normalized description of a location in the database, or in other words, the record of hierarchical containment information.

  The hierarchical containment information is typically accessible using the gazetteer or geo-tagger, while explicit polygonal information is often not available [203]. Thus, we rely on the containment information rather than on explicit polygonal information about the locations. Since a city is located in a country, a mapping can be obtained from a value of a finer granularity to a coarser granularity. The geo-coordinate of a location consists of a pair of latitude and longitude co-ordinates and it is represented as $\{(lon, lat)|lon \in [-90, 90), lat \in [-180, 180)\}$. It is useful to obtain location information from external sources such as Google Maps API. The Google Maps API supports old location names and also accepts spelling variations [204]. The returned geographical coordinates and containment information help us with distinguishable and approximate locations.

  For example, the normalized description of "London" is "London, UK" and its corresponding geo-coordinate is $(-0.1277583, 51.50735)$ by Google maps geocode API. If the input is "London or Paris", the corresponding normalized description is "London, UK or Paris, France" and the corresponding geo-coordinate is $(-0.1277583, 51.50735) \cup (2.352222, 48.85661)$. In a similar way to the dates of writing, we represent the missing locations as NA in the fields of original input, normalized value and geo-coordinates.

The accurate and detailed metadata constitute the basis for social network research and content analysis. For the missing or ambiguous entities in the letter metadata,



(a) A Letter Dated 1763 (b) Annotation (c) An Envelope Dated 1585

Figure 3.2: An example of historic letters: (a) is a page of a letter written by Adolf von Buch on the 22^{nd} of July, 1763 and (c) is a part of an envelope of a letter written by William Herle on the 28^{th} of August, 1585. (b) is an annotation of the metadata appearing in these two samples. (a) http://tei.ibi.hu-berlin.de/berliner-intellektuelle/manuscript?Brief13vonBuchanBeausobre+en\#1 [Last accessed: September 25, 2015]. (c) http://www.livesandletters.ac.uk/herle/images/209v.html [Last accessed: September 25, 2015].

although we record them as missing values in the database, we are still faced with the challenges of ambiguity. We assume that uncertain entities in the letter metadata can be inferred or implied from other entities in the context of the corresponding letter. We will introduce our probabilistic approach to addressing this issue later in Chapter 4.

3.4 Correspondence Networks

The availability of such structured metadata as we introduced in Section 3.3, reveals details relevant to the correspondents and gives us a description of their social network over time. We aim to propose a correspondence network model that closely resembles today's social networks based on the persons, places, and

dates from metadata. It is our hypothesis that such a correspondence network provides a more holistic view of that period of time, its key players, and its circles of acquaintances than what would be perceivable through individual letters alone. In this section, we first present the definition of a letter and then we introduce our correspondence network model with three derived graphs and corresponding measurements.

3.4.1 Correspondence Network Model

A historic letter normally consists of two parts: the metadata and the content. In this section, we begin with the definition of a letter and then give a detailed description of our correspondence network model, which is the basis for the further relation analysis.

Definition 3.1. Letter. A letter is represented as a tuple $l = (S, R, \mathrm{t}, l_s, l_r, c)$. S denotes a list of senders, for each $s \in S$, such that $s \in P$. By analogy, R denotes a list of recipients, again being a list of individuals. $\mathrm{t} \in T$ specifies the date when a letter has been written. l_s and l_r specify the original location and destination of the letters, respectively, with $l_s \in L$ and $l_r \in L$. $c \in C$ denotes the corresponding content of the letter in the form of word sequences.

Although the content of a letter is also included in the definition above, in this chapter, we focus on the study of metadata within a correspondence network structure. The content will be considered later in Chapter 4. In the following part, we introduce the definition of our correspondence network.

Definition 3.2. Correspondence Network. A correspondence network is represented as a multi-edge hypergraph $H = (V, E)$, where nodes $V \subseteq P$ correspond to correspondents (senders/recipients) and edges $E \subseteq 2^V \times 2^V \times \mathbb{N}$ correspond to the letters sent among correspondents.

The edges are directed and each edge consists of a subset of nodes. For each edge $e = \langle H_e, T_e, i \rangle$, $H_e \subseteq V$ is the head of e, which represents the set of recipients of each letter, $T_e \subseteq V$ is its tail, which corresponds to the set of senders of each letter. Note that H_e and T_e are disjoint, i.e., $H_e \cap T_e = \emptyset$, for all $e \in E$.

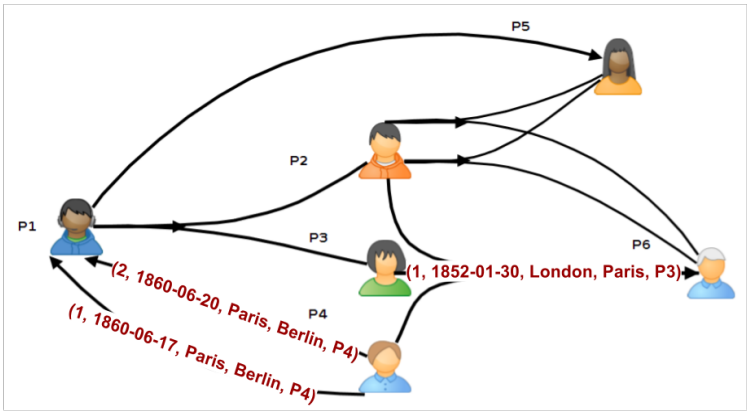


Figure 3.3: A simple correspondence network model: six individuals are correspondents (nodes) and the arrow lines drawn among these individuals represent edges. Each edge originates from the sender(s) and points to the recipient(s), and we only assign three hyperedges in the graph with the set of attributes for the sake of simplicity.

Each edge has associated attributes $\{(d, t, l_s, l_r, aw) \mid d \in \mathbb{N}, t \in T, l_s \in L, l_r \in L, aw \in V\}$. Considering that there might be more than one letter sent between the same sender(s) and recipient(s), d denotes the index number of each edge that distinguishes different letters sent between two sets of nodes. For example, if there are k edges between two set of nodes, the value of d lies within the range $[1, k]$. $t \in T$ corresponds to the date when letters were written. $l_s \subseteq L$ and $l_r \subseteq L$ correspond to the locations of letters being sent and received, respectively. Given that one letter might have multiple senders, but only one of them is the author, we use aw to represent the author of the letter and we define an "author" function $aut : E \rightarrow V$ as an attribute for each hyperedge to mark the author of each letter.

Figure 3.3 shows a simple example of a correspondence model. The nodes (figures) represent correspondents, and hyperedges (arrow lines) represent the letters sent between people [205]. For example, there is a hyperedge which starts from $P1$ as its tail and ends at $P2$ and $P3$ as its head, which illustrates that person $P1$ sent a letter to both person $P2$ and person $P3$.

As we discussed in Section 2.5.2, correspondence networks can be divided into two types of networks, namely *individual* correspondence networks and *group* correspondence networks. An *individual* correspondence network is a network built from one person's private letter collection. It is an *ego-centric* network with a star structure and in which this specific person is placed at the center

of the graph. A *group* correspondence network is a network built from letter collections of groups of people living during the same historic periods and having intersections. This is a *socio-centric* network and can be viewed as a combination of many individual correspondence networks. Unfortunately, the data sources to which we could obtain access only provide one or two individuals' letter collections. Because of the limitation of datasets, our corpus only constitutes a small part of the entire historic correspondence network that we want to study. Therefore in this dissertation, we study correspondence networks at the individual level.

In spite of the fact that our access to various data source is limited, we can still explore three types of relations embedded in the correspondence networks. These relations we specify as the sender-recipient relation, the co-sender relation and the co-recipient relation.

- **Sender-recipient relation**. Two individuals are connected if and only if there is at least one letter between them.

- **Co-sender relation**. Two individuals are connected if and only if they have sent at least one letter together to the same recipient(s).

- **Co-recipient relation**. Two individuals are connected if and only if they have received at least one letter from the same sender(s).

With regard to the different kinds of relationships mentioned above, we "decompose" the complete correspondence network model into three types of networks. Compared to the original complex hypergraph structure, it is rather easier, clearer and more concise to cope with different relations only with the necessary information. In the following sections, we will introduce three graph representations derived from the correspondence network to represent different kinds of relationships.

3.4.2 Sender-Recipient Network

The sender-recipient relation is the most typical type of relation that researchers from the area of digital humanities choose to visualize in the form of networks. The sender-recipient network can provide a general view of the living experiences

of historic persons and their close or distant relations with each other. Hence, in order to describe the relationship between senders and recipients, we introduce the first derived directed graph from the correspondence network model. For the purpose of constructing a sender-recipient network, a corresponding incidence matrix $\mathbf{M}$ $(V \times V)$ based on the correspondence model is represented as:

$$M_{ij} = \begin{cases} 1 & \exists e \in E : v_i \in T_e,\ v_j \in H_e \\ 0 & otherwise. \end{cases}$$

v_i and v_j represent two nodes in the hypergraph. If there is a hyperedge e with v_i in the tail and v_j in the head, then the entry $M_{ij} = 1$, otherwise 0. Given such a matrix, for each entry $M_{ij} = 1$, an edge from v_i to v_j is included in the sender-recipient network. In the following, we introduce the definition of sender-recipient network and a simple graph example is shown in Figure 3.4.

Definition 3.3. (Sender-Recipient network). A sender-recipient network is a directed graph $G_{sr} = (V_{sr}, E_{sr})$, which is composed of a set V_{sr} $(V_{sr} \subseteq V)$ of nodes and a set $E_{sr} \subseteq V \times V$ of directed edges. V_{sr} represents the correspondents and E_{sr} represents the letters sent between correspondents.

In a sender-recipient network, we use a set of quintets $\{(d, t, l_s, l_r, aw) \mid d \in \mathbb{N}, t \in T, l_s \in L, l_r \in L, aw \in V_{sr}\}$ as edge attributes and each quintet is the same as the edge attribute in the hypergraph model. In the case of multiple letters with the same locations and the same date, we use an index number i to differentiate each element in the edge attribute set. t represents the date of writing. Locations l_s and l_r represent the origin and the destination of letters, respectively. aw corresponds to the writer of each letter. For instance, for k letters between two nodes i and j, the corresponding edge attribute is: $\{(1, t^1_{ij}, l^1_i, l^1_j, aw^1), (2, t^2_{ij}, l^2_i, l^2_j, aw^2), ..., (k, t^k_{ij}, l^k_i, l^k_j, aw^k)\}$. In this way, we condense the multi-edges with multi-attributes between two nodes into one single edge. The number of quintets in the set corresponds to the number of letters exchanged between two individuals. Thus we can represent the sender-recipient network by an adjacency matrix A^{sr} with entries that are not simply zero or one, but which are associated with the number of quintets in the attribute set that corresponds to the number of letters between i and j.

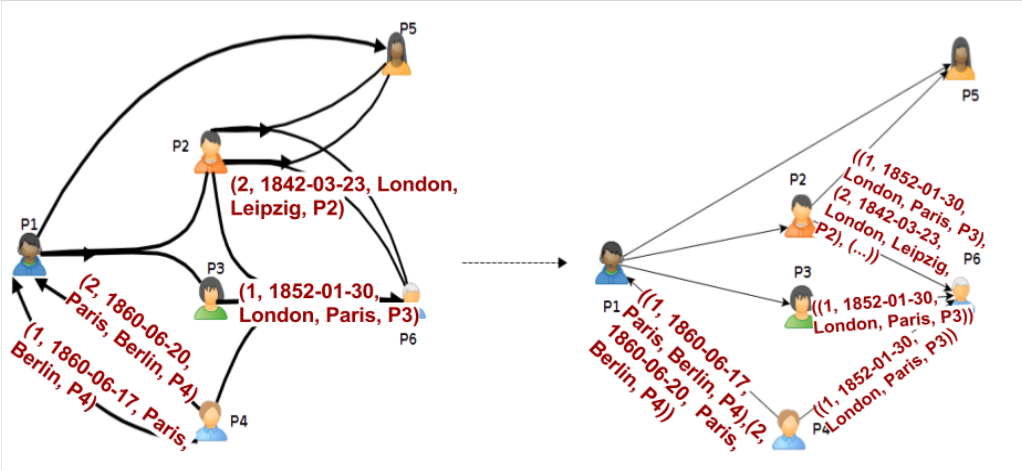


Figure 3.4: A simple sender-recipient network: the network on the left side is a hypergraph correspondence model, and the network on the right side is a sender-recipient network model obtained from the left side. In the sender-recipient network, an edge between each two nodes represents the correspondence(s) and is associated with a set of quintets as the edge attribute, while the arrow of each edge points to the recipient of the corresponding letter.

$$A_{ij}^{sr} = \left|\{(1, t_{ij}^1, l_i^1, l_j^1, aw^1), (2, t_{ij}^2, l_i^2, l_j^2, aw^2), ..., (k, t_{ij}^k, l_i^k, l_j^k, aw^k)\}\right| \tag{3.1}$$

In a sender-recipient network, we define the weight of an edge as the corresponding value in the adjacency matrix A^{sr} that quantifies the relationship between senders and recipients. In order to capture particular features of the sender-recipient network structure, we propose the following two measurements on reciprocity, namely local reciprocity and global reciprocity, and we introduce the degree, betweenness, and closeness centrality measures for weighted networks. These measurements can help us to answer questions such as who is the most important person in the sender-recipient network, and whether the letters between two individuals take place in one direction or not. These measurements will later be applied to the data analysis and reveal interesting features and patterns that contributes to our understanding of the correspondence network.

- **Local Reciprocity**. Newman [206] and Zafarani *et al.* [26] define the (global) reciprocity as the fraction of edges that are reciprocated. However, in the weighted network, this fraction cannot measure the weights carried by

the mutual edges between any two nodes. For instance, in a network in which each pair of individuals have an average of 10 letters between each other, the reciprocity is different from a network in which two individuals have an average of only 1 letter between each other. For the purpose of quantifying the extent that two individuals have reciprocal ties, we calculate the *local reciprocity* for the sender-recipient network using the following formulas:

$$\delta_{ij}^{sr} = \begin{cases} 1 & A_{ij}^{sr} A_{ji}^{sr} > 0 \\ 0 & otherwise. \end{cases}$$

$$R(i,j) := \delta_{ij}^{sr} \frac{min\{A_{ij}^{sr}, A_{ji}^{sr}\}}{max\{A_{ij}^{sr}, A_{ji}^{sr}\}}, \tag{3.2}$$

in this formula, $min\{A_{ij}^{sr}, A_{ji}^{sr}\}$ and $max\{A_{ij}^{sr}, A_{ji}^{sr}\}$ compare the weight of reciprocal edges between two nodes, and the fraction $\frac{min\{A_{ij}^{sr}, A_{ji}^{sr}\}}{max\{A_{ij}^{sr}, A_{ji}^{sr}\}}$ measures the ***balance*** of reciprocal behavior between two nodes. δ_{ij}^{sr} is an alternative notation of Kronecker delta, i.e., a discrete function of two variables. If the two variables are equal, the function equals 1, otherwise 0. Similarly, in our case, if there are a pair of reciprocal edges between nodes i and j, δ_{ij}^{sr} equals 1, otherwise 0. δ_{ij}^{sr} in Equation 3.2 guarantees that the value of the denominator is not 0. For instance, in Figure 3.6, $R(A,B)$ is calculated as $1 \times \frac{4}{5} = \frac{4}{5}$. However, we notice that local reciprocity treats $(A_{ji}^{sr} = 10, A_{ji}^{sr} = 10)$ and $(A_{ij}^{sr} = 1, A_{ji}^{sr} = 1)$ as equal, but actually these two are different in their edge weights (***volume***). In this case, only pairs of nodes with both a high local reciprocity and a high sum of edge weights, in other words, the nodes that achieve the ***balance*** and ***volume*** of reciprocity, are the nodes we expect.

Akoglu *et al.* [207] also noticed the volume and the balance of local reciprocity, and used a weighted ratio $r_w = \frac{min(w_{ij}, w_{ji})}{max(w_{ij}, w_{ji})} log(w_{ij} + w_{ji})$ as the measurement. But this measure simplifies the two factors into a single objective function and does not explain or evaluate the log function explicitly. Hence, we propose a skyline-based approach to capturing nodes with both high balance and high volume in our correspondence network. The skyline operation is a database query for filtering out a set of not dominated points from a large set of datapoints [208]. A point ***dominates*** another if it is ***better*** in all relevant

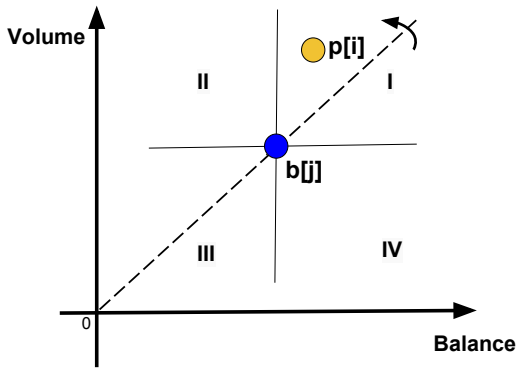


Figure 3.5: a simple example of selecting a not dominated point using azimuth: the x-axis represents the local reciprocity $R(i,j)$ and the y-axis represents the sum of the edge weights $(A_{ij}^{sr} + A_{ji}^{sr})$ between a pair of nodes.

dimensions and strictly ***better*** in at least one dimension. There is no unified definition for "better", which depends on the specific query and selection criteria set by users. In this case, we assume that pairs of nodes that have smaller values in both dimensions of balance and volume, are ***not dominated*** by any other pairs of nodes. Only the pairs of nodes that are not dominated will be captured as potentially interesting pairs.

Therefore, in our sender-recipient network, we consider selecting pairs of nodes with both a high local reciprocity and a high sum of edge weights as the task of finding the *Pareto Front*, namely a set of points that are not dominated on a plot where each point corresponds to a pair of nodes. Take Figure 3.5 as an example. The x-axis represents the local reciprocity $R(i,j)$ and the y-axis represents the sum of the edge weights $(A_{ij}^{sr} + A_{ji}^{sr})$ between a pair of nodes. The $azimuth(b[j], p[i])$ in this figure calculates the azimuth between two points $b[j]$ and $p[i]$ (relative to $b[j]$) in order to analyze the dominated/not dominated relation between each two points $b[j]$ and $p[i]$.

Algorithm 1 shows the pseudocode on finding "Pareto Front" with respect to the preference "high *volume* and high *balance*". In this algorithm, the input is a list of points Pt assigned with two values *balance* and *volume*. First, these points are sorted by $\frac{volume}{balance}$ in an ascending order. Then, a list B with $pt[0]$ as the first element is created. The next step, for loop (line 4–12), is to select all the points satisfying the condition: in this case, if $azimuth(b[j], pt[i]) \in Quadrant$ II, i.e., $b[j]$ and $pt[i]$ are not dominated, then $pt[i]$ is assigned to B. Otherwise if $azimuth(b[j], pt[i]) \in Quadrant$ I, i.e.,

Algorithm 1 Algorithm for finding *Pareto Front*

INPUT: Pt: a list of points in $2D$ (*balance*, *volume*)
num: the number of points given by users
m, k: the thresholds to determine the bounding area of points chosen
OUTPUT: result: a list of *not dominated* points (*Pareto Front*)

1: Sort Pt by $\frac{volume}{balance}$ in an ascending order ($balance \neq 0$)
2: $j \leftarrow 0$
3: $b[0] \leftarrow pt[0]$
4: **for** $i = 1 : (length(Pt) - 1)$ **do**
5: **if** $azimuth(b[j],\ pt[i]) \in Quadrant$ II **then**
6: $B \leftarrow pt[i]$; $j \leftarrow j + 1$
7: **else if** azimuth$(b[j],\ pt[i]) \in Quadrant$ I **then**
8: **while** $j > 1$ & azimuth$(b[j], pt[i]) \in Quadrant$ I **do**
9: delete $b[j]$; $j \leftarrow j - 1$
10: **end while**
11: **if** $azimuth(b[j],\ pt[i]) \in Quadrant$ II **then**
12: $B \leftarrow pt[i]$; $j \leftarrow j + 1$
13: **end if**
14: **end if**
15: **end for**
16: $B \leftarrow InterceptFilter(m, k)$
17: **if** $(length(result) + length(B)) \leq num$ **then**
18: $result.add(B)$
19: delete B from Pt, clear B
20: goto line 2
21: **else**
22: random pick $num - length(result)$ points from B, add the points in $result$
23: **end if**
24: **return** result

$pt[i]$ dominates $b[j]$, then $b[j]$ is deleted and the new $azimuth(b[j-1], pt[i])$ is calculated until the new $azimuth(b[j-1], pt[i])$ does not belong to *Quadrant* I . In the meantime the point satisfying the situation will be added to list B. In this case, we do not consider the condition that $azimuth(b[j], pt[i])$ belongs to *Quadrant* III as $b[j]$ dominates $pt[i]$ in *Quadrant* III. $azimuth(b[j], pt[i])$ in *Quadrant* IV is not also considered, since the points are sorted into an ascending order. This process is iterated until all the points are traversed.

The next step (line 13) is to set the bounds $balance = m$ and $volume = k$ as thresholds on the plot in order to select the desired points by users. The initial value for m and k is 0. In our case, we regard the points within $x \geq m \quad (0.5 \leq m \leq 1)$ and $y \geq k \quad (\frac{max(A_{ij}^{sr} + A_{ji}^{sr})}{2} \leq y \leq max(A_{ij}^{sr} + A_{ji}^{sr}))$

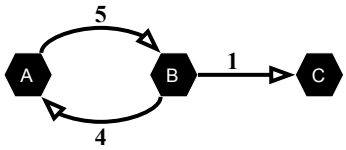


Figure 3.6: a simple example of directed and weighted graph: the number on each edge represents the corresponding edge weight.

as desired points. If the number of points in *result* is less than the number of points *num* that is needed, these points will be removed from the original list Pt and the point selection step will be performed again for the rest of the Pt list until finally the *num* points have been found.

Let the number of points in P be n. The time complexity of sorting is $O(nlogn)$ and the point selection process takes $O(n)$, so the overall time for our algorithm is $O(nlogn)$. The space complexity for our algorithm is $O(n)$ without using extra space.

- **Global Reciprocity**. We also measure the *global reciprocity* for the whole graph in order to find how many pairs of edges in the graph are reciprocal. This is calculated as:

$$R := \frac{\sum\limits_{i,j,i<j} R(i,j)}{\sum\limits_{i,j,i<j} \delta_{ij}^{sr}}, \tag{3.3}$$

$\sum\limits_{i,j,i<j} \delta_{ij}^{sr}$ in Equation 3.3 is a normalization factor that corresponds to the number of reciprocal pairs in the network. For instance, the global reciprocity R in Figure 3.6, is calculated as $\frac{\frac{4}{5}}{1} = \frac{4}{5}$.

- **Weighted Degree**. The degree centrality for a given node in unweighted network is measured as the number of edges incident to the node (cf. Section 2.2.4). Recently degree centrality has also been extended to the sum of edge weights when analyzing weighted networks [209] and is named as *node strength* [210]. Given a node i in an (undirected) weighted graph, the weighted degree centrality is measured as:

$$D_c(i) := \sum_{i \neq j} A_{ij}^{sr} \,, \tag{3.4}$$

where $\sum_{i \neq j} A_{ij}^{sr}$ represents the sum of the weights of all edges incident to node i. In order to assess the popularity and significance, we introduce two following measurements for in-degree and out-degree in directed and weighted networks.

$$D_c^{out}(i) := \sum_{i \neq j} A_{ij}^{sr} \tag{3.5}$$

$$D_c^{in}(i) := \sum_{i \neq j} A_{ji}^{sr} \tag{3.6}$$

In this case, $\sum_{i \neq j} A_{ij}^{sr}$ represents the sum of the weights of all the edges that start from node i and $\sum_{i \neq j} A_{ji}^{sr}$ represents the sum of the weights of all the edges that point at node i.

- **Weighted Shortest Path**. In order to extend closeness and betweenness centrality to weighted networks, we introduce a generalization of shortest paths for weighted networks. The shortest path in unweighted networks is defined as the path that has the minimal path length between two nodes (cf. Section 2.2.3). The shortest path has been generalized to weighted networks on the basis of the idea that weights are considered as costs and the shortest path between two nodes should be the least costly path [211]. However, this measurement is not suitable for our sender-recipient network. The edge weight in the sender-recipient network is the strength and not the cost, so we use the inverse of the edge weight as the cost. The shortest path for a weighted graph is calculated as:

$$w(v_x, v_{x+1}) := \frac{1}{A_{x,x+1}^{sr}} \,, \tag{3.7}$$

$$w(p(i,j)) := \sum_{e \in p(i,j)} w(e)\,, \tag{3.8}$$

$$Sp(i,j) := \underset{p(i,j) \in \mathcal{P}(i,j)}{\operatorname{argmin}} \; w(p(i,j))\,, \tag{3.9}$$

where the cost of an edge $w(v_x, v_{x+1})$ is calculated as the inverse of the edge weight. The cost of a path $w(p(i,j))$ is calculated as the total costs of all the traversed edges on a path. As mentioned in Section 2.2.3, $P(i,j)$ denotes the set of all the paths from i to j. The weighted shortest path from i to j is the path with the minimum cost. Accordingly, the weighted shortest path length is calculated as:

$$d(i,j) := w(Sp(i,j))\}\,, \tag{3.10}$$

where the sum of the cost of the weighted shortest path is the path length, which is also the basis for the extension of betweenness centrality and closeness centrality to weighted networks.

- **Weighted Betweenness**. Betweenness is an indicator for the amount of influence that a person has in a network [209]. The nodes with high betweenness represent individuals who control the information flow between others. In an unweighted network, betweenness centrality for a node i is measured as the extent to which i is on the shortest paths between other nodes [26] (cf. Section 2.2.3). This centrality is extended to weighted networks and calculated as:

$$B_c^w(i) := \sum_{i \neq u \neq v} \frac{|Sp(u,i,v)|}{|Sp(u,v)|}\,, \tag{3.11}$$

where $|Sp(u,v)|$ denotes the number of weighted shortest paths between nodes u and v, and $|Sp(u,i,v)|$ denotes the number of weighted shortest

paths between u and v that pass through i. This measure is also interpreted as the degree to which an individual is important in contacting other individuals within the sender-recipient network.

- **Weighted Closeness**. Closeness captures the average distance between a node and every other node in the network. In Section 2.2.4, it is measured only for the case of unweighted networks as the inverse of the total geodesic distances from a given node to other nodes. This centrality measure is also extended to weighted network [209] and calculated as:

$$Cl_c^w(i) := \frac{1}{\sum_{i \neq j} d(i,j)}, \tag{3.12}$$

where $\sum_{i \neq j} d(i,j)$ calculates the average distance (average shortest path length) between node i and other nodes in the network. We use the inverse of $\sum_{i \neq j} d(i,j)$ to represent the weighted closeness centrality, or in other words, the smaller that the average shortest path is, the more central the node.

In Section 3.6, we will use various datasets to investigate whether the sender-recipient networks of four individuals all follow the power-law degree distribution or not. Furthermore, betweenness and closeness centrality measures will both be employed in the experiments to find the most important nodes (individuals) within the sender-recipient network besides the central node. The reciprocity measure allows us to explore the reciprocal behavior between individuals and will be applied to datasets in order to find out the individuals who have the most frequent reciprocal contacts with the central person, and whether there are some special patterns in terms of different types of correspondents.

3.4.3 Co-Sender Network

Newman [58, 212] examined the macro- and micro-properties of co-authorship (collaboration) networks in 2001. In the following year, Barabási *et al.* [213] explored the dynamics and evolution of co-authorship (collaboration) networks. Since then, co-authorship networks have been used extensively in order to identify the structure of collaborations and the status of individuals in scientific publications [214]. In these networks, two scientists are considered to be connected if

and only if they have coauthored one or more papers together. We have adapted this concept of "connectivity" for the construction of co-sender network in this section, i.e., two individuals are considered to be connected if and only if they have coauthored one or more letters together.

In this section, in order to describe the relationship between multiple senders, we first introduce the second derived graph from the correspondence network model. For the purpose of constructing a co-sender network, a corresponding incidence matrix S derived from the correspondence model is represented as:

$$S_{ij} = \begin{cases} 1 & \exists e \in E : v_i \in T_e \wedge v_j \in T_e \wedge v_i \neq v_j \wedge aut(e) \cap \{v_i, v_j\} \neq \emptyset \\ -1 & \exists e \in E : v_i \in T_e \wedge v_j \in T_e \wedge v_i \neq v_j \wedge aut(e) \cap \{v_i, v_j\} = \emptyset \\ 0 & otherwise. \end{cases}$$

v_i and v_j represent two individuals in the hypergraph. Based on the hypothesis that the letter writer knows other co-senders better than non-authors, we create two types of edges to represent two kinds of senders in the Co-sender Network. If a hyperedge e with two nodes v_i and v_j belongs to the tail of the edge, and either of them also belongs to the set of author(s), then $S_{ij} = 1$. In other words, if two individuals send a letter together and either one of them is the author, then the entry $S_{ij} = 1$ and thus an edge (i, j) is included in the Co-Sender network. Another situation is that two nodes belong to the tail of the specific edge but neither of them belongs to the set of authors, in which case $S_{ij} = -1$. In other words, although two individuals send a letter together, there still exists another sender who is the real author. Therefore, for each entry $S_{ij} = -1$, a different type of edge (i, j) is included in the Co-Sender network. Given such a matrix, a Co-Sender network is defined as follows.

Definition 3.4. Co-Sender network. The co-sender graph is represented as an undirected graph $G_s = (V_s, E_s)$, where V denotes the set of nodes ($V_s \subseteq V$), and $E_s \subseteq V_s \times V_s$ denotes the set of edges representing the co-sender relationship.

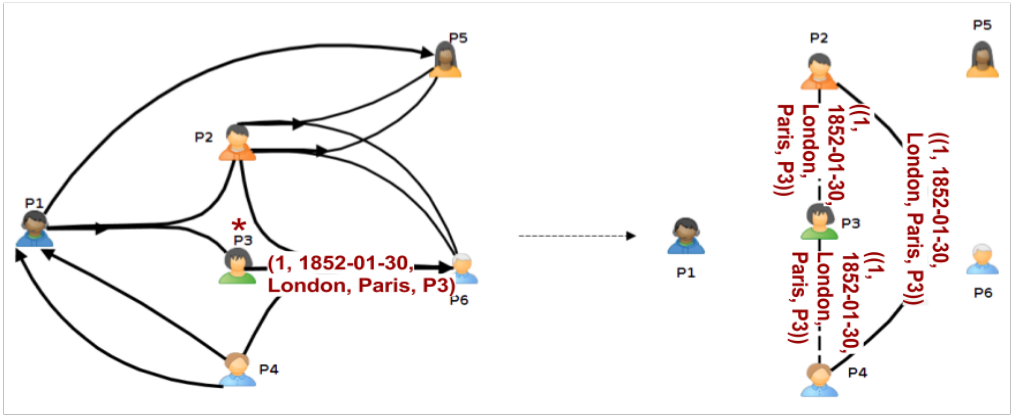


Figure 3.7: A simple Co-Sender network: the network on the left side is the original hypergraph model (the $*$ near $P3$ means that $P3$ is the author of a letter with $P2$ and $P4$ as co-authors), and the network on the right side is the Co-Sender network transformed from the hypergraph model. We use two different types of edges to illustrate the relationship between co-senders. The dashed edge between two nodes implies that either of them is the author of the letter. The solid edge connecting two nodes implies that they are co-senders but neither of them is the author. The triangle consisting of two dashed edges and one solid edge indicates an analogy to the social balance theory, i.e., a co-sender of my co-sender is likely to be my co-sender.

In contrast to the sender-recipient network, the edges in a co-sender network are normally considered to be undirected, since we presuppose that sending letters together is a symmetrical relationship. In order to differentiate author and co-sender in the co-sender network, two functions are derived from the hypergraph model. One is $count : V_s \times V_s \to \mathbb{N}$, which indicates the author-sender relation, i.e., how many times an individual is the author while he/she sent a letter with others. On the other hand, $author : V_s \to \mathbb{N}$ is another function uploaded from the author function aut in the definition of hypergraph model, which indicates how many times an individual is the author of any given letter. In other words, given two nodes i and j, in the following formulas, we use $count(i, j)$ to denote the number of letters written by i and sent together with j, and $author(i)$ to denote the number of letters of which i is the author in the network.

$$count(i,j) := |\{e \in E \mid i \in aut(e) \wedge i \in T_e \wedge j \in T_e\}| = |\{e \in E \mid i \in aut(e) \wedge j \in T_e]\}| \quad (3.13)$$

$$author(i) := |\{e \in E \mid i \in aut(e)\}| \quad (3.14)$$

The set of edges $|E|$ in the hypergraph model (cf. Section 3.4.1) is also applied in the equations above to denote the set of all letters. In a similar way to the sender-recipient network, we use a set of quintets $\{(d, t, l_{si}, l_{sj}, aw) \mid d \in \mathbb{N}, t \in T, l_{si} \in L, l_{sj} \in L, aw \in V_s\}$ as edge attributes and each quintet is the same as the edge attribute in the hypergraph model. We can represent the co-sender network by an adjacency matrix A^s and each entry in the matrix is defined as the number of elements in the corresponding edge attribute set.

$$A^s_{ij} = \left|\{(1, t^1_{ij}, l^1_{si}, l^1_{sj}, aw^1), (2, t^2_{ij}, l^2_{si}, l^2_{sj}, aw^2), ..., (k, t^k_{ij}, l^k_{si}, l^k_{sj}, aw^k)\}\right| \tag{3.15}$$

The number of elements in the attribute set corresponds to the number of letters sent by i and j together. In a Co-Sender network, the weight of an edge is equal to the corresponding value in the adjacency matrix A_s that quantifies the relationship between senders. In the subsequent part of this section, we introduce our measurements with the following questions: who are the most important person(s) in the co-sender network? Who wrote these letters? Who sent the most letters together with whom? What is the average number of senders per letter? How many co-senders in average does an individual have? Is there any potential community of individuals in the co-sender network?

In order to find answers to all these questions, we first propose two probabilities to measure the collaborative behaviors. The Louvain algorithm for community detection is also applied to a co-sender network in order to find collaborative groups of individuals. These measurements will later be applied in Section 3.6.2, and they will reveal interesting features and patterns that contributes to our understanding of collaborations between historic persons.

- **Co-Sender Probability**. It is our hypothesis that people who have sent many letters together are likely to know each other better on average than those who have only sent letters together on an infrequent basis. In order to account for this, we measure the co-sender probability that two individuals sent a letter together as,

$$P(i, j) := \frac{A^s_{ij}}{E_s}\,, \tag{3.16}$$

where A^s_{ij} represents how many times two individuals have sent a letter together and E_s represents the number of all the letters in the co-sender network.

- **Author Probability**. Senders who have written many letters, we assume, know other co-senders better on average than those who have just sent a few letters together. In order to account for this effect, we propose following two probabilities. The first is the probability that an individual is the author of randomly picked letter, which is calculated by the following formula:

$$P(i \in aut(e)) := \frac{author(i)}{|E_s|}, \tag{3.17}$$

where $author(i)$ specifies how many times an individual i is the author and $|E_s|$ denotes the number of all the letters in the co-sender network. The second is a conditional probability that an individual is the author, given that he/she sent a letter together with another person.

$$P(i \in aut(e) | i \in T_e, j \in T_e) := \frac{count(i,j)}{A^s_{ij}} \tag{3.18}$$

In this formula, $count(i,j)$ calculates how many times an individual is the author while another correspondent sent a letter with him/her together, and A^s_{ij} represents how many times these two individuals have sent a letter together.

- **Community Detection**. In order to capture highly connected circles of friends, colleagues or families in historic correspondences, we will apply the Louvain algorithm, which was introduced in Section 2.3.3, in the experiment of the co-sender network for community detection, since its accuracy is comparable to the accuracy of other algorithms but it offers better scalability [49]. We presuppose that the community structure in co-sender networks will provide us with a deeper understanding of the relations between individuals than studying sender-recipient relations alone.

The probabilities proposed above will be employed on one dataset in Section 3.6, in order to obtain an overview of the personal collaboration in a letter collection. Not only that, we will also use this dataset to examine whether there are any embedded groups of individuals in the network using the Louvain algorithm for community detection. In addition, the weighted betweenness centrality, which was introduced in Section 3.4.2, will also help us to find the important node(s) in this network.

3.4.4 Co-Recipient Network

A co-citation network is a network with nodes corresponding to papers and edges corresponding to the co-citation relation between papers [19]. Two papers are considered to be connected if and only if they are cited together in another paper. If two papers are often cited together, they are highly likely to have a common topic. We apply the idea of co-citation network in the area of historic correspondence research: the more often two individuals have received a letter together, the more likely it is that they are related. We adapt this concept of "connectivity" for the construction of co-recipient network in this section, i.e., two individuals are considered to be connected if and only if they have received one or more letters together.

In this section, analogous to the construction of co-sender network in Section 3.4.3, nodes are the correspondents and an edge is constructed when they have received at least one letter together. In order to describe the relation between multiple recipients, we introduce the third derived graph from the correspondence network model. For the purpose of constructing a co-recipient network, an incidence matrix R derived from the correspondence model is represented as:

$$R_{ij} = \begin{cases} 1 & \exists e \in E : v_i \in H_e \wedge v_j \in H_e \wedge v_i \neq v_j \\ 0 & otherwise. \end{cases}$$

v_i and v_j represent two individuals in the hypergraph. If a hyperedge e with two nodes v_i and v_j both belonging to the head of the edge, $R_{ij} = 1$, otherwise $R_{ij} = 0$. In other words, if two individuals have received the same letter, then an edge (i, j) is included in the Co-Recipient network. Given such a matrix, a Co-Recipient network is defined in the following way.

Definition 3.5. Co-Recipient network. The co-recipient graph is represented as an undirected graph $G_r = (V_r, E_r)$, where V_r denotes the set of nodes ($V_r \subseteq V$), and $E_r \subseteq V_r \times V_r$ denotes the set of edges representing the co-recipient relations.

An example of Co-Recipient network is shown in Figure 3.8. We use a set of quintets $\{(d, t, l_{ri}, l_{rj}, aw) \mid d \in \mathbb{N}, t \in T, l_{ri} \in L, l_{rj} \in L, aw \in V_r\}$ as edge attributes and each quintet is the same as the edge attribute in hypergraph model. We can represent the co-recipient network by an adjacency matrix A^r and each entry in the matrix is defined as the number of elements in the corresponding edge attribute set.

$$A^r_{ij} = \left|\{(1, t^1_{ij}, l^1_{ri}, l^1_{rj}, aw^1), (2, t^2_{ij}, l^2_{ri}, l^2_{rj}, aw^2), ..., (k, t^k_{ij}, l^k_{ri}, l^k_{rj}, aw^k)\}\right| \tag{3.19}$$

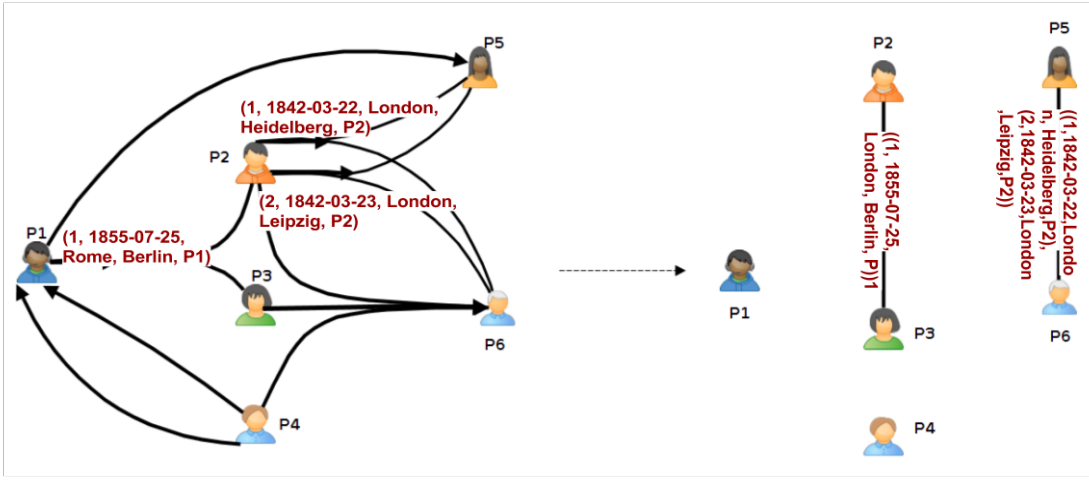


Figure 3.8: A simple co-recipient graph: the network on the left side is the original hypergraph correspondence model, and the network on the right side is the co-recipient network transformed from the left. In the co-recipient graph, the edges between each two nodes illustrate that they are both recipients of at least one letter.

In the subsequent part of this section, we begin our measurements with the following questions: who are the most important person(s) in the co-recipient network? Who received the most letters together with whom? What is the average number of recipients per letter? How many co-recipients in average does an individual have? Is there any potential community of individuals in the co-recipient network?

In order to find answers to these questions, we propose a co-recipient probability to measure the frequency of receiving letters together. The Louvain algorithm for community detection is also applied to the co-recipient network for potential highly connected groups of recipients.

- **Co-Recipient Probability**. We assume that people who have received many letters together are likely to know each other better on average than those who have only received letters together on an infrequent basis. In order to account for this, we measure the co-recipient probability that two individuals received a letter together as:

$$P(ij) := \frac{A^r_{ij}}{|E_r|}\,, \tag{3.20}$$

where A^r_{ij} represents how many times two individuals have received a letter together and $|E_r|$ represents the number of all the letters in the co-recipient network.

- **Community Detection**. In order to capture highly connected co-recipients, we can also apply the Louvain algorithm, introduced in Section 2.3.3, in the co-recipient network to explore the community structure and make a comparison with the co-sender networks. This will help us to identify the relations between individuals more thoroughly.

3.5 Network Dynamics

In this section, we analyze individual relations in the correspondence network using a dynamic approach. Retaining the temporal information in the network provides us with a diachronic view of the interactions among nodes and allows us to perform an extensive and comprehensive analysis of the whole network. It should be noted that we do not abandon the use of static network analysis, but rather provide a temporal view in order to analyze the correspondence network more precisely. We use two types of representations, namely contact sequences and graphlet sequences, to describe the evolving correspondence network. For the contact sequences, we focus on two contact patterns, i.e., inter-contact time and reciprocal time, in order to obtain interesting insights of individual correspondence behavior in the historic times. For the graphlet sequences, we generalize measurements on the changes of nodes and edges in order to obtain fluctuating and persistent patterns of graphlets over time.

3.5.1 Contacts and Graphlets

Most letter repositories only contain the records of the dates of writing, but not the dates when the letter was received. Hence it is difficult for us to determine the temporal duration of each letter. Therefore, we assume that each letter has no duration and each letter was written at a specific point in time. These points in time and the order of letters in the repositories determine the sequence of letters over time.

In this section, in order to capture the individual interaction patterns and global changes of the correspondence network, we describe our correspondence network using two types of representations. The first representation is a sequence of contacts in which each contact is an interaction in the form of letters between individuals, and the second representation is a sequence of graphlets, in which each graphlet contains a group of contacts that take place within a given time interval. The first representation keeps the discrete temporal information of the interaction between each pair of nodes in the correspondence network. The second representation considers the correspondence network as an evolving network, and techniques from static network theory could be applied directly to a sequence of graphlets. In this way, we can keep all the temporal information and explore the correspondence network dynamically.

Our correspondence network can be represented as a set of contacts and a contact is the basic unit of interaction, i.e., a letter sent between two individuals at a certain point in time. We introduce the definition of a contact as follows.

Definition 3.6. Contact. Given a correspondence network H, a contact ct from node i to j is defined as a quadruple $ct = \{(i, j, t, d) \mid t \in T, d \in \mathbb{N}\}$, where t denotes the date when a letter was written and d is the index number to differentiate different edges between any two nodes.

We assume that the whole timespan of a correspondence network is finite, from the start time $t_s \in T$ to the end time $t_e \in T$. In this way, our correspondence network is represented as a set Ct of contacts that happen between a set of nodes V during a finite time interval $[T_s, T_e]$.

In order to capture the global changes of the correspondence network over continuous periods of time, the correspondence network can also be described as a sequence of graphlets. We define two functions $f_e(e,t)$ and $f_v(v,t)$ to illustrate the occurrence of a certain node or edge at a certain time. If there exists an edge e at time t, $f_e(e,t)$ equals 1, otherwise 0; if there exists a node v at time t, $f_v(v,t)$ equals 1, otherwise 0. The set of available points in time assigned with an edge $e \in E$ is represented as $D(e) := \{t \in T \mid f_e(e,t) = 1\}$. Similarly, the set of available points in time of a node $v \in V$ is denoted as $D(v) := \{t \in T \mid f_v(v,t) = 1\}$. We introduce the definition of a graphlet as follows.

Definition 3.7. Graphlet. A graphlet $\mathcal{G}$ of the correspondence network H is defined as the set of occurrences of nodes in V and edges in E during a time interval $[t_i, t_j]$, which is denoted as $\mathcal{G} = (V_{[t_i,t_j]}, E_{[t_i,t_j]})$, where $t_i, t_j \in T$, $V_{[t_i,t_j]} := \{v \in V \mid f_v(v,t) = 1, t_i \leq t \leq t_j\}$ and $E_{[t_i,t_j]} := \{e \in E \mid f_e(e,t) = 1, t_i \leq t \leq t_j\}$.

In order to condense the multiple edges in one graphlet into a single edge, we measure the weight of a given edge e as the number of occurrences of edge e in a graphlet $\mathcal{G}$: $w_{\mathcal{G}}(e) := |\{e \in E \mid f_e(e,t) = 1, t_i \leq t \leq t_j\}|$. In this way, we condense the multiple edges in one graphlet into a single edge. We use *st* to denote the time duration of each graphlet. The time granularity of a graphlet is set according to the time granularity of the data, for example, monthly or yearly intervals.

The adjacency matrix of a graphlet $\mathcal{G}$ within a time interval $[t_i, t_j]$ is denoted as $A^{\mathcal{G}}$. Each element in $A^{\mathcal{G}}$ corresponds to the weight of the edge between two corresponding nodes within a time interval $[t_i, t_j]$.

$$A^{\mathcal{G}}_{ij} := |w_{\mathcal{G}}(e_{ij})| \tag{3.21}$$

3.5.2 Reciprocal Time and Inter-Contact Time

Motivated by our interests in contact behavior of individuals in historic correspondences, we define and study two contact patterns in this section, the first of which is called *inter-contact* time and the second of which is called *reciprocal time*. Inter-contact time measures the time between letters sent continuously from a specific sender to a specific recipient, while reciprocal time represents

the time between a letter sent from a certain sender to a certain recipient and a following letter sent as a reply. We assume that human correspondence is driven by the well-timed responses to received letters. We also assume that continuous and mutual contacts reflect the closeness of individuals. In the following part, we introduce our definition of *inter-contact time* in order to measure the time between two consecutive contacts.

Definition 3.8. Inter-Contact Time. Given a contact (i, j, t_m, d_m), and the following contact (i, j, t_n, d_n), the *inter-contact time* between those two contacts is defined as the time interval $[t_m, t_n]$, where $t_m \leq t_n$, $d_m < d_n$, and $|t_n - t_m| \leq \psi$. The whole set of inter-contact time between two nodes i and j is represented a set of k time intervals $\sigma(i, j) := \{[t_1, t_2], [t_2, t_3], ..., [t_{k-1}, t_k]\}$.

This measure helps us to explore the contact behavior of individuals without having prior knowledge of the content of their letters. However, if the inter-contact time happens to be unreasonably long, e.g., many years between two contacts, we do not consider these two contacts to be consecutive. The unreasonably long inter-contact time might be caused by the loss of letters or the receiving of an unrelated letter from the same writer after a long delay. In order to filter all the inter-contact time intervals that are much longer than expected, we use the method proposed by Tukey [215] to help us find the threshold. Tukey used the "fences" calculated by quantiles as a means of filtering the values that are too far away from the statistical center of the data range. Since these values lay outside the expected range, they are also called outliers. The "fences" are calculated as follows:

$$Q = Quantile(\{|t_n - t_m|\}), \tag{3.22}$$

$$\psi := [Q_1 - k(Q_3 - Q_1), Q_3 + k(Q_3 - Q_1)], \tag{3.23}$$

where *Quantile* denotes a function that calculates the quantiles of the set of the lengths of different inter-contact time intervals and Q denotes the set of four quantiles: Q_1, Q_2, Q_3 and Q_4. ψ denotes the fence that flags possible outliers. k is set to 1.5, since it is widely used in practice [216]. In our case, we use the

upper limit of ψ as the threshold for the length of an inter-contact time within our dataset. In the following, we introduce our definition of *reciprocal time* in order to measure the time between letters and their corresponding responses. This measure helps us to uncover the response pattern embedded in the letter collections in a quantitative way.

Definition 3.9. Reciprocal Time. Given a contact (i, j, t_m, d_m), and the following replying contact (j, i, t_n, d_n), the *reciprocal time* between these two contacts is defined as the time interval $[t_m, t_n]$, where $t_m \leq t_n$, $d_m < d_n$ and $|t_n - t_m| \leq \psi$. The whole set of reciprocal time between two nodes i and j is represented a set of k time intervals $\varepsilon(i, j) := \{[t_1, t_2], [t_2, t_3], ..., [t_{k-1}, t_k]\}$. Similarly, we use Tukey's fences to remove the unexpectedly long reciprocal time interval in ε.

3.5.3 Measurements

In our study we make use of the fact that each edge in the graph is assigned to a point in time to study how the graph evolves over time. The centrality measures for static networks have been adapted to temporal networks based on the prerequisite that each path has a duration [55]. However, in our correspondence network, each edge is only assigned to a point in time instead of a time interval. In view of this situation, we do not focus in this section on the measurements related to the temporal paths. Rather we focus on the fluctuation and persistent patterns of nodes and edges in the graphlet sequences over time. Viswanath *et al.* [217] applied the idea of resemblance to measure the quantitative overlap in edges between two network snapshots. In our case, we adopt their notion and introduce the refresh rate of nodes and edges in our own definitions.

Definition 3.10. Refresh Rate of Nodes. Given two consecutive graphlets $\mathcal{G}_i$ and $\mathcal{G}_j$ in H, the refresh rate of nodes is defined as the proportion of the number of nodes that only appear in $\mathcal{G}_j$ to the number of all the nodes in $\mathcal{G}_i$.

$$U^v_{\mathcal{G}_i} := 1 - \frac{|V_{\mathcal{G}_i} \cap V_{\mathcal{G}_j}|}{|V_{\mathcal{G}_i}|} \tag{3.24}$$

$|V_{\mathcal{G}_i}|$ denotes the number of nodes in graphlet $\mathcal{G}_i$ and $|V_{\mathcal{G}_j}|$ denotes the number of nodes in the following graphlet $\mathcal{G}_j$. The intersection $|V_{\mathcal{G}_i} \cap V_{\mathcal{G}_j}|$ represents

the number of the overlapping nodes in these two graphlets. The value of $U^v_{\mathcal{G}_i}$ varies between 0 and 1. If $U^v_{\mathcal{G}_i} = 0$, the entire set of nodes in graphlet $\mathcal{G}_i$ remains unchanged in graphlet $\mathcal{G}_j$. If $U^v_{\mathcal{G}_i} = 1$, none of the nodes occurring in $\mathcal{G}_i$ exists in $\mathcal{G}_j$. Similarly, we introduce the following definition of refresh rate of edges in order to find the fraction of edges that change from one graphlet to the next.

Definition 3.11. Refresh Rate of Edges. Given two consecutive graphlets $\mathcal{G}_i$ and $\mathcal{G}_j$ in H, the refresh rate of edges is defined as the proportion of the number of edges that only appear in $\mathcal{G}_j$ to the number of all the edges in $\mathcal{G}_i$.

$$U^e_{\mathcal{G}_i} := 1 - \frac{|E_{\mathcal{G}_i} \cap E_{\mathcal{G}_j}|}{|E_{\mathcal{G}_i}|} \tag{3.25}$$

$|E_{\mathcal{G}_i}|$ denotes the number of edges in graphlet $\mathcal{G}_i$ and $|E_{\mathcal{G}_j}|$ denotes the number of edges in graphlet $\mathcal{G}_j$. The intersection $|E_{\mathcal{G}_i} \cap E_{\mathcal{G}_j}|$ represents the number of the overlapping edges in these two graphlets. The value of $U^e_{\mathcal{G}_i}$ varies between 0 and 1. If $U^e_{\mathcal{G}_i} = 0$, the entire set of edges in graphlet $\mathcal{G}_i$ remains unchanged in graphlet $\mathcal{G}_j$. If $U^e_{\mathcal{G}_i} = 1$, none of the edges occurring in $\mathcal{G}_i$ exists in $\mathcal{G}_j$.

Furthermore, we also focus on the nodes and edges that are persistent over time. Tang *et al.* [218] proposed a *temporal-correlation coefficient* to compute the overlap of nodes between any two successive undirected and unweighted graphs. This approach has the prerequisite that nodes are stable in the network over time. Since we are dealing with evolving correspondence network with fluctuating nodes and edges over time, this method is not appropriate for our research. We generalize the idea of refresh rates of nodes and edges in order to measure the persistent nodes and edges in the network. In this case, we focus on the persistent patterns not only in consecutive graphlets, but also in non-consecutive graphlets.

Definition 3.12. Persistent Rate of Nodes. Given any two graphlets $\mathcal{G}_i$ and $\mathcal{G}_j$ in H, the refresh rate of nodes between $\mathcal{G}_i$ and $\mathcal{G}_i$ is defined as the proportion of the number of nodes that are common in both $\mathcal{G}_i$ and $\mathcal{G}_j$ to the number of all the nodes in $\mathcal{G}_i$ and $\mathcal{G}_j$.

$$PS^v_{(\mathcal{G}_i,\mathcal{G}_j)} := \frac{|V_{\mathcal{G}_i} \cap V_{\mathcal{G}_j}|}{|V_{\mathcal{G}_i} \cup \mathcal{G}_j|} \tag{3.26}$$

$|V_{\mathcal{G}_i}|$ denotes the number of nodes in graphlet $\mathcal{G}_i$ and $|V_{\mathcal{G}_j}|$ denotes the number of nodes in graphlet $\mathcal{G}_j$. The intersection $|V_{\mathcal{G}_i} \cap V_{\mathcal{G}_j}|$ represents the number of the overlapping nodes in these two graphlets. The union $|V_{\mathcal{G}_i} \cup V_{\mathcal{G}_j}|$ represents the number of total nodes in these two graphlets. The value of $PS^v_{(\mathcal{G}_i,\mathcal{G}_j)}$ varies between 0 and 1. If $PS^v_{(\mathcal{G}_i,\mathcal{G}_j)} = 1$, the entire set of nodes in graphlet $\mathcal{G}_i$ equals the set of nodes in graphlet $\mathcal{G}_j$. If $PS^v_{(\mathcal{G}_i,\mathcal{G}_j)} = 0$, none of the nodes occurring in $\mathcal{G}_i$ exists in $\mathcal{G}_j$. Similarly, we introduce the following definition of persistent rate of edges to find the fraction of edges which remain persistent from one graphlet to the other.

Definition 3.13. Persistent Rate of Edges. Given any two graphlets $\mathcal{G}_i$ and $\mathcal{G}_j$ in H, the persistent rate of edges between $\mathcal{G}_i$ and $\mathcal{G}_j$ is defined as the proportion of the number of edges that are common in both $\mathcal{G}_i$ and $\mathcal{G}_j$ to the number of all the edges in $\mathcal{G}_i$ and $\mathcal{G}_j$.

$$PS^e_{(\mathcal{G}_i,\mathcal{G}_j)} := \frac{|E_{\mathcal{G}_i} \cap E_{\mathcal{G}_j}|}{|E_{\mathcal{G}_i} \cup E_{\mathcal{G}_j}|} \tag{3.27}$$

$|E_{\mathcal{G}_i}|$ denotes the number of edges in graphlet $\mathcal{G}_i$ and $|E_{\mathcal{G}_j}|$ denotes the number of edges in graphlet $\mathcal{G}_j$. The intersection $|E_{\mathcal{G}_i} \cap E_{\mathcal{G}_j}|$ represents the number of the overlapping edges in these two graphlets. The union $|E_{\mathcal{G}_i} \cup E_{\mathcal{G}_j}|$ represents the number of total nodes in these two graphlets. The value of $PS^e_{(\mathcal{G}_i,\mathcal{G}_j)}$ varies between 0 and 1. If $PS^e_{(\mathcal{G}_i,\mathcal{G}_j)} = 1$, the entire set of edges in graphlet $\mathcal{G}_i$ equals the set of edges in graphlet $\mathcal{G}_j$. If $PS^e_{(\mathcal{G}_i,\mathcal{G}_j)} = 0$, none of the edges occurring in $\mathcal{G}_i$ exists in $\mathcal{G}_j$.

3.6 Experiments

In this section, our objective is to apply the concepts and measurements of correspondence networks to real datasets in order to deepen our understanding of the relations between historic persons and the patterns embedded in their letters. Considering the size of the datasets we have obtained, we begin with the static analysis of four letter collections concerning two types of relations, i.e., the sender-recipient relation and the co-sender relation. We examine the static patterns embedded in these two networks constructed from these datasets. On the other hand, we observe the evolution of the correspondence network and explore the fluctuating and persistent patterns over the years.

3.6.1 Dataset

In this section, we give a brief description about all the letter collections we will use in the experiments. Our datasets consist of the letter metadata from four different institutions, and support the exploration of various relationships and interesting patterns embedded in the correspondence networks.

- **Darwin Correspondence Dataset**. Charles Robert Darwin (1809–1882) is famous *inter alia* for his contributions to the biological theory of natural evolution. During his lifetime, Darwin used letters to exchange information and academic ideas with his friends, family and individuals who were helpful towards his research. These letters provide an access to his academic circles in science, culture and religion. Researchers of the *Darwin Correspondence Project*[1], based in the Cambridge University Library, have collected, transcribed and published most letters written by or to Darwin. The dataset used in this dissertation is a subset of Darwin's correspondences spanning from January $6th$, 1826 to January $28th$, 1882. It consists of metadata of $9,001$ letters: $1,358$ person names and $1,486$ locations.

- **Alfred Russel Wallace Correspondence Dataset**. Alfred Russel Wallace (1823–1913) is one of the $19th$ century's most famous naturalists, who is also recognized as the "father" of evolutionary bio-geography. He is the co-discoverer with Charles Darwin of the process of evolution by natural selection. He made other significant contributions not only to biology, but to subjects such as glaciology, land reform, anthropology, and astrobiology. During his lifetime, he corresponded with many of the leading figures in science, politics, and literature both in Europe and North America. These letters contain discussions, observations, and discoveries on a variety of scientific and social subjects. The *Wallace Letters Online*[2] provides access to the letters written or received by Alfred Russel Wallace. These letters contain important observations, discoveries, and fascinating discussions on a variety of subjects. The dataset used in this dissertation is a subset of Wallace's correspondences spanning from January 11th, 1840 to November 14th, 1913. It consists of metadata of 1,862 letters: 759 person names and 591 locations.

- **Mark Twain Correspondence Dataset**. Samuel Langhorne Clemens, better known by his pen name *Mark Twain* (1835–1910), is one of the America's most famous literary icons. He wrote 28 books and numerous short stories, letters and sketches. The *Mark Twain Project Online*[3] offers access to the most recently discovered letters of Mark Twain. Untrammeled by literary conventions, Mark Twain recorded what was in his mind in these letters. The dataset used in this dissertation constitutes a subset of Mark Twain's correspondences spanning from June $15th$, 1858 to January $8th$, 1878. It consists the metadata of $1,510$ letters: 346 person names and 205 locations.

- **Philipp Melanchthon Correspondence Dataset**. Philipp Melanchthon (1497–1560), a close friend of Martin Luther (1483–1546), is one of the most important reformers in the first half of the 16th century. He published numerous books and commentaries in the areas of science, history, and theology. His contribution to the research and education system gave him the name of "Praeceptor Germaniae" (Teacher of Germany). He is also well known as the first systematic theologian of the Protestant Reformation, and played an important role in theology and church history. His letters are considered as an important source for studying German history during the early modern period. These letters have been (and are still) collected and carefully analyzed over decades by the *Melanchthon Research Center*[4] in Heidelberg, Germany. The dataset used in this dissertation is a subset of Melanchthon's correspondences spanning from November $8th$, 1520 to January $29th$, 1560. It consists of metadata of 124 letters: 101 person names and 76 locations.

Table 3.1 lists the basic information of these letter collections. We can see that Darwin, Wallace and Twain lived within the same time period (ca. 19th century), and their letter collections have some senders or recipients in common, even though they did not all have the same profession. These datasets will supply our static and temporal analysis with their detailed metadata information. However, all three datasets only contain tens of letters with multiple senders or recipients, whereas the letter collection of Melanchthon is the only dataset to which we can obtain

[1] www.darwinproject.ac.uk/ [Last accessed: September 25, 2015].
[2] http://www.nhm.ac.uk/research-curation/scientific-resources/collections/library-collections/wallace-letters-online/index.html [Last accessed: September 25, 2015].
[3] www.marktwainproject.org/homepage.html [Last accessed: September 25, 2015].
[4] www.haw.uni-heidelberg.de/forschung/forschungsstellen/melanchthon/mbw-online.de.html [Last accessed: September 25, 2015].

access that contains over a hundred letters with multiple senders. Due to the lack of an available and appropriate dataset, an analysis of the co-recipient networks lies beyond the scope of our experiments in this chapter. Consequently, in the following experiments, we will focus strictly on the analysis of sender-receiver networks and co-sender network.

Dataset	Time Period	Letters	People	Locations	Letters per person	Intersection
Darwin	1826–1882	9,001	1,358	1,486	6.63	Wallace, Twain
Wallace	1840–1913	1,862	759	591	3.15	Darwin, Twain
Twain	1858–1878	1,510	346	205	4.36	Darwin, Wallace
Melanchthon	1520–1560	124	101	76	1.23	—

Table 3.1: A brief summary of our datasets. Here we give a general overview of our datasets concerning the time period, the respective number of letters, people and locations, the average number of letters by each person, and the overlaps among datasets.

3.6.2 Static Analysis

In this section, we apply measurements that have been proposed and generalized in the previous sections to the letter collections of four historic persons, and we highlight the important individuals, their latent communities, and interesting static patterns embedded in the resulting graphs.

Sender-Recipient Network. As we mentioned in Section 2.5.2, the sender-recipient relation represents the most intuitive relationship existing in the historic correspondences, and is frequently represented in network visualization in the area of digital humanities. In this chapter, this "who is writing to whom" relation is represented as a directed graph, with people being nodes and the letters being edges. We employ various measurements on datasets and present a detailed analysis of the results as follows.

- **Weighted degree distribution**. In order to find nodes other than the central node that are important in the network, we calculate separately the weighted degree centrality for three datasets. The cumulative weighted degree distributions of both in- and out-degree for each dataset are shown in Figure 3.10 on a double logarithmic axis (log-log plot). It is interesting and reasonable to find that both the in-degree and the out-degree distributions of all three datasets follow heavy-tailed distributions. In other words, a few individuals have significantly higher degrees, and most others have much lower

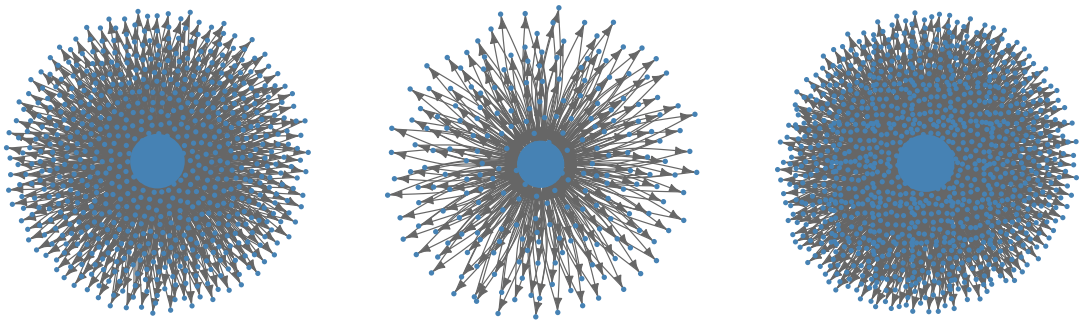
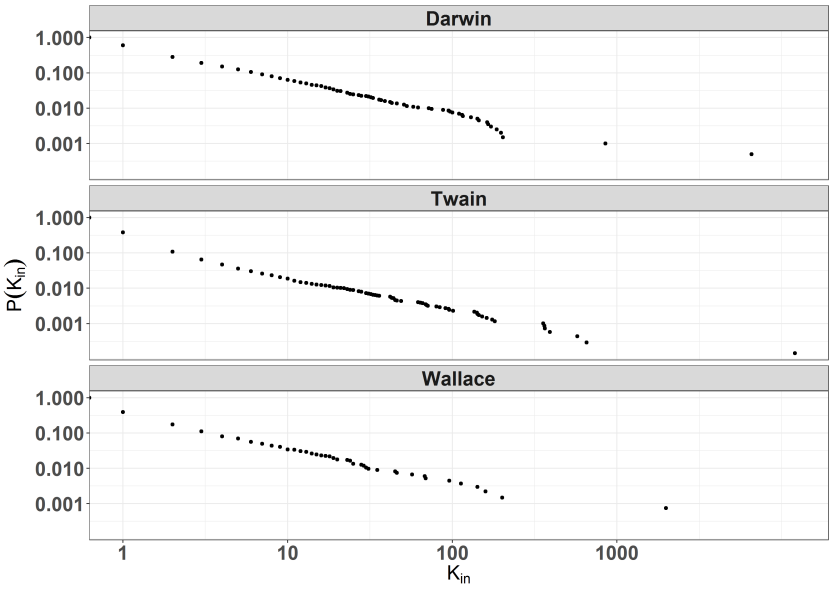


Figure 3.9: Three figures from left to right in the plot correspond to the sender-recipient networks of Darwin, Wallace and Twain, respectively. For each figure, nodes represent the correspondents and directed edges represent the letters sent between each two individuals. We mark the central node in each graph with blue. We have simplified each graph to show only the individuals (nodes) who have at least 2 letters sent to or received by them.

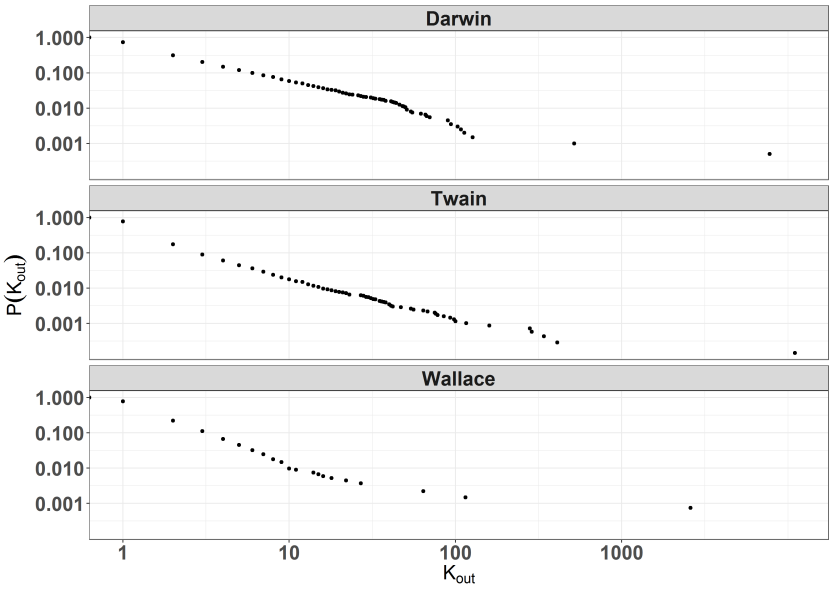
degrees in each graph. We think that there is a tendency whereby individuals, at least in the past, tended to write and keep a lot of letters with only a few persons, even though they have hundreds of unique correspondents.

- **Weighted Degree**. In order to find the important individuals in these datasets, we calculate the in-degree and out-degree for each node in the sender-recipient networks and show the top 4 individuals in Table 3.2. We can see that no matter whether we look to the scientists (Darwin and Wallace) or to the writer (Twain), at least one of their family members ranks high on the lists of in-degree and out-degree scores. For example, on the top-4 lists of in-degree and out-degree scores of Wallace's sender-recipient network, *William Greenell Wallace* is his son and *Violet Isabel Wallace* is his daughter. The frequent occurrences of family members on the top-lists of in-/out-degree is a reasonable expectation since these letter collections were originally collected and preserved by the family members of these historic persons.

  In addition to family members, most others on the lists are friends, colleagues, and publishers of Darwin, Wallace, and Mark Twain, respectively. However, it is surprising to find on the top in-/out-degree lists of Mark Twain, four occurrences are his business managers (Ralph W. Ashcroft, Franklin G. Whitmore, and Charles L. Webster, who appears twice). That is because Mark Twain invested a great deal of his writing profits in new inventions and technologies, but failed a lot. He blamed his failures on investment advisors



(a) In-Degree Distribution



(b) Out-Degree Distribution

Figure 3.10: The weighted in-degree and out-degree distributions of our datasets are shown on the double logarithmic axis (log-log plot). The horizontal axis represents the in/out-degree of each node in the graph, and the vertical axis represents the corresponding cumulative degree distribution.

In-Degree		Out-Degree	
Darwin			
Joseph Dalton Hooker	851	Joseph Dalton Hooker	517
Charles Lyell	203	Asa Gray	127
John Murray	198	Robert Francis Cooke	113
Thomas Henry Huxley	186	George Howard Darwin	108
Wallace			
William Greenell Wallace	201	Charles Robert Darwin	64
Raphael Meldola	159	Charles Lyell	27
Violet Isabel Wallace	142	Joseph Dalton Hooker	22
Edward Bagnall Poulton	113	William Turner Thiselton-Dyer	18
Twain			
Olivia Langdon Clemens	575	William Dean Howells	410
Charles Luther Webster	391	Charles Luther Webster	288
Franklin G. Whitmore	365	Orion Clemens	280
Henry Huttleston Rogers	363	Joseph Hopkins Twichell	160

Table 3.2: Top 4 individuals ranked by in-degree and out-degree scores in Darwin, Twain and Wallace's datasets, respectively. The person names in blue represent the individuals who appear on both the top lists of in-degree and out-degree scores, and the person names in orange represent the individuals who are the family members of these historic persons.

and fired one after another of them [219]. Furthermore, in each dataset, there is at least one individual who ranks highly both on the top in-degree and out-degree lists. For example, *Joseph Dalton Hooker*, as a good friend and confidant of Charles Darwin, kept in contact with Darwin most frequently in the sender-recipient network. These values also indicate the high tendency of reciprocal interactions between Darwin and Hooker. In the following part, we will measure the reciprocity in the sender-recipient network of these historic persons.

- **Reciprocity**. Our focus upon reciprocity here concerns the tendency towards forming mutual contacts between a pair of individuals by their responses between each other. In a highly reciprocal relation, both individuals have interests in preserving their relationship. Conversely, in a low reciprocal relation, one person seems more active in keeping up their relationship than the other. In order to investigate who kept most reciprocal contacts with the central person in the network, we calculate the local reciprocity for each node in the graph and select the ones with both high volume and balance, as shown in Figure 3.11 and Table 3.3 below.

  In Figure 3.11, we can find that most family members of all these historic persons are relatively higher in volume but lower in balance compared to

other people in the dataset. We think that this is because family members can choose to respond directly to them in person, which is more convenient and faster than writing letters. Similar to the ranking of weighted degree in Table 3.2, most individuals in Table 3.3 are the colleagues, friends, publishers, or business managers of these historic persons. Among them, the individuals who have the same profession as the central person in the graph rank relatively higher on the list than those individuals in other professions.

Local Reciprocity		
Darwin	**Wallace**	**Twain**
Joseph Dalton Hooker	Charles Robert Darwin	William Dean Howells
Asa Gray	Charles Iyell	Frederick A. Duneka
George Howard Darwin	Joseph Dalton Hooker	Charles Luther Webster
Alfred Russel Wallace	William Turner Thiselton-Dyer	Joseph Twichell
John Murray	Annie Wallace	Orion Clemens

Table 3.3: Top 5 correspondents ranked by local reciprocity scores in Darwin, Twain and Wallace's datasets. The person names in blue represent the individuals who are the family members of these historic persons.

- **Weighted Betweenness**. Betweenness centrality allows us to find the nodes (individuals) that are important in connecting two other nodes in the network. We use the measurement of weighted betweenness centrality introduced in Section 3.4.2 and rank all the nodes accordingly. The top 5 nodes are shown in Table 3.4. It is interesting yet also unsurprising that the family members of these historic persons occupy positions of high ranks for betweenness scores. For instance, in the top list of Darwin, three persons are from Darwin's family. This is because most letters in each collection are letters written/received by a specific historic person or the family members of this person.

Weighted Betweenness		
Darwin	**Wallace**	**Twain**
George Cupples	William Greenell Wallace	Orion Clemens
Emma Darwin	Henry Rider Haggard	Olivia Langdon Clemens
Francis Darwin	Frank Evers Beddard	Clara Clemens
George Howard Darwin	Annie Wallace	Andrew Chatto
Joseph Dalton Hooker	Frances Sims	Bret Harte

Table 3.4: Top 5 individuals ranked by the scores of weighted betweenness in the datasets of Darwin, Wallace, and Twain, respectively. The person names in blue represent the individuals who are the family members of these historic persons.

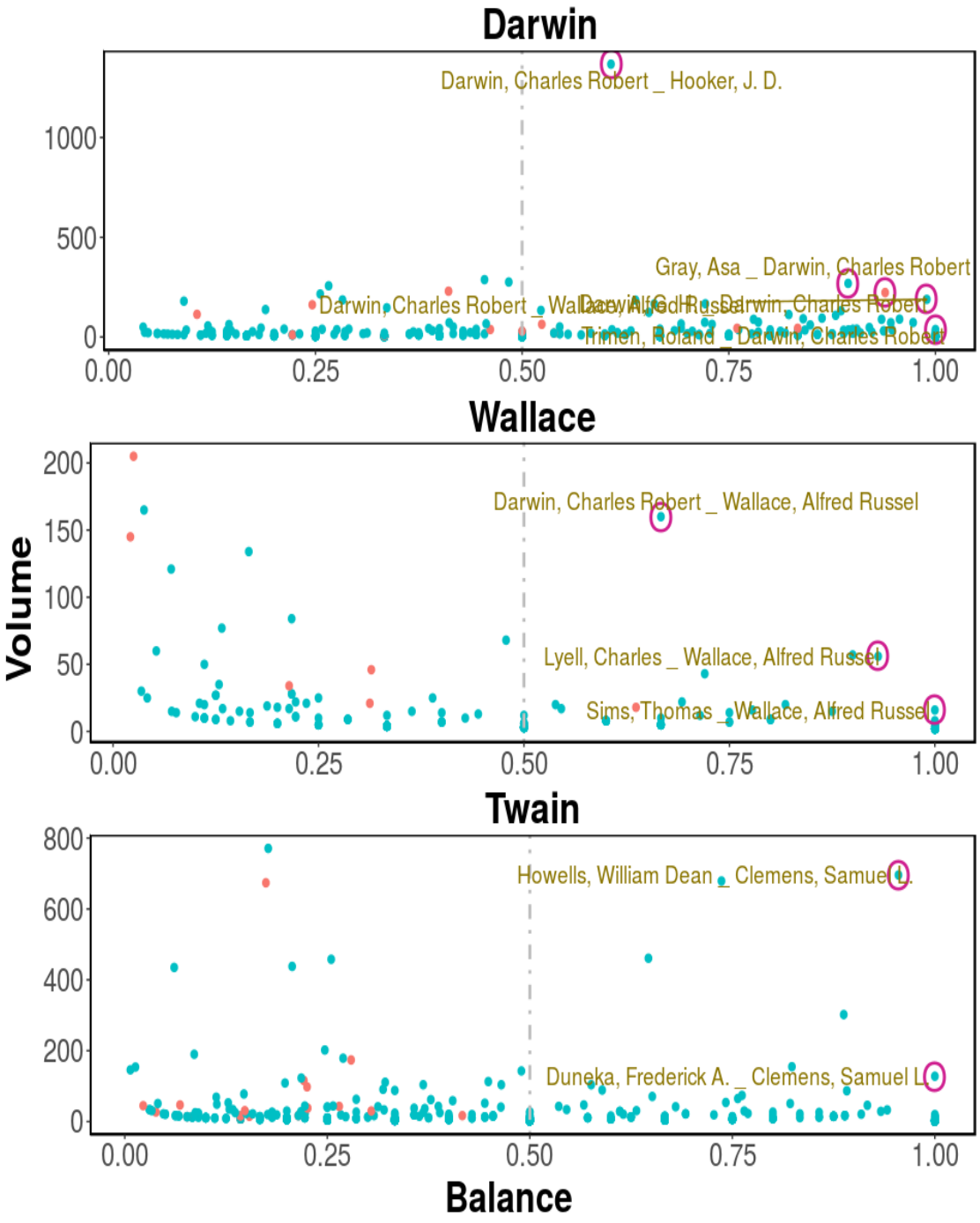


Figure 3.11: Three plots from top to the bottom on the figure correspond to the letter collections of Darwin, Wallace and Twain, respectively. On each plot, the x-axis represents the local reciprocity (balance) and the y-axis represents the sum of edge weights (volume). Each point on each plot corresponds to a pair of nodes in the sender-recipient graph. Among these points, the red points correspond to the pairs of these three individuals and their corresponding family members, and the points with red circle are the points selected by our algorithm. The dashed vertical lines on each plot filter out the points that are not satisfying our requirements.

- **Weighted Closeness**. Closeness centrality helps us to find how close a node is to other nodes in the network. We use the measurement of weighted closeness centrality introduced in Section 3.4.2 and rank all the nodes accordingly. The top 5 nodes are shown in Table 3.5. In general, the rank of closeness centrality is similar to the rank of weighted degree centrality in Table 3.2.

Weighted Closeness		
Darwin	**Wallace**	**Twain**
Joseph Dalton Hooker	William Greenell Wallace	Olivia Langdon Clemens
Charles Lyell	Raphael Meldola	Charles Luther Webster
John Murray	Violet Isabel Wallace	Franklin G. Whitmore
Thomas Henry Huxley	Edward Bagnall Poulton	Henry Huttleston Rogers
William Darwin Fox	Charles Robert Darwin	William Dean Howells

Table 3.5: Top 5 individuals ranked by the scores of weighted closeness in the datasets of Darwin, Wallace, and Twain, respectively.

Since our sender-recipient graphs of these three historic persons are all star-structure and small scale, we can only obtain limited knowledge about their real sender-recipient relation and about the way that nodes play their roles. In the following paragraphs, we analyze a dataset in the frame of the co-sender network to explore the latent collaboration relations in the historic correspondences.

Co-Sender Network. It is our hypothesis that most people who have sent multiple letters together might know one another quite well. We construct the co-sender network based on the letter collection of Philipp Melanchthon. Table 3.6 presents a summary of the basic properties of the co-sender network of Melanchthon.

Melanchthon's Co-Sender Network	
Total letters	124
Total senders	101
Total authors	25
Senders per letter	0.81
Number of communities	16
Size of the largest community	18
Size of the second largest community	16
Diameter	6

Table 3.6: A summary of the basic properties in the co-sender network of Philipp Melanchthon.

In Table 3.6 we can find that there are 25 authors in the co-sender network of Melanchthon, but only 24 letters were written by authors other than Melanchthon. Thus in this case we do not calculate the conditional probability that an author

sent a letter together with another person proposed in Section 3.4.3. Instead, we calculate the co-sender probability for pairs of senders in the network. Moreover, having measured the centrality for each node in the co-sender network, we will next highlight the results and their implications, after which we will undertake a detailed analysis of the communities obtained in the following part.

- **Co-Sender Probability**. In order to find out which pair of individuals always preferred to send letters together in Luther's circle of correspondents, we calculate the co-sender probability proposed in Section 3.4.3 for each pair of nodes in the network. As shown in Table 3.7, Melanchthon seems more likely to collaborate with Johannes Bugenhagen, Justus Jonas and Caspar Cruciger. Interestingly, these three individuals were all colleagues of Melanchthon and are not on the list of top 10 people who kept most frequent contact with Melanchthon. Table 3.7 also indicates the frequent collaborations between Johannes Bugenhagen and four other people (Martin Luther, Caspar Cruciger, Georg Maior and Justus Jonas) in the network, which can be regarded as an indirect and supplementary piece of evidence in support of the potential relations between nodes in the correspondence network.

	Pair of Senders	**Betweenness**
1	Martin Luther – Philipp Melanchthon	Justus Jonas
2	Johannes Bugenhagen – Philipp Melanchthon	Caspar Cruciger
3	Justus Jonas – Philipp Melanchthon	Nikolaus von Amsdorf
4	Caspar Cruciger – Philipp Melanchthon	Martin Luther
5	Johannes Bugenhagen – Martin Luther	Benedikt Pauli
6	Justus Jonas – Martin Luther	Franz Burchard
7	Philipp Melanchthon – Georg Maior	Johannes Brenz
8	Johannes Bugenhagen – Caspar Cruciger	Friedrich Myconius
9	Johannes Bugenhagen – Georg Maior	Tilemann Plettener
10	Johannes Bugenhagen – Justus Jonas	Anton Lauterbach

Table 3.7: Top 10 pairs of senders ranked by co-sender probability and top 10 senders, other than Melanchthon, ranked by weighted betweenness scores in the co-sender network of Philipp Melanchthon.

Table 3.7 also shows the top 10 individuals ranked by the betweenness scores in the co-sender network of Philipp Melanchthon. When compared to the top list of co-senders, there are obvious changes in people and their positions on the list. For instance, Martin Luther, who ranks in the highest position on the co-sender list, ranks in the fourth position in the list of betweenness score

and Justus Jonas ranks in the first position instead. It is also surprising to see that Nikolaus von Amsdorf ranks highly on this list, as he is actually not a friend of Melanchthon, and they had serious conflicts in their beliefs. This person will be further mentioned in the following analysis of communities.

- **Community Structure**. Since we are dealing with letter collections in which most letters belong to one single person, we can only obtain a giant component while studying the sender-recipient relation, with this person being the central node in the sender-recipient network. In order to investigate the potential various groups of individuals, we apply the Louvain algorithm, which was introduced in Section 2.3.3, to the co-sender network of Philipp Melanchthon. The main community structure is shown in Figure 3.12. We can find that the smallest community contains 3 nodes, and the most frequent size of a community is 4. The largest community with Melanchthon inside contains 18 nodes, all of which correspond to Melanchthon's family members, friends or colleagues who have the same profession as him, e.g., Justus Jonas and Caspar Cruciger. The second largest community contains 16 nodes, and it is interesting to find that some of the people in this community are not Melanchthon's friends and they had conflicts with him before, e.g., Nikolaus von Amsdorf and Andreas Karlstadt. Moreover, compared to other communities, doctors are more involved in this community, e.g., Thomas Eschaus and Augustin Schurff.

3.6.3 Temporal Analysis

In this section, we present our experimental results for our temporal study of correspondence networks with three empirical datasets. We first give a brief summary of the basic properties of our correspondence network from a temporal point of view. Table 3.8 shows the percentage of the letters in our datasets with precise dates. Most letters in our datasets have been dated to a granularity of days, and we thus restrict our analysis only to letters that have precise dates. Then we examine the distributions of inter-contact time and reciprocal time in order to obtain individual contact patterns embedded in their correspondence behaviors. Furthermore, we apply the measurements proposed in Section 3.5.3 to reveal the fluctuating and persistent patterns in the networks. We find that the

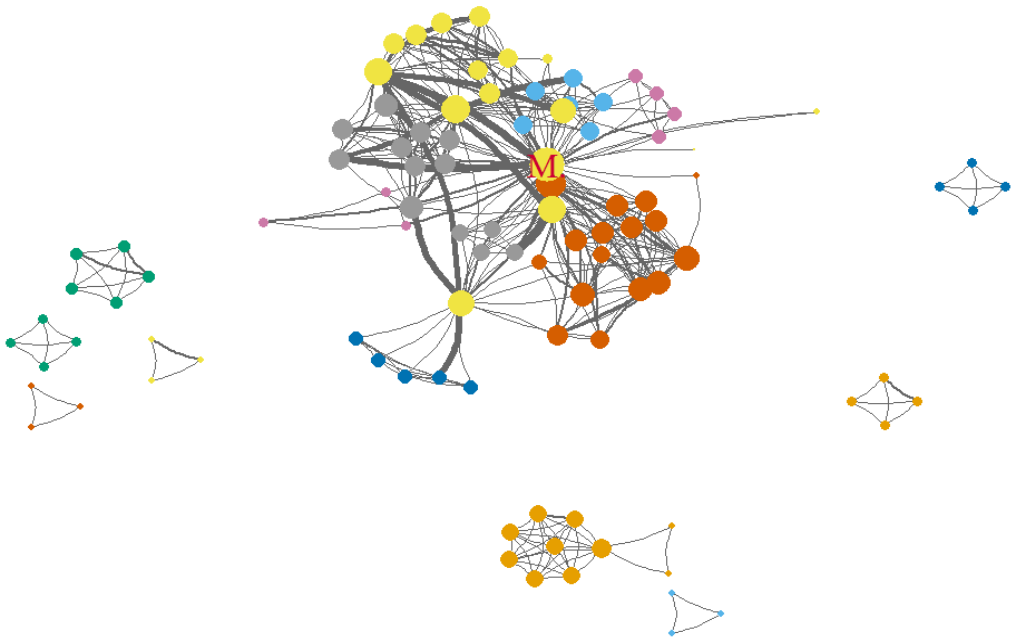


Figure 3.12: Community structures in the co-sender network of Philipp Melanchthon. Here we use different colors to represent different communities. Melanchthon is marked with the letter M in red. The nodes in yellow represent the largest communities in the graph. The size of each node corresponds to the degree of the node, and the width of each edge corresponds to the weight of the edge.

correspondence behaviors of historic persons reflect both stable and fluctuating signs, or in other words, their contact circles expanded significantly after they rose to fame, even though only a few correspondents kept contact with them over time.

	Letters	**Letters with Precise Dates**
Darwin	14,342	13,543 (94,42%)
Wallace	4,588	4,234 (92.3%)
Twain	23,654	19,595 (82.9%)

Table 3.8: The composition of letters with precise dates in three datasets, respectively.

Figure 3.14 shows the total number of letters sent and received per year by Darwin, Wallace, and Twain, respectively. And Figure 3.15 shows the number of letters sent or received by these three individuals, respectively. During their lifetimes, there are a few significant daily fluctuations hidden behind these numbers that cannot be neglected. The frequency of interactions changes significantly before or after important life events, such as the letter-writers' moving between countries or birthdays. For instance, Wallace received 14 letters on his birthday January

$8th$, 1913. Darwin sent 5 letters on February $22th$, 1868 regarding research on sexual selection. The number of letters of all three individuals exploded after they became famous, and kept a highly fluctuation pattern afterwards. For instance, before Charles Darwin went to Cambridge in 1827, his recorded letters were only a few in number and most of them were sent to friends and relatives. After he first published his theory of natural selection with Wallace in 1858, Darwin's communication with other individuals substantially increased. Alfred Russel Wallace served as President of the Anthropology Department of the British Association in 1866. Twenty years later, in November 1886, Wallace began a ten-month trip to the United States to give a series of popular lectures. These two facts might explain the sudden increase of the total number of letters that happened around these two years in Figure 3.14.

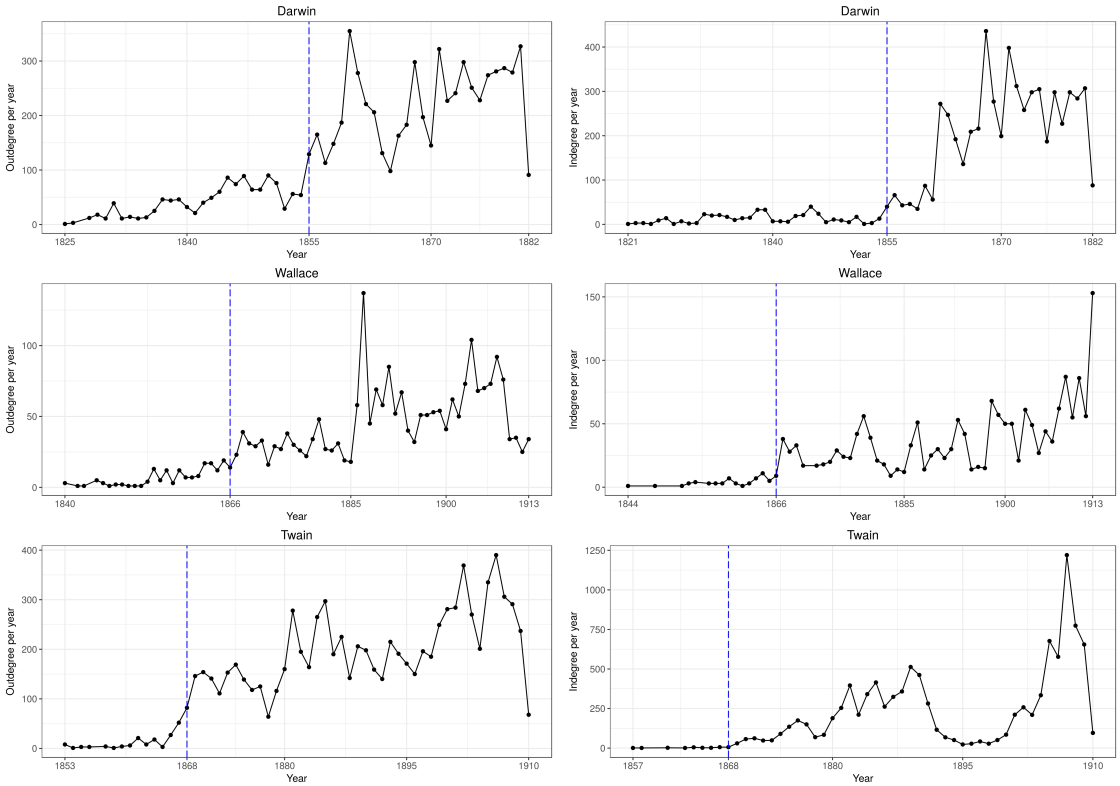


Figure 3.13: The figure on the left side describes the out-degree of Darwin, Wallace and Twain per year, respectively. This corresponds to the letters sent by them. The figure on the right side describes the in-degree of Darwin, Wallace and Twain per year, respectively. This corresponds to the letters received by them. On each figure, the x-axis represents the year, and the y-axis represents the number of letters sent or received by year. The blue dashed line on each plot marks the year since when the number of their letters had significantly increased.

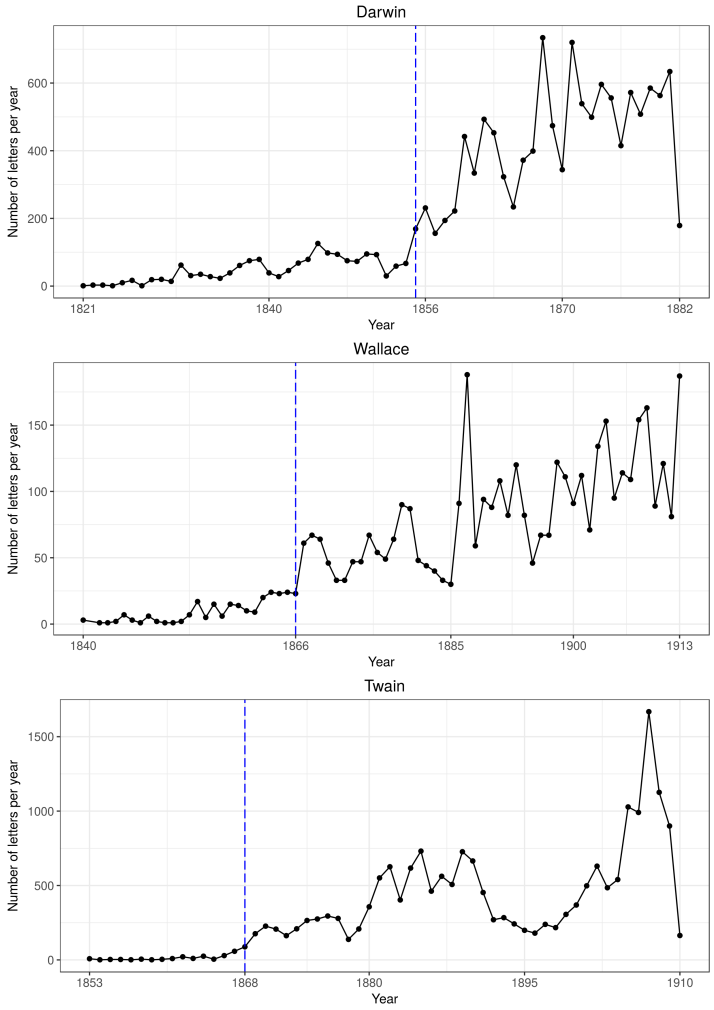


Figure 3.14: Each plot, from top to the bottom, describes the total Letters sent and received by Darwin, Wallace and Twain, respectively. On each plot, the x-axis represents the year, and the y-axis represents the number of letters by year. The blue dashed line on each plot marks the year since when the number of their letters had significantly increased.

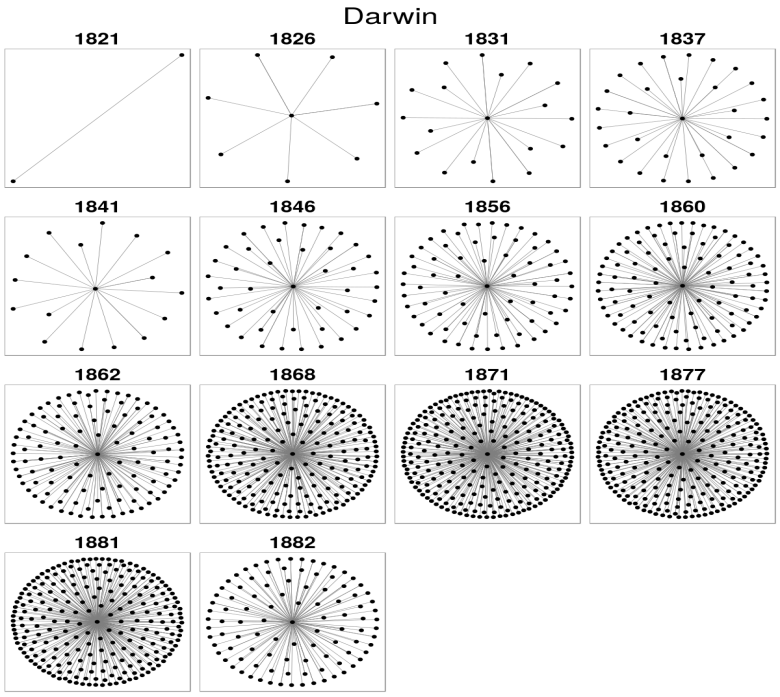
- **Inter-Contact Time and Reciprocal Time**. According to the definitions of inter-contact time and reciprocal time in Section 3.5.2, we calculate the distributions of inter-contact time between two consecutive letters and reciprocal time between a letter and its reply in three datasets. Each time distribution is measured at the individual level; in other words, we only observe the behavior of one historic person in each dataset. In order to avoid unreasonably long intervals in time, we employ the notion of the Tukey's fence, which was introduced in Section 3.5.3 regarding the sets of inter-contact time and reciprocal time. The results show that the correspondence behaviors of all three historic persons exhibit exponential distributions, or in other words, among the contact sequences of one individual, most contacts occurred within

a relatively short time period. For instance, consecutive letters (54% for Darwin, 40% for Wallace, 73% for Twain) in three datasets were sent within less than a period of 50 days.

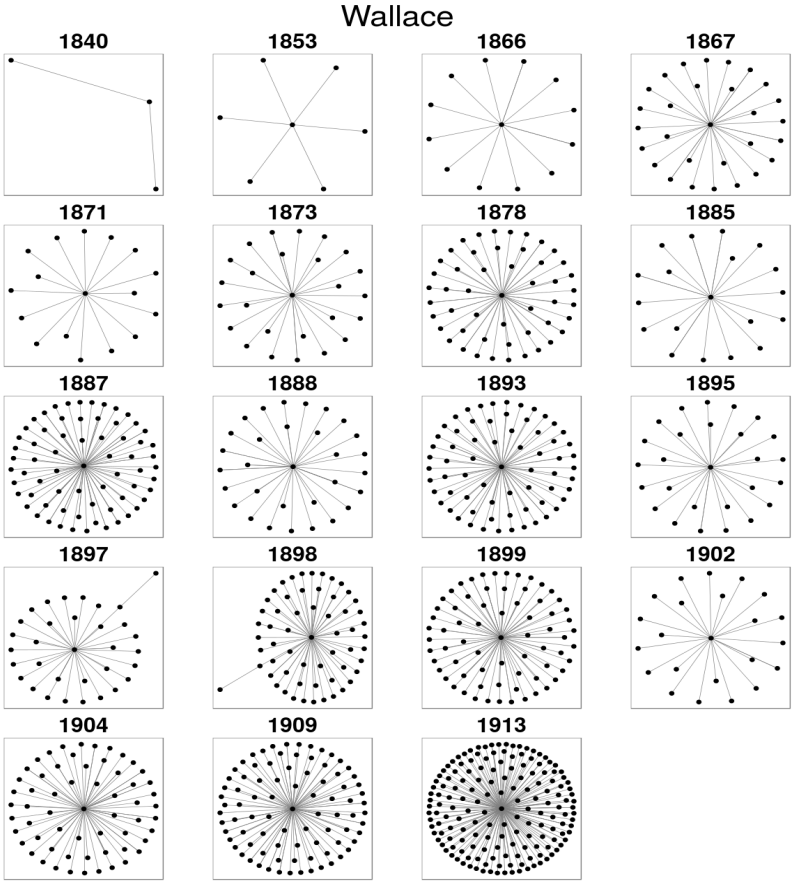
- **Fluctuating Pattern**. In order to examine the changes of individual interactions over time, we calculate the refresh rates of nodes in two consecutive (yearly) graphlets for three datasets, respectively. As shown in Figure 3.16, during the early life of these historic persons, there are similar patterns of dramatic fluctuations of nodes in the correspondence networks. In contrast with their late life, the fluctuations of nodes continue to be at a high level. We suppose that when they rose to fame, more and more people tended to contact with them. Although they had a large number of correspondents, they kept frequent contacts with only a few of them. For instance, in the last year of Wallace's lifetime, he had 163 correspondents, but only 9 people, e.g., Edward Bagnall Poulton and David Prain, who were Wallace's colleagues and with whom he had collaborations, in common with the people one year earlier.

- **Persistent Pattern**. Now we turn our focus to the stable patterns in correspondences networks over time. We calculate the persistent rates of nodes in any two (yearly) graphlets for three datasets, respectively. The patterns to which we pay most attention are not only persistent nodes in consecutive graphlets, but also the nodes recurring over many years. In Wallace's dataset to which we have access, Wallace only kept contact with Henry Walter Bates from 1845 to 1847, as he had become friends with Henry Walter Bates during that time and they went to South America together for an expedition in 1848. From 1849 to 1850 he only kept contacts with Samuel Stevens, who was a natural history agent in London. Wallace and Bates were his clients, and Steven supported their expedition to South America. This Bates also reoccurred in the graphlet of 1857 with Darwin. We suppose that these letters in 1857 should be related to the expedition and the species they sent back.

  In addition, in Twain's dataset, the persistent rate between 1853 and 1862 ranks most highly (57.1%), as he left home for a visit or traveled in these two years and kept most contact with his family members. From 1861 to



(a) This figure corresponds to the simplified correspondence network of Darwin.



(b) This figure corresponds to the simplified correspondence network of Wallace.

Figure 3.15: A visualization example of Darwin's and Wallace's correspondence networks in the form of graphlet sequences over the years.

1863 the persistent rate continues at 44.4%, during which time he kept most contact with his brother Orion Clemens. This time period is a critical point in Twain's lifetime. The American Civil War began in 1861 and halted river trade, and thus Mark Twain ended his career on the river and started writing for various newspapers using the pen name "Mark Twain".

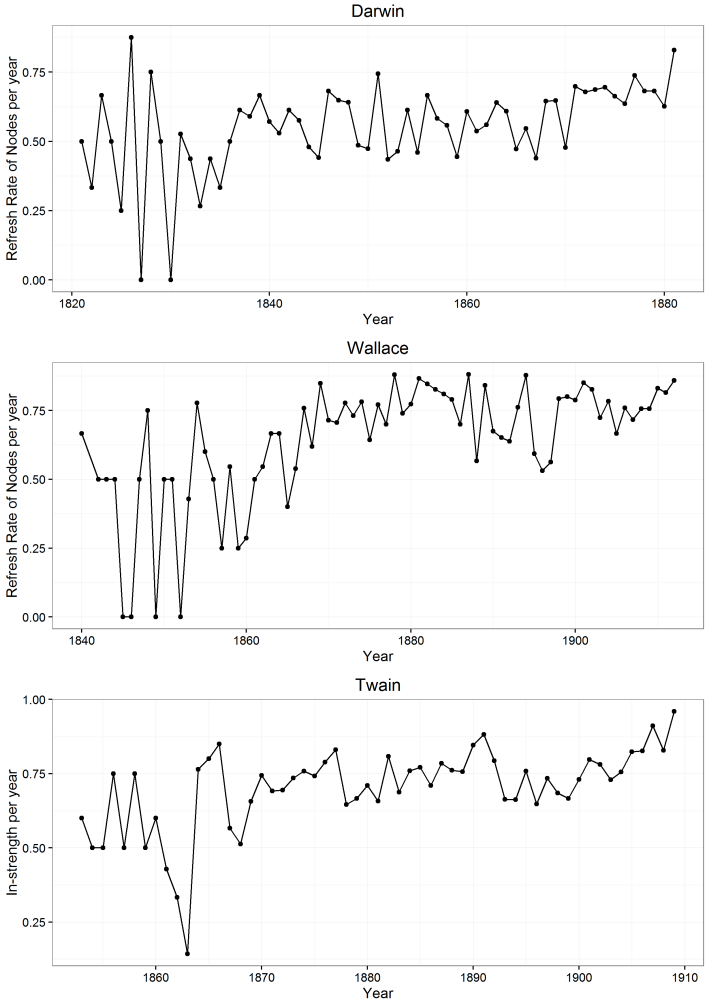


Figure 3.16: This figure shows the refreshing rates of nodes in the graphlet sequences of Darwin, Wallace and Twain, respectively. The x-axis represents years, and the y-axis represents the refresh rates of nodes per year.

In summary, the correspondence activities of all three historic persons reflect both stable and fluctuating patterns. The correspondents of these historic persons expanded significantly after they rose to fame and kept a highly fluctuation pattern afterwards. In other words, although these historic persons had many correspondents, they only kept continuous contact with only a few individuals over time.

3.7 Summary of the Chapter

Historic correspondences provide new insights into the history of intellectual-exchange. However, previous research on historic correspondences only remained in its early stage, such as the construction of repositories or network visualization. Few projects have moved beyond the above stage to exploit the detailed modeling of correspondence networks, let alone to develop new concepts and formal algorithms derived from modern social network analysis. In this chapter, we aim to analyze historic correspondence in the form of social networks on the basis of letter metadata. After a description of different entities in the metadata, we proposed a ***correspondence network model***, which integrates three types of personal relationships as well as temporal and geographical information. As a further contribution of this chapter, we ***generalized and developed*** measurements regarding different relations derived from the model. Furthermore, we investigated the ***dynamic structures and the evolving patterns*** of correspondence network with definitions and measurements from a temporal point of view.

We then applied our models and measurements to empirical datasets, in order not only to obtain an overview of the correspondence networks in the given historic periods, but also to explore latent relations between historic persons and interesting (static and dynamic) patterns embedded in the networks. Based on the experimental results of the letter collections of three historic persons, we highlighted the ***important individual nodes***, the ***latent collaborative communities***, and the ***"changing yet stable" contact behaviors*** of historic persons.

In contrast with previous research into correspondence networks, our comprehensive model combines personal, temporal and geographical information into a formal network structure. This model will be further extended to integrate the textual information in the following chapter. Furthermore, it is appropriate for us to derive our model into different graphs regarding different relations, and it is convenient to adapt our model to study the correspondence network from a geographic or temporal point of view, since we retain all these information in our network model.

For future work, the application of our correspondence network model on a large scale of available datasets is a strong desideratum, in order that more network

concepts and computational methods can be developed or applied to the historic network research. Not only that, more investigations into the temporal and geographical dimensions such as regional influence and temporal communities will be implemented with the intention of enriching the exploration of correspondence networks. In this chapter, when we extract entities from the letter metadata, we find that the letter collections present us with uncertain data which are hard to refine without the help of letter contents. To bridge this gap, we will integrate in the following chapter our correspondence network model with letter content information and combine network analysis with statistical techniques, in order to provide more precise measurements of the correspondence network and generate a powerful way of tackling uncertain entities in the metadata of historic correspondences.

I shall never be a heretic; I may err in dispute, but I do not wish to decide anything finally; on the other hand, I am not bound by the opinions of men.

— Martin Luther

CHAPTER 4

CORRESPONDENCE NETWORKS: CONTENT ANALYSIS

4.1 Overview and Objectives

In Chapter 3, we focused on the metadata of historic correspondences and proposed our correspondence network model. In this chapter, we will shift our focus from correspondence metadata to correspondence ***content*** by taking letter texts into consideration. The contents of correspondences, in the form of letters, are unstructured texts that contain the sender's topics discussed between the recipient and himself/herself. By detecting and analyzing the topics embedded in historic letters, we will discover how historic persons understood and recorded the world around them, and how they described history from a personal point of view. In this chapter, we will concentrate on ***correspondent-specific*** (i.e., author-only or sender-recipient pair) topic extraction and exploration in combination with unsupervised statistical approaches and network analysis techniques of historic texts. We assume that the letter content is closely related to the letter's date of writing, and the correspondents involved in this contact (sender and recipient). Thus, we think it is highly desirable to detect and analyze topics in a temporal and person-specific context. We will thereby extend our correspondence model to provide a stage on which personal, temporal, and topical information in the metadata and the contents of letters are interconnected.

The first objective in this chapter is to discover ***common or distinctive topic(s)*** between the correspondents in the correspondence network. It is our hypothesis

that the specific topic correlates with the specific individuals in the correspondence network. Therefore we propose a measurement named ***topic participation score*** to explore the academic and social interests of either each person or a sender-recipient pair in the network. This measurement is a combination of statistical techniques and correspondence network structures that can effectively reveal the topic-person relations embedded in networks.

The second objective is to investigate different ***trends of topics*** over time. We extend the topic participation score to measure the topic trends in order to find the most popular topic(s). Through explicitly observing the dynamic pattern and the topic shifts of individuals, we will effectively analyze and correlate dynamic changes in topics with the life events of historic persons.

The third objective is to refine ***uncertain entities*** existing in the metadata of historic correspondences. We propose a probabilistic framework for the refinement of the uncertain entities in the correspondence metadata, i.e., anonymous person names, missing dates and incomplete place names. We combine topic similarity, network structures, and correspondence metadata to deal with the data uncertainty issue in the corpus.

This chapter is organized in the following way. Section 4.2 will first lay out the issues addressed in this chapter, namely the task of effectively extracting and exploring the topics, data uncertainty and natural language processing for historic letters. Section 4.3 presents the representation of the contents of letters, the extension of our correspondence network model and corresponding measures on person-topic relations. Section 4.4 illustrates how we investigate the trends of topics and person-topic relations over time, and Section 4.5 describes our probabilistic framework for the issue of data uncertainty. We will use a collection of historic correspondences for experimental evaluation. Details of both the dataset and the results of our experiments are presented in Section 4.6, followed by a summary of this chapter in Section 4.7.

4.2 Problem Statements

The main issues to be addressed in this chapter are summarized by the following problem statements.

- **Effective Extraction and Exploration of Topics**. A key challenge and prerequisite for all further tasks addressed in this chapter is to extract and explore the topics effectively. A few studies have been managed in comparing and evaluating the performance of various statistical approaches such as the Latent Dirichlet Allocation (LDA, cf. Section 2.6.2) on historic texts. In the area of the digital humanities, LDA is widely used as an unsupervised statistical approach to detect latent topics in a large historic corpus. However, it requires a given number of topics to estimate topic and word distributions. No studies in the area of digital humanities have explained how to find the desired number of topics that are embedded in large historic texts. Although researchers from the area of computer science have proposed various approaches (cf. Section 2.6.2) to finding the "right" number of topics, the question of which method is best remains an open issue. No extensive and comprehensive work has been carried out to compare and evaluate the performances of these approaches in terms of different domains. Therefore, in this chapter, we employ three different statistical approaches introduced in Section 2.6.2 in order to find the desired number of topics for our dataset.

  Furthermore, relations between topics and other entities associated with texts, e.g., senders and recipients, have not been exploited in the previous research. Our hypothesis is that the combination of network structures and statistical techniques can generate effective measurements for topic exploration. We assign topic distribution to our correspondence network as edge attributes. In this way, we can associate the topics with nodes (correspondents) and other edge attributes such as dates of writing. In this chapter, we propose the measurement named topic participation score to explore the relation between person and topic and we further extend this score to measure the trends of topics over time.

- **Data Uncertainty**. The second challenge we face is data uncertainty. Collections of historic correspondences typically carry some degree of uncertainty

due to aging texts and untimely preservation. Ambiguously named entities such as anonymous writers or missing dates are also inevitable barriers in the analysis of historic texts. The accurate and detailed metadata constitute the basis for social network research and content analysis. For the missing or ambiguous entities in the letter metadata, although we record them as missing values in the database, we are still faced with the challenges of ambiguity. Without the refinement of unknown or imprecise data, tasks such as literature study or information retrieval, can only be done on a coarse-grained level. In this chapter, we will focus on those uncertain entities that exist in the given correspondence metadata, i.e., person names, place names, and dates. We assume that the similarity between the contents of different letters contributes to refining ambiguous entities in the letter metadata. We also assume that uncertain entities in the letter metadata can be inferred or implied from the other entities in the metadata of the corresponding letter.

Consequently, we propose a probabilistic framework, within which we integrate the metadata and the content of a letter into a network structure. We will hereby calculate and compare the similarity between letters with uncertain entity and letters embedded in correspondence networks, based on the document-topic probability distributions generated by topic modeling techniques. We will thereafter choose the candidate letters in the network and calculate the joint distribution of the entities in the metadata sequences, in order to find the most probable mention(s) for the ambiguous entity.

- **Natural Language Processing for Historic Letters**. The third challenge we face is the natural language processing for historic letters. Historic letters are often mixed with multiple historic languages: for example, in a single letter written in the early modern period (ca. 1450–1750), sentences mainly written in early modern German were often mixed with a single word or phrase written in Latin. Furthermore, due to the absence of a standard language, historic language variants differ from each other in their spelling, morphology, syntax, and lexical semantics. For instance, the person name “Martin Luther” is spelled in Martin Luther’s letter collection as “Martinus Luder”, “Martinus Luther”, “Martinus”, “D.Martinus”, “Mart. Luther”, “Martinus Lutherus”, “Martino Luther”, "M.L.", as well as with other variations. Modern NLP techniques such as Part-of-Speech (POS) tagger or Named En-

tity Recognizer (NER) are not feasible to use when they are directly applied to historic correspondences. Thus, in this chapter we use separate lemmatizers for different historic languages (i.e., Latin and Early New High German) to reduce inflectional forms of a word to a base or dictionary form of a word, in other words, a lemma.

Moreover, we have chosen to focus on the individuals mentioned in the metadata, i.e., sender(s) and recipient(s), and to admit no any further individuals mentioned in the investigated letters, for the simple reason that there are so many variations for person names in the letter contents and we do not have access to letter collections with the annotations of these entities. We do not remove texts in either languages, since it would remove a significant amount of the corpus and thus limit the hypothesis concerning topics over the years making up the early modern period. In this chapter, we will compare the impact of different lexical features (e.g., with/without stopwords, with/without stemming, with/without POS tagging, to name but a few) in these experiments regarding effective topic extraction and the refinement of uncertain data.

4.3 Letter Representation and Correspondence Network Extension

Our correspondence network, as detailed in Section 3.4.1, represents individuals as nodes and letters as edges. The temporal and geographical information in the metadata of each letter is integrated into the network as edge attributes. We assume that the letter contents that will be analyzed in this chapter can be mapped onto our correspondence network. In this way, we can effectively observe the node (person)-topic distribution and thereby discover the temporal patterns of individual interactions regarding specific themes. By capturing the topics embedded in letters, we can explore how topic(s) in the corpus evolved over time as well as how an individual's topic(s) evolved over time. Not only that, we can become more aware of the dynamic interactions between individuals in terms of different topics.

We will extend the model of the correspondence network by including the content of letters as a part of edge attributes and mapping the topic distribution of each

letter onto the network. Now we will turn formally to define the concepts and representations of the letter content.

4.3.1 Letter Content Representation

In this section, we adopt the form of word sequence to represent the text of a letter in order to keep the word order and collocations in the original texts. We will use the words "letter" and "text" interchangeably hereafter.

Definition 4.1. Letter Content C. Let $C = \{c_1, c_2, ..., c_m\}$ be a collection of letter contents and W be the set of distinct words occurring in C. The *content* of each letter $c \in C$ is represented as a sequence of words denoted by $\{w_1, w_2, ..., w_n\}$, where w_n is the *n-th* word in the text.

Definition 4.2. Topic θ. A *topic* θ in a letter collection C is represented by a topic model θ, i.e., a probabilistic distribution of words $\{p(w|\theta)\}_{w \in W}$, such that $\sum_{w \in W} p(w|\theta) = 1$. We assume that there are k topics in C and we denote the set of all topics as Z.

In the letter collection C, the topic distribution for a letter c is $\{p(\theta|c)\}_{\theta \in Z}$. In other words, $p(\theta|c)$ is the probability of the topic θ occurring in letter c. $p(w|\theta)$ is the probability of the word w belonging to topic θ. Let $n_{w,c}$ be the number of the occurrences of word w in letter c. We denote the number of the occurrences of a word w in c assigned to topic θ $n_{w,c,\theta}$, and we denote $n_{c,\theta}$ the number of the occurrences of all words in c assigned to topic θ: $n_{c,\theta} := \sum_w n_{w,c,\theta} = \sum_w n_{w,c} p(\theta|c)$.

where $n_{w,c}$ denotes the number of occurrences of a certain word w in letter c and $\sum_w n_{w,c}$ denotes the sum total of the word frequencies in c.

Term Frequency Inverse Document Frequency (*tf-idf*). The term frequency (tf) measures how frequently a word occurs in a text [220]. We denote term frequency as $tf_{w,c}$, which corresponds to $n_{w,c}$, that is to say, the number of occurrences of a word w in a letter c. The term frequency $tf_{w,c}$ is divided by $\sum_w n_{w,c}$ for normalization. This is represented as:

$$tf_{w,c} := \frac{n_{w,c}}{\sum_w n_{w,c}}. \tag{4.1}$$

The inverse document frequency of words, decreases the weights of commonly used words and increases the weights of words that are not used frequently in a collection of documents [221]. Inverse document frequency (idf) for a word w is calculated as follows:

$$idf_w = \lg \frac{|C|}{|\{c \mid w \in c\}|}, \tag{4.2}$$

where $|C|$ denotes the total number of letters in the corpus and $|\{c \mid w \in c\}|$ denotes the number of texts that contain at least one occurrence of a certain word w. This formula can be combined with the term frequency, which we represent as *tf-idf*, to measure how important a word is to a text in a corpus [220]. This is calculated through the following formula:

$$tf\text{-}idf_{w,c} := tf_{w,c} \cdot idf_w = \frac{\frac{n_{w,c}}{\sum_w n_{w,c}}}{\lg \frac{|C|}{|\{c|w \in c\}|}}. \tag{4.3}$$

In other words, $tf\text{-}idf_{w,c}$ reaches its highest when w occurs frequently within a small number of documents, and conversely becomes lower when w occurs fewer times in a single text, or occurs within many texts, and indeed it reaches its lowest when w occurs in nearly all documents. For instance, there is a text containing 100 words in which the word "luther (Luther)" occurs twice. Hence, the term frequency of "luther" is as follows.

$$tf_{luther,c} := \frac{n_{luther,c}}{\sum_w n_{luther,c}} = \frac{2}{100} = 0.02 \tag{4.4}$$

Then we assume that there are 100 documents, in 10 of which the word "luther" occurs. So the inverse document frequency of "luther" is as follows.

$$idf_{luther} = \lg \frac{|C|}{|\{c \mid luther \in c\}|} = lg(\frac{100}{10}) = lg10 = 1 \tag{4.5}$$

Thus, tf-idf is calculated through the following equation:

$$tf\text{-}idf_{luther,c} := tf_{luther,c} \cdot idf_{luther} = 0.02 * 1 = 0.02 \tag{4.6}$$

In order to know the proportions of each topic in the whole corpus, we propose a measurement with respect to the topic coverage of a corpus through the following definition.

Definition 4.1. Corpus Topic Coverage. Given a corpus C and a topic θ, the *Corpus Topic Coverage* is defined as $P(\theta|C)$ the proportion of the corpus assigned to this topic.

This measurement can help us to identify the most important topics in the whole corpus, and is calculated as follows:

$$P(\theta \mid C) := \frac{\sum_c n_{c,\theta}}{n_{w,C}} = \frac{\sum_c \sum_w n_{w,c,\theta}}{\sum_c \sum_w n_{w,c}} = \frac{\sum_c \sum_w n_{w,c} p(\theta|c)}{\sum_c \sum_w n_{w,c}}, \tag{4.7}$$

where $n_{w,C}$ denotes the total number of words in C and $\sum_c n_{c,\theta}$ denotes the sum of the words assigned to a given topic θ in each letter. For example, in Table 4.1, there is a small corpus of 10 letters. The coverage of topic 1 in the corpus is thus computed as follows.

$$\begin{aligned}\sum_c n_{c,\theta=1} := \sum_c \sum_w n_{w,c,\theta=1} = \sum_c \sum_w n_{w,c} p(\theta = 1|c) = 0.193 * 155 \\ + 0.288 * 87 + 0.251 * 155 + 0.382 * 438 \\ + 0.224 * 175 + 0.190 * 358 + 0.255 * 403 \\ + 0.219 * 513 + 0.294 * 210 + 0.259 * 195 \approx 696.17\end{aligned}$$

$$\begin{aligned}n_{w,C} := \sum_c \sum_w n_{w,c,\theta} = \sum_c \sum_w n_{w,c} = 155 + 87 + 155 + 438 + 175 + 358 + 403 \\ + 513 + 210 + 195 = 2689\end{aligned}$$

$$P(\theta = 1 \mid C) := \frac{\sum_c n_{c,\theta=1}}{n_{w,C}} = \frac{696.173}{2689} \approx 0.259$$

One can see in this example that the coverage of topic 1 in this corpus equals ca. 25.89% and the coverages of the other three topics equal ca. 23.62%, 25.83%, and 24.66% respectively. Therefore, among these 10 letters, topic 1 is chosen as the leading topic. By this measure, we can obtain a general overview of the major topics in a letter collection. This measure will be further extended to measure person-topic relation in the following section.

ID	1	2	3	4	Word Count	Sender	Recipient
1	0.193	0.241	0.271	0.295	155	A	B
2	0.288	0.223	0.281	0.208	87	B	A
3	0.251	0.188	0.334	0.227	155	A	B
4	0.382	0.257	0.157	0.204	438	A	B
5	0.224	0.189	0.438	0.149	175	B	A
6	0.190	0.207	0.281	0.322	358	A	C
7	0.255	0.255	0.259	0.231	403	C	A
8	0.219	0.248	0.306	0.226	513	A	C
9	0.294	0.271	0.175	0.260	210	A	C
10	0.259	0.218	0.165	0.357	195	B	C

Table 4.1: A simple example of 10 letters with metadata information (i.e., the letter ID, the number of words per letter, sender, and recipient) and corresponding letter-topic distribution (four topics numbered by 1, 2, 3, and 4).

4.3.2 Correspondence Network Extension

We now extend the definition of the correspondence network model in Section 3.4.1 with the letter content information. By combining word distribution and network structures, we can discover new types of interesting patterns, e.g., we can explore the person who is correlated with a certain topic.

Definition 4.2. Correspondence Network. For the correspondence network, we define a "content" function con: $E \rightarrow c$ by which to assign the letter content in the form of a word sequence, as an edge attribute for each edge. We also define a "topic" function top: $E \rightarrow \{\theta_1, \theta_2, ...\theta_k\}$ in order to assign the topic distribution of each letter to the corresponding edge as an edge attribute. Thus, the updated set of attributes for an edge is $\{(d, t, l_s, l_r, aw, c, \{\theta_1, \theta_2, ...\theta_k\}) \mid d \in \mathbb{N}, t \in T, l_s \in L, l_r \in L, aw \in V, c \in C, \theta \in Z\}$.

In order to know which particular topic (or topics) an individual is interested in and to what extent different topics are covered in his/her letters, we introduce the following definition of *individual topic participation score*.

Definition 4.3. Individual Topic Participation Score. Given a correspondence network H, the *participation* of an individual (a certain node v) in topics is defined as a probability distribution $p(\theta|v)$ over his letters with regard to different latent topics.

We use $\tau(v)$ to denote the set of edges that are incident to node v. Therefore, the *topic participation score* of a node v is measured as the proportion of the number

of words assigned to a topic θ divided by the total sum of words in the letters sent or received by an individual. This score is calculated as follows:

$$p(\theta \mid v) := \frac{\sum_{e \in \tau(v)} n_{\theta,e}}{\sum_{e \in \tau(v)} \sum_{w \in c = con(e)} n_{w,c}} = \frac{\sum_{e \in \tau(v)} \sum_{c = con(e)} n_{c,\theta}}{\sum_{e \in \tau(v)} \sum_{w \in c = con(e)} n_{w,c}}, \tag{4.8}$$

where $n_{\theta,e}$ denotes the number of words assigned to a topic θ that is mapped onto edge e, and $n_{w,c}$ denotes the total number of words in c. Take Table 4.1 as an example. If we would like to calculate individual A's participation score in topic 1, based on the 9 letters that A is involved in, this would be calculated as follows.

$$\begin{aligned}\sum_{e \in \tau(A)} n_{\theta=1,e} := \sum_{e \in \tau(A)} \sum_{c = con(e)} n_{c,\theta=1} &= 0.193 * 155 + 0.288 * 87 \\ &+ 0.251 * 155 + 0.382 * 438 + 0.224 * 175 \\ &+ 0.190 * 358 + 0.255 * 403 + 0.219 * 513 \\ &+ 0.294 * 210 \approx 645.633\end{aligned}$$

$$\sum_{e \in \tau(A)} \sum_{w \in c = con(e)} n_{w,c} := 155 + 155 + 438 + 358 + 513 + 210 + 87 + 175 + 403 = 2494$$

$$p(\theta = 1 \mid A) := \frac{\sum_{e \in \tau(A)} n_{\theta=1,e}}{\sum_{e \in \tau(A)} \sum_{w \in c = con(e)} n_{w,c}} = 645.633/2494 \approx 0.259$$

A's participation score for topic 1 equals ca. 25.89%, and participation scores for the other three topics are ca. 23.76%, 26.56% and 23.79%, respectively. In other words, individual A mentioned topic 3 most in his/her letters. By this measure, we can find the key focus of an individual within any given letter collection.

In order to know which kinds of topics are present in correspondences between a specific pair of sender-recipients and to what extent these topics are covered in their letter exchange, we propose a sender-recipient pair topic participation score as follows.

Definition 4.4. Sender-Recipient Pair Topic Participation Score. Given a correspondence network H, we define the *participation* of a sender-recipient pair

(e.g., nodes i and j) in topics as a probability distribution of topics $p(\theta|i,j)$ embedded in the letters sent between them.

We overload the function $\tau(i)$ to $\tau(i,j)$ in order to represent the set of edges that lie in between i and j. Therefore, the topic participation score of a sender-recipient pair on a given topic θ is measured as the fraction of the number of words assigned to a given topic divided by the total sum of words in the letters (edges) sent between i and j. This proportion is calculated as follows:

$$p(\theta \mid i,j) := \frac{\sum_{e\in\tau(i,j)} n_{\theta,e}}{\sum_{e\in\tau(i,j)} \sum_{w\in c=con(e)} n_{w,c}} = \frac{\sum_{e\in\tau(i,j)} \sum_{c=con(e)} n_{c,\theta}}{\sum_{e\in\tau(i,j)} \sum_{w\in c=con(e)} n_{w,c}}, \tag{4.9}$$

where $n_{\theta,e}$ denotes the number of words assigned to a topic θ that is mapped onto edge e, and $n_{w,c}$ denotes the total number of words in c. Similarly, if we take Table 4.1 as an example, the participation of individuals A and B as a pair in topic 1 is calculated as follows.

$$\begin{aligned}\sum_{e\in\tau(A,B)} n_{\theta=1,e} &:= \sum_{e\in\tau(A,B)} n_{\theta=1,e} \\ &= 0.193 * 155 + 0.288 * 87 + 0.251 * 155 \\ &+ 0.382 * 438 + 0.224 * 175 \approx 300.558\end{aligned}$$

$$\sum_{e\in\tau(A,B)} \sum_{w\in c=con(e)} n_{w,c} := \sum_{e\in\tau(A,B)} \sum_{w\in c=con(e)} n_{w,c} = 155 + 87 + 155 + 438 + 175 = 1010$$

$$p(\theta = 1 \mid A,B) := \frac{\sum_{e\in\tau(A,B)} n_{\theta=1,e}}{\sum_{e\in\tau(A,B)} \sum_{w\in c=con(e)} n_{w,c}} = 300.558/1010 \approx 0.298$$

From this we can see that the pair of A and B's participation score for topic 1 equals ca. 29.76%, and the scores for other three topics are ca. 22.93%, 26.09% and 21.22%, respectively. In this example, topic 1 is thus chosen as the most important topic mentioned in the interactions of A and B.

Definition 4.5. Representative Topics for An Individual. The representative topics *Rep* for an individual i is defined as the topic or topics in which A's participation scores $p(\theta|i)$ are higher than a certain threshold η.

$$Rep(i) := \{\theta \mid p(\theta|i) \geq \eta\} \tag{4.10}$$

Similarly, we separately define the representative topics for a sender-recipient pair and for a letter as follows.

Definition 4.6. Representative Topics for a Sender-Recipient Pair. The representative topics *Rep* for a sender-recipient pair (i, j) is defined as the topic or topics in which the participation scores $p(\theta|i, j)$ are higher than a certain threshold η.

$$Rep(i, j) := \{\theta \mid p(\theta|i, j) \geq \eta\} \tag{4.11}$$

Definition 4.7. Representative Topics for a Letter. The representative topics *Rep* for a letter (an edge e) is defined as the topics in which the topic proportion of this letter is higher than a certain threshold η.

$$Rep(e) := \{\theta \mid p(\theta|e) \geq \eta\} = \{\theta \mid p(\theta|c) \geq \eta\}. \tag{4.12}$$

Although the previous studies such as Rosen-Zvi *et al.* [199] and McCallum *et al.* [138] (cf. Section 2.6.3) also explored person-topic relation, they only focused on one specific relation, i.e., author-topic relation only or author-recipient-topic only. Compared to their approaches, our measurement of topic-person relation has the following advantages: *a)* our measurement not only allows the integration of network structures, but also uses an intuitive way to explore topic-person relation, *b)* our measurement is not limited to deal with the author-topic or author-recipient-topic relation, a variety of relations are also taken into consideration, e.g., sender-topic, sender-recipient-topic, multiple senders, and multiple sender-recipients, *c)* our measurement will be extended to explore the trends of topics over time and the dynamic person-topic relation in the following section.

4.4 Temporal Study of Person-Topic Relations

In Section 3.5.1, we present a correspondence network as a set of contacts Ct, which exist between a set of nodes V during a time interval $[T_s, T_e]$ for the temporal study. A *contact* from node i to node j is represented as a quadruple $ct = \{(i, j, t, d) \mid t \in T, d \in \mathbb{N}\}$, where t denotes the time when a letter was sent and d denotes the index number used to differentiate different edges between any two nodes. In Section 3.5.1, in order to capture the global changes of network topology over time, we also represent a correspondence network as a sequence of graphlets, and the whole timespan of the network is split into a corresponding sequence of time intervals. These two representations help us to observe person-person relation over time within the frame of correspondence networks.

In this section, we take a closer look at person-topic relation over time. In Section 4.3.2, we extend our correspondence network to integrate textual information and network structures so that we can explore the relationship between individuals not only from their roles in the network, but also from the aspect of their letters. In this section, we can track both trends of topics and person-topic dynamic patterns over time. It is interesting to know which topic (or topics) is (or are) always popular and whether it correlates with a specific historic event.

4.4.1 Topic Trends

The trends of topics indicate the popularity of recurring topics that in turn provides key insights into the interest of scholars and their social events, or into the interests of a particular historic period, especially during the early modern period. We track various trends of topics and identify the topics that occur with an increasing or decreasing popularity. We investigate these topics and make efforts to correlate them with corresponding historic events occurred during a given historic period.

Griffiths and Steyver [127] used the mean of the topic distribution of texts directly to describe the trends of topics over the years without giving a clear definition of "the trend in a topic". Although this approach is easy and straightforward, it omits the word distribution for each topic and simply used the mean of the topic distribution instead, which seems a bit coarse for measuring topic trends. In order

to describe the changes in the prevalence of topics over time, we have divided the timespan of the whole corpus into a series of time intervals, and calculated the topic coverage of all letters within these different time intervals using the measurement *corpus topic coverage* (cf. Section 4.3.1). Compared to Griffiths' approach, our approach is also intuitive and can be easily extended to explore the dynamic person-topic relations, which are introduced in the following section. Hereby we define the trend of a topic as follows.

Definition 4.8. Topic Trend. Given a correspondence network H, we define the trend TD of a certain topic θ as the linear fitting of the coverage of this topic in the corpus over time: $TD(\theta) := at_k + b$, where $a \in \mathbb{R}, b \in \mathbb{R}, t_k \in T$.

The coverage of topic θ of letters sent within a given time interval $[t_i, t_j]$ is calculated based on the *corpus topic coverage* measure proposed in Section 4.3.1 using the following equation:

$$P(\theta \mid e_{[t_i,t_j]}) := P(\theta \mid C_{[t_i,t_j]}) = \frac{\sum_{c \in C_{\left[t_i,t_j\right]}} \sum_w n_{w,c} p(\theta|c)}{\sum_{c \in C_{\left[t_i,t_j\right]}} \sum_w n_{w,c}}, \tag{4.13}$$

where $P(\theta \mid e_{[t_i,t_j]})$ represents the proportion of a specific topic θ embedded in the letters sent during a certain time interval $[t_i, t_j]$, and $|e_{[t_i,t_j]}|$ denotes the number of letters sent during $[t_i, t_j]$.

The linear analysis of topic trends indicates whether there is a topic that either rose or fell in popularity over the given time interval. In order to find the most popular topic(s) and the most unpopular one(s) over time, we define the *rising* topics and the *fading* topics based on the slopes of topic trends as follows.

Definition 4.9. Rising Topics. Given a topic θ and its trend $TD(\theta)$, we define θ as a rising topic if the slope of $TD(\theta)$ is higher than a given threshold η.

$$Rp := \{\theta \mid a \geq \eta, a \in TD(\theta)\} \tag{4.14}$$

Similarly, we define the fading topics as follows.

Definition 4.10. Fading Topics. Given a topic θ and its trend $TD(\theta)$, we define θ as a fading topic if the slope of $TD(\theta)$ is lower than a given threshold η.

$$Fd := \{\theta \mid a \leq \eta, a \in TD(\theta)\} \tag{4.15}$$

These rising and fading topics respectively reflect the emerging interests and falling interests of historic persons in different topics.

4.4.2 Dynamic Person-Topic Relation

The distribution of topics signals the focus between the pairs of correspondents and helps us to gain insights into the network structures and the connections between individuals and historic events. In this section, we use the topic participation score for each node and pair of nodes in the correspondence network, in order to analyze how the interests of historic figures have changed over time. Not only that, we are also interested in the shift of topics during letter interactions and groups of individuals who share similar interests in specific topics over time.

In order to describe the changes in topic participation score for each individual or pair of individuals over time, we have divided the timespan of the whole corpus into a series of time intervals. Having considered the topic participation score for a pair of nodes, we then calculate the corresponding topic participation score for each individual within a given time interval $[t_i, t_j]$ using the following equation:

$$p(\theta \mid v)_{[t_i,t_j]} := \frac{\sum_{e_{\left[t_i,t_j\right]} \in \tau(v)} n_{\theta,e}}{\sum_{e_{\left[t_i,t_j\right]} \in \tau(v)} \sum_{w \in c = con(e)} n_{w,c}} = \frac{\sum_{e_{\left[t_i,t_j\right]} \in \tau(v)} \sum_{c = con(e)} n_{c,\theta}}{\sum_{e_{\left[t_i,t_j\right]} \in \tau(v)} \sum_{w \in c = con(e)} n_{w,c}}, \tag{4.16}$$

and the corresponding topic participation score for each sender-recipient pair within a given time interval $[t_i, t_j]$ is calculated using the following equation:

$$p(\theta \mid i,j)_{[t_i,t_j]} := \frac{\sum_{e_{\left[t_i,t_j\right]} \in \tau(i,j)} n_{\theta,e}}{\sum_{e_{\left[t_i,t_j\right]} \in \tau(i,j)} \sum_{w \in c = con(e)} n_{w,c}} = \frac{\sum_{e_{\left[t_i,t_j\right]} \in \tau(i,j)} \sum_{c = con(e)} n_{c,\theta}}{\sum_{e_{\left[t_i,t_j\right]} \in \tau(i,j)} \sum_{w \in c = con(e)} n_{w,c}}. \tag{4.17}$$

In this way, we can find out whether there is a representative topic between two individuals over time or whether there is any obvious difference in topics between different kinds of correspondents (e.g., family members, friends, the royal family, enemies, among others).

4.5 Data Uncertainty in Correspondence Networks

As we have already introduced in Section 3.3, the existing data uncertainty in historic correspondences; now in this section, our goal is to refine any uncertain entities in the metadata by proposing a probabilistic framework in combination with network structures, topic distribution, and the composition of letter metadata.

In this section, we are specifically interested in the named entities in the metadata. This specific interest derives from the fact that, for each letter in our corpus, the entities in the metadata have already been recognized and separately annotated with different entity types by historians and linguists. Not only that, we also regard the issue of data uncertainty as a special task of entity disambiguation, i.e., the disambiguation of the named entities in the metadata of letters. In this section, our focus is not to provide an extension of topic models but rather to develop a probabilistic framework for metadata uncertainty in historic correspondences.

As we mentioned in Section 2.7, although ambiguous, imprecise, and unknown entities in historic texts are inevitable, there is a dearth of literature which considers and extrapolates uncertain data in the digital humanities in meaningful ways. Furthermore, there is a notable lack of literature in the area of social networks concerning how they deal with data uncertainty issue in details. Motivated by the challenge of ambiguity, we develop a novel approach to refining different types of uncertain entities using our correspondence network structure, co-occurrence of entities in the letter metadata, and topic distributions.

4.5.1 Probabilistic Approach to Data Uncertainty

In order to achieve the aim of entity refinement, we first formally define the data uncertainty as a probabilistic issue. This definition is not only beneficial for bridging the data uncertainty that exists in historic letters with the correspondence network, but can also help us to facilitate many tasks such as information integration and information retrieval. Next, we propose a probabilistic framework by

decomposing the data uncertainty issue into three probabilities. We will calculate the similarities between different entities embedded in graphs, respectively. In order to verify the effectiveness of our probabilistic framework, we will conduct experiments over a correspondence network and a small test set of letter collection in Section 4.6.6.

In Section 3.3.1, we represent a letter as a tuple $\{S, R, l_s, l_r, t, c\}$. In this chapter, we denote m a subset of a letter: $m = \{S, R, l_s, l_r, t\}$, which corresponds to the different types of entities in the metadata. In Section 3.3.2, our correspondence network is defined as a hypergraph H, in which nodes represent individuals and directed edges represent letters. Each edge consists of a subset of nodes and is associated with different entities as attributes. Here we extend the definition of a hyperedge as a combination of incident nodes and edge attributes.

Definition 4.11. Hyperedge. Given a hypergraph $H = (V, E)$ where nodes V correspond to individuals and hyperedges E correspond to the letters sent among correspondents, we represent each hyperedge as a sequence of nodes and edge attributes denoted by $e = \{(d, H_e, t_e, l_s, l_r, t, aw, c, \{\theta_1, \theta_2, ...\theta_k\}) \mid d \in \mathbb{N}, t \in T, l_s \in L, l_r \in L, aw \in V, c \in C, \theta \in Z, H_e \subseteq P, T_e \subseteq P\}$.

In this definition, H_e denotes the head of e, which represents the set of recipients of each letter. T_e denotes the tail of e, which represents the set of senders of each letter. H_e and T_e are disjoint, i.e., $H_e \bigcap T_e = \emptyset$, for all $e \in E$. We define an attribute function att for each edge $att : E \rightarrow Y$. Y is the set of three types of entities in the metadata of all the letters in G, i.e., persons, locations and dates: $Y = \{P, L, T\}$.

We choose a set of letters L_u with ambiguous metadata M_u. We denote the *metadata* of each letter l_u of the given set as m_u. We construct a correspondence network H from the letters with definite and precise metadata. We assume that each letter $l_u \in L_u$ belongs to the same letter collection as H, otherwise any given letter l_u does not have any intersection with H, which makes the disambiguation of the entities meaningless. The set of entities in the metadata of L_u, denoted by Y_u, has the same entity types as Y in H. We can describe the data uncertainty approach more formally with the following notations in Table 4.2.

Variable	Description
H	Correspondence network
L	Set of letters in H
L_u	Set of letters
$l_u \in L_u$	Letter in the set of letters L_u
M	Metadata of all the letters in H
$m \in M$	Metadata of a single letter in H
M_u	Metadata of all the letters in L_u
m_u	Metadata of a single letter in L_u
$Y \supseteq M$	Set of all the entities in M
$y \in m$	Entity in the metadata m of a letter in H
$y' \in m$	Entity (not y) in m
$Y_u \supseteq M_u$	Set of all the entities in M_u
$y_u \in m_u$	Entity in Y_u
$y'_u \in m_u$	Entity (not y_u) in m_u
R	Relation type
att	Attribute function for each edge
typ	Function to match entities of the same type

Table 4.2: Notations for Data Uncertainty Approach

Definition 4.12. Data Uncertainty. Given a letter l_u, the corresponding *metadata* m_u and a correspondence network H, the goal is to find similar letters (edges) in H to the given letter and identify the most likely candidate entity y from the metadata of those letters for an ambiguous entity y_u in m_u. We assume that m_u contains at least one entity type that is not ambiguous.

Firstly in this section, we employ the adapted symmetrical KL divergence (cf. Section 2.7.4) and topic distribution of each letter, in order to filter out irrelevant candidates. As there could be hundreds of entities in the correspondence network H, it is extremely time-consuming to implement the probabilistic framework for all relevant and irrelevant entities.

To improve the efficiency, we select the top letters in H with the highest similarity scores compared to the letter l_u with uncertain entities as candidates. Then we compute the most likely mapping entity y as follows:

$$\underset{y \in Y}{\operatorname{argmax}} P(y \mid y_u, m_u) := \underset{y \in Y}{\operatorname{argmax}} \frac{P(y_u, m_u, y)}{P(y_u, m_u)} \propto \underset{y \in Y}{\operatorname{argmax}} P(y_u, m_u, y)\,, \qquad (4.18)$$

where the denominator $P(y_u, m_u)$ can be ignored since it does not influence the result. Here we assume that y_u and m_u are conditionally independent given y. The joint probability of an entity y_u whose context is the *metadata* m_u referring to a likely entity y can be rewritten as:

$$P(y_u, m_u, y) := P(y)P(y_u|y)P(m_u|y). \tag{4.19}$$

The main focus now is to estimate the following three components of $P(y_u, m_u, y)$:

- $P(y)$: the probability of entity y.
- $P(y_u|y)$: the probability of observing y_u given a corresponding candidate entity y.
- $P(m_u|y)$: the probability of observing metadata m_u as the context for entity y.

4.5.2 Probability Estimation

Our probabilistic framework for the data uncertainty is established under the parameters laid out by $P(y)$, $P(y_u|y)$ and $P(m_u|y)$. In this section, we present the details of the parameter estimation.

- $P(y)$. We estimate $P(y)$ using the following equation:

$$p(y) := \frac{|e|y \in att(e)|}{|E|}\,, \tag{4.20}$$

  where $|e|y \in att(e)|$ denotes the number of edges in the network that carry entity y as an attribute and $|E|$ denotes the number of all the edges in H.

- $P(y_u|y)$. We assume that the probability $P(y_u|y)$ of observing an ambiguous entity y_u given a candidate entity y, is always the same and we thus define it as a constant η where $0 < \eta \leq 1$.

- $P(m_u|y)$. $P(m_u|y)$ captures the probability of observing metadata m_u as the context for entity y. Since we are dealing with the correspondence network that contains edges associated with different types of entities, we assume that the *metadata* m_u consist of the same types of entities as each edge in the

correspondence network, and the observation of different types of entities y'_u in m_u given each candidate entity y in H is independent. Thus we have the following equation,

$$P(m_u|y) := \prod_{y'_u \in m_u} P(y'_u \mid y)\,, \tag{4.21}$$

where $P(m_u|y)$ can be represented as the product of the probabilities of different types of entities y'_u, given y. If entity y'_u is equal to entity y, $P(y'_u|y) = 1$. For other cases, we compute $P(y'_u \mid y)$ as follows:

$$P(y'_u|y) := \frac{\sum_{\{e|y \in att(e)\}} R(y'_u, y)}{|\{e|y \in att(e)\}|}. \tag{4.22}$$

$R(y'_u, y)$ is a relation function concerning the situations in which y'_u and y co-occur in the metadata of a certain letter and is measured by the matching between y'_u and corresponding entity y' :

$$R(y'_u, y) := \begin{cases} 1, & \text{if } y'_u \text{ equals } y' \text{ or } y' \subset y'_u \\ 0, & \text{otherwise.} \end{cases} \tag{4.23}$$

In the following part, we give a detailed description of the situations that $R(y'_u, y) = 1$.

y'_u	y'	y	Relation	$R(y'_u, y)$
1516-05-01	1516-05-01	y_u's candidate in H	$y'_u = y'$	1
1517-05-10	1516-05-01	y_u's candidate in H	$y'_u \neq y'$	0
1516-??-??	1516-05-01	y_u's candidate in H	$y'_u \supset y'$	1
1516-??-??	1517-05-12	y_u's candidate in H	$y'_u \not\supset y'$	0
1516-05-10 or 1516-08-10	1516-05-10	y_u's candidate in H	$y'_u \supset y'$	1
1516-05-10 or 1516-08-10	1517-05-15	y_u's candidate in H	$y'_u \not\supset y'$	0
before 1516	1515-10-21	y_u's candidate in H	$y'_u \supset y'$	1
after 1516	1515-10-21	y_u's candidate in H	$y'_u \not\supset y'$	0

Table 4.3: Examples of matching of temporal expressions. y is the candidate entity in H for the ambiguous entity y_u. y and y' are different types of entities associated with a certain edge e, which corresponds to the entity information in the metadata m of letter l. y_u and y'_u are different types of entities in the metadata m_u of letter l_u. In this example, y'_u corresponds to the temporal expression in m_u, and y' corresponds to the precise date in H.

Temporal Expression. y'_u denotes a temporal expression regarding the date of writing in the metadata m_u, and y' denotes the corresponding date associated with a candidate edge e in H.

- Given y'_u and y', y'_u equals y', $R(y'_u, y) = 1$.
  For instance, in Table 4.3, the first row, y'_u equals y', such that $R(y'_u, y) = 1$.
- Given y'_u and y', with y' being more fine-grained than y'_u and $y' \subset y'_u$, then $R(y'_u, y) = 1$.
  For instance, in Table 4.3, the third row, we regard y'_u as a date range from the beginning to the end of the year 1516. y' is more fine-grained than y'_u and $y' = \text{1516-05-01} \subset y = [\text{1516-01-01}, \text{1516-12-31}]$. Thus, $R(y'_u, y) = 1$.
- Given y'_u and y', with y'_u being a combination of multiple temporal expressions and $y' \subset y'_u$, such that $R(y'_u, y) = 1$.
  For instance, in Table 4.3, the fifth row, y'_u is a combination of two dates and y' equals one of the dates, such that $y' \subset y'_u, R(y'_u, y) = 1$.

Geographical Expression. y'_u denotes a geographical expression concerning the origin or destination of a letter in the metadata, and y' denotes the corresponding location name associated with a candidate edge e in H.

- Given y'_u and y', y'_u equals y', then $R(y'_u, y) = 1$.
  For instance, in Table 4.4, the first row, y'_u equals y, such that $R(y'_u, y) = 1$.
- Given y'_u and y', with y' being more fine-grained than y'_u and $y' \subset y'_u$, then $R(y'_u, y) = 1$.
  For instance, in Table 4.4, the third row, y' as a city in Germany, is more fine-grained than y'_u and $y' =$ "Wittenberg, Germany" $\subset y'_u =$ "Germany". Thus, $R(y'_u, y) = 1$.
- Given y'_u and y', with y'_u being a combination of multiple geographical expressions and $y' \subset y'_u$, such that $R(y'_u, y) = 1$.
  For instance, in Table 4.4, the fifth row, y'_u is a combination of two locations and y' equals one of the locations, such that $y'_u \supset y', R(y'_u, y) = 1$.

y'_u	y'	y	Relation	$R(y'_u, y)$
Wittenberg, Germany	Wittenberg, Germany	y_u's candidate in H	$y'_u = y'$	1
Paris, France	Wittenberg, Germany	y_u's candidate in H	$y'_u \neq y'$	0
Germany	Wittenberg, Germany	y_u's candidate in H	$y'_u \supset y'$	1
Erfurt, Germany	Wittenberg, Germany	y_u's candidate in H	$y'_u \neq y'$	0
Venice and Meißen	Venice, Italy	y_u's candidate in H	$y'_u \supset y'$	1
Venice and Meißen	Wittenberg, Germany	y_u's candidate in H	$y'_u \not\supset y'$	0

Table 4.4: Examples of matching of geographical expressions. y is the candidate entity in H for the ambiguous entity y_u. y and y' are different types of entities associated with a certain edge e, which corresponds to the entity information in the metadata m of letter l. y_u and y'_u are different types of entities in the metadata m_u of letter l_u. In this example, y'_u corresponds to the geographical expression in m_u, and y' corresponds to the precise location name in H.

Person Name. y'_u denotes a person name belonging to the sender or the recipient of a letter in the metadata, and y' denotes the corresponding name of a correspondent associated with a candidate edge e in H.

- Given y'_u and y', with y'_u equals y', such that $R(y'_u, y) = 1$.
  For instance, in Table 4.5, the first row, y'_u equals y, $R(y'_u, y) = 1$.
- Given y'_u and y', with $y'_u \subset y'$, such that $R(y'_u, y) = 1$.
  For instance, in Table 4.5, the third row, y'_u is part of y', y'_u = Martin $\subset$ y' = Martin Luther. Thus, $R(y'_u, y) = 1$.
- Given y'_u and y', with y'_u being a combination of multiple person names and $y' \subset y'_u$, such that $R(y'_u, y) = 1$.
  For instance, in Table 4.5, the fifth row, y' is a combination of two person names and y'_u equals one of the names, thus, $y'_u \supset y', R(y'_u, y) = 1$.
- Given y'_u and y', with y'_u being an abbreviation of a person's name and $y'_u \subset y'$, such that $R(y'_u, y) = 1$.
  For instance, in Table 4.5, the seventh row, y'_u is an abbreviation and $y'_u \subset y'$, $R(y'_u, y) = 1$.

$R(y'_u, y)$ is 0 when entity y'_u does not co-occur with y in the metadata of a certain letter, or in other words, the probability of observing y'_u, given y, is 0. However, this has the potential to lead to the product of probabilities $P(y'_u|y)$ being 0. In order to avoid this problem, we further smooth $P(y_u \mid y)$ by using Jelinek-Mercer smoothing [222] as follows:

$$P(m_u|y) := \prod_{y'_u \in m_u} \left(\lambda \cdot P(y'_u \mid y) + (1 - \lambda) \cdot \frac{|\{y'_u \mid y'_u \in Y_u\}|}{|\{Y_u \mid typ(Y_u) = typ(y'_u)\}|} \right), \quad (4.24)$$

y'_u	y'	y	Relation	$R(y'_u, y)$
Martin Luther	Martin Luther	y_u's candidate in H	$y'_u = y'$	1
Martin Luther	Georg Spalatin	y_u's candidate in H	$y'_u \neq y'$	0
Martin	Martin Luther	y_u's candidate in H	$y'_u \subset y'$	1
Georg	Martin Luther	y_u's candidate in H	$y'_u \not\subset y'$	0
Martin Luther and Martin Bucer	Martin Luther	y_u's candidate in H	$y'_u \supset y'$	1
Martin Luther and Martin Bucer	Georg Spalatin	y_u's candidate in H	$y'_u \not\supset y'$	0
M. Luther	Martin Luther	y_u's candidate in H	$y'_u \subset y'$	1
G. Spalatin	Martin Luther	y_u's candidate in H	$y'_u \not\subset y'$	0

Table 4.5: Examples of matching of person names. y is the candidate entity in H for the ambiguous entity y_u. y and y' are different types of entities associated with a certain edge e, which corresponds to the entity information in the metadata m of letter l. y_u and y'_u are different types of entities in the metadata m_u of letter l_u. In this example, y'_u corresponds to the person name in m_u, and y' corresponds to the precise person name in H.

where $\lambda \in (0, 1)$ is a parameter that balances the two parts. $\frac{|y'_u|}{|\{Y_u | typ(Y_u)=typ(y'_u)\}|}$ calculates the frequency of y'_u in all entities that have the same type as y'_u in the entity set Y_u.

4.6 Experiments

This section begins with a description of our experiment pipeline, before we turn to present an illustration of the impact of different text-processing techniques applied in the preprocessing of historic letters. This is critical for further experiments such as topic extraction and exploration of person-topic relations. After this, we will present the results of applying our approach (viz., the topic participation score) to historic letters with the aim of analyzing the relationships between correspondents and topics, as well as the correlation between topic trends and historic events. Exploration of the dynamics of topics and the connections between topics and associated individuals can reveal not only latent relationships between letters and individuals, but also latent relationships between individuals. Moreover, we present an evaluation of our data uncertainty approach and show the effectiveness of the approach in terms of refining imprecise entities in the metadata of letters.

We first briefly describe our experimental processing ***pipeline***. Figure 4.1 shows the pipeline of our experiments in this chapter. Our pipeline includes five components, namely data acquisition (input), data cleaning, text pre-processing, network analysis and corresponding output.

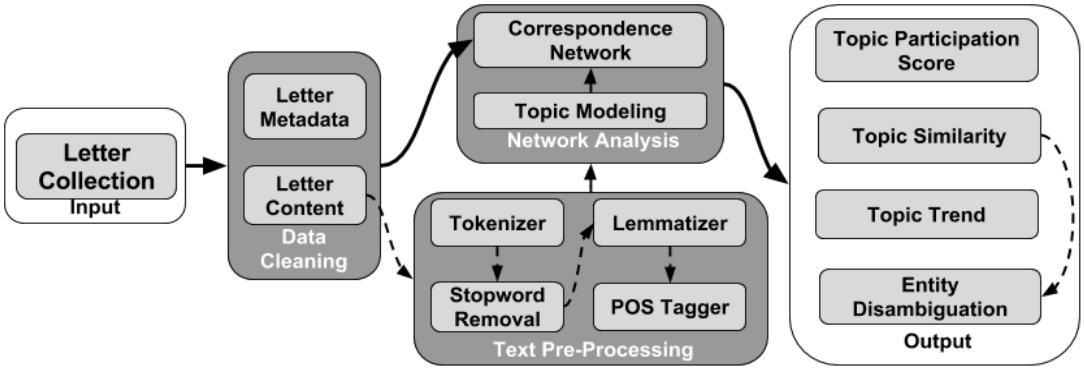


Figure 4.1: The pipeline for experiments in exploration of the correspondence network of our dataset. The two boxes in white corresponds to the input and the output of the pipeline. The three boxes in dark grey represent the different components of the pipeline, and two different curves (dashed and solid curves) are used to represent the different combinations within or between multiple components. For example, the letter metadata and the letter content are both integrated into network analysis, the arrowed curve between these two components, namely network analysis and corpora cleaning, is solid. However, only the letter content is imported into the text pre-processing component, so the arrowed curve between letter content and text-processing is dashed.

The first component in the pipeline is data acquisition (***input***), which is used to read epistolary collections and convert the data in different formats of digitized archives to unified unstructured texts. The second component ***data cleaning*** involves the data splitting and filtering techniques, in order to separate the metadata from letter texts and store them in a database. Then, the third component ***text pre-processing*** is applied to the separated letter texts instead of metadata. This component contains several natural language processing tools such as tokenizer, stopword removal tool, lemmatizer, and Part-of-Speech (POS) tagger. The tokenizer converts texts into tokens and stopword removal tool filters out extremely common words that would provide little information in the further experiments. Lemmatizer and POS Tagger process the tokens and add the lemma and POS features to each token. This component enables us not only to annotate and analyze the texts, but also to provide a comparative basis for further topic extractions.

The fourth component ***network analysis*** in the pipeline combines the results of previously components with topic modeling technique (i.e., LDA), in order to produce the topic-related output (e.g., topic participation score) and refine the uncertain entities (e.g., missing dates or incomplete person names) in the letter metadata. All outputs will be stored in a database for further analyses.

4.6.1 Dataset

Martin Luther (1483–1546) was a German professor of theology, composer, priest, monk and is now heralded as the father of the Protestant Reformation. Luther is the first writer of the classical German language; his translation of the Bible marks the fundamental act of the construction of literary German [223]. In his letters, the major events of his lifetime are reflected, as well as the way he developed in his own religious life. His letters are valuable because of the records of his commitment within the field of Bible translation from Hebrew and ancient Greek into Early New High German.

Our corpus currently contains $2,671$ letters collected by the Theologische Fakultät at Heidelberg University. These letters spanning from 1501 to 1546 are private letters that belong to the early modern theologian Martin Luther (1483–1546). His letters are digitized and transcribed by historians from the Theologische Fakultät at Heidelberg University in full texts and the corresponding letter metadata are recorded in a database. We have divided the documents of hundreds of letters into different texts so that each letter is a separate file including the text of this letter. Most of the letters in the corpus were written in two languages: Latin and Early New High German (ENHG). Latin constitutes approximately 54.2% of the letters in the corpus and Early New High German constitutes approximately 45.8%. The letters often contain elaborate annotations of historians, which do not belong to the original texts, in the opening and closing phrases of letters. We exclude such annotation from content extraction. The following is a sample of a digitized letter written in Latin by Martin Luther at Wittenberg in 1516 (after extraction and annotation deletion). The Arabic numbers that appear in the sample correspond to the order of annotations following the text.

> *Venerabili Patri religiosoque viro, Georgio Leiffer, Eremitae Augustiniano Erfurdiano, Patri suo in Domino.*
> *Ihesus.*
> *Salutem in Domino et paraclito1 eius. Optime Pater et dulcis Frater in Domino, audio Fraternitatem tuam procellis tentatam agitari et variis fluctibus inquietari, sed benedictus Deus pater misericordiarum et Deus totius consolationis2, qui providit tibi optimum, quantum in hominibus potest haberi, paraclitum et consolatorem, R. Patrem Magistrum Bartholomaeum; tantum curae tuae fuerit, sensu et sentimento*

proprio abiecto illius verbis locum dare in corde tuo. Certus enim sum et ex mei et tui experientia doctus, imo et omnium, quos unquam vidi inquietos, scio, quod sola prudentia sensus nostri causa sit et radix universae inquietudinis nostrae. Oculus enim noster nequam est3 valde, et ut de me loquar, hui! in quantis me miseriis vexavit, et usque modo vexat extreme... F. Martinus Lutherus Augustinianus.

4.6.2 Preprocessing

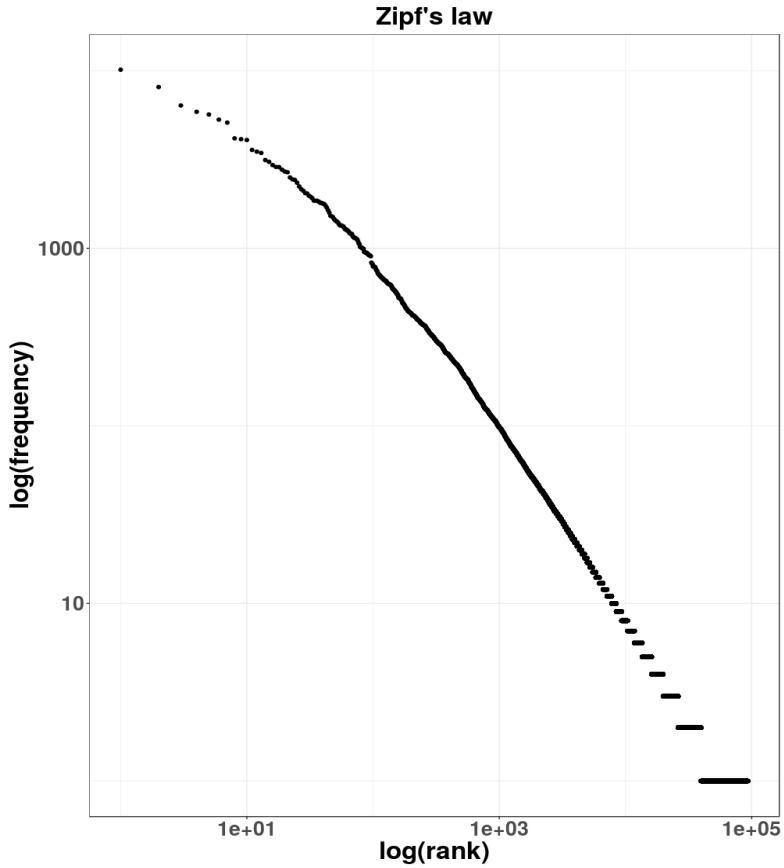
We used all 2,671 letters with each letter constituting a single text in the corpus. Although it is common to apply techniques such as stemming or POS Tagging on texts during the text preprocessing step, we will compare the impact of different processing techniques by following each of the four steps. These preprocessing techniques help to reduce the size of the vocabulary of the corpus and allow the topic modeling techniques to operate on a cleaner dataset instead of raw unprocessed texts.

1. **"Original" Texts**. In this step, we tokenize the texts, lowercase all words, and additionally remove extra whitespaces, numbers and punctuation. As we explore the historic correspondences on the level of texts instead of on sentences, we remove the punctuation contained in the letters. Furthermore, most Arabic numbers in the letters, as we show in the quotation above, correspond to the order of the annotations of the text. Thus, we have also removed the Arabic numbers from the letters. In this way, we produce a vocabulary of 93,367 words, which have occurred a total of 549,523 times in the corpus. In Figure 4.2, we find among the top 20 words in the corpus, most of which are conjunction words, articles, pronouns or prepositions. Furthermore, we can see in Figure 4.2 how these words are distributed across letters and we can find that the word frequency decreases very rapidly with word rank. This distribution is long-tailed and follows Zipf's law [224].

2. **Stopwords Filtering**. Stopwords are words that constitute a large proportion in the corpus but provide little substantial information. Removing these words can significantly reduce the number of parameters that must be matched with and thus decreases the complexity of statistical models [133]. Although there are a few stoplists for Latin available on the Internet, a standard stoplist for ENHG is rarely available or adequate. In order to solve this

rank	word	freq	rank	word	freq
1	und	10162	11	auch	3590
2	vnd	8114	12	mit	3511
3	non	6387	13	sed	3450
4	der	5883	14	ist	3148
5	das	5682	15	wir	3075
6	die	5318	16	cum	2949
7	est	5115	17	daß	2879
8	quod	4170	18	von	2871
9	nicht	4126	19	qui	2776
10	ich	4082	20	dem	2710

(a) The top 20 high frequency words in Martin Luther's letter collection.



(b) the word rank (x-axis) and word frequency (y-axis) of Martin Luther's letter collection in a log-log scale.

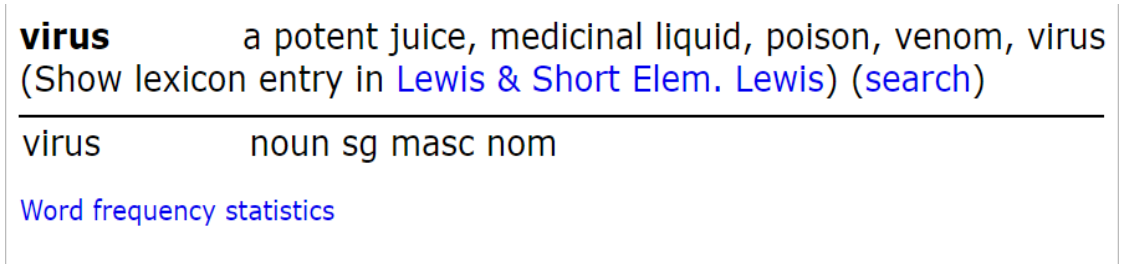
Figure 4.2: A summary of the word distribution in Martin Luther's letter collection.

problem of inadequate resources, we append the existing stoplist for Latin with two kinds of words: first, words that have occurred fewer than two times in the entire corpus; and second, words that consist of either a single character or two letters. The infrequent words, in our case, some of which are errors made in transcription and some of which are foreign words (e.g., words in Greek), are seldom used in the corpus. We are also careful not to remove words that might provide significant information for historic analysis. We have hence created a stopword list of 910 words. Table 4.6 lists a part of it, and we exclude any words that belong to this list in order to reduce the size of the vocabulary in corpus.

Furthermore, we have employed the technique of *tf-idf* weighting here to further reduce the size of the vocabulary and to address the issue of data sparseness. In this experiment, we use the median of *tf-idf* as a cut-off point and we only include words that have a *tf-idf* value larger than the median. In this way, we remove the most common words in the corpus and words with a low frequency. This reduce the number of unique words in the corpus from $93,526$ to $50,411$.

3. **Lemmatization**. Similar to a stemmer, a lemmatizer can reduce reflectional forms and sometimes derivationally related forms of a word to a common base form. The difference between lemmatizers and stemmers is that stemming is a crude heuristic process that chops off the ends of words, while lemmatization is usually more professional and reliable in its use of vocabulary and its morphological analysis of words [225]. However, there is no available ENHG stemmer, and if we use a stemmer for modern German on our letters written in ENHG, it adds an unwanted level of ambiguity to the historic texts. In other words, it tends to produce an output that does not look like words, and can therefore be confusing. Consequently, in our case, instead of using stemmers, we use the following two lemmatizers for Latin and ENHG, separately.

   Latin. Perseus is a website run by Tufts University[1] that offers texts and text-related services. One of these services is Morpheus, which takes a Greek or Latin word as input, and returns a morphological analysis in the form of XML documents.



(a) Lemma output of the Latin word virus by Morpheus in Perseus.

```
vnnd
        +[exlex] und
        +[errid] 54743
        +[xlit] l1=1 lx=1 l1s=vnnd
        +[morph/safe] 0
        +[moot/word] und
        +[moot/tag] KON
        +[moot/lemma] und
```

(b) Lemma output of the ENHG words vnnd by CAB

Figure 4.3: An example of lemma analysis provided by two API services Perseus and CAB.

ENHG. Deutsches Textarchiv (DTA) provides full-text web service "Cascaded Analysis Broker" for error-tolerant linguistic analysis (DTA::CAB)[2]. The CAB web-service provides error-tolerant linguistic analysis for historic German texts, including the normalization of historic orthographic variants to "canonical" modern forms, POS tags, and lemmas.

4. **POS Tagging**. POS taggers annotate the words in a text with Part-of-Speech. By using POS taggers for historic languages (i.e., Latin and ENHG), we can identify and extract words that belong to a certain word class and exclude all other words. In this case, we use two separate POS taggers for Latin and ENHG, and reduce our dataset to only three word classes: noun, verb and adjective.

   Latin. We use Stuttgart Treetagger for Latin POS tagging. Stuggart Treetagger is a tool for annotating texts with POS and lemma information[3]. It

[1] http://www.perseus.tufts.edu/ [Last accessed: February 3, 2017].
[2] http://www.deutschestextarchiv.de/cab/ [Last accessed: February 3, 2017].
[3] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ [Last accessed: February 3, 2017].

has been successfully used to tag 21 languages and is adaptable to other languages if a lexicon and a manually tagged training corpus are available.

ENHG. We still use DTA[2] for ENHG POS tagging.

stopwords
adhic aliqui aliquis ante apud atque aut autem cum cur
deinde dum ego enim ergo est etiam etsi fio haud
hic iam idem igitur ille infra inter interim ipse ita
magis modo mox nam nec necque neque nisi non nos
possum quae quam quare qui quia quicumque quidem
quilibet quis quisnam quisquam quisque quisquis quo

Table 4.6: An example of stopwords excluded from our corpus. There are 910 words in total in our stop-list.

	Vocabulary Sizes	Total Occurrence of Words
orig	93,367	549,523
orig + stop	50,291	113,234
orig + stop + lemma	39,667	75,958
orig + stop + lemma + pos	38,914	71,797

Table 4.7: A brief summary of the dataset preprocessed by different techniques in 4 steps.

Table 4.7 shows the basic statistics of the dataset preprocessed in four steps. *a)* 'orig' indicates letters with tokenization, lowercasing as well as removal of extra whitespaces, numbers and punctuation. *b)* 'stop' indicates letters with additional stopword removal and minimal word length restriction. *c)* 'lemma' indicates letters with additional lemmatization applied. *d)* The label 'pos' is used to indicate the use of POS tagging on letters and we reduce our dataset to only three word classes: noun, verb and adjective.

4.6.3 LDA Experiments

In this section, we employ three metrics (Griffiths [127], CaoJuan [126], and Arun [124]), which have already been introduced in Section 2.6.2, in order to select separately the preferable number of topics for LDA modeling. However, owing to the fact that our dataset is relatively small and most letters are related to a specific area (theology), only the Griffiths' metric provides a helpful and useful

result. Therefore in this section, we only present the results of the Griffiths' metric for the four different preprocessing steps, respectively.

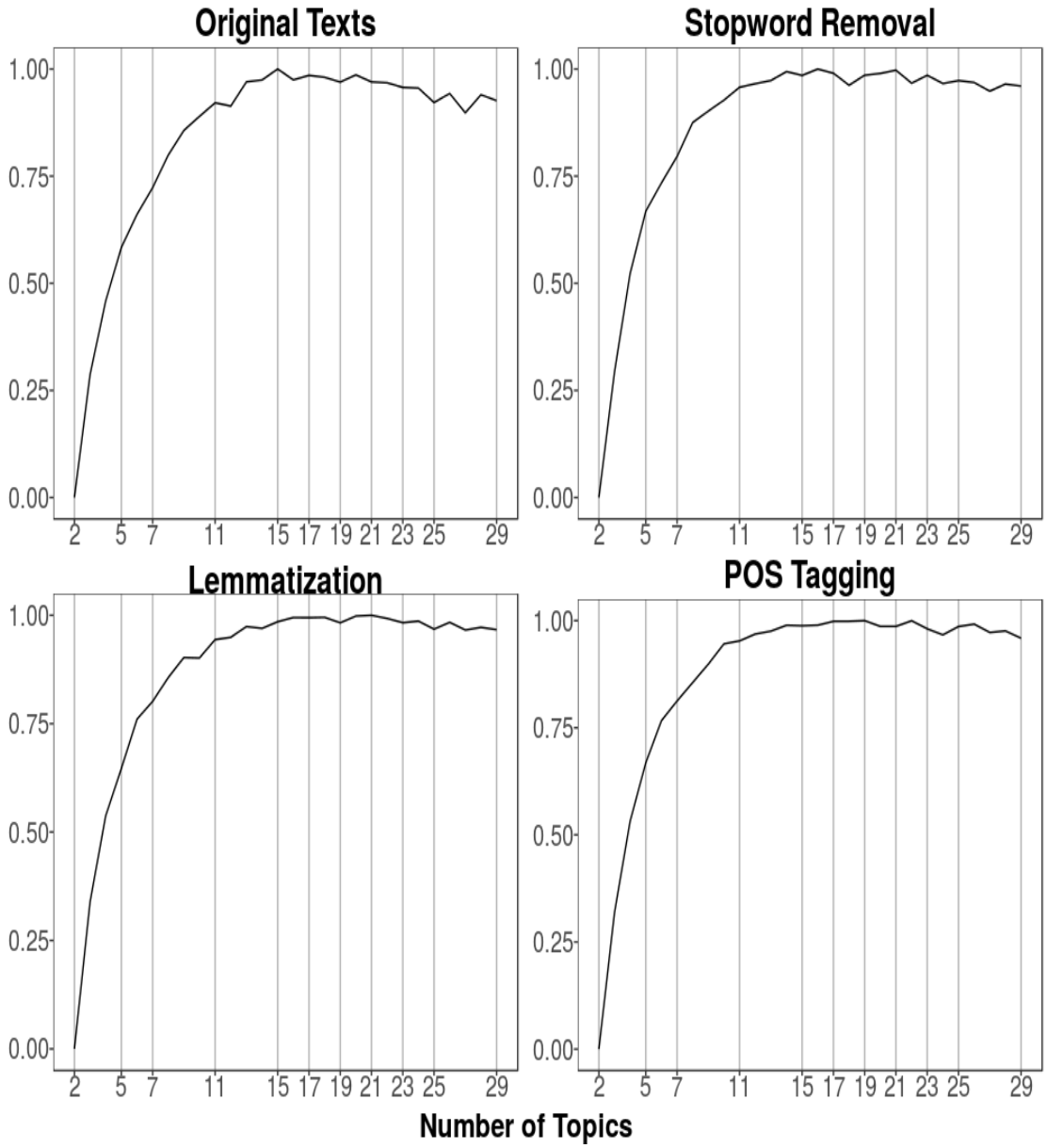


Figure 4.4: Heuristic selection of number of topics in four different preprocessing steps. In each plot, the x-axis corresponds to the number of topics from 2 to 29, and the y-axis corresponds to the different values of log-likelihood. Our objective is to find the number of topics with the maximum log-likelihood in four plots, respectively.

Heuristic Selection of Number of Topics. Table 4.8 shows the number of "desired" topics selected by the Griffiths' metric at the four different preprocessing

steps mentioned above. In each step, the Dirichlet hyperparameters are fixed at $\beta = 0.1$, i.e., the parameter of the Dirichlet for words over topics, and $\alpha = 50/k$, i.e., the parameter of the Dirichlet for topics over texts, for all runs of the LDA algorithm. The value of β is relatively small, so that it can be expected to result in a fine-grained topic decomposition of the corpus. Figure 4.4 shows the estimates of $p(w|k)$ computed for the number of topics from 2 to 29, and we choose the number of topics with the highest log-likelihood value as "desired" ones. For instance, in the "lemmatization" plot, given the fixed hyperparameters and a choice of numbers of topics ranging from 2 to 29, we see that $p(w|k)$ initially increases as function of k, reaches a peak at $k = 21$, and thereafter decreases. Therefore we select 21 as the desired number of topics for the following experiments. This number helps us to generate a topic model that is rich enough to fit the information available in the corpus, yet not so complex as to create noise [127].

Preprocessing Techniques	Desired Number of Topics
orig	15
orig + stop	16
orig + stop + lemma	21
orig + stop + lemma + pos	19

Table 4.8: "Desired" number of topics at each preprocessing step.

Comparison of LDA outputs of four preprocessing steps. In order to compare the effect of different preprocessing techniques, we present the respective results of LDA in Table 4.9. We note that "stopwords filtering" has a significant effect by removing a lot of short common function words (e.g., wir, von, mit, among others). These words do not have significant meanings in the letter, but are frequently used in the corpus. Lemmatization helps us to understand the meaning of the words in each topic more easily and more clearly by normalizing each word to the lemma form. But the effect of POS tagging is not so obvious. Sometimes Stuttgart Treetagger mistakenly interprets an instance of a historic person name as "adjective". Consequently, in this case, we use the output of lemmatization in the following experiments, on the grounds that it reduces the vocabulary size of the corpus and addresses the issue of data sparseness.

4.6.4 Interpreting Topics

In this section, we interpret the meaning of each topic by summarizing the meaning of the top 20 words of the 21 topics in Table 4.10, respectively. The labeling of

Topics	"Original" Texts	Stopwords Filtering	Lemmatization	POS Tagging
1	wir euch von	aur valedominoora viteb	littera census aur	gratiampacem patronus menio
	mit der nit auch	libello quasimodogeniti spei	siquid pentecoste viteb	advenio perfero abominatio
	haben uns doctor	quinta stipendium marchio andrer	nuncius scheda ioanni iubeojubeo	faber eloquentia arbeiten dolus
2	quam nobis nos	fratris augusti regis	kurf legat velum	junkfrau occupatio simon
	atque tua christi	gemain mortui gratiampacemdomino	trier dominatio nirgend	kunig verschreiben iubilo
	christo quae tuae cum	coniugium filiam filia mandato	adam obruo aprilis jsrael	nobilitas foedus official sepes
3	wyr habens myt	libellis mariae pastorem	domina agricola arbor	aprilis guberno exploro
	magister decembris	decembris iohanne istic pastore	afferoattulo ingratus amitto	ordinatio compereocomperio fraternitas
	nycht mane redit sectas	certiores nurumbergensis sathan	phil luth stephanus aristoteles	nicolaus kosen leimbach doctrine
4	ecclesiae christi euangelii	tue satana presertim	bischof buch disputation	possumpotiorpoto conditio percunia
	ecclesia eum deus	vbique vnus comiti commendo	untertan mai universität	iter obruo ionae
	cum anno dei esse	ferdinandus illustriss iona	kais anbeten büchlein geleit	gallus turcam martine mulier
5	fehlt coelum adresse	chf expectamus hir	gratiapax innocens christophoro	antonius pascha cancellarius
	vox domini peccatorum vgl	ignotus quinta syn	aegre quorsum domina	georgii francisco afferoattulo
	phil ludere übergeschrieben	agricola got oculi pastoribus	rediit valeora asinaasinus matthias	palatinus bruck clarissimooptimo diaconus

Table 4.9: An example of LDA experimental output. Each column corresponds to the 10 most probable words of one topic generated at one preprocessing step and each row corresponds to the most likely 5 topics selected in four different preprocessing steps, respectively.

a topic is a subjective task and requires experts, e.g., historians, to interpret correctly the semantic meaning described by the list of most likely words. Some topics contain words with significant meanings and thus could be definitively interpreted by us. Take topic 12 in Table 4.10 as an example. This topic contains words such as "sabbatho", "extorque", "gravo", and "promissum", which are related to Christianity, extort, and hope, respectively. Thus, we label this topic "Christianity, Extort and Promise". Due to our limited historic knowledge and the statistical nature of the topic modeling techniques, not all these topics could be definitively interpreted by us. A few topics (e.g., topic 7) contain words with meanings that are completely unrelated to the meanings of other words. It is not so easy for us to understand and label this topic without the help of the context of the dataset.

We calculate the topic coverage of the whole corpus and find that topic 17 (ca. 7.16%), topic 15 (ca. 6.86%), topic 4 (ca. 5.12%), topic 21 (ca. 5.10%), and topic 18 (ca. 4.95%) are the top 5 topics in the corpus. Topic 17 is the most prominent topic and contains words such as "hochgelehrt", "vniuerisitet", "pfarrer", and "saint", which are more likely to be related to education and clerics. Topic 15 contains words such as "gevatter", "eltern", "frau", "schulen", and "kloster", which are more related to family and education. Topic 4 is related to the theme of disputation and reformation, which is in accordance with the famous academic disputation in 1517 and the subsequent Protestant Reformation in which Luther plays a significant role [226]. Topic 21 is more related to marriage and wealth, and topic 18 is more related to suffering and sacrifice. Since most letters in the corpus were written or received by Martin Luther, the top 5 topics in which Luther participated most are the same as the top 5 topics in the corpus.

We also calculate the topic participation score for each pair of sender-recipient in the correspondence network, in order to examine closely the relation between the estimated topics and the pairs of sender-recipients. We ask at this stage whether there are some similar topics between Luther's friends, Luther's family members, and/or Luther's opponents. With the help of various biographies of Martin Luther, we divide the correspondents who had the most frequent contact with Luther into four categories: friends and colleagues, royal correspondents, family members, and foes.

- **Luther's friends and associates**. Table 4.11 shows the top 3 topics between Luther and eight people who were his colleagues and who had most contacts with him in the correspondence network. These individuals are all Luther's good friends or close associates. In a way different from the topic coverage of the whole corpus, almost all 8 individuals focused on topic 21 and topic 18 in their top 3 topics. Topic 21 is mainly concerned with "marriage, family, and wealth", which could be connected to Martin Luther's marriage in 1525. His marriage to a former nun is quite meaningful, as it sets the seal of approval on clerical marriage [229]. His marriage was supported by many of his friends, e.g., Georg Spalatin, Nikolaus Amsdorf and Justus Jonas, but it was also opposed by some of his friends such as Philipp Melanchthon, who saw it as the downfall of the Reformation [227]. We suppose that Luther might not only invite his friends and associates to the wedding by sending them letters, but also that he could explain the reason for his marriage to others in the letter contents.

  Topic 18 mainly refers to "suffering and sacrifice" and this topic could be related to the Bible and Luther's own life experiences. Luther was a Professor of Bible Studies at the University of Wittenberg at the time when he posted his famous 95 Theses. He spent most of his life translating and refining the translation of Bible into German [230]. This commitment was well expressed in his letters with his colleagues. Not only that, he challenged papal authority by his famous *Ninety-Five Theses* and was excommunicated by Pope Leo. We suppose that the life experience of Martin Luther might teach him something relevant to suffering and sacrifice for his belief, and how he might have shared these thoughts with friends or associates in the form of letters.

- **Royal Correspondents**. Interestingly, 11 people from the royal family appear in the top 30 person-list who had most contacts with Luther in the correspondence network. Table 4.12 shows the top 3 topics between Luther and his 8 royal correspondents, who had most frequent letter contacts with him compared to other royal people. These individuals are electors, dukes, or princes from royal families. These preserved letters not only reflect their interests and attitudes towards Martin Luther, but also in turn represent Luther's respect towards them. In a way different from the topic coverage of the whole corpus, all 8 individuals focused on topic 17 and topic 15 in their

top 3 topics. Their common interests in "education, clerics and family" are not only closely associated with their social status as governors, but are also connected with their religious beliefs in *Lutheranism*. For instance, Johann Friedrich I (Sachsen) (1503–1554), Kurfürst Johann von Sachsen (1468–1532) and Herzog Albrecht von Preußen were all followers of Lutheranism. Friedrich III. (Sachsen) (1463–1525) and Landgraf Philipp von Hessen (1504–1567) both protected Martin Luther from his opponents.

- **Luther's family**. Table 4.13 displays the top 3 topics between Luther and his family members: his wife Katharina Luther, his mother Margarete Luther, and his son Johannes Hänschen Luther. The focus of Luther's family is slightly different from the focus of Luther's colleagues, friends and royal correspondents. It is reasonable that Luther's family members care most about education and family (topic 15). They also focused on the topic of disputation and reformation (topic 4) in their contacts with Luther. Topic 4 reflects a critical point in Luther's life and career: as we mentioned above, in 1517 Martin Luther presented his famous Ninety-Five Theses, and in 1518 he put his views upon Reformation on display during the Heidelberg Disputation. His theses and opinions spread throughout Germany but gained an unanticipated notoriety [231]. In 1521, Pope Leo excommunicated Martin Luther from the Catholic Church. We suppose that this strike might be an enduring painful experience for Luther and that he might intend to clarify his belief in the letters with people from royal families, but also to share his feelings more with his family members.

- **Luther's foes**. Surprisingly, Luther preserved his letters with his opponents, e.g., Johannes Eck and Andreas Karlstadt. Table 4.13 also shows the top 3 topics between Luther and these two foes, respectively. Andreas Karlstadt and Johann Eck were both interested in topic 1 (message, sacramental, patron) in their letters with Luther. This might be related to the Leipzig Debate between these two characters and Luther. *Sacrament* was one of the major divergences between Andreas Karlstadt and Martin Luther. Moreover, their major focuses in their letters with Luther are associated with negative sentiments. For example, topic 18 involves words concerning suffering and topic 12 involves words such as "extorqueo" that are closely associated with Luther's excommunication.

Label	Topic 1 Message, Sacramental, Patron	Topic 2 Lady, Theology, Blasphemy	Topic 3 Colleague, Guards, Escape
1	littera	kurf	domina
2	census	legat	agricola
3	aur	velum	arbor
4	siquid	trier	afferoattulo
5	pentecoste	dominatio	ingratus
6	viteb	nirgend	amitto
7	nuncius	adam	phil
8	scheda	obruo	luth
9	ioanni	aprilis	stephanus
10	iubeojubeo	jsrael	aristoteles
11	purgatorium	kardinal	marchio
12	dormeodormio	landvoigt	occupatio
13	exemplarexemplare	papatu	penuria
14	patronus	simon	evado
15	sacramentalis	conjunx	sämtlich
16	adiuvo	dom	xviii
17	terci	domina	eisleben
18	vendo	michaelem	libetest
19	donatus	angenehm	gratuitus
20	bodenstein	blasphemo	ioachimo

Label	Topic 4 Disputation, Reformation	Topic 5 Lady, Occupation	Topic 6 Education, Migration
1	bischof	gratiapax	stipendium
2	buch	innocens	pascha
3	disputation	christophoro	institutio
4	untertan	aegre	raptim
5	mai	quorsum	juventus
6	universität	domina	concionatores
7	kais	rediit	humanissime
8	anbeten	valeora	mane
9	büchlein	asinaasinus	migro
10	geleit	matthias	pomeranum
11	kapiteln	postridie	praefectus
12	wahl	resipisco	intercedo
13	concili	catharinae	nurmbergae
14	richter	francisci	augustinensis
15	babst	wandel	laetus
16	reformation	weller	octobr
17	aufheben	witt	$\tau\widetilde{\eta}$cf
18	exemplar	durchstrichen	currus
19	sinken	extremumextremus	veter
20	mainz	occupatissimus	ferinus

	Topic 7	Topic 8	Topic 9
Label	Occupation	Excommunication, Justification	Society, Clerics
1	paternitas	doct	irre
2	georgii	dictio	generalis
3	exaudio	moritz	ecclesiastes
4	vittembergae	anathema	influentia
5	archiepiscopus	hebraice	mher
6	gratiapax	leimbach	urbanus
7	multitudo	amicissimo	fidelisincero
8	obruo	dives	laurentius
9	solor	fessus	friderico
10	vesper	dorotheae	diaconus
11	albertus	geonest	francisci
12	occupatio	grammaticus	gratie
13	doct	mons	sapientie
14	stipendium	verax	contentus
15	tristis	costacostum	fastidium
16	equus	elisabeth	occupatissimus
17	calend	hinschius	theologie
18	impar	innocens	arbeiten
19	soe	iubeojubeo	burgerlich
20	amantissimo	iustificatio	laetolaetor

	Topic 10	Topic 11	Topic 12
Label	Government, Superiority	Insanity, Visit	Christianity, Extort, Promise
1	celsitudo	insanio	sabbatho
2	guberno	visitator	iohannis
3	franciscus	aldenburgensi	florenos
4	vbique	assum	extorqueo
5	cancellarius	cursus	gravo
6	clarissimooptimo	valechristo	iter
7	collegium	fera	übergeschrieben
8	deliberatio	francisco	magnifice
9	illustriss	visito	visitator
10	palatinus	xxii	datio
11	venerabilidomino	benedomino	menio
12	principianhalt	bibliabiblium	spiritualiter
13	visito	compereocomperio	exemplaris
14	aboleo	fanaticus	linco
15	adolesco	reminiscor	velum
16	formula	stehlen	iubilo
17	henricus	agricola	andrer
18	preposito	consistorium	victus
19	restitutio	eisleben	promissum
20	vittenberge	libens	spiro

Label	Topic 13 Christian calendar, Friends, Solace	Topic 14 Ceremony, Christ, Leader	Topic 15 Education, Family
1	jonas	venerabilichristo	frau
2	november	aula	hans
3	bucer	pacemchristo	kloster
4	epiphanie	aulicus	gestreng
5	heben	saxonie	schulen
6	rihel	euangeliste	ehrsam
7	transmitto	valechristo	gevatter
8	elisabeth	absolvo	seer
9	oecolampadius	visitatio	türken
10	simon	bulla	fehlen
11	schrecken	hausman	pfarrherr
12	bruck	vaco	eltern
13	ecclesiastes	ionas	edel
14	enders	exemplar	schwager
15	nutzen	spalatine	hallen
16	affinis	cygnee	jude
17	ferdinandus	pacemdomino	fuhrsichtig
18	sick	typus	kette
19	vertrösten	hausmanns	zinsen
20	bucers	supplicatio	son

Label	Topic 16 Royalty, Offering	Topic 17 Education, Clerics	Topic 18 Suffering, Sacrifice
1	valeora	adern	pestis
2	junkfrau	hochgelehrt	conditio
3	venerabilidomino	pfarrer	martine
4	quasimodogeniti	churf	schola
5	irrito	pfarre	enders
6	dominica	johann	baptismumbaptismus
7	gratiampacemdomino	lektion	sacrificium
8	fortuna	ahnen	lipsenses
9	theol	saint	istic
10	conforto	würdigen	passio
11	contemptor	ern	communico
12	kopfen	gulden	rarus
13	kupfern	ken	isthic
14	larua	vniuersitet	auctoritas
15	official	handlung	carlstadio
16	prodeoprodio	torgau	virvirorvirumvirus
17	punio	botschaft	capellanus
18	garen	belangen	infans
19	lücke	georg	peccatorpeccatum
20	nequam	mgr	calculumcalculus

	Topic 19	Topic 20	Topic 21
Label	Turkish, Church	Deception, Clerics, Time	Marriage, Wealth, Family
1	antonius	carlstadii	pauper
2	gallus	spalatine	nuptia
3	turcam	vittenberge	filia
4	vinum	pomerano	mulier
5	turca	cras	lexlexis
6	rusticus	subito	parens
7	vado	numburgensis	gratiampacemdomino
8	lauterbach	prandium	maritus
9	exuro	proora	pecunia
10	december	rector	augustus
11	ferio	signo	coniugium
12	nequitianequities	venerabilidomino	regina
13	redemptio	eleutherius	nuncius
14	altar	expendo	sumptus
15	turcas	canonicumcanonicus	benedictio
16	walch	gestern	abominatio
17	asinus	ruhe	comitium
18	dominicadominicum	verhören	voveo
19	vadovador	aurifaber	hospes
20	pecunia	dolus	iter

Table 4.10: This table contains 20 of the most likely words in the 21-topic decomposition of the Martin Luther's letter collection. We assign each topic with a label as a summary of the meaning of words in that topic. Each column corresponds to one topic and each row corresponds to a word in this topic. Due to our limited historic knowledge and the statistical nature of topic modeling, not all these topics could be definitively interpreted by us.

Luther's Colleagues	1st Topic	2nd Topic	3rd Topic
Georg Spalatin	14	**21**	**18**
Nikolaus Amsdorf	19	**21**	**18**
Justus Jonas	19	**21**	**18**
Philipp Melanchthon	10	**21**	3
Nikolaus Hausmann	14	**21**	**18**
Wenzeslaus Link	**18**	**21**	16
Johann Lang	**18**	**21**	8
Anton Lauterbach	19	**21**	**18**

Table 4.11: Ranked by topic participation scores, this table lists top 3 topics between Luther and his top 8 colleagues, who kept contact with him in the form of letters.

4.6.5 Network Dynamics

In this section, our objective is to find out how these topics changed dynamically and how they are connected to specific individuals. The dynamic topic-person

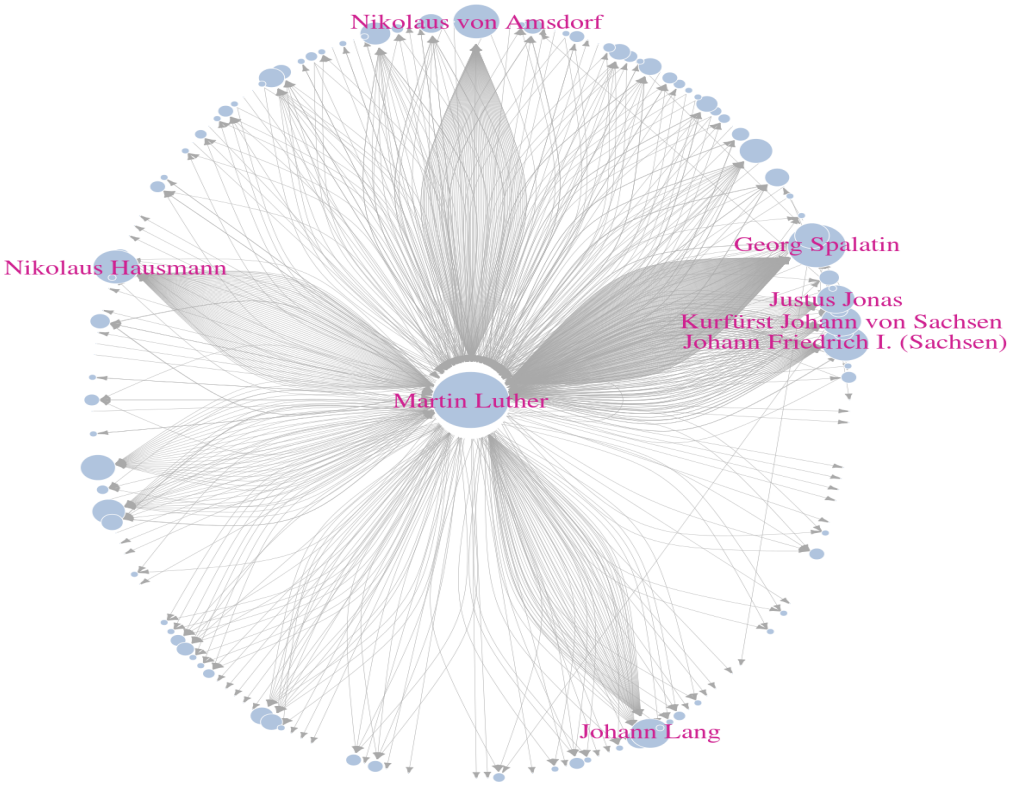


Figure 4.5: A simple example of Luther's correspondence network. The nodes represent people and the multi-edges between nodes represent letters between them. The names of people who had most contacts (letters) with Martin Luther (central node) are labeled in red, and people (nodes) who had only one letter with Luther are removed from this graph. We omit edge attributes in this graph for the sake of simplicity.

Luther's Royal Correspondents	1st Topic	2nd Topic	3rd Topic
Johann Friedrich I (Sachsen)	**17**	**15**	**4**
Kurfürst Johann von Sachsen	**17**	**15**	**4**
Herzog Albrecht von Preußen	**17**	**15**	**4**
Fürst Georg von Anhalt	10	**15**	**17**
Landgraf Philipp von Hessen	**17**	**15**	9
Friedrich III (Sachsen)	**17**	2	**15**
Herzog Georg von Sachsen	**17**	**15**	**4**
Fürst Johann von Anhalt	13	**15**	**17**

Table 4.12: Ranked by topic participation scores, this table lists top 3 topics between Luther and his top 8 royal correspondents, who kept contact with him in the form of letters.

Luther's Families	1st Topic	2nd Topic	3rd Topic
Katharina Luther	**15**	17	**9**
Margarethe Luther	**15**	**4**	10
Johannes Hänschen Luther	21	**4**	**9**
Luther's Foes	**1st Topic**	**2nd Topic**	**3rd Topic**
Johann Eck	18	**1**	9
Andreas Karlstadt	12	**1**	**17**

Table 4.13: Ranked by topic participation scores, this table lists top 3 topics between Luther and his 3 family members and 2 foes, who kept contact with him in the form of letters.

relation would constitute a new window through which to explore and digest the historic correspondence collection. We first trace the trend of each topic and analyze the rising and fading trends for the whole corpus. Then we separately explore the topic participation of a single individual and pairs of sender-recipients over time. In the following experiments, we discover the emerging events and the exchange of ideas by correlations between topics and individuals.

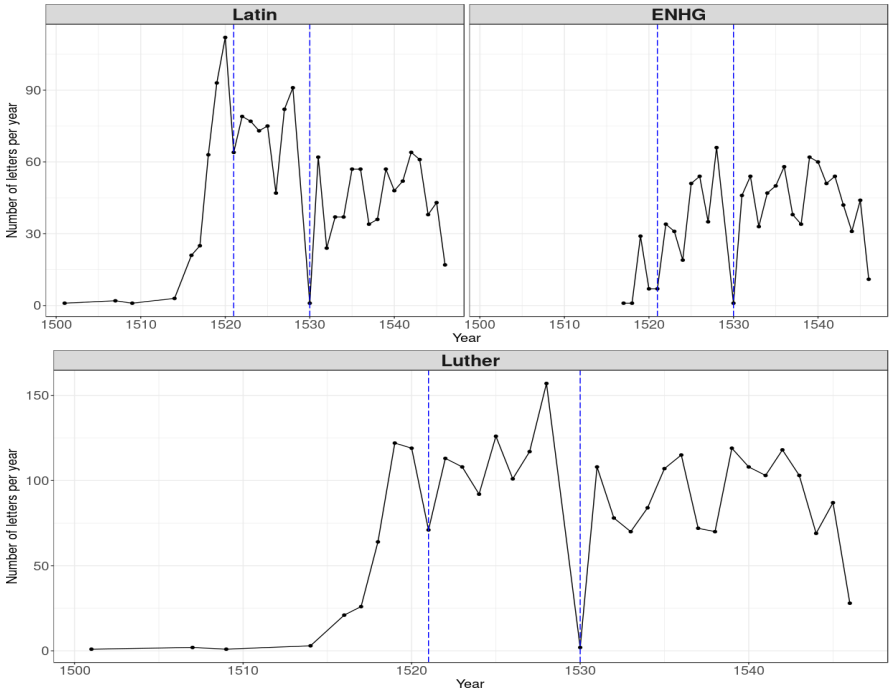


Figure 4.6: Martin Luther's Letter Collection: the upper two plots correspond to the number of letters per year written in Latin (left) and written in Early New High German (right). The lower plot corresponds to the number of letters per year written in both of these languages. The vertical dashed lines on each plot represents the sharp decrease of the number of letters around a certain year.

We divide the 2,671 letters of Luther's collection by year and use the yearly data to analyze the trends of topics and dynamic patterns of person-topic relations. Figure 4.6 shows the number of letters per year from 1501 to 1546. There is a steady increase in the quantity of letters in Luther's early life (1501–1520). However, in the following decade, the number of letters drops off sharply twice in 1521 and 1530. These two points of decline might be due to his excommunication by the Pope and his condemnation as an outlaw by the Emperor in 1521. Luther only kept contact with several best friends and the Emperor in 1521. Because of his identity as a public outlaw, he was absent for the editing of the Augsburg Confession in 1530. This might explain why there is a significant decrease in Luther's output of letters at that time. We also make a distinction between letters written in Latin and Early New High German (ENHG) in Figure 4.6, and it is obvious that letters written in ENHG occurred after 1515, whereas before 1515 the letters collected were all written in Latin. This might be associated with Luther's influence on the development of ENHG. In 1522 Martin Luther translated the New Testament from the original Hebrew into ENHG and continued to work on his translation of the Bible until his death.

Topic trends. A first impression of how the 21 topics changed over time is shown in Figure 4.7 by 21 panels with one trend line each. We use linear fitting to describe the trends of different topics, i.e., how each topic evolved over time. These trends are useful indicators of relative topic popularity in this dataset and offer potential correlations with certain social events. Each plot in Figure 4.7 corresponds to a trend, and we sort these trends by their slopes of linear fitting, from the smallest to the largest value for comparison. Out of 21 trends, 14 trends have negative slopes of linear fitting, and only 7 have positive slopes. According to a predefined threshold and the slopes of linear fitting for each topic trend, topic 15 and 17 are regarded as *rising* topics and shown in Figure 4.8.

These two topics became both relatively popular before 1517, but in around 1520 and 1528, these two topics had an opposite level of popularity. We suppose that Luther is a person who would like to avoid talking about his frustrations in the letters. In 1520, he was condemned as heretical by Pope Leo X. He tried to explain his views to Pope and his Nuncio, but he failed. However, he mentioned topic 15 (education and family) most and topic 17 (education and clerics) much less frequently. On the other hand, in 1528, his daughter died, but he talked

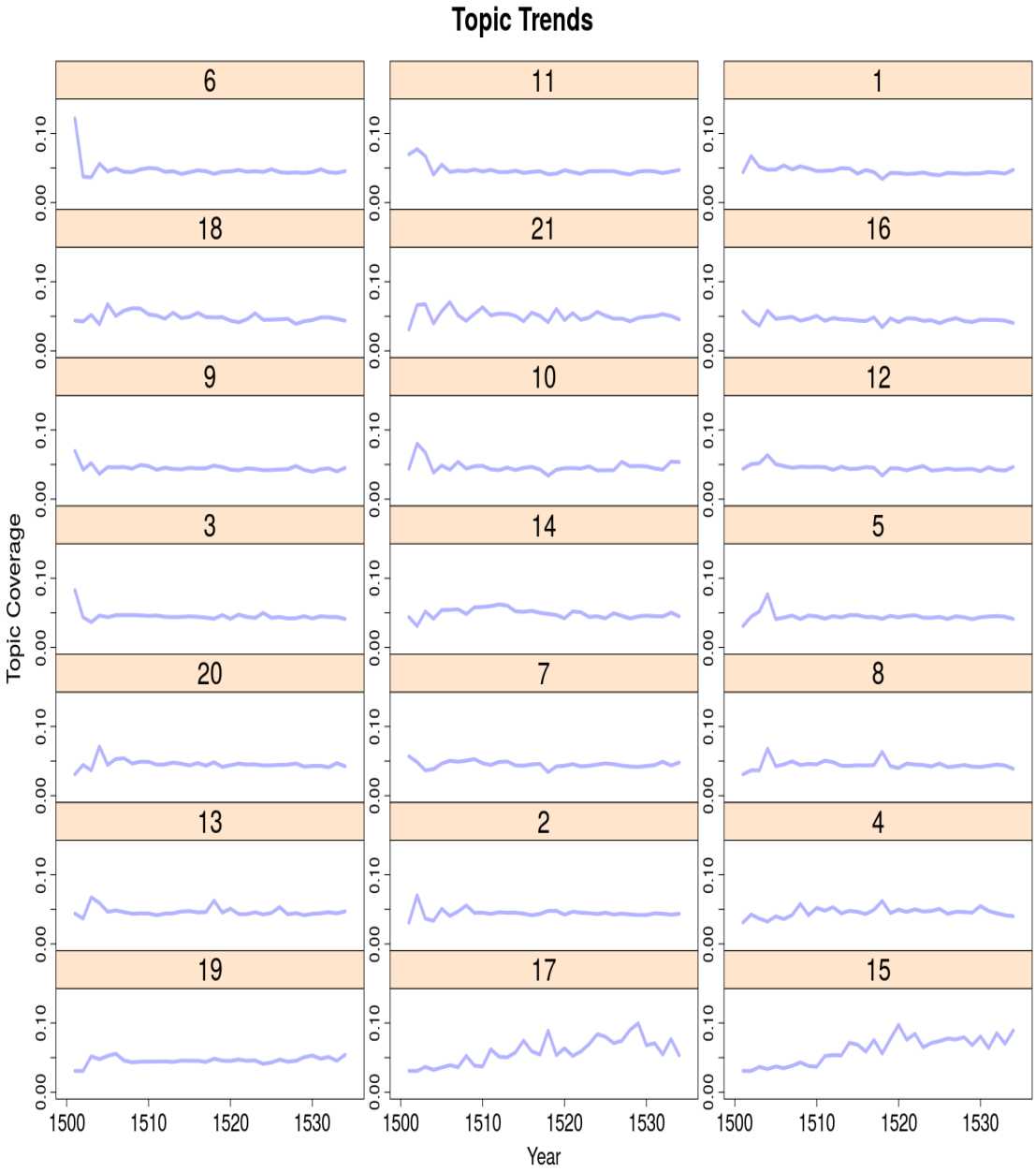


Figure 4.7: A general view of topic trends in Martin Luther's letter collection. These trends are sorted by slope. In this figure, the x-axis represents the ordered years from 1501 to 1546 and the y-axis represents the proportion of each topic within a certain year.

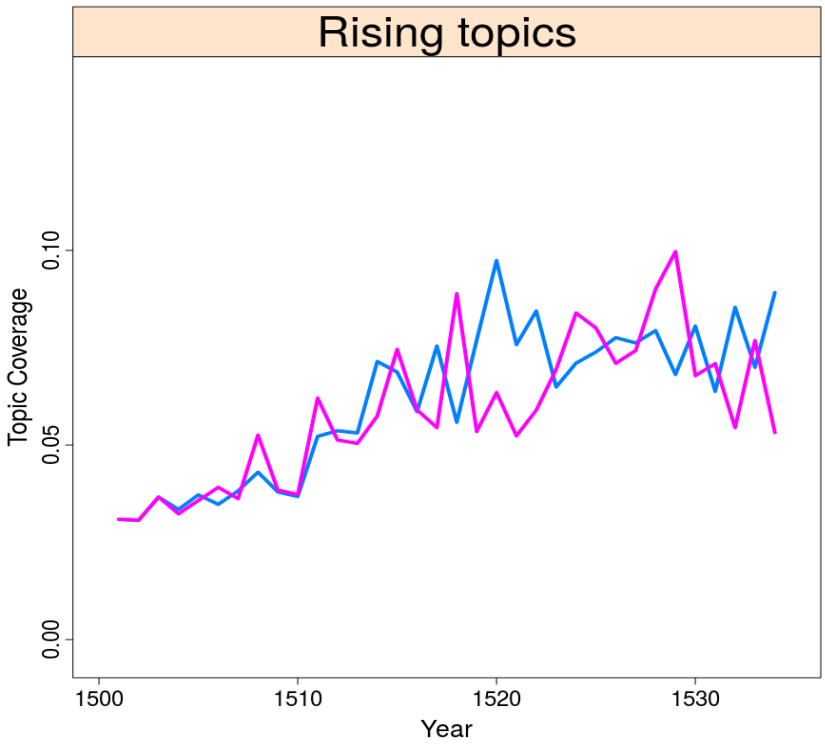


Figure 4.8: The trends of two most popular topics (topic 15 and 17) in Martin Luther's letter collection. In this figure, the x-axis represents the ordered years from 1501 to 1546 and the y-axis represents the proportion of a topic within a certain year. The blue line represents topic 15 and the red line represents topic 17.

about topic 17 most and topic 15 much less often. We suppose that he might be a person who once gets struck by a certain event, then shifts the focus to other events immediately.

Dynamic person-topic relation. We calculate the topic participation scores of Luther and pairs of sender-recipients per year, respectively. In accordance with the topic participation scores in the static correspondence network, topic 15 and topic 17 relevant to education, family and clerics are the most frequent subjects that are both mentioned by Luther and discussed with others in his letters over the years. Moreover, the shifts of Luther from topic to topic could be correlated with specific events in his life. For instance, Luther talked about topic 6 (education and migration) most in 1501, since in this year he entered the University of Erfurt. In 1519, Luther shifted his major focus to topic 4 (disputation and reformation), a result that can be correlated with the famous historic event of the *Leipzig Debate* in 1519. As we have mentioned above, he was invited to this disputation by the two of his opponents, Andreas Karlstadt and Johann Eck. This debate marks a

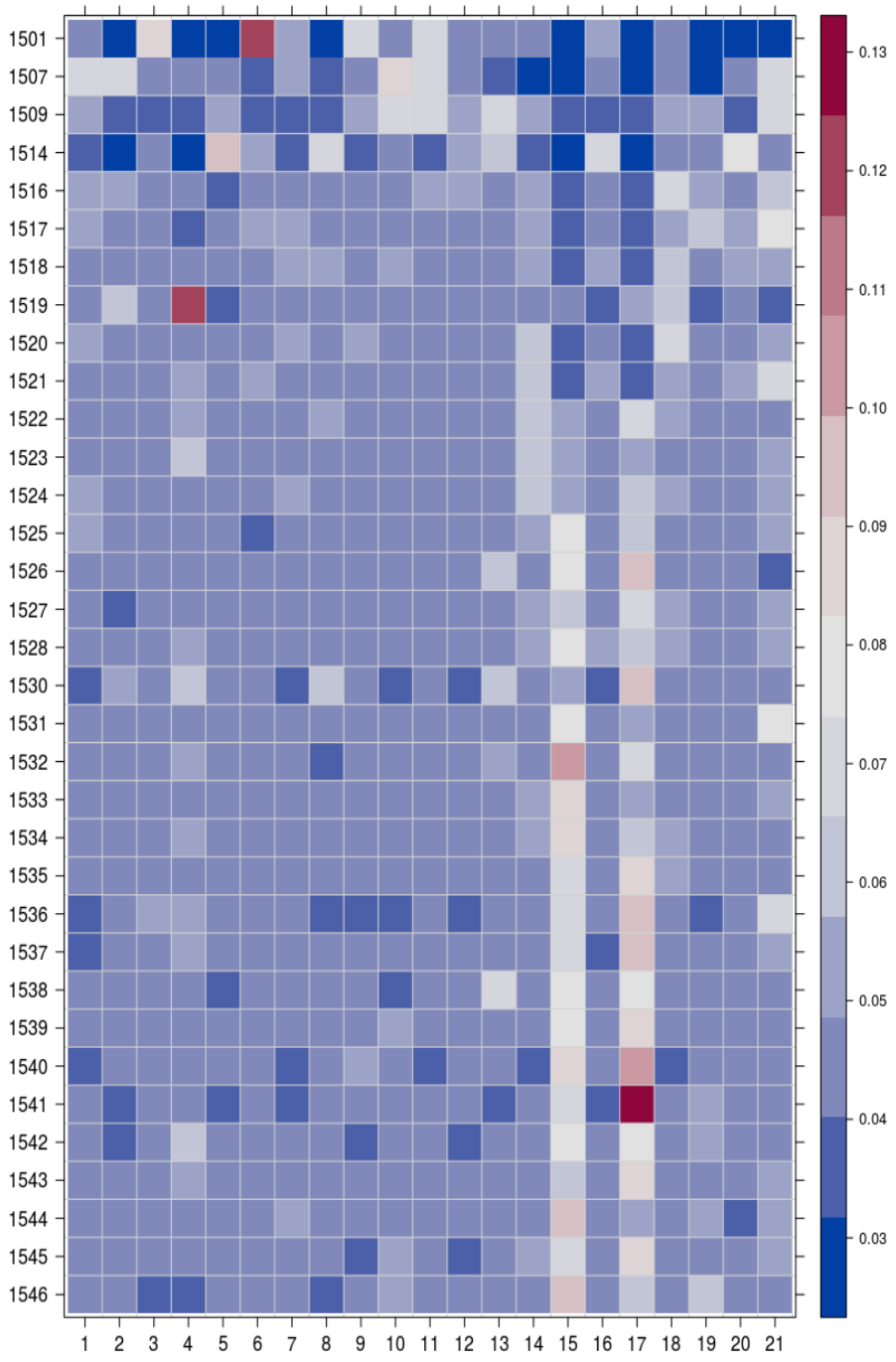
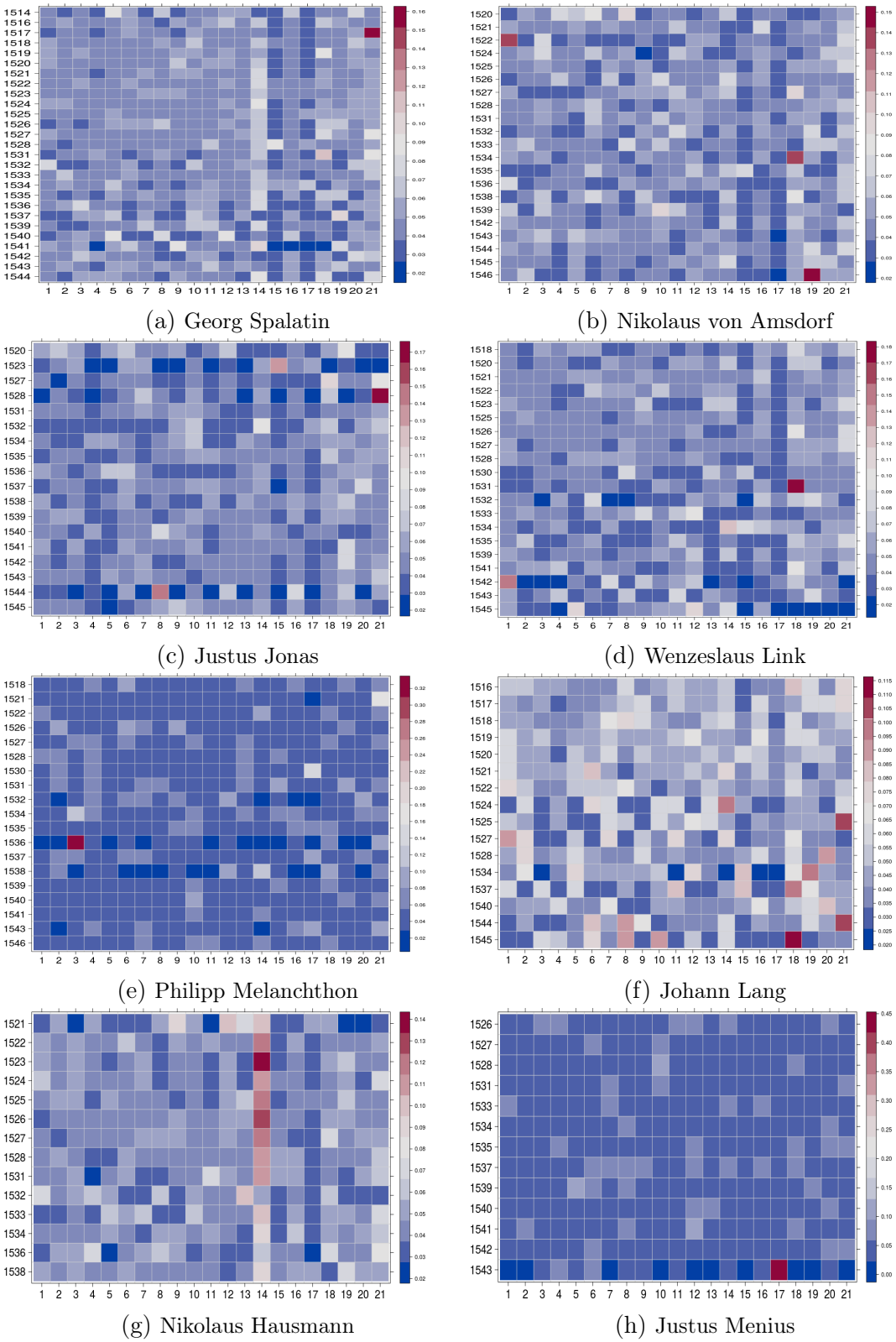


Figure 4.9: The topic participation score of Martin Luther over the years. The horizontal axis represents different topics from 1 to 21, and the vertical axis represents the ordered years from 1501 to 1546. We use different colors to represent the scores of each topic. The redder the color, the higher the score of a certain topic within a certain year.



(a) Georg Spalatin

(b) Nikolaus von Amsdorf

(c) Justus Jonas

(d) Wenzeslaus Link

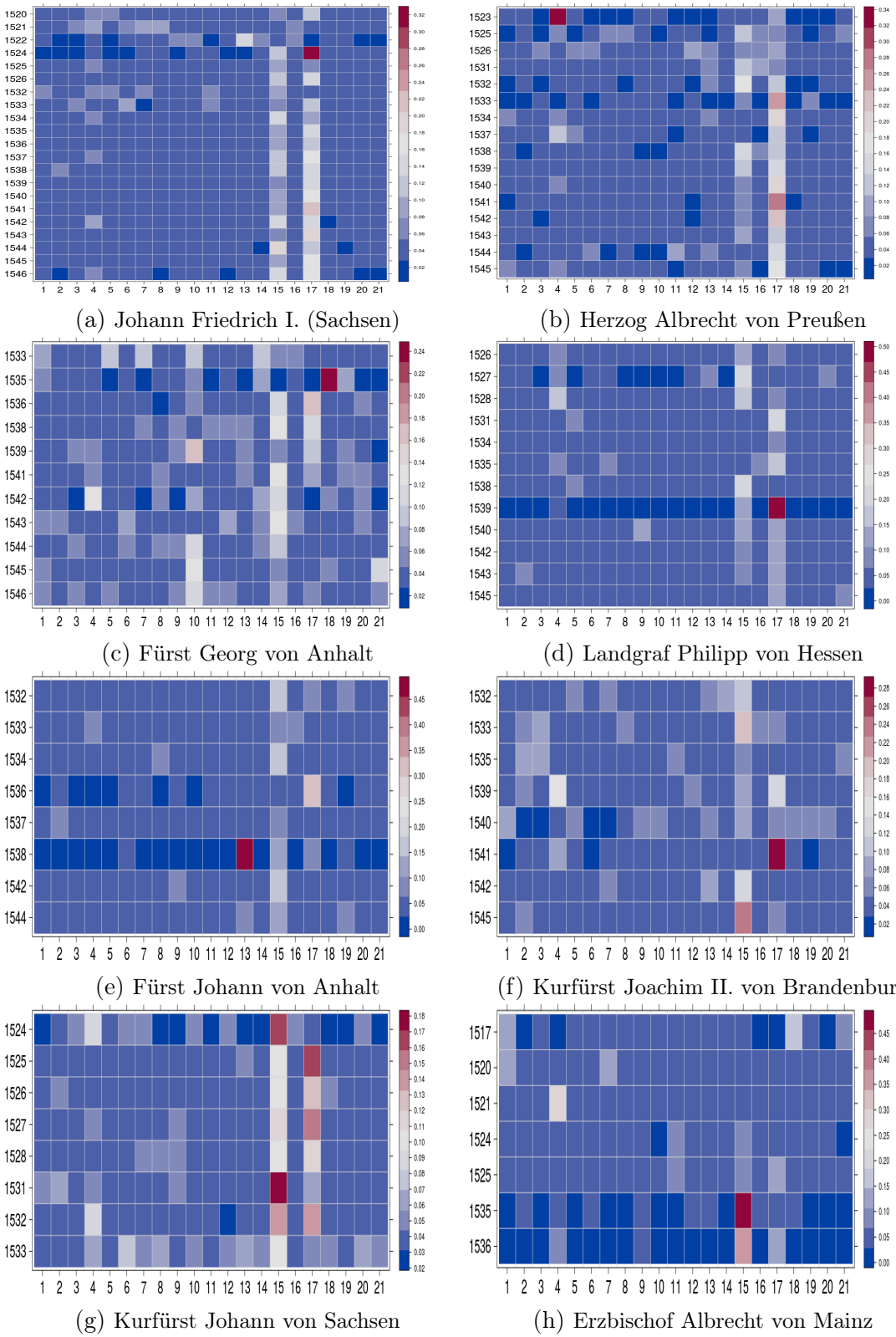
(e) Philipp Melanchthon

(f) Johann Lang

(g) Nikolaus Hausmann

(h) Justus Menius

Figure 4.10: Eight plots of topics between Luther and his eight colleagues or friends who had most frequent contacts with him over the years. The horizontal axis represents the 21 topics in the dataset, and the vertical axis represents the ordered years from 1501 to 1546 that these people had letter exchanged with Luther. The redder the color in the plot, the higher the score of a certain topic within a certain year.



(a) Johann Friedrich I. (Sachsen)

(b) Herzog Albrecht von Preußen

(c) Fürst Georg von Anhalt

(d) Landgraf Philipp von Hessen

(e) Fürst Johann von Anhalt

(f) Kurfürst Joachim II. von Brandenburg

(g) Kurfürst Johann von Sachsen

(h) Erzbischof Albrecht von Mainz

Figure 4.11: Eight plots of topics between Luther and his eight royal correspondents who had most frequent contacts with him over the years. The horizontal axis represents the 21 topics in the dataset, and the vertical axis represents the ordered years from 1501 to 1546 that these people had letter exchanged with Luther. The redder the color in the plot, the higher the score of a certain topic within a certain year.

critical point in Luther's life and career, as to some extent it is the event which led to the further excommunication of Luther by Pope Leo X. In 1541 and 1542, Luther cared most about topic 17, and this can be correlated with the famous event of the *Diet of Regensburg*, a theological debate between the Protestants and the Catholics called in order to restore religious unity [228]. Although Luther did not attend this debate, attendants recorded the conclusion of the debate and sent it to Luther in the form of letters.

Furthermore, the topics to which Luther's friends and royal correspondents paid attention were not stable over the years but changing from time to time, as is shown in Figure 4.10 and Figure 4.11. The evolving interests on topics not only reveal the academic focus between Luther and his correspondents over time, but could also be associated with certain historic events. Figure 4.10 lists the topic participation scores of the top 8 individuals who were Luther's friends or colleagues and kept most frequent contact with Luther over the years. Compared to the scores and the individuals in Table 4.11 above, Anton Lauterbach was replaced by Justus Menius, who exchanged a smaller total number of letters but had more continuous contacts with Luther over the years. We also find that the topic shifts from year to year are closely connected to the significant events in Luther's life.

For instance, in 1522, Nikolaus von Amsdorf cared most about topic 1 (message, sacramental, patron) in the letters with Luther. This is because at that time Luther secretly returned to Wittenberg for a series of theological and social reforms. At this year Luther also finished and published his German translation of the New Testament with the help of his friends and colleagues. Therefore, topic 1 is reasonably associated with the contents of Reformation and the Bible. Take Philipp Melanchthon as another example. In 1536, he corresponded most of all about topic 3 (colleagues, guards, and escape) with Luther, as Luther agreed to the Wittenberg Concord on the Lord's Supper and tried to resolve differences with other reformers. Topic 3 is reasonably correlated with the people and the content involved in the Concord and the Reformation.

Figure 4.10 shows the topic participation scores of the top 8 individuals who were from royal families and kept most frequent contacts with Luther over the years. Compared to the scores and the people in Table 4.12 above, Friedrich III (Sachsen) and Herzog Georg von Sachsen was replaced by Kurfürst Joachim II

von Brandenburg and Erzbischof Albrecht von Mainz with a smaller total number of letters but more continuous contact. In a similar way to the examples above, we can also find the connection between topics and the life events of Luther. For instance, Landgraf Philipp von Hessen, in 1539, focused on Topic 17 (education and clerics) most of all. Philipp von Hessen dealt with the divergence between Roman Catholics and Protestants at that time, which is why he paid attention to (theology) education and wrote to Luther for his advice. Consequently, based on the topic participation scores of 16 individuals over the years in the dataset, we illustrate the effectiveness of the measurement *topic participation scores* in trend analysis, exploration of person-topic relations, and correlations between topics and historic events.

4.6.6 Data Uncertainty

In order to evaluate the effectiveness of our approach to disambiguating uncertain data, we begin with the settings of experiments, before turning to evaluate the effectiveness of our approach and study the impact of parameters upon the performance of our probabilistic framework.

Experimental Settings. Considering that the benchmark dataset for the disambiguation of uncertain entities in the metadata of historic letters is not publicly available, we create a gold standard dataset for this experiment. As the accessible letter collections with definite metadata are limited, our dataset in the following experiments contains 110 letters of Martin Luther, which have been chosen on the basis of the sender(s), recipient(s), and the number of letters between them. We also check the metadata of these letters manually in order to ensure that each entity is definite. The metadata of these letters consist of 14 person names, 22 locations and 101 dates.

As letters in our dataset were all sent by Martin Luther and were mostly sent from Wittenberg, Germany, it is not very meaningful for us to predict or disambiguate of the sender and the origin in this experiment. Instead, missing, incomplete or ambiguous values in entities such as recipients, destinations and dates are taken into consideration. In order to obtain an unbiased estimation of the performance of our approach on the dataset, we use a 10-fold cross validation to split the

dataset into 10 equal partitions. Our dataset then produces a test set of 11 letters and a training set of 99 letters. This process is duly repeated 9 times. For each iteration, every letter falls either into the training set or into the testing set, but not both. In other words, each letter is used for both training and testing, and each letter is in the testing set exactly once.

We manually annotate 11 letters with more coarse-grained entities or missing values in the metadata as the test set for each iteration. For instance, for person names, we remove the first or the last name; for dates of writing, we replace the dates with the corresponding year; for locations, we move up in the spatial hierarchy, or in other words, we replace a city or town name by the corresponding country, respectively. In Table 4.14, we show the types of entities in the test set that are measured in each iteration of our 10-fold cross validation.

	Recipient	Destination	Date
Missing	6	6	6
Ambiguous	5	5	5
Total	11	11	11

Table 4.14: A description of our test set for each iteration of 10-fold cross validation. We list three types of entities, namely recipient, destination, and date. We manually annotate 11 instances for each type of entity as a combination of the missing or ambiguous situations.

Evaluation Measure. Considering that no previous work deals with the evaluation of uncertain data in historic correspondences, we calculate the similarity between different letters using topic distributions and adapted symmetric KLD (cf. Section 2.7.4) distance measure as a comparison to our approach. Based on the topic distribution for each letter and the distance between two letters calculated as similarity scores and ranked in descending order, we construct a list of candidate letters with a sequence of candidate entities for each missing or ambiguous entities in the test set, respectively. These candidate letters help us to filter out the irrelevant letters in the dataset in order to improve the efficiency of our approach.

Then we apply our probabilistic approach to these candidate letters and evaluate the performance of our approach by using the accuracy, which is a solid evaluation measure used in most studies of entity disambiguation in the area of information retrieval [166]. We locally aggregate the correctly predicted or refined results of

entities of a certain type, and then calculate an average over all the entities within this type for 10 iterations. Moreover, the accuracy of topic similarity is calculated as the mean of the number of correctly predicted or refined entities, divided by the total number of the entities that belong to a specific type for 10 iterations. In addition, we also take the output of $P(y)$ (cf. Section 4.5.2), which calculates the occurrence probability of each candidate entity, as a straightforward probabilistic approach for comparison. The accuracy of this output is evaluated using the same method as for the topic similarity.

Results. For each iteration of 10-fold cross validation, we construct a correspondence network based on the training set of 99 letters. The candidate letters are chosen from the network based on their topic similarity with each letter in the test set, respectively. In each iteration process, the parameter λ (cf. Formula 4.24) varies from 0.1 to 0.9 with an increment of 0,1, and the probability $P(y_u \mid y)$ (cf. Section 4.5.2) is set to 0.5 in all experiments. In each iteration process, for each value of λ, only the top 15 ranked candidate letters in the constructed network are included in the final training dataset. This resulted in 165 outputs in total (15 candidates for each of the 11 letters with uncertain entities).

In Figure 4.12, we show the performance of our approach within the process of 10 cross-fold validation and varied parameter λ from 0.1 to 0.9 concerning different types of uncertain entities, respectively. We find that the performance of our approach is sensitive to a smaller λ varied from 0.1 to 0.5, but not very sensitive to a larger λ from 0.6 to 0.9. In other words, a smaller λ reflects that the occurrences of this entity play a more important role in the validation test set, whilst on the contrary, a higher λ reflects a more important impact of our probabilistic approach on the test set. Figure 4.12 also shows that λ in the range of 0.6 to 0.9 yields the better accuracy on average, and therefore we use 0.6 as the default value for the further analysis in our experiments.

	Recipient	Destination	Date
TS	8.26%	7.71%	38.84%
PAC	16.36%	12.72%	32,73%
PA	**51.66%**	**56.43%**	**60.64%**

Table 4.15: Accuracy for different types of entities (recipient, destination and date) achieved by 10-fold cross validations using different approaches. TS represents topic similarity approach, PAC represents occurrence probability P(y), and PA represents our probabilistic approach.

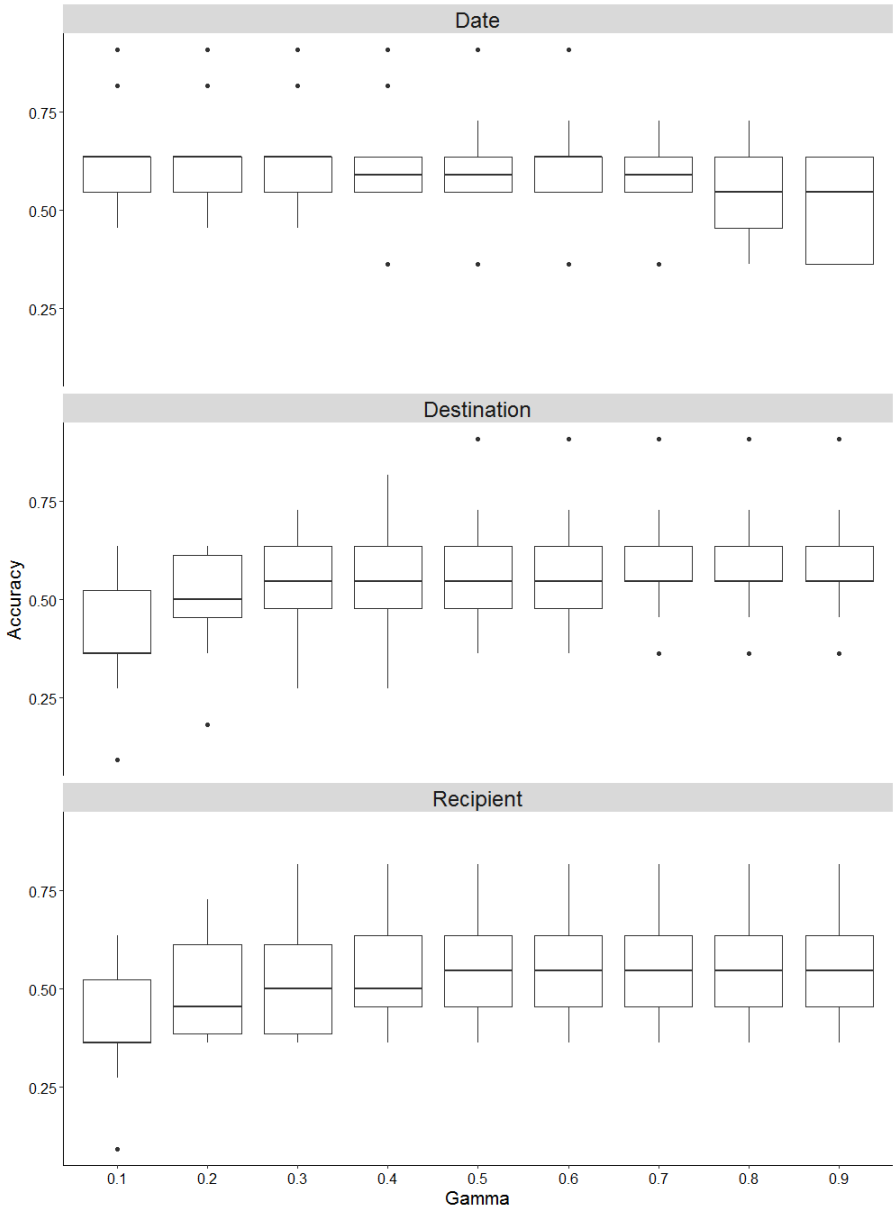


Figure 4.12: Boxplot visualization of the accuracy of our probabilistic approach with the parameter λ varying from 0.1 to 0.9 in each iteration. The x-axis corresponds to the parameter λ and y-axis corresponds to the respective accuracy of 10 iterations for each λ. We evaluate our approach on three types of entities: the temporal expressions (dates), geographical expressions (destinations), and the person name expressions (recipients).

In order to analyze the effectiveness of our approach to different types of entities, we show the corresponding accuracy and the comparison with the topic similarity approach and the occurrence probability approach in Table 4.15. Although the performance of our approach is not extremely high, it still greatly outperforms the other two approaches concerning all three types of entities. It achieves over 50% accuracy for all three types of entities and shows the advantage of our probabilistic framework. If we can get access to the repositories in which letters were written by multiple individuals instead one single person, we believe that the accuracy of our approach will be significantly improved.

4.7 Summary of the Chapter

Historic correspondences establish and deepen relationships between correspondents by exchanging information and creating knowledge. Although researchers have been highly aware of the importance of historic correspondences for some time, most related studies still focus on the visualization of the metadata and the literary study of the content of the letters instead of the models and explorations in combination of these two. Network visualizations have been frequently employed in correspondence network research, yet the embedded relationships between correspondents and other entities (e.g., topics) are seldom explored, not to mention the remaining data uncertainty issue in the letters.

In this chapter, we aim to analyze historic correspondences in the form of social networks on the basis of letter metadata and contents. We first extended our correspondence network with ***content*** information and developed measurements such as ***topic participation scores*** in order to quantify the relationships between correspondents and topics. Our measurement not only allows the integration of network structures, but also provides an intuitive way to explore topic-person relation. This measurement is not limited to deal with the specific author-topic or author-recipient-topic relation, a variety of relations are also taken into consideration, e.g., sender-topic, sender-recipient-topic, multiple senders, and multiple sender-recipients. Not only that, we extended this measurement to explore the ***trends of topics*** in terms of the corpus and individuals over time. We analyzed the rising and fading trends of topics over time and correlated these trends with significant historic events.

As a further contribution of this chapter, we proposed a probabilistic framework for the refinement of the ***uncertain entities*** in the letter metadata. We think that uncertain entities in the letter metadata can be inferred or implied from the other entities in the metadata of the corresponding letter. We thus developed a novel probabilistic approach in combination with the network structures, and co-occurrences of entities in the letter metadata to handle this issue.

We then applied our approaches to the empirical dataset, in order not only to analyze the relationships between correspondents and topics embedded in letters dynamically, but also to conduct experiments in a bid to highlight the advantages of our data uncertainty approach compared to other methods. The experiments following this stage all indicated the ***applicability and effectiveness*** of our approaches: our topic participation scores successfully connect the topics with specific historic events or historic persons. Our data uncertainty approach is also significantly better in terms of accuracy than the other two methods we have used for comparison.

In contrast with previous research on correspondence networks, our comprehensive network model combines the metadata and the content information of letters into a formal network structures. Furthermore, in comparison with previous research on data uncertainty in digital humanities, we develop a novel probabilistic framework in combination with the network structures, entity co-occurrences in the letter metadata, and the similarity between letter contents.

For future work, the extraction of entities such as person names, location names and dates embedded in letter contents using the combination of named entity recognizer and manual annotation is a strong desideratum, in order that more network measurements can be developed or applied to the historic correspondences. This will in turn lead to more precise analysis of correlation between historic persons, individual relationships, and historic events. The topic participation scores will be further extended to measure the influence of individuals, while conveying their topics to others in their letters. Furthermore, our probabilistic framework for the refinement of uncertain entities will be further improved by network measurements and evaluated on large-scale datasets.

Letters are among the most significant
memorial a person can leave behind them.
— Johann Wolfgang von Goethe

CHAPTER 5
CONCLUSIONS AND FUTURE WORK

Hundreds of years ago, before Facebook or Twitter were introduced, there were correspondences, in the form of letters, linking different individuals across national or intercontinental borders. Colleagues discussed their work and opinions through letters, friends shared interests and activities through letters, and family members conveyed their feelings and experiences to each other through letters. Not only that, historic correspondence also served a purpose that was much greater than the mere exchange of information. It also functioned as "an instrument of cultural exchange and transmission" [232]. Historic correspondences created networks to disseminate information for different social classes; it shaped various communities, and it transmitted political, scientific, social, and even commercial information and opinions among people in their daily lives.

Although an increasing number of projects work on the digitization of historical correspondences [83, 84, 85], the full editions of letters or the free access to large amounts of online letters are still not available. For instance, in the British Library, only around five percent of the whole paper archives have been digitized [1]. Access to a significant online letter resource might cost substantial amount of euros. Therefore, considering the size and the access of publicly available data, personal correspondence has been one of the most widely used sources in the historical correspondence research. Since such correspondences ordinarily took place between individuals and were not intended for public exposure, it may provide a clearer understanding of the characteristics of the author and the recipient of a letter. Their letters are often specifically focused and sometimes

address issues that are covered in earlier letters. These letters offer insights into the opinions or experiences of a person, the observation of him or her on social or academic events, and the message he or she chooses to convey to the recipient informally or persuasively. Not only that, personal correspondence can be official when letters are used to convey the messages of an agency or a government authority. Official letters are more formal in their language and tone, and may provide different topics related to individual writers from informal letters.

With the availability of personal correspondences, most historic correspondence studies focus on the visualization of individual correspondence networks and the interpretations accordingly. Few of them pay attention to the formal modeling and approaches of correspondence networks, not to mention the data uncertainty issue such as missing person names, ambiguous location names and incomplete dates that exists in historic correspondences. Different from the previous related research, the objective of our interdisciplinary research, spanning Computer Science, Linguistics, and History, is to reassemble and interpret the correspondence networks about people and their times comprehensively. Based on personal letter collections, we reposition historic persons within their own certain time periods, and explore the network of his/her correspondences involving his or her friends, family, colleagues, enemies, and other categories of relations.

Particularly in this dissertation, we presented our correspondence network models and corresponding approaches to explore the latent relations embedded in the networks. In this final chapter of the dissertation, we first present a summary of the key aspects of our work in Section 5.1. Then, in Section 5.2, we discuss a range of open issues and provide an outlook for future research.

5.1 Summary

We began this dissertation with an introduction of the motivation, major challenges, and our main contributions in Chapter 1. In Chapter 2 we presented the background and related work, which laid the foundations for the rest of the dissertation. To name but a few of these foundations, we demonstrated the basic concepts of graph principles and topic modeling techniques, as well as an overview of key studies in historic correspondence networks. We pointed out that the most

important prerequisite for a comprehensive correspondence network analysis is a formal modeling.

For this reason, in Chapter 3, we proposed a ***correspondence network model***. We develop a comprehensive model that integrates both the metadata and the content information of historic letters into a hypergraph representation. We provide ***valuable and in-depth insights*** into the multiple relations embedded in the historical correspondences and to measure these relations extensively from both static and dynamic points of view. This model constitutes our most important contribution and basis for the problems, extensions, measurements and evaluations in the whole dissertation. As a second contribution of this dissertation, we developed specific measurements for specific types of personal relationships, namely ***local reciprocity*** and ***global reciprocity*** based on our correspondence network model.

Furthermore, in order to measure the evolution of the network and the contact patterns of individuals over time, we introduced two concepts, namely contact sequences and graphlets, in order to observe the network from both the local and the global points of view. Then, we constructed networks on the empirical datasets. The results obtained not only indicate that ***meaningful and interpretable relationships*** between correspondents can be extracted and analyzed using our correspondence networks, but also provide new insights on dynamic patterns embedded in correspondence networks.

The other major contributions of this dissertation were presented in Chapter 4, in which we extended our correspondence model with topics extracted from letter contents. We developed ***topic participation scores*** to measure individual-specific (author-only or sender-recipient pair) topics in combination with network structures and topic modeling techniques. Furthermore, we measured the ***rising*** and the ***fading trends*** of topic over time using topic participation scores, in order to provide new insights into the correlations between topic trends and the social or biographical events of historic persons.

For the last major contribution of this dissertation, we proposed a novel ***probabilistic framework*** for the refinement of uncertain data such as ambiguous person names or missing dates in the metadata of historic letters. We developed

a novel probabilistic framework in combination with topic similarity, network structures, and entity co-occurrences in the letter metadata to refine the uncertain entities. We then applied our measurements to empirical datasets in order to evaluate the effectiveness of the models. The results obtained indicate not only the correlation between topic trends and specific events of historic persons, but also the ***high quality and effectiveness*** of our statistical approaches.

5.2 Future Work

In the previous chapters, we have developed models and corresponding measurements for correspondence networks based on the letter metadata and letter contents. We have also demonstrated the quality and usefulness of our approaches. However, there are still some interesting points in which our work could be extended, and we briefly introduce such issues for further work in the following paragraphs.

- **Linked Data**. The first explorative subject for future work will be to discover further the relationships among correspondents, organizations and social events with the help of external sources. We have already stored external sources, such as personal profile and organization information extracted from Wikipedia, in our databases, and these sources can be further integrated into the network modeling. Such a model extension will not only help to investigate more embedded relations, but also motivate measurements on the multi-type edge weights and community detections in the correspondence networks.

- **Regional Influence**. The second subject of required further exploration will be the measurement of regional influence of certain correspondents. Although our network model integrates geographical information (namely origins and destinations of letters), due to the limitation of available data sources, we have not investigated the latent relations between locations and correspondents in our current study. It is our hypothesis that a letter, which contains significant information, is conveyed from one person to another over space and thus involves possibly connections with historic events. This influence

of letters and persons over space and time will be further explored once we can obtain enough data with sufficient and detailed location information.

- **Entities in the Letter Contents**. The third point will be the extraction and exploration of entities embedded in letter contents. There are many named entities such as person names, location names and dates embedded in letter contents, and these entities can provide us with more precise information about personal relationships and historic events. However, due to the significant differences between historic and modern languages, named entity recognizer (NER) for modern languages cannot be applied to historic texts directly. Fortunately, researchers currently also pay attention to the construction of online dictionaries and annotated corpora of historic correspondences [14, 82]. Once we obtain the access to these dictionaries and corpora, we can extract the named entities from texts and integrate them into the correspondence networks. It is our hypothesis that the combination of entities from texts and metadata will assist us with tasks such as data uncertainty and event detection.

- **Topic Sequences**. The fourth subject of exploration will be the further extension of topic participation scores. Each person in our correspondence network is associated with topic distributions, namely topic participation scores. With these scores we can find the major topics embedded in each letter of each individual or pairs of correspondents, respectively. With the help of temporal information (date of writing), we can assign each person with a time-ordered sequence of major topics in each of his or her letter. It is our hypothesis that people involved in the same region, organization, or event are more likely to share similar topic sequences. In addition, the topic sequences can be further used as features for clustering of nodes in the correspondence networks. Since most correspondence networks are star-structured, this sequence-based topic clustering approach can help to discover the latent communities in the networks.

- **Data Uncertainty**. The fifth potential subject of exploration will be the further improvement of the data uncertainty approach. In Chapter 4, we developed a probabilistic framework for the refinement of the uncertain entities in the letter metadata and provided good evaluation results. However,

one could try to improve further the precision of the disambiguation by more elaborate network measurements into the probability estimation. Besides, we can further evaluate the models when applied to large-scale datasets. More datasets from various sources can provide more detailed evaluations of the reliability and flexibility of the models.

BIBLIOGRAPHY

[1] P. Claus. *History: An Introduction to Theory, Method and Practice*. Routledge, London and New York. 2012.

[2] B. Lightman. *Correspondence Networks*. A Companion to the History of Science. John Wiley & Sons, Hoboken. 2016.

[3] Circulation of Knowledge and Learned Practices in the 17th-century Dutch Republic. http://ckcc.huygens.knaw.nl/ [Last accessed: November 7, 2015].

[4] G. D. L. Camiciotti, D. Pallotti. *Letter Writing in Early Modern Culture, 1500–1750*. Journal of Early Modern Studies. 3: 7–8. 2014.

[5] Extensible Markup Language (XML). https://www.w3.org/XML/ [Last accessed: July 8, 2016].

[6] Text Encoding Initiative (TEI) Correspondence SIG. http://www.tei-c.org/Activities/SIG/Correspondence/ [Last accessed: December 14, 2016].

[7] Early Modern Letters Online. http://emlo.bodleian.ox.ac.uk/ [Last accessed: November 7, 2015].

[8] Centre for Editing Lives and Letters. http://www.livesandletters.ac.uk/ [Last accessed: November 7, 2015].

[9] Darwin Correspondence Project. http://www.darwinproject.ac.uk/ [Last accessed: November 7, 2015].

[10] A. Bergs, L. Brinton. *English Historical Linguistics– An International Handbook (Volume 2)*. De Gruyter Mouton, Berlin. 2012.

[11] T. Nevalainen, H. Raumolin-Brunberg. *Sociolinguistics and Language History – Studies based on the Corpus of Early English Correspondence*. Brill/Rodopi. Amsterdam. 1996.

[12] D. R. van Voorhis. *A Prophet of Interior Lutheranism: the Correspondence of Johann Arndt*. PhD Dissertation. University of St Andrews, United Kingdom. 2008.

[13] H. Schneider. *Reading Han Fei as "Social Scientist": a Case-Study in "Historical Correspondence"*. Comparative Philosophy. 4(1): 90–102. 2013.

[14] M. Piotrowski. *Natural Language Processing for Historical Texts*. Morgan & Claypool Publishers, San Rafael. 2012.

[15] E. Saykol, A. K. Sinop, U. Gudukbay, O. Ulusoy, A. E. Cetin. *Content-based Retrieval of Historical Ottoman Documents Stored as Textual Images*. IEEE Transactions on Image Processing. 13(3): 314–325. 2004.

[16] M. Rosen-Zvi, T. Griffiths, M. Steyers, and P. Smyth. *The Author-Topic Model for Authors and Documents*. Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI). 487–494. Banff Park Lodge, Canada. 2004.

[17] L. Pearl, M. Steyvers. *Detecting Authorship Deception: A Supervised Machine Learning Approach using Author Writeprints*. Literary and Linguistic Computing. 27(2): 183–196. 2012.

[18] Oxford Online English Dictionary: Network. http://www.oed.com/view/Entry/126342?rskey=YbCsUY&result=1&isAdvanced=false#eid [Last accessed: August 26, 2016].

[19] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, Oxford. 2010.

[20] Oxford Online English Dictionary: Social Network. http://www.oed.com/view/Entry/183739?redirectedFrom=social+network#eid139354802 [Last accessed: August 26, 2016].

[21] J. Ugander, B. Karrer, L. Backstrom, C. Marlow. *The Anatomy of the Facebook Social Graph*. CoRR, abs/1111.4503. 2011.

[22] Facebook Climbs to 1.59 Billion Users and Crushes Q4 Estimates With 5.8B Revenue. http://techcrunch.com/2016/01/27/facebook-earnings-q4-2015/ [Last accessed: April 2, 2017].

[23] A. Java, X. Song, T. Finn, B. Tseng. *Why We Twitter: Understanding Microblogging Usage and Communities*. Proceedings of the 9th WebKDD and 1st SNA-KDD Workshop on Web Mining and Social Network Analysis. 56–65. San Jose, USA. 2007.

[24] Twitter MAU Were 302M for Q1, Up 18% YOY. http://www.benzinga.com/news/earnings/15/04/5452400/twitter-mau-were-302m-for-q1-up-18-yoy [Last accessed: April 2, 2017].

[25] Oxford Online English Dictionary: Graph. http://www.oed.com/view/Entry/80819?rskey=Tc1Gdh&result=1&isAdvanced=false#eid [Last accessed: August 26, 2016].

[26] R. Zafarani, M. A. Abbasi, H. Liu. *Social Media Mining: An Introduction*. Cambridge University Press, Cambridge. 2014.

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

[28] E. D. Kolaczyk. *Statistical Analysis of Network Data: Methods and Models*. Springer, New York. 2009.

[29] R. Albert, H. Jeong, A. Barabási. *Diameter of the World-Wide Web*. Nature. 401: 130–131. 1999.

[30] T. V. Canh. *Learning Social Links and Communications from Interaction, topical, and Spatio-Temporal Information*. PhD Dissertation. Heidelberg University, Germany. 2014.

[31] M. E. J. Newman, M. Girvan. *Finding and Evaluating Community Structure in Networks*. Physical Review E. 69(2): 026113. 2004.

[32] L. C. Freeman. *Centrality in Social Networks: Conceptual Clarification*. Social Networks. 1: 215–239. 1978.

[33] L. Katz. *A New Status Index Derived from Sociometric Analysis*. Psychometrika. 39–43. 1953.

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

[35] S. Fortunato. *Community Detection in Graphs*. Physics Reports. 486(3): 75–174. 2010.

[36] M. Girvan, M. E. J. Newman. *Community Structure in Social and Biological Networks*. Proceedings of the National Academy of Sciences (PNAS). 99(12): 7821–7826. 2002.

[37] J. W. Moon, L. Moser. *On Cliques in Graphs*. Israel Journal of Mathematics. 3(1): 23–28. 1965.

[38] M. A. Porter, J. Onnela, P. J. Mucha. *Communities in Networks*. Notices of the American Mathematical Society. 56(9): 1082–1097. 2009.

[39] S. C. Johnson. *Hierarchical Clustering Schemes*. Psychometrika. 32(3): 241–254. 1967.

[40] M. E. J. Newman. *Detecting Community Structure in Networks*. The European Physical Journal B-Condensed Matter and Complex Systems. 38(2): 321–330. 2004.

[41] S. P. Lloyd. *Least Square Quantization in PCM*. IEEE Transactions on Information Theory. 28(2): 129–137. 1982.

[42] W. E. Donath, A. J. Hoffman. *Lower Bounds for the Partitioning of Graphs*. IBM Journal of Research and Development. 17(5): 420–425. 1973.

[43] M. Ester, H. Kriegel, J. Sander, X. Xu. *A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD). 96(34): 226–231. Portland, USA. 1996.

[44] G. Palla, I. Derényi, I. Farkas, T. Vicsek. *Uncovering the Overlapping Community Structure of Complex Networks in Nature and Society*. Nature. 435(7043): 814–818. 2005.

[45] S. M. van Dongen. *Graph Clustering by Flow Simulation*. PhD Dissertation. University of Utrecht, Netherlands. 2000.

[46] M. E. J. Newman. *The Structure and Function of Complex networks*. SIAM review. 45: 167–256. 2003.

[47] C. Berge, E. Minieka. *Graphs and hypergraphs*. North-Holland Publishing Company, Amsterdam. 1973.

[48] V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre. *Fast Unfolding of Communities in Large Networks*. Journal of Statistical Mechanics: Theory and Experiment. 10. 2008.

[49] A. L. Barabási. *Network Science*. Cambridge University Press, Cambridge. 2016.

[50] P. Holme. *Modern Temporal Network Theory: a Colloquium*. The European Physical Journal B. 88(234). 2015.

[51] R. K. Pan, J. Saramäki. *Path lengths, Correlations, and Centrality in Temporal Networks*. Physical Review E. 84(1): 016105. 2011.

[52] J. Huan, Z. Zhuang, J. Li, C. L. Giles. *Collaboration Over Time: Characterizing and Modeling Network Evolution*. Proceedings of the 2008 International Conference on Web Search and Data Mining (WSDM). 107–116. Palo Alto, USA. 2008.

[53] V. Colizza, A. Barrat, M. Barthélemy, A. Vespignani. *Prediction and Predictability of Global Epidemics: the Role of the Airline Transportation Network*. Proceedings of the National Academy of Sciences (PNAS). 103(7). 2006.

[54] G. Miritello, E. Moro, R. Lara. *Dynamical Strength of Social Ties in Information Spreading*. Physical Review E. 83(4): 045102. 2011.

[55] P. Holme, J. Saramäki. *Temporal Networks*. Physics Reports. 519(3): 97–125. 2012.

[56] J. K. Tang. *Temporal Network Metrics and Their Application to Real World Networks*. PhD Dissertation. University of Cambridge, United Kingdom. 2011.

[57] A. Barabási, R. Albert. *Emergence of Scaling in Random Networks*. Science. 286: 509–512. 1999.

[58] M. E. J. Newman. *The Structure of Scientific Collaboration Networks*. Proceedings of the National Academy of Sciences (PNAS). 98(2): 404–409. 2001.

[59] L. E. C. Rocha, N. Masuda. *Random walk centrality for temporal networks*. New Journal of Physics. 16(063023). 2014.

[60] M. Hanke, R. Foraita. *Clone Temporal Centrality Measures for Incomplete Sequences of Graph Snapshots*. BMC Bioinformatics. 18:261. 2017.

[61] Y. Hulovatyy, H. Chen, T. Milenković. *Exploring the structure and function of temporal networks with dynamic graphlets*. Bioinformatics. 31:i171–i180. 2015.

[62] B. B. Xuan, A. Ferreira, A. Jarry. *Computing Shoretest, Fastest, and Foremost Journeys in Dynamic Networks*. International Journal of Foundations of Computer Science. 14(267). 2003.

[63] P. Holme. *Network Reachability of Real-World Contact Sequences*. Physical Review E. 71(4): 046119. 2005.

[64] G. Kossinets, J. Kleinberg, D. Watts. *The Structure of Information Pathways in a Social Communication Network*. Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Las Vegas, USA. 435–443. 2008.

[65] P. Holme. *Scale-free Topology of E-mail Networks*. Physical Review E. 66(3): 035103. 2002.

[66] H. Wu, J. Cheng, S. Huang, Y. Ke, Y. Lu, Y. Xu. *Path problems in temporal graphs*. Proceedings of the VLDB Endowment. 7(9). 2014.

[67] N. Santoro, W. Quattrociocchi, P. Flocchini, A. Casteigts, F. Amblard. *Time-Varying Graphs and Social Network Analysis: Temporal Indicators and Metrics*. arXiv preprint arXiv:1102.0629. 2011.

[68] A. Casteigts, P. Flocchini, W. Quattrociocchi, N. Santoro. *Time-Varying Graphs and Dynamic Networks*. International Journal of Parallel, Emergent and Distributed Systems. 27(5): 387–408. 2011.

[69] V. Kostakos. *Temporal Graphs*. Physica A: Statistical Mechanics and its Applications. 388(6): 1007–1023. 2011.

[70] A. Clauset, N. Eagle. *Persistence and Periodicity in a Dynamic Proximity Network*. Proceedings of the DIMACS Workshop on Computational Methods for Dynamic Interaction Networks. New Brunswick, USA. 2007.

[71] A. L. Barabási. *The Origin of Bursts and Heavy Tails in Humans Dynamics*. Nature. 435: 207–212. 2005.

[72] A. Johansen. *Probing Human Response Times*. Physica A: Statistical Mechanics and its Applications. 338(1–2): 286–291. 2004.

[73] K. I. Goh, A. L. Barabási. *Burstiness and Memory in Complex Systems*. Europhysics Letters Association. 81(4): 48002. 2008.

[74] J. P. Eckmann, E. Moses, D. Sergi. *Entropy of Dialogues Creates Coherent Structures in E-mail Traffic*. Proceedings of the National Academy of Sciences (PNAS). 101(40): 14333–14337. 2004.

[75] E. Valdano, C. Poletto, A. Giovannini, D. Palma, L. Savini, V. Colizza. *Predicting Epidemic Risk from Past Temporal Contact Data*. PLOS Computational Biology. 11(3): e1004152. 2015.

[76] V. Neiger, C. Crespelle, E. Fleury. *On the Structure of Changes in Dynamic Contact Networks*. Signal Image Technology and Internet Based Systems. 731–738. 2012.

[77] S. Scellato, C. Mascolo, M. Musolesi, V. Latora. *Distance Matters: Geo-social Metrics for Online Social Networks*. Proceedings of the ACM Workshop on Online Social Networks (WOSN). Boston, USA. 2010.

[78] J. G. Oliveira, A. Barabási. *Human Dynamics: Darwin and Einstein Correspondence Patterns*. Nature. 437(7063): 1251. 2005.

[79] R. D. Malmgren, D. B. Stouffer, A. S. Campanharo, and L. A. N. Amaral. *On Universality in Human Correspondence Activity*. Science. 325(5948): 1696–1700. 2009.

[80] Oxford Online Dictionary: Historic. https://en.oxforddictionaries.com/definition/historic [Last accessed: August 26, 2016].

[81] Oxford Online Dictionary: Correspondence. https://en.oxforddictionaries.com/definition/correspondence [Last accessed: August 26, 2016].

[82] Corpora of Early English Correspondence. http://www.helsinki.fi/varieng/CoRD/corpora/CEEC/index.html [Last accessed: November 7, 2015].

[83] Corpus Epistolicum Recentioris Aevi. http://www.uni-mannheim.de/mateo/camenahtdocs/cera_e.html [Last accessed: November 7, 2015].

[84] Mapping the Republic of Letters. http://republicofletters.stanford.edu/ [Last accessed: November 7, 2015].

[85] Electronic Enlightenment: Letters and Lives Online. http://www.e-enlightenment.com/index.html [Last accessed: November 7, 2015].

[86] Frühneuzeitliche Ärztebriefe des deutschsprachigen Raums. http://www.medizingeschichte.uni-wuerzburg.de/akademie/index.html [Last accessed: November 7, 2015].

[87] Online Archive of Italian Literary Correspondences in Early Modern Age. http://www.archilet.it/HomePage.Aspx [Last accessed: November 7, 2015].

[88] Open Correspondence: The Correspondence Network of the Nineteenth-Century Literary World. http://www.opencorrespondence.org/ [Last accessed: November 7, 2015].

[89] Vernetzte Korresponzenden: Visualisierung von mehrdimensionalen Informationsstrukturen in Briefkorporan. http://kompetenzzentrum.uni-trier.de/de/projekte/projekte/briefnetzwerk/ [Last accessed: November 7, 2015].

[90] E. Moreton, N. O'Leary, P. O'Sullivan. *Visualising the Emigrant Letter*. Revue Européenne des Migrations Internationales. 30(3): 49–69. 2014.

[91] D. A. Kronick. *The Commerce of Letters: Networks and "Invisible Colleges" in Seventeenth-and Eighteenth-century Europe*. The Library Quarterly. 71.1: 28–43. 2001.

[92] Critical Edition of the Correspondence of Martin Opitz von Boberfeld (1597–1639). http://www.hab.de/de/home/wissenschaft/projekte/martin-opitz-von-boberfeld-1597-1639.html [Last accessed: November 7, 2015].

[93] Thomas Bodley. http://www.livesandletters.ac.uk/bodley/bodley.html [Last accessed: November 7, 2015].

[94] Bess of Hardwick's Life. http://www.bessofhardwick.org/background.jsp?id=142 [Last accessed: November 7, 2015].

[95] The Electronic Capito Project. http://www.itergateway.org/capito/ [Last accessed: November 7, 2015].

[96] Carolus Clusius Project. http://clusiuscorrespondence.huygens.knaw.nl/ [Last accessed: November 7, 2015].

[97] The Newton Project. http://www.newtonproject.sussex.ac.uk/prism.php?id=153\#hd15 [Last accessed: November 7, 2015].

[98] The Spenser Letters. http://www.english.cam.ac.uk/ceres/haphazard/letters/lettersindex.html [Last accessed: November 7, 2015].

[99] Thomas Gray Archive. http://www.thomasgray.org/ [Last accessed: November 7, 2015].

[100] Sir Hans Sloane's Correspondence Online. http://sloaneletters.com/ [Last accessed: November 7, 2015].

[101] Françoise de Graffigny's Correspondence. http://french.chass.utoronto.ca/graffigny/ [Last accessed: November 7, 2015].

[102] The Correspondence of Hugo Grotius. http://grotius.huygens.knaw.nl/years [Last accessed: November 7, 2015].

[103] D. Hansen, B. Shneiderman, M. A. Smith. *Analyzing Social Media Networks with NodeXL: insights from a connected world*. Elsevier, Morgan Kaufmann, Burlington. 2010.

[104] The Cullen Project: The Consultation Letters of Dr William Cullen. http://www.cullenproject.ac.uk/ [Last accessed: November 7, 2015].

[105] The correspondence of Constantijn Huygens. http://resources.huygens.knaw.nl/briefwisselingconstantijnhuygens/en [Last accessed: November 7, 2015].

[106] Corpus of Ioannes Dantiscus' Texts and Correspondence. http://dantiscus.ibi.uw.edu.pl/ [Last accessed: November 7, 2015].

[107] Inventory of the Correspondence of Johann Valentin Andreae (1586–1654). http://www.hab.de/en/home/research/projects/inventory-of-the-correspondence-of-johann-valentin-andreae-1586-1654.html [Last accessed: November 7, 2015].

[108] The Linnaean Correspondence. http://linnaeus.c18.net/ [Last accessed: November 7, 2015].

[109] The Correspondence of William Orange. http://resources.huygens.knaw.nl/wvo/en [Last accessed: November 7, 2015].

[110] The Correspondence of Oswald Myconius. https://myconius.unibas.ch/ [Last accessed: November 7, 2015].

[111] William Dugdale: A Catalogue of his Correspondence. http://www.xmera.co.uk/dugdale_cat/index.php [Last accessed: November 7, 2015].

[112] Edition of the Letters Philipp Jakob Spener (1635–1705). http://www.edition-spenerbriefe.de/en/page/willkommen.php?lang=DE [Last accessed: November 7, 2015].

[113] "Everything on Paper Will Be Used Against Me:" Quantifying Kissinger. http://blog.quantifyingkissinger.com/all-posts/ [Last accessed: November 7, 2015].

[114] D. Bamman, A. Anderson, N. A. Smith. *Inferring Social Rank in An Old Assyrian Trade Network*. Digital Humanities Conference (DH). Lincoln, USA. 2013.

[115] E. L. Moreton. *The Emigrant Letter Digitised: Markup and Analysis*. Dissertation. University of Birmingham. United Kingdom. 2016.

[116] C. V. D. Heuvel. *Mapping Knowledge Exchange in Early Modern Europe: Intellectual and Technological Geographies and Network Representations*. International Journal of Humanities and Arts Computing. 9.1: 95–114. 2015.

[117] Oxford Online Dictionary: Topic. https://en.oxforddictionaries.com/definition/topic [Last accessed: August 26, 2016].

[118] D. M. Blei, A. Y. Ng, M. I. Jordan. *Latent Dirichlet Allocation*. Journal of Machine Learning Research. 3(1): 993–1022. 2003.

[119] A. Srivastava, M. Sahami. *Text Mining: Classification, Clustering, and Applications*. CRC Press, Boca Raton. 2009.

[120] D. M. Blei. *Probabilistic Topic Models*. Communications of the ACM. 55(4): 77–84. 2012.

[121] S. Geman, D. Geman. *Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images*. IEEE Transactions on Pattern Analysis and Machine Intelligence. 6(6): 721–741. 1984.

[122] M. Jordan, Z. Ghahramani, T. Jaakkola, L. Saul. *An Introduction to Variational Methods for Graphical Models*. Machine Learning. 37: 183–233. 1999.

[123] M. Wainwright, M. Jordan. *Graphical Models, Exponential Families, and Variational Inference*. Foundations and Trends in Machine Learning. 1(1–2): 1–305. 2008.

[124] R. Arun, V. Suresh, C. E. V. Madhavan, M. N. N. Murthy. *On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations*. Advances in Knowledge Discovery and Data Mining. Springer, Heidelberg. 391–402. 2010.

[125] Y. W. Teh, M. I. Jordan, M. J. Beal, D. M. Blei. *Sharing Clusters among Related Groups: Hierarchical Dirichlet Processes*. Advances in Neural Information Processing Systems. 1385–1392. 2004.

[126] C. Juan, X. Tian, L. J. Tao, Z. Y. Dong, T. Sheng. *A Density-based Method for Adaptive LDA Model Selection*. Journal Neurocomputing. 72(7-9): 1775–1781. 2009.

[127] T. L. Griffiths, M. Steyvers. *Finding Scientific Topics*. Proceedings of the National Academy of Sciences 101 (Suppl_1). 5228–5235. 2004.

[128] Introduction to Topic Modeling with Paper Machines. https://devo-evo.lab.asu.edu/methods/?q=system/files/week5TopicModeling.pdf [Last accessed: December 14, 2016].

[129] D. J. Newman, S. Block. *Probabilistic Topic Decomposition of An Eighteenth Century American newspaper*. Journal of the American Society for Information Science and Technology. 57(6): 753–767. 2006.

[130] E. Meeks. *Comprehending the Digital Humanities*. Digital Humanities Specialist. 2011.

[131] R. K. Nelson. *Mining the Dispatch*. http://dsl.richmond.edu/dispatch/pages/intro [Last accessed: December 14, 2016].

[132] C. Blevin. *Topic Modeling Martha Ballard's Diary*. http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/ [Last accessed: December 14, 2016].

[133] D. Mimno. *Computational Historiography: Data Mining in a Century of Classics Journals*. Journal on Computing and Cultural Heritage. 5(1): 1–19. 2012.

[134] L. C. Freeman. *The Impact of Computer based Communication on the Social Structure of an Emerging Scientific Specialty*. Social Networks. 6(3): 201–221. 1984.

[135] J. Shetty, J. Adibi. *The Enron Email Dataset Database Schema and Brief Statistical Report*. Information Sciences Institute Technical Report. University of Southern California, USA. 2004.

[136] R. Rowe, G. Creamer, S. Hershkop, S. J. Stolfo. *Automated Social Hierarchy Detection Through Email Network Analysis*. Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. San Jose, USA. 109–117. 2007.

[137] J. Diesner. *Communication Networks from the Enron Email Corpus "It's Always About the People. Enron is no Different"*. Computational & Mathematical Organization Theory. 11(3): 201–228. 2005.

[138] A. McCallum, A. Corrada-Emmanuel, X. Wang. *Topic and Role Discovery in Social Networks*. Proceedings of the 19th International Joint Conference on Artificial Intelligence (IJCAI). Edinburgh, Scotland. 786–791. 2005.

[139] P. A. Longley, M. F. Goodchild, D. J. Maguire, D. W. Rhind. *Geographic Information:Systems and Science*. John Wiley and Sons Publishing, Hoboken. 2001.

[140] B. Plewe. *The Nature of Uncertainty in Historical Geographic Information*. Transactions in GIS. 6(4): 431–456. 2002.

[141] G. P. Rees. *Uncertain Date, Uncertain Place: Interpreting the History of Jewish Communities in the Byzantine Empire using GIS*. Digital Humanities Conference (DH). Hamburg, Germany. 2012.

[142] F. Binder, B. Entrup, I. Schiller, H. Lobin. *Uncertain about Uncertainty: Different Ways of Processing Fuzziness in Digital Humanities Data*. Digital Humanities Conference (DH). Lausanne, Switzerland. 2014.

[143] I. Bhattacharya, L. Getoor. *Collective Entity Resolution in Relational Data*. ACM Transactions on Knowledge Discovery from Data (TKDD). 1(1):5. 2007.

[144] M. Andert, F. Berger, J. Ritter, P. Molitor. *Optimized Platform for Capturing Metadata of Historical Correspondences*. Literary and Linguistic Computing. 2014.

[145] Oxford Online Dictionary: Entity. https://en.oxforddictionaries.com/definition/entity [Last accessed: August 26, 2016].

[146] R. Grishman, B. Sundheim. *Message Understanding Conference-6: a Brief History*. Proceedings of the 16th Conference on Computational Linguistics (COLING). Copenhagen, Denmark. 1: 446–471. 1996.

[147] N. Chinchor. *Overview of MUC-7*. Proceedings of the 7th Message Understanding Conference (MUC-7). Fairfax, USA. 1998.

[148] D. Carmel, M. Chang, E. Gabrilovich, B. Hsu, K. Wang. *Entity Recognition and Disambiguation Challenge (ERD)*. ACM SIGIR Forum. 48(2):63-77. 2014.

[149] J. Hoffart. *Discovering and Disambiguating Named Entities in Text*. PhD Dissertation. Saarland University, Germany. 2015.

[150] D. Nadeau, S. Sekine. *A Survey of Named Entity Recognition and Classification*. Journal of Linguisticae Investigationes. 30(1). 2007.

[151] C. Grover, S. Givon, R. Tobin, J. Ball. *Named Entity Recognition for Digitized Historical Texts*. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC). Marrakech, Morocco. 2008.

[152] L. Borin, D. Kokkinakis, L. Olsson. *Naming the Past: Named Entity and Animacy Recognition in 19th Century Swedish Literature*. Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH). 1–8. Prague, Czech Republic. 2007.

[153] A. X. Chang, V. I. Spitkovsky, C. D. Manning, E. Agirre. *A Comparison of Named-Entity Disambiguation and Word Sense Disambiguation*. Proceedings

of the 10th International Conference on Language Resources and Evaluation (LREC). 860-867. Portorož, Slovenia. 2016.

[154] Y. Li, C. Wang, F. Han, J. Han, D. Roth, X. Yan. *Mining Evidences for Named Entity Disambiguation*. Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 1070–1078. Chicago, USA. 2013.

[155] D. A. Smith, G. Crane. *Disambiguating Geographic Names in a Historical Digital Library*. Proceedings of the 5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL). 127–136. Darmstadt, Germany. 2001.

[156] D. Vrandečić, M. Krötzsch. *Wikidata: a Free Collaborative Knowledgebase*. Communications of the ACM. 57(10): 78–85. 2014.

[157] B. Aleman-Meza, C. Halaschek-Wiener, I. B. Arpinar, C. Ramakrishnan, A. P. Sheth. *Ranking Complex Relationships on the Semantic Web*. IEEE Internet Computing. 9(3): 37–44. 2005.

[158] J. Hoffart, M. A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, G. Weikum. *Robust disambiguation of named entities in text*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP). Edinburgh, United Kingdom. 782–792. 2011.

[159] R. Bunescu, M. Paşca. *Using Encyclopedic Knowledge for Named entity Disambiguation*. Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL). 6: 9–16. Trento, Italy. 2006.

[160] A. Alhelbawy, R. J. Gaizauskas. *Graph Ranking for Collective Named Entity Disambiguation*. Annual Meeting of the The 52nd Annual Meeting of the Association for Computational Linguistics (ACL). (2): 75–80. Baltimore, USA. 2014.

[161] E. Minkov, W. W. Cohen, A. Y. Ng. *Contextual search and name disambiguation in email using graphs*. Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 27–34. Seattle, USA. 2006.

[162] B. Malin, E. Airoldi, K. M. Carley. *A network Analysis Model for Disambiguation of Names in Lists*. Computational & Mathematical Organization Theory. 11(2): 119–139. 2005.

[163] L. Hermansson, T. Kerola, F. Johansson, V. Jethava, D. Dubhashi. *Entity Disambiguation in Anonymized Graphs Using Graph Kernels*. Proceedings of the 22nd ACM International Conference on Information & Knowledge Management (CIKM). 1037–1046. San Francisco, USA. 2013.

[164] I. Bhattacharya, L. Getoor. *Entity Resolution in Graphs*. Mining Graph Data. 311. 2006.

[165] F. H. Levin, C. A. Heuser. *Evaluating the Use of Social Networks in Author Name Disambiguation in Digital Libraries*. Journal of Information and Data Management (JIDM). 1(2): 183–197. 2010.

[166] W. Shen, J. Han, J. Wang. *A Probabilistic Model for Linking Named Entities in Web Text with Heterogeneous Information Networks*. Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data (SIGMOD). 1199–1210. Snowbird, USA. 2014.

[167] W. H. Gomaa, A. A. Fahmy. *A Survey of Text Similarity Approaches*. International Journal of Computer Applications. 68(13). 2013.

[168] Y. Jiang, G. L. Li, J. H. Feng, W. S. Li. *String Similarity Joins: an Experimental Evaluation*. Proceedings of the VLDB Endowment. 7(8): 625–636. 2014.

[169] V. I. Levenshtein. *Binary Codes Capable of Correcting Deletions, Insertions, and Reversals*. Soviet Physics Doklady. 10: 707–710. 1966.

[170] S. Needleman, C. D. Wunsch. *A General Method Applicable to the Search of Similarities in the Amino Acid Sequence of Two Proteins*. Journal of Molecular Biology. 48: 443–453. 1970.

[171] M. Jaro. *Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida*. Journal of the American Statistical Association. 84: 414–420. 1989.

[172] W. Cohen, P. Ravikumar, S. Fienberg. *A Comparison of String Metrics for Matching Names and Records*. KDD Workshop on Data Cleaning and Object Consolidation. 3: 73–78. Washington, USA. 2003.

[173] W. Winkler. *String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage*. Proceedings of the Section on Survey Research Methods. 354–359. 1990.

[174] P. Jaccard. *Étude comparative de la distribution florale dans une portion des Alpes et des Jura*. Bulletin de la Société Vaudoise des Sciences Naturelles. 37: 547–579. 1901.

[175] A. Huang. *Similarity Measures for Text Document Clustering*. Proceedings of the 6th New Zealand Computer Science Research Student Conference (NZCSRSC). 49–56. Christchurch, New Zealand. 2008.

[176] R. Mihalcea, C. Corley, C. Strapparava. *Corpus-based and Knowledge-based Measures of Text Semantic Similarity*. Proceedings of the 21st National Conference on Artificial Intelligence (AAAI). 1:775–780. Boston, USA. 2006.

[177] S. Dennis, T. Landauer, W. Kintsch, J. Quesada. *Introduction to Latent Semantic Analysis*. Slides: 25th Annual Meeting of the Cognitive Science Society. Boston, USA. 2003.

[178] T. K. Landauer, S. T. Dumais. *A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge*. Psychological Review. 104. 1997.

[179] T. K. Landauer, P. W. Foltz, D. Laham. *An Introduction to Latent Semantic Analysis*. Discourse Processes. 25(2–3): 259–284. 1998.

[180] P. W. Foltz, W. Kintsch, T. K. Landauer. *The Measurement of Textual Coherence with Latent Semantic Analysis*. Discourse Processes. 25(2–3): 285–307. 1998.

[181] K. W. Boyack, D. Newman, R. J. Duhon, R. Klavans, M. Patek, J. R. Biberstine, B. Schijvenaars, A. Skupin, N. Ma, K. Börner. *Clustering More than Two Million Biomedical Publications: Comparing the Accuracies of Nine Text-based Similarity Approaches*. PlOS One. 6(3). 2011.

[182] A. Islam, D. Inkpen. *Semantic Text Similarity using Corpus-based Word Similarity and String Similarity*. ACM Transactions on Knowledge Discovery from Data (TKDD). 2(2). 2008.

[183] T. Hofmann. *Unsupervised Learning by Probabilistic Latent Semantic Analysis*. Machine Learning. 42: 177–196. 2001.

[184] C. Leacock, M. Chodorow. Combining Local *Context and WordNet Sense Similarity for Word Sense Identification*. WordNet, An Electronic Lexical Database. The MIT Press, Cambridge. 1998.

[185] Z. Wu, M. Palmer. *Verb Semantics and Lexical Selection*. Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL). 133–138. Las Cruces, USA. 1994.

[186] R. Resnik. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI). 1: 448–453. Montreal, Canada. 1995.

[187] D. Lin. *Extracting Collocations from Text Corpora*. First Workshop on Computational Terminology (Computerm). 57–63. Montreal, Canada. 1998.

[188] J. J. Jiang, D. W. Conrath. *Semantic Similarity based on Corpus Statistics and Lexical Taxonomy*. Proceedings of the 10th International Conference on Research in Computational Linguistics (ROCLING/IJCLCLP). Taipei, Taiwan. 1997.

[189] T. K. Landauer, D. S. McNamara, S. Dennis, W. Kintsch. *Handbook of Latent Semantic Analysis*. Psychology Press, Taylor & Francis, Oxford. 2014.

[190] L. Lee. *Measures of distributional similarity*. Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL). 25–32. Maryland, USA. 1999.

[191] J. Han, M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Burlington. 2006.

[192] V. Rus, M. Lintean, R. Banjade, N. Niraula, D. Stefanescu. *SEMILAR: The Semantic Similarity Toolkit*. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. 163–168. Sofia, Bulgaria. 2013.

[193] B. Fuglede, T. Fleming. *Jensen-Shannon Divergence and Hilbert Space Embedding*. Proceedings of The IEEE International Symposium on Information Theory (ISIT). 31. Chicago, USA. 2004.

[194] M. Koppel, J. Schler, S. Argamon. *Computational Methods in Authorship Attribution*. Journal of the American Society for Information Science and Technology. 60(1): 9–26. 2009.

[195] E. Stamatatos. *A Survey of Modern Authorship Attribution Methods*. Journal of the American Society for Information Science and Technology. 60(3):538–556. 2009.

[196] R. Zheng, J. Li, H. Chen, Z. Huang. *A Framework for Authorship Identification of Online Messages: Writing-style Features and Classification Techniques*. Journal of the American Society for Information Science and Technology. 57(3): 378–393. 2006.

[197] T. C. Mendenhall. *The Characteristic curves of Composition*. Science. 11(11): 237–249. 1887.

[198] F. Mosteller, D. L. Wallace. *Inference and Disputed Authorship: the Federalist*. Adison-Wesley. Reading. 1964.

[199] M. Rosen-Zvi, T. Griffiths, M. Steyvers, P. Smyth. *The Author-Topic Model for Authors and Documents*. Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI). 487–494. Banff, Canada. 2004.

[200] D. Mimno, A. McCallum. *Topic Models Conditioned on Arbitrary Features with Dirichlet-multinomial Regression*. Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence (UAI). 411–418. Helsinki, Finland. 2008.

[201] Y. Seroussi, I. Zukerman, F. Bohnert. *Authorship Attribution with Topic Models*. Computational Linguistics. 40(2): 269–310. 2014.

[202] J. Pustejovsky, K. Lee, H. Bunt, L. Romary. *ISO-TimeML: An International Standard for Semantic Annotation*. The 7th International Conference on Language Resources and Evaluation (LREC). Valletta, Malta. 2010.

[203] J. Strötgen, M. Gertz. *Domain-Sensitive Temporal Tagging*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, San Rafael. 2016.

[204] Google Map API. https://developers.google.com/maps/web/?hl=de [Last accessed: April 2, 2017].

[205] G. Gallo, G. Longo, S. Pallottino, S. Nguyen. *Directed Hypergraphs and Applications*. Discrete Applied Mathematics. 42(2–3): 177–201. 1993.

[206] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, Oxford. 2010.

[207] L. Akoglu, P. O. V. Melo, C. Faloutsos. *Quantifying Reciprocity in Large Weighted Communication Networks*. The 16th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). 85–96. Kuala Lumpur, Malaysia. 2012.

[208] S. Borzsony, D. Kossmann, K. Stocker. *The Skyline Operator*. Proceedings of 17th International Conference on Data Engineering. 421–430. Heidelberg, Germany. 2001.

[209] M. E. J. Newman. *Scientific Collaboration Networks. II. Shortest Paths, weighted networks, and Centrality*. Physical Review E. 64(1): 016132. 2004.

[210] A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. Vespignani. *The Architecture of Complex weighted networks*. Proceedings of the National Academy of Sciences (PNAS). 101(11): 3747–3752. 2004.

[211] E. W. Dijkstra. *A Note on Two Problems in Connexion with Graphs*. Numerische Mathematik 1. 269–271. 1959.

[212] M. E. J. Newman. *Scientific collaboration networks. I. Network construction and fundamental results*. Physical review E. 64(1): 016131. 2001.

[213] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, T. Vicsek. *Evolution of the Social Network of Scientific Collaborations*. Physical A: Statistical Mechanics and its Applications. 311(3): 590–614. 2002.

[214] C. Biscaro, C. Giupponi. *Co-Authorship and Bibliographic Coupling Network Effects on Citations*. PLOS ONE. 9(6): e99502. 2014.

[215] J. W. Tukey. *Exploratory Data Analysis*. Addison-Wesley, Boston. 1977.

[216] D. C. Hoaglin. *John W.Tukey and Data Analysis*. Statistical Science. 311–318. 2003.

[217] B. Viswanath, A. Mislove, M. Cha, K. P. Gummadi. *On the Evolution of User Interaction in Facebook*. Proceedings of the Second ACM Workshop on Online Social Networks (SIGCOMM). 37–42. Barcelona, Spain. 2009.

[218] J. Tang. S. Scellato, M. Musolesi, C. Mascolo, V. Latora. *Small-World Behavior in Time-Varying Graphs*. Physical Review E. 81(5): 055101. 2010.

[219] L. Sonneborn. *Mark Twain (Who Wrote That?)*. Chelsea House Publications, New York. 2010.

[220] C. D. Manning, P. Raghavan, H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, Cambridge. 2008.

[221] J. Silge, D. Robinson. *Term Frequency and Inverse Document Frequency (tf-idf) Using Tidy Data Principles*. https://cran.r-project.org/web/packages/tidytext/vignettes/tf_idf.html [Last accessed: April 2, 2017].

[222] F. Jelinek, R. L. Mercer. *Interpolated Estimation of Markov Source Parameters from Sparse Data*. Proceedings of the First International Workshop on Pattern Recognition in Practice. Elsevier (Science), Amsterdam. 1980.

[223] S. T. Coleridge. *Specimens of the Table Talk of Samuel Taylor Coleridge*. John Murray, London. 1836.

[224] D. M. W. Powers. *Applications and Explanations of Zipf's Law*. Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning (NeMLaP/CoNLL). 151–160. Sydney, Australia. 1998.

[225] Stemming and Lemmatization. https://nlp.stanford.edu/IR-book/html/htmledition/stemming-and-lemmatization-1.html [Last accessed: April 2, 2017].

[226] P. F. Grendler. *Renaissance Education Between Religion and Politics*. Taylor & Francis Group, London. 2006.

[227] L. D. Mansch, C. Peters. *Martin Luther: The Life and Lessons*. McFarland & Company, Jefferson. 2016.

[228] Reformation. https://www.luther2017.de/en/reformation/ [Last accessed: February 3, 2017].

[229] L. Berhard. *Martin Luther: An Introduction to his Life and Work*. Fortress Press, Philadelphia. 1987.

[230] M. A. Mullett. *Martin Luther*. Routledge, London. 2004.

[231] Martin Luther. http://www.tlogical.net/bioluther.htm [Last accessed: February 3, 2017].

[232] F. Bethencourt, F. Egmond. *Cultural Exchange in Early Modern Europe, vol.3: Correspondence and Cultural Exchange in Europe, 1400-1700*. Cambridge University Press, Cambridge. 2007.