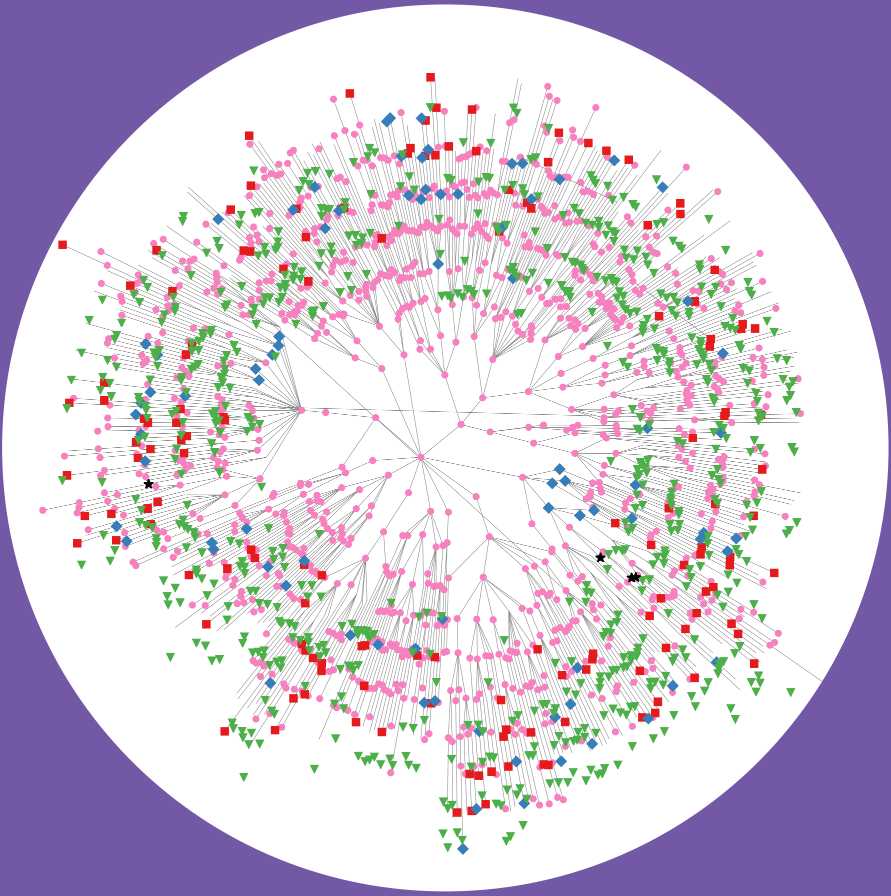
Department of Computer Science

Collective Entity Resolution Methods for Network Inference

Eric Malmi



Aalto University

DOCTORAL DISSERTATIONS

Aalto University publication series
DOCTORAL DISSERTATIONS 73/2018

Collective Entity Resolution Methods for Network Inference

Eric Malmi

A doctoral dissertation completed for the degree of Doctor of Science (Technology) to be defended, with the permission of the Aalto University School of Science, at a public examination held at the lecture hall AS1 of the school on 20 June 2018 at 12.

Aalto University
School of Science
Department of Computer Science
Data Mining Group

Supervising professors
Professor Aristides Gionis,
Aalto University, Finland

Preliminary examiners
Professor Lise Getoor,
University of California Santa Cruz, USA

Professor Gunnar W. Klau,
Heinrich Heine University Düsseldorf, Germany

Opponents
Associate Professor David F. Gleich,
Purdue University, USA

Aalto University publication series
DOCTORAL DISSERTATIONS 73/2018

ISBN 978-952-60-7949-3 (printed)
ISBN 978-952-60-7950-9 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)
http://urn.fi/URN:ISBN:978-952-60-7950-9

Unigrafia Oy
Helsinki 2018

Finland

NORDIC ECOLABEL
441 697
Printed matter

Aalto University

Abstract

Aalto University, P.O. Box 11000, FI-00076 Aalto www.aalto.fi

Author
Eric Malmi

Name of the doctoral dissertation
Collective Entity Resolution Methods for Network Inference

Publisher School of Science

Unit Department of Computer Science

Series Aalto University publication series DOCTORAL DISSERTATIONS 73/2018

Field of research Data Mining

Manuscript submitted 29 November 2017 **Date of the defence** 20 June 2018

Permission to publish granted (date) 7 February 2018 **Language** English

☐ **Monograph** ☒ **Article dissertation** ☐ **Essay dissertation**

Abstract

Data has become an essential resource, which is used to guide decision making across many levels of society. To fully leverage the abundance of data sources, the various sources need to be integrated, which poses difficult computational challenges. Entity resolution techniques address these challenges by trying to identify data records referring to the same underlying entity. Often, relational information about the records (for example, a friendship network between the users of a social networking service) is available, but this information is ignored by the traditional entity resolution techniques. The goal of this thesis is to develop novel collective entity resolution methods which match records by leveraging relational information and produce an entity network. The developed methods are applicable to a wide array of applications - from bioinformatics to ontologies - but the initial motivation for this work has been the problem of integrating genealogical data to infer large-scale genealogical networks (family trees).

This thesis makes the following methodological contributions: First, we develop novel methods for linking vital records, such as birth records, to infer genealogical networks. An experimental evaluation of the inferred networks shows that even fully automatic methods can produce fairly accurate networks, and moreover, the estimated link probabilities provide a reliable way to quantify the certainty of the inferred family relationships. Second, we propose methods with theoretical guarantees for aggregating the edges of directed acyclic graphs in the case that the correspondance between input-graph nodes is known. Third, if the correspondance is unknown, an alignment between the nodes has to be found. We study the resulting network alignment problem and propose methods for aligning multiple networks and for aligning networks actively by leveraging human experts.

The proposed vital-record linking methods have been employed to automatically link a dataset of five million historical birth records from Finland. To visualize the resulting network and to enable the exploration of the inferred links, we have developed an online tool called AncestryAI, which has been used so far by thousands of genealogists in Finland. In the final part of the thesis, we demonstrate the usefullness of the inferred genealogical network for the field of computational social science by presenting a longitudinal analysis on assortative mating, that is, the tendency to marry someone with a similar socioeconomic status. This phenomenon is quantified by comparing the socioeconomic statuses of the automatically inferred spouses. We find evidence that assortative mating existed in Finland (1735-1885), but interestingly, we do not observe any monotonically decreasing or increasing trend in the strength of assortative mating.

Keywords entity resolution, network alignment, machine learning, genealogy, computational social science

ISBN (printed) 978-952-60-7949-3 **ISBN (pdf)** 978-952-60-7950-9

ISSN (printed) 1799-4934 **ISSN (pdf)** 1799-4942

Location of publisher Helsinki **Location of printing** Helsinki **Year** 2018

Pages 180 **urn** http://urn.fi/URN:ISBN:978-952-60-7950-9

Aalto-yliopisto

Tiivistelmä

Aalto-yliopisto, PL 11000, 00076 Aalto www.aalto.fi

Tekijä
Eric Malmi

Väitöskirjan nimi
Kollektiivisia tietueiden linkitysmenetelmiä verkostojen päättelyyn

Julkaisija Perustieteiden korkeakoulu

Yksikkö Tietotekniikan laitos

Sarja Aalto University publication series DOCTORAL DISSERTATIONS 73/2018

Tutkimusala Tiedonlouhinta

Käsikirjoituksen pvm 29.11.2017 **Väitöspäivä** 20.06.2018

Julkaisuluvan myöntämispäivä 07.02.2018 **Kieli** Englanti

□ **Monografia** ☒ **Artikkeliväitöskirja** □ **Esseeväitöskirja**

Tiivistelmä

Datasta on tullut tärkeä resurssi, joka ohjaa päätöksentekoa monilla yhteiskunnan tasoilla. Eri datalähteet tulisi kyetä yhdistämään, jotta niitä voisi hyödyntämää tehokkaasti, mikä muodostaa haastavan laskennallisen ongelman. Tietueiden linkitysmenetelmät vastaavat tähän ongelmaan yrittäen tunnistaa samaan entiteettiin viittaavat tietueet. Usein tietueista on saatavissa relationaalista tietoa, kuten esimerkiksi sosiaalisen verkoston käyttäjien välinen ystävyysverkosto, mutta perinteiset linkitysmenetelmät jättävät nämä relationaaliset tiedot huomiotta. Tämän työn tavoitteena on kehittää uusia kollektiivisia tietueiden linkitysmenetelmiä, jotka hyödyntävät relationaalista tietoa ja tuottavat entiteettiverkoston. Kehitettyjä menetelmiä voidaan soveltaa moniin kohteisiin, kuten bioinformatiikkaan tai ontologioihin, mutta työn alkuperäisenä tavoitteena on ollut laajojen sukuverkostojen eli sukupuiden päättely.

Väitöskirjassa esitellään seuraavat metodologiset kontribuutiot: 1. Henkilötietojen, kuten kastetapahtumien, linkittämiseen on kehitetty uusia menetelmiä, joilla voidaan päätellä sukuverkostoja. Pääteltyjen verkostojen analyysi osoittaa, että jopa täysin automaattiset menetelmät voivat tuottaa melko tarkkoja verkostoja. Tämän lisäksi menetelmien tuottamat todennäköisyydet tarjoavat luotettavan tavan määrittää löydettyjen perhesuhteiden epävarmuus. 2. Verkostojen yhdistämiseen on kehitetty teoreettisesti perusteltuja menetelmiä, jotka soveltuvat suunnattuihin asyklisiin verkostoihin, joiden solmujen vastaavuudet on tunnettu. 3. Mikäli solmujen vastaavuus on tuntematon, tulee solmut ensin kohdistaa toisiinsa. Työssä on tarkasteltu näin syntyvää verkostojen kohdistamisongelmaa ja kehitetty menetelmiä useiden verkostojen kohdistamiseen automaattisesti sekä kahden verkoston kohdistamiseen interaktiivisesti ihmisasiantuntijoita hyödyntäen.

Henkilötietojen linkittämiseen kehitettyjä menetelmiä on sovellettu viiden miljoonan Suomesta kerätyn yli sata vuotta vanhan kastetapahtuman linkittämiseen. Tuloksena saadun verkoston visualisointiin ja pääteltyjen sukulaisuussuhteiden etsintään on kehitetty AncestryAI-niminen verkkotyökalu, jota tuhannet sukututkijat Suomessa ovat tähän mennessä käyttäneet. Pääteltyä sukuverkostoa voidaan soveltaa myös laskennallisen yhteiskuntatieteen alalle. Työn loppuosassa esitellään pitkittäistutkimus liittyen assortatiiviseen parinmuodostukseen, eli ihmisten taipumukseen valikoida puoliso, jolla on samanlainen sosioekonominen asema. Ilmiötä mitataan vertailemalla automaattisesti pääteltyjen puolisoiden sosioekonomista asemaa. Tulokset osoittavat ilmiön esiintyneen Suomessa vuosina 1735-1885, mutta hieman yllättäen ilmiön voimakkuudessa ei havaita monotonista heikkenemistä tai vahvistumista tarkastelujakson aikana.

Avainsanat tietueiden linkitys, verkostojen kohdistaminen, koneoppiminen, sukututkimus, laskennallinen yhteiskuntatiede

ISBN (painettu) 978-952-60-7949-3 **ISBN (pdf)** 978-952-60-7950-9

ISSN (painettu) 1799-4934 **ISSN (pdf)** 1799-4942

Julkaisupaikka Helsinki **Painopaikka** Helsinki **Vuosi** 2018

Sivumäärä 180 **urn** http://urn.fi/URN:ISBN:978-952-60-7950-9

Preface

The first thing that caught my attention when I joined the data mining group at Aalto was that during the group meetings people were often smiling and having fun. These weekly gatherings quickly became something that I would look forward to attending week after week throughout the 4.5 years that I was part of the group. I would like to thank all the past and present group members for creating such a nice and also scientifically stimulating atmosphere. In particular, thank you to Aris Gionis, our group leader and my supervisor, with whom it has been a great privilege to work. Aris taught me a lot about math and research, he was exceptionally responsive to my emails, and he always had time to meet when I needed advice. He was also very open to all of my non-mainstream research ideas—from generating rap lyrics to analyzing 400-year old church records—helping me to turn these into successful research projects.

I'm grateful for the pre-examiners of this thesis, Lise Getoor and Gunnar W. Klau, for their valuable comments on the manuscript. Their own works have been central sources of inspiration to the methods developed in this thesis. Also, I should thank the Nokia Foundation, the Finnish Foundation for Technology Promotion (TES), and the Emil Aaltonen Foundation for their generous financial support for my research.

This thesis would look very different if it wasn't for my friend and collaborator, Arno Solin, who told me about the HisKi dataset back in 2013. I usually test ideas by waiting until the next day to see if I'm still as excited about the idea. In many cases, I'm not. However, with the idea of inferring family trees, I got even more excited the next day. Thank you to Arno for the valuable advice and feedback on genealogy, data visualization, typography, machine learning, and other topics.

Moreover, I received a lot of valuable feedback and advice from the numerous genealogists who commented my work, and this knowledge would have been hard or impossible to acquire otherwise. In particular, I would like to thank Pekka Valta, Juha Mäkelainen, and Matti Juhala, who spent dozens of hours sharing their knowledge about genealogy and providing comments on AncestryAI. Likewise, I'm very thankful to the Genealogical Society of Finland, P. T. Kuusiluoma, Teppo Ylitalo, and Jouni Malinen, who provided me with access to the HisKi data and encouraged me to pursuit this effort.

The collaboration with many genealogists from the early phases of my thesis project was essential for the success of the thesis and this collaboration was primarily made possible by the efforts of the Aalto and CS department communications teams who helped to promote my work outside academia. I would like to thank especially Tiina Aulanko-Jokirinne for her active and enthusiastic involvement.

I had the pleasure to work with a very strong group of co-authors, namely Sanjay Chawla, Evimaria Terzi, Nikolaj Tatti, Arno Solin, and Marko Rasa. Marko also built the web interface of AncestryAI and he's probably the person who has taught me the most about programming through various hobby projects we've done together.

During my studies, I also had the opportunity to visit and intern at several different places. I would like to thank Aris for helping to arrange these visits, Sanjay for hosting me at QCRI, Timo Smura and Hannu Verkasalo for hosting me at Verto Analytics, Daniele Pighin and Enrique Alfonseca for hosting me at Google, and Przemyslaw Grabowicz and Krishna Gummadi for hosting me at the Max Planck Institute for Software Systems. These visits greatly widened my perspective on research and other aspects of life.

Doing a PhD wouldn't have been nearly so pleasant experience if it wasn't for all the friends, colleagues, and collaborators who would listen and give ideas for my research, hang out at conferences, play foosball, and explore some of the remotest places on Earth (physically as well as virtually[1]). These people include but are not limited to: Antonis Matakos, Kiran Garimella, Michael Mathioudakis, Han Xiao, Sanja Scepanovic, Polina Rozenshtein, Suhas Thejaswi, Orestis Kostakis, Preethi Lahoti, Jeffrey Lijffit, Antti Ukkonen, Luiza Sayfullina, Alexander Grigorievsky, Jaakko Luttinen, Juho Kokkala, Pyry Takala, Mikael Kuusela, Anis Nasir, Oskar Kohonen, Pekka Parviainen, Pauli Miettinen, Mikko Tolonen, Michael Briga, Virpi Lummaa, Claudia Wagner, and Ingmar Weber.

To my parents and two sisters, I feel really blessed for having such a supportive and unconditionally loving family. Of the many things I've learned from you, what has probably inspired me academically the most are my mother's diligence and attention to detail and my father's courage and openness to other cultures. To my beloved daughter, Lumi, your enthusiastic and determined exploration of the world around you exemplifies what research is at its best and I'm really proud to be your dad. And to my beloved wife, Maria, thank you for bringing order to my life when my mind was too occupied with research, for being my practice audience and my Finnish copy-editor, and for your love expressed in so many ways.

Finally, I would like to thank God for everything.

[1] Thanks to Antonis for introducing me to *GeoGuessr.com*.

Zürich, May 3, 2018,

Eric Malmi

Contents

List of Publications

This thesis consists of an overview and of the following publications which are referred to in the text by their Roman numerals.

I Eric Malmi, Marko Rasa, and Aristides Gionis. AncestryAI: A tool for exploring computationally inferred family trees. In *Proceedings of the 26th International Conference on World Wide Web Companion (WWW 2017 Companion, Demo Track)*, Perth, Australia, pages 257–261, 2017.

II Eric Malmi, Aristides Gionis, and Arno Solin. Computationally Inferred Genealogical Networks Uncover Long-Term Trends in Assortative Mating. In *Proceedings of The 2018 Web Conference (WWW 2018)*, Lyon, France, 10 pages, 2018.

III Eric Malmi, Arno Solin, and Aristides Gionis. The blind leading the blind: Network-based location estimation under uncertainty. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2015)*, Porto, Portugal, pages 406–421, 2015.

IV Eric Malmi, Nikolaj Tatti, and Aristides Gionis. Beyond rankings: comparing directed acyclic graphs. *Data Mining and Knowledge Discovery*, 29, pages 1233–1257, 2015.

V Eric Malmi, Sanjay Chawla, and Aristides Gionis. Lagrangian relaxations for multiple network alignment. *Data Mining and Knowledge Discovery*, 31, pages 1331–1358, 2017.

VI Eric Malmi, Aristides Gionis, and Evimaria Terzi. Active Network Alignment: A Matching-Based Approach. In *Proceedings of the International Conference on Information and Knowledge Management (CIKM 2017)*, Singapore, pages 1687–1696, 2017.

Author’s Contribution

Publication I: “AncestryAI: A tool for exploring computationally inferred family trees”

Malmi developed the idea, designed the proposed inference method, and had the main responsibility in writing the article. The family-tree layout algorithm was developed and documented by Rasa.

Publication II: “Computationally Inferred Genealogical Networks Uncover Long-Term Trends in Assortative Mating”

Malmi developed the idea, wrote the first version of the manuscript, designed the proposed methods, and ran the experiments.

Publication III: “The blind leading the blind: Network-based location estimation under uncertainty”

Malmi and Solin jointly developed the idea. Malmi had the main responsibility in writing the article. He also designed the proposed methods and ran the experiments.

Publication IV: “Beyond rankings: comparing directed acyclic graphs”

Malmi, Tatti, and Gionis jointly developed the idea and wrote the paper. Malmi ran the experiments and designed the generalization of Kendall-tau for DAGs and the greedy method for DAG aggregation. The proof for Proposition 1 was derived by Tatti.

Publication V: "Lagrangian relaxations for multiple network alignment"

Malmi developed the idea together with Chawla, who proposed to combine the facility-location formulation with the multiple-network-alignment formulation. Malmi wrote the first version of the paper, derived the proofs for the theoretical results, designed the proposed methods, and ran the experiments.

Publication VI: "Active Network Alignment: A Matching-Based Approach"

Malmi developed the idea, ran the experiments, and had the main responsibility in writing the article. The TOPMATCHINGS method was designed by Malmi, while the GIBBSMATCHINGS method was designed jointly by Malmi, Terzi, and Gionis.

List of Abbreviations

CSS	computational social science
DAG	directed acyclic graph
ER	entity resolution
MLN	Markov logic networks
PSL	probabilistic soft logic

1. Introduction

This chapter presents the motivation and scope of this thesis, summarizes the main contributions of the thesis, and finally outlines the organization of the subsequent chapters and articles.

1.1 Motivation and Scope

It has been estimated[1] that data scientists spend 50 to 80 percent of their time on *data wrangling*, that is, preparing data for analysis, and only a minority of their time on the actual data analysis. Data wrangling encloses various tasks, but if the data originates from multiple sources, one important task is to identify the records or mentions in the data referring to the same entities—a problem known as *entity resolution*. While in practice this problem is often addressed by manual data curation, it can also be tackled by computational methods, which can save the data scientist from a significant amount of manual work if the number of records to be *matched* is high. Developing novel computational methods for entity resolution is the broad area that this thesis contributes to.

Traditionally, entity-resolution methods are presented with multiple questions which ask whether two records refer to the same entity, and the methods address these questions independently. However, it is common to have relational information about the records, which could improve the accuracy of the entity-resolution process if properly accounted for. Consider the following example:

> Person p goes to a conference and meets persons a and b who work at the same university. After the conference, p goes to Twitter and wants to start following a and b, but quick Twitter searches on a and b return multiple user profiles in both cases. After some investigative work, p discovers that there is one pair of search results, $\hat{a}$ and $\hat{b}$, whose names match with a and b but who also follow each other. Thus p infers that $\hat{a}$ and $\hat{b}$ must be the persons

[1] `https://www.nytimes.com/2014/08/18/technology/for-big-data-scientists-hurdle-to-insights-is-janitor-work.html`

that p met at the conference, and starts following them.

While perhaps not entirely realistic, the above example illustrates effectively the problem of *collective entity resolution*. In the example, the matching decisions are made between a recalled name and a Twitter profile returned when querying by the recalled name. If the matching decisions were made independently for a and b, it could be challenging to infer the correct Twitter profile among all the *candidate matches* returned by a search. However, the additional *relation information* (a and b are from the same university and thus more likely to follow each other than two random persons would be) can be used to help with the matching decisions. Methods that leverage such relational information are known as *collective entity resolution* methods (Bhattacharya and Getoor, 2007).

Relational information can come in many forms but one general way of modeling it is to assume that, in addition to being given two sets of records, we are given *networks*, connecting the nodes within each dataset. In this model, the nodes correspond to records and they are associated with attributes, such as name and age, and the records related to each other are connected by an edge. The above example can be presented within this model by considering one small input network, consisting of two adjacent nodes with name attributes, and one large network, consisting of the profiles of all Twitter users and the corresponding follower graph. The task is to align the nodes of the small input network to the nodes of the large network.

This problem is known as *network alignment* and it is an instance of collective entity resolution, although network alignment has been studied mostly in the context of aligning biological networks (Clark and Kalita, 2014; Elmsallati et al., 2016; Guzzi and Milenković, 2017; Meng et al., 2016). In our case, the problem is initially motivated by a task of aligning multiple partial *genealogical networks* (*family trees*). Most of the existing network-alignment methods are designed for aligning only two networks, which is why we take an existing *pairwise* network aligner with a good experimental performance and extend it to handle *multiple* networks. Furthermore, in some cases the attributes of the input-network nodes and the edges between the nodes are corrupted by so much noise that fully automatic alignment methods are insufficient—especially in the case of genealogical networks, where the node attributes are recorded by interpreting historical hand-written documents (see Figure 1.1 for an example of such a document). In such cases, it is necessary to involve human experts in the alignment process, which, unfortunately, is typically slow and costly. To tackle this problem, we propose an *active* network alignment method that tries to minimize the required human involvement by asking the most informative questions from the humans.

In addition to using network information as *input*, it is sometimes also desirable to *output* a network. This output network describes the inferred relations between the entities. To this end, the proposed multiple network alignment method has been designed to infer an *entity network* in addition to aligning the

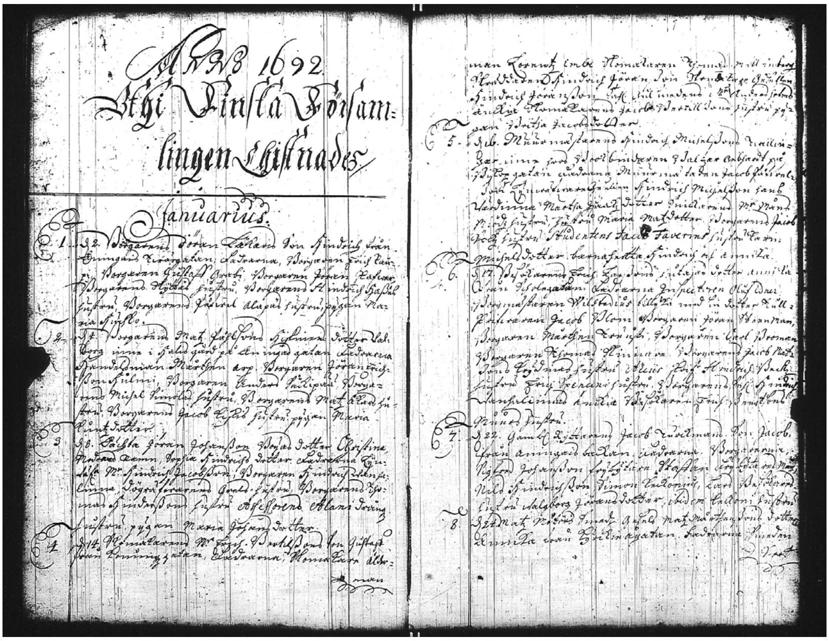

Anno 1692
Uthi Finska Försam-
lingen Christnade
Januarius.

Figure 1.1. A scanned image from the Finnish parish registers from the year 1692. Source: Finland's Family History Association (FFHA).

input-network nodes (records) to the entities. Moreover, we study the problem of *directed acyclic graph (DAG) aggregation*,[2] which tries to find a centroid DAG which is as similar to the input DAGs as possible. This problem is encountered, for example, when aggregating pairwise preferences from multiple people.

The aforementioned network-alignment and DAG-aggregation methods have various applications from genealogical networks to social networks (Zhang and Yu, 2015), protein-protein interaction networks (Singh et al., 2008), and information cascades (Malmi et al., 2015). However, we also propose a collective entity resolution method specifically for the problem of inferring genealogical networks. This method takes as input a large collection of vital records and tries to link each birth record to the birth records of the parents. The linking task is formulated as an optimization problem, which tries to link records with similar attributes and to also capture an intuition that people tend to have children with the same spouse. The method is applied to a dataset of 5 million historical birth records from Finland, from which it infers a large-scale genealogical network, containing a connected component of 2.6 million people. The accuracy of the linking method is assessed based on a human-constructed genealogical network.

[2] Terms *network* and *graph* are used interchangeably in this thesis. The reason for using both of these terms is that the former is predominantly used in the field of bioinformatics, where particularly the network-alignment problem has been studied extensively, while the latter term is more popular in theoretical computer science literature, which this thesis also builds upon.

Finally, we apply the inferred genealogical network to study a phenomenon known as *assortative mating* or *social homogamy*, referring to a tendency of people to marry a spouse with a similar socioeconomic status. This analysis addresses two questions: (i) can we detect assortative mating in an automatically inferred genealogical network, and (ii) how has the intensity of this phenomenon evolved over a time period of 150 years. Overall, we argue that large-scale genealogical networks, which can be inferred with the methods proposed in this thesis, can open up fundamentally new type of analysis opportunities in the field of *computational social science*. This field typically leverages data from online social media services, but since these services have generally existed for less than 15 years, they do not allow studying phenomena that take multiple human generations to occur or evolve. In contrast, genealogical data can cover multiple centuries or even millennia (Schich et al., 2014).

1.2 Contributions

The main contributions of this thesis are summarized as follows:

- Publication V generalizes Natalie (El-Kebir et al., 2015; Klau, 2009)—an accurate *pairwise* network alignment method—to handle *multiple* networks.
- Publication VI develops an *active* network alignment method, which leverages human experts (oracles) in the network alignment process. This method yields high alignment accuracies with fewer oracle queries than its competitors.
- Publication IV formalizes the problem of *aggregating directed acyclic graphs* and proposes algorithms with approximation guarantees. This problem asks to integrate networks with a known node correspondence but inconsistent edges.
- Publications I, II, and III propose methods for linking vital records to infer *large-scale genealogical networks*. These methods are applied to link a dataset of 5 million historical birth records from Finland and evaluated using a genealogical network constructed by a human genealogist.
- Publication I presents an open web service called AncestryAI (`http://ancestryai.cs.hut.fi/`), which allows making searches and exploring the automatically inferred genealogical network, thus facilitating the job of a genealogist.
- The inferred genealogical network has been used to analyze *assortative mating* (the tendency of people to marry someone with a similar status) over a period of 150 years in Finland. This analysis, presented in Publica-

tion II, reveals interesting long-term patterns in the strength of assortative mating.

Most of the source code and datasets used in this thesis can be accessed at: `https://github.com/ekQ/`. The code and data not available online, are available upon request from the author.

1.3 Organization

This thesis follows the format of an article-based dissertation, meaning that it consists of a set of articles, appended at the end of the thesis, and a compilation part (Chapters 1 to 6) that summarizes the studied problems and the main findings of the articles.

The compilation part is organized as follows: Chapter 2 provides an overview of the related work on entity resolution, which is the main topic of this thesis. Chapter 3 (based on Publications I, II, and III) presents two methods for linking vital records in order to infer genealogical networks. Additionally, it introduces a location-estimation method, which can be used for various types of networked data sources but which is primarily used to geolocate the vital records. Chapter 4 (Publications IV, V, and VI) discusses the problems of aggregating and aligning networks, whereas Chapter 5 (Publication II) presents an analysis of an inferred genealogical network and discusses the future analysis opportunities. Finally, conclusions are drawn in Chapter 6.

The publications are appended in a thematic order and they consist of two journal publications (IV, V) and four papers published in peer-reviewed conference proceedings (I, II, III, VI), the first of which is a demo paper.

2. Entity Resolution

Entity resolution (ER) is the problem of identifying which records refer to the same underlying entity. It is encountered in many applications where data from different sources need to be integrated or when a single dataset is cleaned from duplicate entries. The ER problem itself goes by various duplicate names, including *record linkage*, *data matching*, *duplicate detection*, and *reference reconciliation*.

The ER problem has been studied actively for several decades due to a large number real-world applications where it is encountered (Christen, 2012; Herzog et al., 2007). In this chapter, we present a non-comprehensive overview of the studies most related to the topics of this thesis.

2.1 Steps of the Entity-Resolution Process

Christen (2012) distinguishes five steps in the entity-resolution process, which are illustrated in Figure 2.1. Next we give a brief summary of these steps, assuming a context of matching two databases A and B.

Data pre-processing involves various steps needed to make the two databases comparable. For instance, the databases may have slightly differing schemas, where A includes person names in a single field, whereas B has a separate field for first name and last name. Furthermore, when dealing with the names of persons or other entities, it is often necessary to *normalize* each name to a standard form. This is particularly important when dealing with historical records that were kept by multiple different people. For example, a person whose name is *Eric*, might also be recorded as *Erik*, *Eerikki*, or *Erichus*.

Often it is infeasible to perform a detailed comparison for all $A \times B$ record pairs, which is why the number of pairs is reduced using *blocking* or *indexing*. This can be done by using a set of blocking keys, such as last name and postal code, and retrieving the records with matching blocking keys, using an inverted index. In more advanced ER approaches, we may have a vector-space representation for each record, which allows us to retrieve the k nearest records using a data structure such as k-d trees (Bentley, 1975). If our goal is to match record $a \in A$

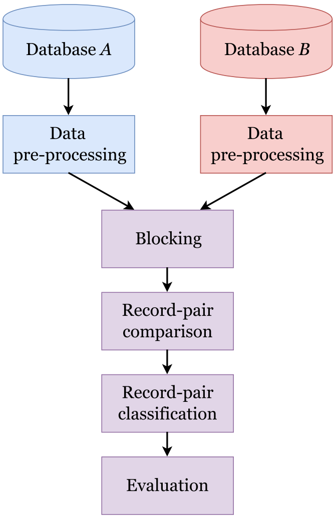


Figure 2.1. The steps of the entity resolution process according to Christen (2012).

to a record in B, we use the term *candidate matches of* a to refer to the set of possible matches in B obtained from the blocking step.

The record pairs resulting from the blocking step are then compared in more detail in the *comparison* step. For example, for a pair of locations, we may want to compute their geographical distance, and for a pair of names, a string similarity measure. Levenshtein distance, which measures the number of edits needed to transform one string to another, is a popular method for comparing general strings, but methods designed specifically for comparing person names have also been developed. One example of the latter is the Jaro–Winkler distance (Winkler, 1990), which is also used in this work.

In the next step, the record pairs are classified into *matching*, *non-matching*, and sometimes also into *possibly matching* pairs. This can be done in an unsupervised manner, using simple heuristics (for instance, compute the sum of similarity values and set a threshold for this sum) or using some prior knowledge about the problem to classify the record pairs probabilistically, using, for example, the Fellegi–Sunter method, which is described in the next section. Alternatively, if labeled pairs of records known to be matches or non-matches are available, supervised classification methods can be used. Often the record pairs are classified independently, which may lead into suboptimal results if we know, for instance, that neither A nor B contains duplicates, implying that we want to avoid matching two records to one. The main focus of this thesis is *collective* entity resolution, which considers multiple matching decisions jointly to improve the matching quality.

In the final step, the quality of the matched records is evaluated. If the ER problem is viewed as a classification task, standard evaluation measures for classifiers, such as *precision* and *recall*, can be used. Typically, the main challenge is obtaining a sufficiently large ground-truth dataset to compute these measures. In our work, we use several different approaches to overcome this challenge: Some datasets, like the multiplex dataset used in Publication V, contain a unique identifier for each record. In that case, we hide the identifier when performing ER and use it only in the evaluation phase. In other cases, we generate a semi-synthetic dataset by using a set of existing records A and then perturbing the attributes of these records to create another dataset B, keeping track of the original records that the instances of B correspond to. This approach is used to evaluate network-alignment methods on genealogical networks in Publication V and Publication VI. In Publication II, our task is to link birth records and evaluate the resulting genealogical network by comparing it to a ground-truth network constructed by a human. This ground-truth network covers a sizable fraction of the entities mentioned in the birth records, but unfortunately, the nodes of the network do not contain references to the corresponding birth records. This means that merely to evaluate the links inferred between the birth records, we need to solve yet another ER task of matching a ground-truth network to the inferred network. We take a conservative approach and consider a ground-truth-network node to be matched to a birth record only if they have the same first and last name, the birth dates are exactly the same, and there is only one matching node for the given birth record. This give us a set of ground-truth links whose both adjacent nodes are matched to a birth record, and these ground-truth links can then be used to evaluate the inferred links between birth records.

2.2 Fellegi–Sunter Method

The Fellegi–Sunter method for probabilistic record linkage (Fellegi and Sunter, 1969) is one of the most well-known and widely used ER methods (Christen, 2012). It considers two databases A and B, and tries to classify whether a record pair $(a,b) \in A \times B$ is a *matched* pair $(a,b) \in M$, an *unmatched* pair $(a,b) \in U$, or a *potentially matched* pair.

The decision on which class a record pair belongs to is based on a comparison vector $\boldsymbol{\gamma}$, which consists of binary comparisons $\gamma_1, \gamma_2, \ldots, \gamma_n$, such as "first names are the same," "cities are different," and "street name missing on one record." The likelihood of the comparison vector $\boldsymbol{\gamma}^{a,b}$ for record pair (a,b), given that the records are matched, is denoted by

$$m\left(\boldsymbol{\gamma}^{a,b}\right) = p\left(\boldsymbol{\gamma}^{a,b} \middle| (a,b) \in M\right),$$

and by

$$u\left(\boldsymbol{\gamma}^{a,b}\right) = p\left(\boldsymbol{\gamma}^{a,b} \middle| (a,b) \in U\right),$$

given that the records are unmatched. Fellegi and Sunter (1969) then propose the following linkage rule

$$d\left(\boldsymbol{\gamma}^{a,b}\right)=\begin{cases}(a,b)\text{ is } \textit{matched}, & \text{if } m\left(\boldsymbol{\gamma}^{a,b}\right)\Big/u\left(\boldsymbol{\gamma}^{a,b}\right)\geq T_\mu,\\ (a,b)\text{ is } \textit{potentially matched}, & \text{if } T_\lambda < m\left(\boldsymbol{\gamma}^{a,b}\right)\Big/u\left(\boldsymbol{\gamma}^{a,b}\right) < T_\mu,\\ (a,b)\text{ is } \textit{unmatched}, & \text{if } m\left(\boldsymbol{\gamma}^{a,b}\right)\Big/u\left(\boldsymbol{\gamma}^{a,b}\right)\leq T_\lambda,\end{cases}$$

where T_μ and T_λ are positive constants. The main result of Fellegi and Sunter (1969) is that this linkage rule, which is based on the ratio $m\left(\boldsymbol{\gamma}^{a,b}\right)\Big/u\left(\boldsymbol{\gamma}^{a,b}\right)$, is optimal in the sense that among all the linkage rules with a fixed false-positive rate μ and a fixed false-negative rate λ, this rule has the lowest probability of assigning a record pair into the *potentially matched* category. In other words, this linkage rule has the best discriminatory power for classifying record pairs into matches and non-matches. However, this result assumes that we can accurately estimate likelihood terms $m(\boldsymbol{\gamma})$ and $u(\boldsymbol{\gamma})$, which is not straightforward.

More specifically, there are two challenges when estimating the likelihood terms:

1. Records with the same name are much more likely to be a match if the name is rare (for example, *Jezebel*) compared to if the name is a popular one (for example, *Mary*). To encounter this, Fellegi and Sunter propose to adjust the likelihoods based on name frequency.

2. It is common to assume that the different comparison vector terms are conditionally independent (Christen, 2012), that is

$$m\left(\boldsymbol{\gamma}^{a,b}\right)=\prod_i p\left(\gamma_i^{a,b}\Big|(a,b)\in M\right),$$

   and similarly for $u\left(\boldsymbol{\gamma}^{a,b}\right)$. However, this independence assumption does not often hold in practice.

Instead of categorizing record pairs into the three categories used in the Fellegi–Sunter method, we take a slightly different approach in Publications I and II; we find it more natural to estimate $p\left(M_a=b\middle|\boldsymbol{\gamma}^{a,C_a}\right)$, that is, the probability of $b\in B$ being the matching record for $a\in A$ (assuming that B does not contain duplicates). Term $\boldsymbol{\gamma}^{a,C_a}$ denotes the set of comparison vectors between a and its candidate matches $C_a\subseteq B$. The candidate matches are obtained from a blocking step, which uses name and potentially other blocking keys. M_a is a random variable corresponding to the matching record of a and thus it takes values from $C_a\cup\emptyset$, where $\emptyset$ refers to the case that the candidate matches C_a do not contain a matching record for a. Later, in Section 3.1, we show that, using the Bayes' rule, we can derive the following expression for the matching probability

$$p\left(M_a=b\middle|\boldsymbol{\gamma}^{a,C_a}\right)=\frac{p(M_a=b)\,m\left(\boldsymbol{\gamma}^{a,b}\right)\Big/u\left(\boldsymbol{\gamma}^{a,b}\right)}{\sum_{c\in B\cup\emptyset}p(M_a=c)\,m\left(\boldsymbol{\gamma}^{a,c}\right)\Big/u\left(\boldsymbol{\gamma}^{a,c}\right)}. \tag{2.1}$$

Interestingly, this approach is closely connected to the Fellegi–Sunter method since it incorporates the same likelihood ratio $m\left(\boldsymbol{\gamma}^{a,b}\right) \big/ u\left(\boldsymbol{\gamma}^{a,b}\right)$ used by the Fellegi–Sunter method, however, normalizing it by summing over the candidate matches of a. This approach naturally addresses the first challenge regarding name popularity, since a record with a rare name will have only a few candidate matches with the same name, giving the few candidates a higher matching probability due to a smaller denominator in Equation (2.1).

To address the second challenge, regarding the independence assumption, we make an observation that likelihood ratios can be estimated with probabilistic discriminative classifiers when training data is available (Cranmer et al., 2016). This allows us to use any discriminative classifier, such as Support Vector Machines (Cortes and Vapnik, 1995) or XGBoost (Chen and Guestrin, 2016), as long as the classifier outputs probabilities. These discriminative classifiers typically do not make the conditional independence assumption but capture some of the correlations between features. Using an off-the-shelf classifier also makes it easier to add new comparison-vector features, since a separate probability distribution does not need to be estimated for each new feature. There are several previous works which also employ discriminative classifiers for entity resolution (Christen, 2008; Cochinwala et al., 2001; Lian and Xie, 2016; Tay et al., 2016), but we are not aware of previous works using discriminative classifiers for estimating the ratio $m\left(\boldsymbol{\gamma}^{a,b}\right) \big/ u\left(\boldsymbol{\gamma}^{a,b}\right)$.

2.3 Collective Entity Resolution

While traditional ER methods typically classify record pairs independently, collective entity resolution methods aim at matching multiple records jointly. This makes it possible to capture *transitivity constraints*, that is, if we know that $a = b$ and $b = c$, then, by transitivity, records a and c should also refer to the same entity. Moreover, by jointly matching records known to be related, information from one node can be propagated to another node. A classic example of collective entity resolution is the problem of author disambiguation in bibliographic datasets (see, for example, Roy et al., 2013): given a collection of paper citations, disambiguate the authors listed in the citations in order to discover the unique authors (entities) and assign the correct set of papers to each author. In this example, if a group of matching author names appears in multiple citations, it is likely that these author names refer to the same entities.

Next we go through a few popular works on collective entity resolution and then present two general frameworks for collective entity resolution, namely, methods based on *first-order logic* and methods based on *network alignment*.

Bhattacharya and Getoor (2006) propose a generative method based on Latent-Dirichlet Allocation for collective ER. This method assumes that entities form groups and we observe co-occurrence data which is generated as mixtures of groups. This setup fits a rather specific class of collective ER tasks but it is suited particularly well for author disambiguation tasks and thus is shown to perform well on those.

Another work by Bhattacharya and Getoor (2007) proposes a more general method for clustering the records referring to the same entity, leveraging relational information represented as a *reference graph*. Similarity between two clusters is defined as a linear combination of their attribute similarity and the similarity of their neighborhoods. The method merges clusters agglomeratively in a greedy fashion, updating the similarity scores after each merge.

While the aforementioned method (Bhattacharya and Getoor, 2007) models data as a reference graph, Dong et al. (2005) consider a *dependency graph* whose nodes correspond to matching decisions and edges to dependencies between these decisions. This method supports matching multiple record types, such as *authors*, *papers*, and *venues*, jointly. The authors propose an iterative approach that propagates information between the nodes.

A Markov logic network (MLN) (Richardson and Domingos, 2006) is a popular model, which combines first-order logic and undirected graphical models. It can be applied to collective ER (Singla and Domingos, 2006) by using first-order logic to encode the transitivity requirement for entities and rules such as *duplicate papers should have the same venue* and *duplicate papers should have the same title* (Singla and Domingos, 2006). Furthermore, the MLN framework allows learning weights for different rules based on training data, which makes it possible to automatically learn, for example, that having a matching title is a stronger evidence for a duplicate record than having a matching venue. The goal is to find a solution which maximizes the sum of the weights of satisfied rules. Probabilistic soft logic (PSL) is a related approach which similarly uses first-order logic rules as a template language for graphical models (Kimmig et al., 2012). Recently, also PSL has been applied to collective ER (Kim et al., 2017; Kouki et al., 2017). An advantage of PSL compared to MLN is that in PSL the rules take soft truth values in the range $[0,1]$, which makes it easier to incorporate similarities between attributes (Kimmig et al., 2012; Kouki et al., 2017).

From the perspective of collective entity resolution, the main focus of this thesis is on the problem of matching two or more networks of records with the goal of preserving the neighborhood relations between records as well as matching records with similar attributes. This problem is known as *network alignment*, which is discussed next.

2.4 Network Alignment

The network-alignment problem asks to find a matching between the nodes of a source network $G_s = (V_s, E_s)$ and a target network $G_t = (V_t, E_t)$. Each node should be matched to at most one node in the other network, and a high-quality matching should satisfy the following properties:

(i) The aligned nodes should have as similar attributes as possible.

(ii) The neighborhoods of the aligned nodes should be structurally similar.

This problem is related to the *subgraph isomorphism* problem which is NP-complete. However, subgraph isomorphism considers only the second property, asking whether a graph and a subgraph are strictly isomorphic, whereas network alignment typically tries to maximize a structural similarity score while aligning nodes with similar attributes. Often, the structural similarity score is based on the number of conserved edges (i.e. adjacent nodes aligned to adjacent nodes), but some work has also been done on conserving higher-order structures, such as triangles (Mohammadi et al., 2017).

Network alignment can be seen as an instance of collective entity resolution where the goal is to match two networks whose nodes correspond to records and edges to relations between these records. This entity-resolution problem is encountered, for example, when aligning social networks (Goga et al., 2015; Zhang and Yu, 2015), ontologies (Noy et al., 2000; Sarasua et al., 2012), or genealogical networks (Kouki et al., 2016, 2017; Malmi et al., 2017a,b). However, probably most of the studies on network alignment have been conducted outside the entity-resolution domain, namely, in biology, on the problem of aligning protein-protein interaction networks (see, for example, Chindelevitch et al., 2010; Clark and Kalita, 2014; Elmsallati et al., 2016; Flannick et al., 2006, 2009; Guzzi and Milenković, 2017; Hashemifar and Xu, 2014; Hu et al., 2013; Klau, 2009; Kuchaiev et al., 2010; Kuchaiev and Pržulj, 2011; Liao et al., 2009; Malod-Dognin and Pržulj, 2015; Sahraeian and Yoon, 2013; Singh et al., 2008).

Network-alignment methods can be divided into *local* and *global* methods (Meng et al., 2016; Elmsallati et al., 2016). Local methods match small subgraphs independently, meaning that they do not necessarily produce a one-to-one alignment, whereas global methods often formulate a single global optimization problem, which tries to capture the two properties mentioned in the beginning of this section and which requires that the solution is a matching. In this thesis, we only study global methods.

Another way to categorize network-alignment methods is based on whether they are designed for *pairwise* alignment or for *multiple* network alignment. Most existing methods are pairwise, but several methods have also been proposed for multiple network alignment, including IsoRankN (Liao et al., 2009), Græmlin (Flannick et al., 2006, 2009), SMETANA (Sahraeian and Yoon, 2013), and multiMAGNA++ (Vijayan and Milenković, 2017). Among the pairwise methods, Natalie (El-Kebir et al., 2015; Klau, 2009) has performed well in several comparisons (Bayati et al., 2013; Clark and Kalita, 2014; El-Kebir et al., 2015). Natalie formulates the network-alignment problem as a quadratic integer program and solves it using a Lagrangian relaxation (Fisher, 1981), which reduces the original network-alignment problem into multiple *maximum-weight bipartite matching* problems (El-Kebir et al., 2015; Klau, 2009).

In Publication V, we propose several extensions of Natalie to the multiple network alignment problem. These extensions outperform a naïve extension of Natalie as well as IsoRankN.

2.5 Active Learning for Entity Resolution

In *supervised learning*, labeled training data is needed to learn model parameters. The goal of an *active learning* system is to pick the data samples to be labeled in order to achieve a higher accuracy with fewer labeled training samples (Settles, 2010). The selected samples are *queried* by asking an *oracle* (for example, a human expert) to reveal their correct label. Such systems are particularly useful in scenarios where labeled data is scarce and possibly costly to obtain. A common approach for selecting the samples to query is to select the most *uncertain* samples.

Many entity resolution methods are based on training a classifier to predict whether a pair of records is a match or not, but often labeled record pairs are not available in large amounts, which is why several active learning approaches have been developed for entity resolution problems (Christen, 2012; Firmani et al., 2016; Fisher et al., 2016; Sarawagi and Bhamidipaty, 2002; Verroios et al., 2017). The focus of this thesis is on collective entity resolution where the matching decisions are interdependent. This brings a new dimension to the query selection, since knowing whether records a and b are a match or not can directly give us information about the correct matches for the records related to a and b.

Specifically, we focus on the problem of *active network alignment*, where the goal is to query the most informative nodes in order to maximize the alignment accuracy of the remaining nodes. We assume that a similarity function between record attributes is given (or that a record pair classifier has already been trained) and, therefore, active network alignment is strictly speaking not an instance of active learning as we are not learning a model. Rather, it is an instance of *active inference*, which focuses on minimizing user interaction cost while maximizing the benefit and performance of the system (Bilgic and Getoor, 2010).

Actively selecting the nodes to query is a challenging computational problem since quantifying the uncertainty of a node to be matched cannot be done simply based on node-similarity scores since those do not capture relational information. Active network alignment methods have been previously studied mainly in the context of active ontology alignment (Jiménez-Ruiz et al., 2012; Paulheim et al., 2013; Sarasua et al., 2012; Shi et al., 2009). These approaches usually focus on queries which ask whether two nodes are the same or not, whereas we consider queries which ask for the best-matching candidate node for the query node. We argue that the latter type of *relative* queries can be easier for a human expert to answer. The active network problem is discussed in more detail in Section 4.3

and Publication VI.

2.6 Entity Resolution with Genealogical Data

Performing entity resolution for birth, marriage, and other vital records was discussed already in 1946 by Halbert L. Dunn who proposed an idea of creating a *Book of Life* for each person, starting from the person's birth and ending in the person's death (Dunn, 1946). More recently, the problem of inferring genealogical networks, also known as *population reconstruction* (Bloothooft et al., 2015), has received fairly lot of attention (see, for example, Efremova et al., 2015; Christen, 2016; Christen et al., 2015; Kouki et al., 2016; Ranjbar-Sahraei et al., 2015) as the number of indexed genealogical datasets has grown thanks to online resources curating genealogical data such as *Geni.com* and the *WikiTree* project. Outside the entity resolution domain, the inference of family relationships has been studied by Backstrom and Kleinberg (2014), who present a method for spouse inference based on the social network of a person.

Efremova et al. (2015) look into the problem of linking records from multiple genealogical datasets. Similar to our methods discussed in Chapter 3, they cast the problem into supervised binary classification tasks. They find name popularity, geographical distance, and co-reference information to be important features, whereas in our approach, we can avoid having to explicitly model name popularity, since the probability of a candidate parent is normalized over the set of all candidate parents and for popular names this set will be large, thus downweighting the candidate probability.

Christen (2016) proposes a collective method for linking birth, death, marriage, and census records. The linking results obtained using Scottish data are concluded to be inferior to a linkage constructed by a domain expert due to many intrinsically difficult linking cases. We argue that even in such challenging scenarios automatic linking methods can be useful if they are probabilistic. Probabilistic methods, such as the one presented in Chapter 3, can handle some of the matching decisions automatically based on matching probabilities.

Kouki et al. (2016) propose a classification-based approach for integrating multiple partial views of a genealogical network. Recently, they have extended this work to use collective inference based on PSL (Kouki et al., 2017). The partial views are ego-centric networks with first-degree relationships (spouses, children, etc.) and second-degree relationships (grandparents, aunts, etc.). The problem is thus closely related to the problem of linking vital records, which is discussed in Chapter 3 and Publication II. A key advantage of their approach is the flexibility provided by the PSL framework; if a user wants to add a new relational rule to the model, the inference method does not need to be updated as long as the rule can be presented using first-order logical syntax. On the other hand, we show that one of the methods proposed in this work not only provides accurate matches but also reliably quantifies the certainty of the matches—an important feature of a practical entity-resolution system. Kouki et al. (2017) perform their experiments on datasets that are two orders of magnitude smaller than the vital-record dataset used in this work, but they show experimentally that the PSL approach scales almost linearly with the number of record pairs. Performing an experimental comparison between the PSL approach and the proposed methods remains an exciting future direction.

3. Linking Vital Records

Linking historical vital records, including birth, marriage, and death records, is a common and one of the earliest applications of entity resolution (Christen, 2012; Dunn, 1946). The difficulty of matching records that are dated before the adoption of national identification numbers lies in duplicate names, spelling variations, errors in the data, and missing records. Therefore, robust computational methods, which can leverage various attributes of the records, are needed.

By linking the birth record of each person to the birth records of the parents, it is possible to infer large-scale genealogical networks. The results can help a genealogist to speed up the process of reconstructing his or her own family history by identifying some ancestors automatically and by providing the most probable candidate ancestors for the more uncertain cases to narrow down the search.

In this chapter, we introduce two methods for linking birth records. In Section 3.1, we present a non-collective approach, which links the records independently and outputs probability distributions. A collective extension, which links the records jointly and output only a *maximum a posteriori* (MAP) estimate, is presented in Section 3.2. In Section 3.3, we propose a method for inferring the geolocations of historical place names mentioned in vital records, which allows us to measure the geographical distance between two record location—an important feature for determining whether the names mentioned in the records refer to the same person. Finally, in Section 3.4, we present a brief experimental evaluation of the proposed linking methods.

Problem definition. We define a *genealogical network* as a directed graph whose nodes correspond to people and edges to family relationships between the people. Two type of edges are considered: *father* edges, going from a father to a child, and *mother* edges, going from a mother to a child. Each node can have at most one biological father and mother, but the parents are not necessarily known. Because of the temporal ordering of the nodes, this graph is a *directed acyclic graph* (DAG). Each person in the graph is represented by the person's birth record. Given a set of birth records V, the objective of the genealogical network inference is to link each birth record $a \in V$ to the birth records of the person's mother $M_a \in V$ and father $F_a \in V$.

To avoid having to evaluate a quadratic number of parent–child combinations, the inference method gets as input a set of mother candidates $C^{\mathrm{m}} \subseteq V$ and a set of father candidates $C^{\mathrm{f}} \subseteq V$ for each person. These candidate sets are the results of a *blocking* step, which defines the candidates as the people who were born between 10 and 70 years before the child and whose normalized first and last name are equal to the normalized first and last name of a parent mentioned in the child's record.[1]

Since the true parents are often ambiguous, the output should ideally be a probability distribution over different parent candidates, including the case '$\emptyset$' that a parent is not among the candidates.

3.1 Non-Collective Approach

The goal is to find a birth record that matches to a parent mentioned in another birth record. These parent matches are assumed to be independent so that, for instance, the inferred father of a person does not affect the inference of the person's mother. Nevertheless, we assume that there are no duplicate birth records, meaning that the sum of probabilities over the candidate mothers C_a^{m} or over the candidate fathers C_a^{f} of person a should equal to $1 - p(\emptyset)$, where $p(\emptyset)$ is the probability that the true parent is not among the retrieved candidates.

Similar to the Fellegi–Sunter method presented in Section 2.2, we first construct a *comparison vector* $\boldsymbol{\gamma}^{a,m}$ for each child–parent pair (a, m). The vector consists of attribute similarity features based on, for example, names and locations mentioned in the records. The set of comparison vectors between a and the candidate mothers of a is denoted by $\boldsymbol{\gamma}^{a,C_a^{\mathrm{m}}}$. Since likelihood $p\left(\boldsymbol{\gamma}^{a,m} \middle| M_a = m'\right)$ depends only on whether $m = m'$ or not, the joint likelihood of the comparison vectors is given by

$$\begin{aligned}
p\left(\boldsymbol{\gamma}^{a,C_a^{\mathrm{m}}} \middle| M_a = m\right) &= \prod_{m' \in C_a^{\mathrm{m}} \cup \emptyset} p\left(\boldsymbol{\gamma}^{a,m'} \middle| M_a = m\right) \\
&= p\left(\boldsymbol{\gamma}^{a,m} \middle| M_a = m\right) \prod_{m' \in (C_a^{\mathrm{m}} \cup \emptyset) \setminus m} p\left(\boldsymbol{\gamma}^{a,m'} \middle| M_a \neq m'\right) \\
&= \frac{p\left(\boldsymbol{\gamma}^{a,m} \middle| M_a = m\right)}{p\left(\boldsymbol{\gamma}^{a,m} \middle| M_a \neq m\right)} \prod_{m' \in C_a^{\mathrm{m}} \cup \emptyset} p\left(\boldsymbol{\gamma}^{a,m'} \middle| M_a \neq m'\right) \\
&= \alpha \frac{p\left(\boldsymbol{\gamma}^{a,m} \middle| M_a = m\right)}{p\left(\boldsymbol{\gamma}^{a,m} \middle| M_a \neq m\right)},
\end{aligned}$$

where the product term α is constant with respect to m. For fathers, the likelihood is identical apart from replacing M_a by F_a and C_a^{m} by C_a^{f}.

[1] The names are normalized based on a clustering of Finnish/Swedish names obtained from the Genealogical Society of Finland. The clustering has been automatically extended by assigning non-clustered names to the nearest existing cluster or to a new cluster based on Jaro-Winkler name similarities (Winkler, 1990). The developed name normalization tool is available at: `https://github.com/ekQ/historical_name_normalizer`

Now the probability of candidate parent m can be derived using the Bayes' rule as follows

$$\begin{aligned} p\left(M_a = m \middle| \boldsymbol{\gamma}^{a,C_a^{\mathrm{m}}}\right) &= \frac{p\left(M_a = m\right) p\left(\boldsymbol{\gamma}^{a,C_a^{\mathrm{m}}} \middle| M_a = m\right)}{\sum_{m' \in C_a^{\mathrm{m}} \cup \varnothing} p\left(M_a = m'\right) p\left(\boldsymbol{\gamma}^{a,C_a^{\mathrm{m}}} \middle| M_a = m'\right)} \\ &= \frac{p\left(M_a = m\right) \frac{p(\boldsymbol{\gamma}^{a,m} | M_a = m)}{p(\boldsymbol{\gamma}^{a,m} | M_a \neq m)}}{\sum_{m' \in C_a^{\mathrm{m}} \cup \varnothing} p\left(M_a = m'\right) \frac{p(\boldsymbol{\gamma}^{a,m'} | M_a = m')}{p(\boldsymbol{\gamma}^{a,m'} | M_a \neq m')}}, \end{aligned} \tag{3.1}$$

where $p\left(M_a = m\right)$ is the prior probability of m. The prior probabilities are set uniformly over all candidates apart from the case '$\varnothing$' (none of the candidates is the correct parent) whose probability is learned from ground-truth data. As discussed earlier in Section 2.2, the above Equation (3.1) is related to the famous Fellegi–Sunter method, which classifies record pairs based on the ratio of comparison vector likelihoods. Here the same likelihood ratio is normalized by summing over the candidate parents.

In our first method, introduced in Publication I, we make an independence assumption between the different elements of the comparison vector $\boldsymbol{\gamma}^{a,m}$, that is

$$p\left(\boldsymbol{\gamma}^{a,m} \middle| M_a = m\right) = \prod_i p\left(\gamma_i^{a,m} \middle| M_a = m\right).$$

Five attribute similarity features are included in the comparison vector: a Jaro–Winkler name similarity for first names, last names and patronyms, and the age difference as well as the birth place distance between the child and the parent. For each feature i, we estimate the distribution of similarity values for both matching pairs of records $p\left(\gamma_i^{a,m} \middle| M_a = m\right)$ and for non-matching pairs of records $p\left(\gamma_i^{a,m} \middle| M_a \neq m\right)$ based on ground-truth data. Because of the attribute-independence assumption, this method is called NAIVEBAYES.

In real life, the similarity attributes are often not independent, which is why we also consider a second method that does not make this assumption. This method is based on the observation that likelihood ratios can be approximated with probabilistic discriminative classifiers (Cranmer et al., 2016). The most straightforward way to do this is to approximate

$$\frac{p\left(\boldsymbol{\gamma}^{a,m} \middle| M_a = m\right)}{p\left(\boldsymbol{\gamma}^{a,m} \middle| M_a \neq m\right)} \approx \frac{s\left(\boldsymbol{\gamma}^{a,m}\right)}{1 - s\left(\boldsymbol{\gamma}^{a,m}\right)},$$

where $s\left(\boldsymbol{\gamma}^{a,m}\right) \in [0, 1]$ is the output of a probabilistic binary classifier trained to separate matching comparison vectors from non-matching ones on a balanced dataset. We use XGBoost (Chen and Guestrin, 2016) as the classifier s since it has recently been successfully employed to another record linkage task (Lian and Xie, 2016; Tay et al., 2016). This approach is called BINCLASS and it uses a set of 20 features which are described in Publication II.

In addition to not having to assume feature independence, a key advantage of BINCLASS is that it is very easy to add new features to the comparison vector and retrain the model, whereas in NAIVEBAYES we have to estimate two

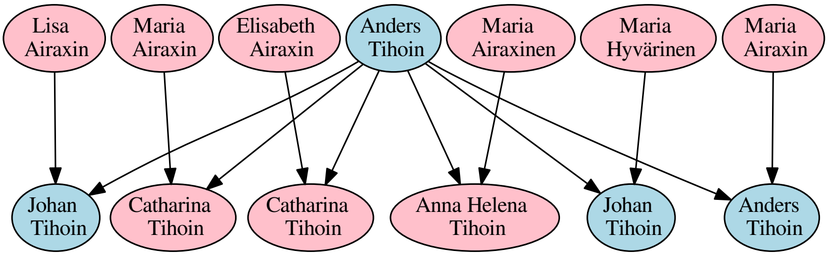


Figure 3.1. A sample subnetwork inferred by NAIVEBAYES. If each child is matched to the most probable father–mother pair independently, the number of spouses per person can be unrealistically high, which is the case for *Anders Tihoin*. Figure taken from Publication II.

new likelihood distributions for each new feature, manually selecting a suitable distribution based on the type of the feature. Moreover, we can add features such as the year of birth, which alone would not be helpful when inferring the most likely parent, but which could be informative jointly with other features. On the other hand, handling missing data is more straightforward with NAIVEBAYES, since if feature j is missing from the record of child a or parent m, we can simply set $p\left(\gamma_j^{a,m} \middle| M_a = m\right) \big/ p\left(\gamma_j^{a,m} \middle| M_a \neq m\right) = 1$, whereas BINCLASS requires us to impute the missing value.

3.2 Collective Approach

NAIVEBAYES and BINCLASS, presented in the previous section, assume that family links can be inferred independently, which can lead to some unlikely outcomes. In particular, the number of spouses per person can become unrealistically high as illustrated in Figure 3.1, which shows a subgraph of the network inferred by NAIVEBAYES. A person called *Anders Tihoin* has been inferred as the father of six children, while each child has been assigned a different mother. Although possible, it is very likely that at least the mothers of *Catharina* (the left one), *Anna Helena*, and *Anders* are actually the same person, since the mother name attribute of these children is almost the same *(Maria Airaxin(en))*.

To address this problem, we propose to minimize the number of mother–father pairs in addition to maximizing the probability of the inferred links. Let $y_{m,f} \in \{0,1\}$ indicate whether at least one child has been assigned to the mother–father pair (m,f) and $x_{a,m}, x_{a,f} \in \{0,1\}$ whether person a is linked to mother m and to father f, respectively. The collective genealogical network inference problem can now be written as

$$\begin{aligned}\max_{x,y} \quad & \Big[-\lambda \sum_{m,f} y_{m,f} + \sum_{a,m} \log p\left(M_a = m \middle| \boldsymbol{\gamma}^{a,C_a^{\mathrm{m}}}\right) x_{a,m} \\ & + \sum_{a',f} \log p\left(F_{a'} = f \middle| \boldsymbol{\gamma}^{a',C_{a'}^{\mathrm{f}}}\right) x_{a',f} \Big] \qquad (3.2)\end{aligned}$$

$$\text{such that} \quad \sum_a x_{a,m} x_{a,f} \le y_{m,f}, \quad \text{for all } m, f, \tag{3.3}$$

$$\sum_m x_{a,m} = 1, \quad \sum_f x_{a,f} = 1, \quad \text{for all } a, \tag{3.4}$$

$$x_{a,m}, x_{a,f}, y_{m,f} \in \{0,1\}, \quad \text{for all } a, m, f, \tag{3.5}$$

where $\lambda \geq 0$ controls the penalty induced by each extra parent pair (or the discount for merging two parent pairs into one).

This optimization problem is an instance of an *uncapacitated facility-location problem*, where parent pairs correspond to *facilities*, child nodes to *demand sites*, and the parameter λ to a *facility opening cost*. The uncapacitated facility-location problem is NP-hard for general graphs, and thus we employ a greedy approach, which starts with the children with the highest maximum link probability. Then it goes through the children one-by-one, assigning each child to the most probable parent pair unless one of the already used parent pairs is more beneficial due to discount λ (for more details, see Algorithm 1 in Publication II). Probabilities p, which are part of the input to the algorithm, are computed with BINCLASS. Note that since these probabilities do not necessarily satisfy the properties of a metric, the approximation guarantees for methods, such as the one proposed by Jain and Vazirani (2001), do not necessarily hold in this setting.

This method is called COLLECTIVE. It outputs a genealogical network, where some of the links inferred by BINCLASS have been rewired to reduce the number of spouses. A limitation of COLLECTIVE is that it does not output link probabilities. In theory, the link probabilities (that is, the marginal probabilities of random variables corresponding to each parent) could be estimated using a Markov chain Monte Carlo (MCMC) method that would sample genealogical networks by proposing swaps to the parent assignments. However, to reach a reasonable level of precision, we would need to sample a very large number of networks, making the MCMC approach impractical.

3.3 Record Geolocation

If a person has two equally likely parent candidates, except that the first is born 500 kilometers from the person's own birth place and the second is born in a neighboring village, it is very likely that the latter is the true parent, since people's mobility was more restricted in the past centuries. Therefore, it is important to be able to estimate the geographical distances between the locations mentioned in the vital records.

The birth records found in the Finnish parish registers, which are our main data source, contain always the name of the parish (roughly corresponding to a city), most of the time the name of the village, and occasionally also the name of the house where the child was born. To geolocate parishes and villages, we can query them from a database of geolocated contemporary Finnish place names

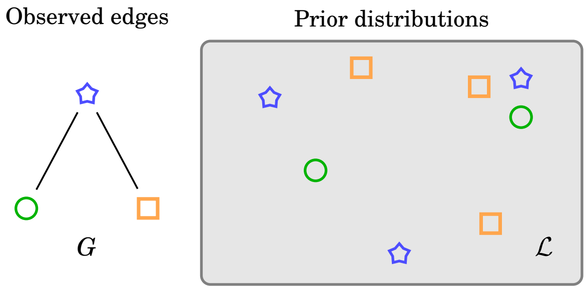


Figure 3.2. A simple example of the GeoLocation problem with three items to be located, and a set of discrete candidate locations. Figure taken from Publication III.

collected by the National Land Survey of Finland. However, there are two challenges: (i) place names may have changed, and (ii) there are often many potential matches with a duplicate name, especially for villages.

In fact, geolocating the historical parish and village names can be seen as a *geographical collective entity resolution* problem (Sehgal et al., 2006), where a set of input parish names with associated member village names need to be matched against another dataset of geolocated city and village names. The key information to be leveraged from the input data is that a parish and its member villages should be located nearby.

To model this problem, we formulate the following general network-based location estimation problem.

Problem 1 (GeoLocation)**.** *Consider a graph $G = (V, E)$ over items V, and a set of candidate locations $\mathcal{L}$. For each item $u \in V$ we are given a prior distribution $p(\ell(u))$ over its candidate locations. The goal is to infer a mapping $\boldsymbol{\ell}$ of items to their candidate locations in order to maximize the likelihood $p(E \mid \boldsymbol{\ell})$ of observing the edges of the graph given the inferred locations.*

The graph edges E are assumed to be independent and an edge probability $p(\{u,v\} \in E \mid \ell(u), \ell(v))$ is assumed to depend only on the distance between items u and v so that the probability decreases the farther the items are. Figure 3.2 presents a simple toy instance of this problem with three items (○, ☆, and □), each having two to three candidate locations. The maximum-likelihood estimates of the locations are found in the upper-right corner, where the distances between adjacent items are minimized.

Problem 1 covers several different geolocation tasks, and previously, a similar formulation has been applied to geolocating users of a social network (Backstrom et al., 2010; Jurgens, 2013; McGee et al., 2013). Our main contribution is to generalize the methods by Backstrom et al. (2010) and by Jurgens (2013) by incorporating prior distributions into the formulation and keeping track of distributions rather than point estimates when iteratively updating location estimates. This generalization outperforms the two previous methods, both in the case of geolocating social-network users and in the case of geolocating vital

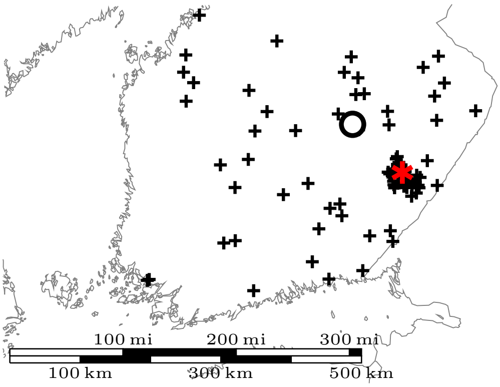


Figure 3.3. Location of a historical parish called *Sääminki* as identified by our method (✱) and the location of an island called Sääminki in contemporary Finland (○). Candidate locations for the member villages of Sääminki are shown by black crosses. Figure taken from Publication III.

records, as shown in Publication III.

As an anecdotal evidence for the usefulness of this method, we show the estimated location of a parish called *Sääminki* in Figure 3.3. A city with this name does not exist in contemporary Finland but there used to be a municipality called Sääminki that nowadays belongs to the city of Savonlinna. The estimated parish location correctly falls into Savonlinna, whereas the database of contemporary place names only contains an island called Sääminki, located outside Savonlinna. The island might be incorrectly inferred as the parish location if the location estimation is done by naïvely querying the database.

3.4 Experimental Evaluation

Next we evaluate the three genealogical network inference methods introduced earlier in this chapter: NaiveBayes, BinClass, and Collective. These methods are applied to a dataset[2] of 5.0 million birth records from Finland, 1650–1917. Additionally, a set of 18 731 ground-truth parent–child links has been obtained by matching a large, manually constructed genealogical network to the birth records. These links are split into training and test data to fit the models and to evaluate their performance, respectively.

The accuracy of the top-1 links inferred by different methods are shown in Table 3.1. BinClass clearly outperforms NaiveBayes and a random baseline that randomly picks one of the candidate parents. Collective further improves the accuracy of BinClass from 61.6% to 65.1%.

To evaluate the accuracy of the link probabilities estimated by BinClass, we bin the ground-truth links by their probability and compute accuracy within each bin. The results presented on the left-hand side of Figure 3.4 show a strong

[2] The dataset can be queried at: `http://hiski.genealogia.fi/hiski?en`

Table 3.1. Accuracy of the links inferred with different methods.

Method	RANDOMCAND	NAIVEBAYES	BINCLASS	COLLECTIVE
Accuracy	12.5%	56.9%	61.6%	**65.1%**

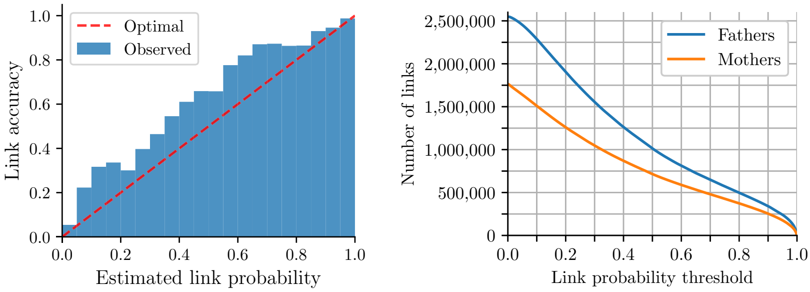


Figure 3.4. Left: The link-probability estimates by BINCLASS correlate strongly with the accuracy of the links binned by their probability. Right: The number of inferred family links with the estimated link probability above a given threshold. Figures taken from Publication II.

correlation between the estimated and the actual link probabilities.[3] This means that we can leverage the link probabilities to assess the reliability of a link and, for instance, to filter out all links below a certain probability threshold as done later in Section 5.1. However, the increased precision obtained by filtering out links comes with the cost of decreased recall. This trade-off is illustrated on the right-hand side of Figure 3.4, which shows the number of links above a given probability threshold. Even though the number of links considerably decreases when increasing the threshold close to 100%, we are still left with 253 814 child–mother links and 341 010 child–father links when using a threshold of 90% as done in Section 5.1.

[3] A limitation of the evaluation results presented here is that our ground-truth child–parent links contain only children whose parent is known, whereas in practice, every collection of birth records contains some people whose parent records are not in the collection. Therefore, the measured link accuracies shown in Figure 3.4 (left) are not representative if there are lots of missing records in the collection of birth records to be linked. Nevertheless, if we have an estimate for the fraction of records with a missing parent record, it can be incorporated into the probability of the case '$\varnothing$,' which is otherwise estimated only based on the fraction of children whose ground-truth parent is included in the dataset but not among the retrieved candidates.

4. Network Aggregation and Alignment

In the previous chapter, we studied the problem of matching (child, mother, father) triplets found in birth records, whereas in this chapter, we consider a more general problem of matching networks of records. In other words, we get as input a set of graphs whose nodes correspond to records and our goal is to integrate these graphs to produce a single entity graph.

In Section 4.1, we consider the problem of aggregating directed acyclic graphs (DAGs). This problem assumes that both the input graphs and the output graph are DAGs and that the correspondence (the correct alignment) between the nodes of the input graph is already known. Therefore, the main problem is to resolve the discordances between the input graph edges and output a DAG which is close to all of the input graphs.

Section 4.2 does not assume that the node correspondence is known but studies the problem of multiple network alignment which aims at inferring an entity graph and a mapping from the input graph nodes to the entity graph nodes. Section 4.3 addresses the network alignment problem in an interactive setting, where the algorithm is allowed to query a human expert to align some of the nodes, which might be necessary if the input data is corrupted by a lot of noise.

Finally, in Section 4.4, we discuss the challenges and some solutions related to parameter tuning when employing the aforementioned methods.

4.1 Directed-Acyclic-Graph Aggregation

Directed acyclic graphs (DAGs) are graphs with directed edges but without directed cycles. They can be used to model many kind of data and objects, such as rankings, user preferences, and information cascades. The problem of DAG aggregation asks to find a summarizing centroid DAG given a set of input DAGs whose nodes have already been aligned, for instance, based on a unique identifier associated with each node. Next, we briefly describe the related work, introduce a distance measure to compare DAGs, and, finally, provide a formal definition of the DAG-aggregation problem based on the proposed distance measure.

Background. The problem of aggregating rankings has been studied for more than two centuries, since it is encountered in many applications such as designing a voting system (Borda, 1781) or aggregating the results of multiple search engines (Dwork et al., 2001). Kendall-tau (Kendall, 1938) is a popular distance measure for rankings, but since the rankings encountered in practical applications are not total orders, Kendall-tau has been extended for *partial rankings*, that is, rankings with ties (Fagin et al., 2006). Aggregating rankings using Kendall-tau leads to an NP-hard problem (Dwork et al., 2001).

Comparing and aggregating DAGs. We propose an extension of the Kendall-tau measure for an even more general class of objects, namely DAGs. DAGs provide a natural way for modeling, for instance, *preference graphs*, which are constructed by collecting a set of pairwise preferences and drawing an edge from v to u if a user has indicated a preference of v over u. Such graphs are likely to be DAGs or near-DAGs since preference relations tend to be transitive, and, in general, they cannot be modeled as partial rankings without losing some information.

While the standard Kendall-tau measure penalizes only *discordant* pairs of items (i precedes j in one ranking, while j precedes i in the other ranking), the proposed extension for DAGs penalizes also *potentially* discordant pairs (i precedes j in one DAG, while there is no edge between i and j in the other DAG, or there is no edge between i and j in either of the graphs). The resulting distance measure, denoted by K, satisfies a relaxed triangle inequality. The exact definition of K and the proof for the inequality are presented in Publication IV.

Using the distance measure K, the DAG aggregation problem is defined as follows.

Problem 2 (DAG AGGREGATION)**.** *Given a set of M directed graphs $G_1,\ldots,G_M$, find a DAG C minimizing*

$$\sum_{i=1}^{M} K(G_i, C).$$

The goal is to find a centroid DAG which minimizes the distance to the input DAGs. This problem is NP-hard which can be shown by a reduction from a known NP-hard problem called FEEDBACK ARC SET as done in Publication IV.

To solve this problem, we propose two algorithms. The first algorithm simply picks the *median* input graph that minimizes the sum of distances to all input graphs. Remarkably, the fact that K satisfies a relaxed triangle inequality gives this simple approach a constant approximation-ratio guarantee. The second approach is a greedy method which does not have an approximation guarantee but which yields a better experimental performance in the experiments presented in Publication IV.

The experiments are conducted using synthetic data, information-cascade data from a music-listening service, and preference-graph data on music artists. The proposed DAG-aggregation methods can be used not only to aggregate but, conveniently, also to cluster DAGs as a part of a k-means type of algorithm. This clustering algorithm alternates between assigning DAGs to their closest clusters

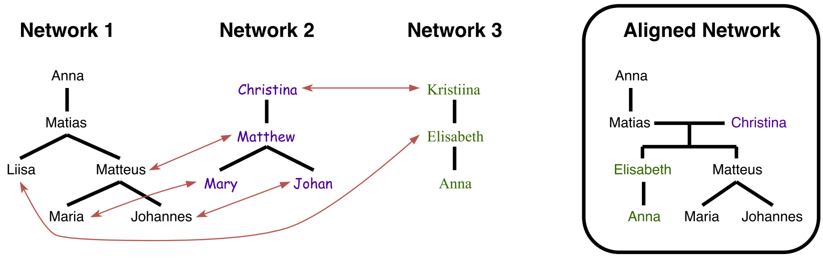


Figure 4.1. An example multiple network alignment problem, where the input graphs are only partially overlapping. The red arrows show the correct alignment between input graph nodes and the figure on the right shows the underlying aligned network.

(as defined by the proposed distance measure) and updating the centroids of the cluster (by aggregating the assigned DAGs).

In summary, the proposed DAG-aggregation methods are useful when integrating graphs (more specifically, directed acyclic graphs) with a known node correspondence (that is, the entity resolution step has already been done) but with inconsistent edges. Next we consider a problem where neither the node correspondence nor the edges of the integrated graph are known.

4.2 Multiple Network Alignment

A majority of existing network-alignment methods assume that the number of networks to be aligned is two. These pairwise network-alignment methods aim to find an alignment or a mapping from the nodes of a source graph to the nodes of a target graph, satisfying two properties: (i) nodes with similar attributes should be aligned, and (ii) adjacent source-graph nodes should be aligned with adjacent target-graph nodes. The graphs may be only partially overlapping which is why some nodes can be left unaligned or aligned with a “gap” node.

Often there is a need to align more than two input graphs, for instance, when finding functional orthologs across the protein–protein interaction networks of multiple species (Singh et al., 2008; Liao et al., 2009), when aligning the social networks of more than two online social networking services (Zhang and Yu, 2015), or when aligning multiple genealogical networks as illustrated in Figure 4.1.[1] If the input graphs are only partially overlapping, any naïve adaptation of a pairwise method to multiple networks based on fixing one of the input graphs as the target network will fail, since none of the input graphs necessarily contains a node for each underlying entity. Furthermore, even if one input graph covers all entities, it may still be missing some edges between them, making it necessary to consider multiple networks jointly in order to have a complete picture of the underlying entity network.

[1] Aligning multiple networks is also required when aligning two sources with multiple network layers. This distinct but related problem is known as *multimodal* network alignment (Nassar and Gleich, 2017).

To model the multiple network alignment problem, we propose an extension of NATALIE (El-Kebir et al., 2015; Klau, 2009) based on formulating the multiple network alignment problem as an integer program which aims at inferring the underlying entity network and assigning the input-graph nodes to the entity nodes. More formally, we are given k input graphs $G_1 = (V_1, E_1)$, ..., $G_k = (V_k, E_k)$, which we assume to be partial manifestations of an underlying *entity graph* $G_e = (U, E_e)$. We further assume that:

- Entities U are represented by a subset of all input graph nodes $V = \bigcup_{i=1}^{k} V_i$, that is, $U \subseteq V$.
- The edges between the entities E_e contain the input graph edges and potentially some missing edges, that is, $E_e = \left(\bigcup_{i=1}^{k} E_i\right) \cup E_m$, where E_m is the set of missing edges.

The goal is to infer the missing edges E_m and to find an assignment $\mathcal{X} : V \to U$ from the input-graph nodes to the entities. The assignment should try to satisfy the following properties:

(P1) Nodes are assigned to entities with as similar attributes as possible.

(P2) Adjacent nodes are assigned to adjacent entities.

To avoid having to introduce extensive amounts of notation, we next provide a high-level overview of the objective function designed to capture the above properties. The exact formulation of the optimization problem and its solution are provided in Publication V.

The objective function for the proposed multiple network alignment problem formulation is given by

$$\min_{\mathcal{X}, E_m} f|U(\mathcal{X})| + \sum_i d_i(\mathcal{X}) - g|\mathcal{S}(\mathcal{X})| + \gamma|E_m|, \tag{4.1}$$

where f, g, and γ are parameters, $|U(\mathcal{X})|$ is the number of inferred entities in assignment $\mathcal{X}$, $d_i(\mathcal{X})$ is the dissimilarity between the ith node and the entity it has been assigned to, and $|\mathcal{S}(\mathcal{X})|$ is the number of "squares," that is, adjacent nodes assigned to adjacent entities. The first term penalizes for using too many entities, since a trivial solution would be to use all input nodes as entities and assign each node to itself, the second term captures property (P1), the third term captures property (P2), and the fourth term penalizes for adding too many edges to the entity graph, since a trivial solution would be to add all possible edges to optimize property (P2). Additionally, the optimization problem comes with constraints ensuring that each node is assigned to exactly one entity and that the nodes of an input graph are assigned to distinct entities. The problem formulation is illustrated in Figure 4.2, which shows three input graphs, the inferred entity graph, and the inferred alignment from input graphs to the entity graph.

To solve the minimization problem, we propose an alternating optimization procedure that repeatedly optimizes the assignment $\mathcal{X}$ and the missing edges

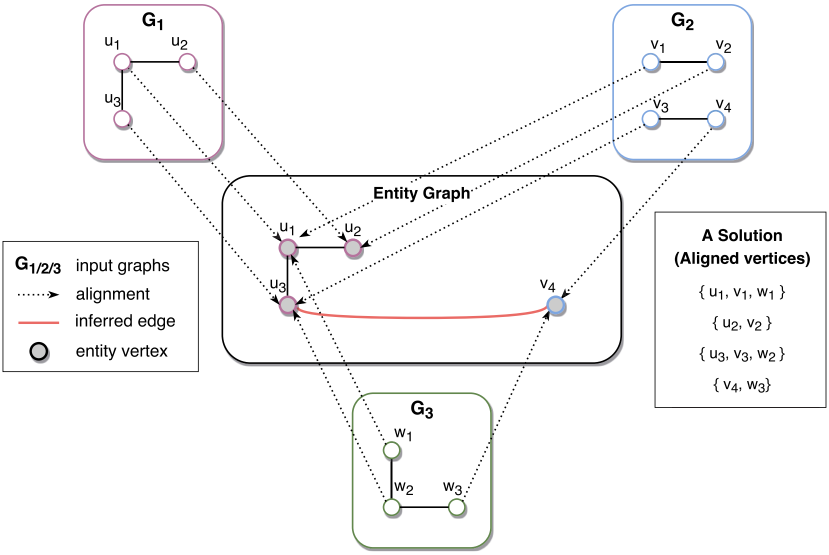


Figure 4.2. An illustration of the proposed multiple network alignment approach. The goal is to infer the underlying entity graph and find an alignment from input-graph nodes to entity nodes. The entity nodes are a subset of the input nodes. A missing edge between entities u_3 and v_4 has been inferred since two pairs of adjacent input nodes have been assigned to these entities.

E_m. Optimizing the assignment is shown to be NP-hard but an approximate solution can be found using a Lagrangian relaxation, which provides a lower and an upper bound for the optimum. In some cases the bounds collapse, guaranteeing that an optimal assignment has been found, but in other cases, they do not as illustrated in Figure 4.3, which plots the bounds for two synthetic instances of the multiple network alignment problem.

In summary, the proposed network alignment method extends a popular pairwise network alignment method called NATALIE (El-Kebir et al., 2015; Klau, 2009) by adding support for multiple networks. Additionally, it aims to infer the edges of the underlying entity graph. The proposed method yields high alignment accuracies, but we found that, in practice, a simpler approach yields similar accuracies while scaling better. The simpler approach splits the multiple network alignment problem of k input graphs into $k-1$ pairwise network alignment problems and updates the edges of a target graph after each pairwise alignment.

In the next section, we address the question of how to leverage human input for network alignment.

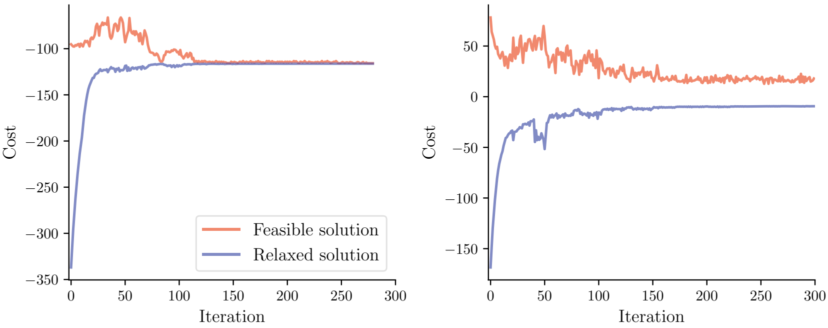


Figure 4.3. The proposed Lagrangian relaxation approach iteratively refines a feasible solution, which yields an upper bound, and a relaxed solution, which yields a lower bound. In the iteration shown on the left, the algorithm finds the global minimum and converges in 280 steps, whereas, on the right, a duality gap remains. Figure taken from Publication V.

4.3 Active Network Alignment

In case the attributes of the nodes to be aligned are corrupted with a lot of noise or there are many ambiguous nodes with similar attributes, it may be infeasible to rely on a fully automatic network alignment method. Thus it may be necessary to use human experts to help with the alignment process. However, since interaction with humans is often time consuming and costly, the amount of interaction should be minimized by presenting the humans with only the most informative queries. The goal of this section is to propose a method for designing queries for the problem of pairwise network alignment in order to obtain a high alignment accuracy with as few queries as possible. We call this problem *active network alignment*.

A fairly recent survey on ontology matching, which is an application of network alignment, lists “user involvement” as one of the future challenges for ontology matching (Shvaiko and Euzenat, 2013). Only a few previous works have studied the active network alignment problem but most of them are found, in fact, in the ontology matching literature (Jiménez-Ruiz et al., 2012; Paulheim et al., 2013; Sarasua et al., 2012; Shi et al., 2009). To the best of our knowledge, the existing works on active network alignment mainly focus on the following type of *absolute* queries: are nodes v and u the same or not. Since comparative judgments are generally easier for humans (Laming, 2003), we consider instead the following *relative* queries:

Given a node v in the source network G_s, and a set of candidate matches $\mathcal{C}_v$ in the target network G_t, which node from $\mathcal{C}_v$ should be matched to v?

The goal of active network alignment is to select informative query nodes v for which to ask the human expert to reveal the correct alignment. More specifically, the problem is defined as follows.

Problem 3 (ActiveNetworkAlignment)**.** *Given a source network* $G_s = (V_s, E_s)$

and a target network $G_t = (V_t, E_t)$, select the node $v \in V_s$ for which to ask an oracle to reveal the correct matched node $u \in V_t$ so that the alignment accuracy for the remaining nodes in V_s is maximized.

To solve Problem 3, we propose the following four-step approach.

Step 1. Cast the original network alignment problem into a maximum-weight bipartite matching problem. The corresponding weighted bipartite graph is denoted by $H = (V_s, V_t, E_h)$.

Step 2. Sample a set of ℓ high-quality matchings, $\mathcal{M}_\ell$, in H.

Step 3. For each node $v \in V_s$, compute the marginal probability distribution over the candidate matches of v based on the set of sampled matchings, $\mathcal{M}_\ell$, and estimate the *certainty* we have about the correct match for v based on the marginal distribution.

Step 4. Identify the node $\hat{v} \in V_s$ with the *least certainty*, and query the oracle to select the best match for $\hat{v}$ among the set of candidate matching nodes $\mathcal{C}_{\hat{v}} \subseteq V_t$.

Next we discuss these steps in more detail. In Step 1, the goal is to construct a bipartite graph H whose weights capture both attribute similarities and structural similarities between the nodes to be aligned. After constructing H, an alignment can be found by computing the maximum-weight matching in H, which can be done in polynomial time, using, for instance, the Hungarian algorithm (Kuhn, 1955). The idea of constructing a bipartite graph and then solving the final alignment by finding the maximum-weight matching is employed by several popular network-alignment methods, including NATALIE (El-Kebir et al., 2015; Klau, 2009), NETALIGNMP++ (Bayati et al., 2013), and ISORANK (Singh et al., 2008).

An intuitively appealing approach for finding an uncertain node to query based on the weighted bipartite graph H would be to find a node with many candidate matches and a uniform distribution of weights for these candidates. This approach would not, however, capture the interdependence between the nodes to be aligned that is caused by the requirement of finding a one-to-one matching (two nodes from the source graph cannot be matched to the same target graph node). In other words, even if a source node v has k candidate matches with equal weights and the correct alignment for v hence appears to be very uncertain, it could be that $k-1$ of the candidate matches need to be reserved for other source graph nodes with a very high probability, leaving only a single likely candidate match for v. Such a scenario is illustrated in Figure 4.4.

To capture the aforementioned node interdependence, we sample a set of matchings, $\mathcal{M}_\ell$, in Step 2 and then quantify the certainty of each node based on these matchings. The sampling can be done, for example, using Gibbs sampling with the idea of computing an initial matching and then creating samples by randomly picking two source graph nodes and swapping their assignments with a probability depending on the edge weights in H (Volkovs and Zemel, 2012).

To quantify node certainty in Step 3, we first compute the marginal probability

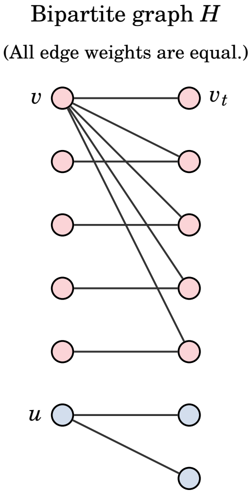


Figure 4.4. An example of a bipartite graph H corresponding to a network alignment problem. Considering merely the number of candidate matches would suggest that v is the most uncertain node and should thus be queried. However, because of the requirement of finding a one-to-one mapping, we can infer that v is unambiguously matched to v_t. Thus u should be queried since it has two ambiguous candidate matches.

$p(M(v) = u \mid H)$ for each $v \in V_s$ and $u \in \mathcal{M}_v \subseteq V_t$ by computing the fraction of matchings in $\mathcal{M}_\ell$ where v has been matched to u. Intuitively, the more uniform the marginal distribution the more uncertain is the correct match for v. The uniformity can be quantified, for instance, based on negative entropy, but we found using simply the maximum of the marginal distribution to yield a good experimental performance.

Querying the most uncertain node is a standard active-learning strategy (Settles, 2010). Therefore, in Step 4, we select the node with the lowest certainty score and ask a human expert (an oracle) to return the correct match for the node. In this work we assume—somewhat unrealistically—that the oracle always returns the correct match.

When applying the proposed four-step approach to a network-alignment task, we can either recompute H after each query or issue multiple queries in parallel. If the human queries are the bottleneck of the alignment process, then parallel querying allows us to speed up the alignment process with the cost of a small decrease in the alignment accuracy as shown experimentally in Publication VI.

The four steps are further illustrated in Figure 4.5 which shows a small network alignment problem and the process of selecting the node to query. Letters and colors are used to denote candidate matches. Indicatively, in an application of aligning social networks, one can think of A = `Andres`, C_1 = `Andrew`, C_2 = `Andreas`, while B = `Brendon`, B_1 = `Brenden`, B_2 = `Brendan`, etc. Since network alignment aims at matching adjacent source nodes to adjacent target nodes, we can deduce that if C is matched to C_1 or C_2, then B should be uniquely matched to B_1, whereas if C is matched to C_3, then B should be uniquely matched to B_2 and A to A_2. Thus, following the algorithm recommendation of querying node C, to a large extent, determines the rest of the alignment.

4.4 Parameter Tuning

A practical challenge when applying the proposed DAG-aggregation or the network-alignment methods is setting the method parameters. In this section, we briefly discuss some strategies for tackling this challenge.

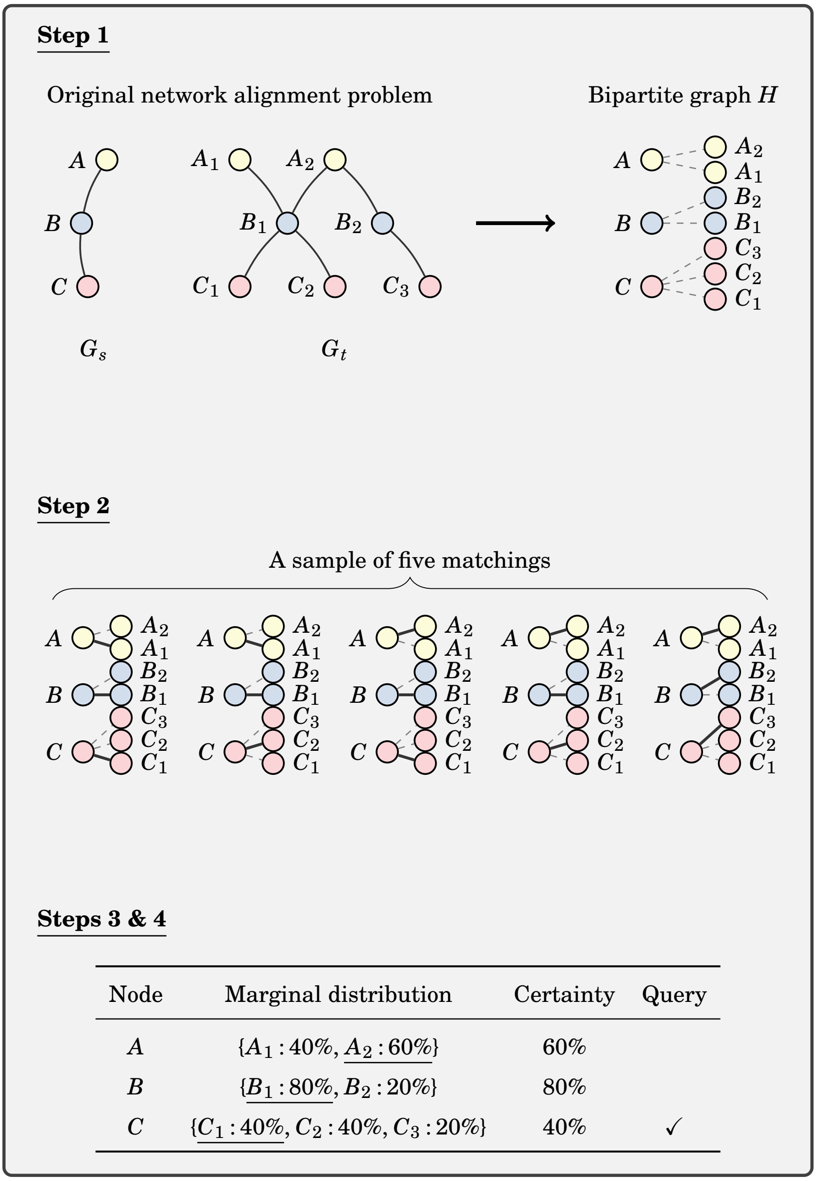
The proposed Kendall-tau measure for comparing DAGs requires two parameters, p and q, which determine the penalty for two types of potentially discordant pairs. A limitation of the measure is that the optimal parameter values are problem dependent as shown in Publication IV, Section 7.1.2. A simple heuristic for tuning the values is provided in Publication IV, but the selection of the optimal p and q values remains an open problem.

The proposed multiple network alignment method requires setting the three parameters in Equation (4.1): f, g, and γ. Parameter γ controls the tendency to add a new edge to the entity graph and it is natural to set this parameter to $\gamma = \frac{g}{2}$, which implies that an edge is added if at least one pair of adjacent nodes is aligned to the corresponding entities. Parameter g controls the balance between the two desired properties (P1) and (P2). These two properties are orthogonal and therefore the optimal balance is problem dependent. Parameter f controls the penalty of introducing a new entity.[2] Tuning method parameters is a common challenge with network-alignment methods in general. Typically it is addressed by simply using a set of default parameters, by optimizing the parameters based on known alignments (see, for example, Flannick et al., 2009), or by computing multiple alignments with different parameter settings (see, for example, Bayati et al., 2013).

When doing active network alignment, one has to set two types of parameters: the parameters of the underlying non-active aligner and a temperature parameter β, which is introduced in Publication VI. The temperature parameter is needed when sampling matchings with Gibbs sampling in Step 2, and it controls how closely the samples are concentrated around the maximum-weight matching. In Publication VI, we optimize β using grid search on a separate development set, containing known alignments, but we also find evidence that the optimal β is problem dependent. A potential way to tackle this problem

[2] Compared to the pairwise version of NATALIE, this parameter has a similar function as the similarity value between the input-graph nodes and a "gap" node. This similarity has to be defined by the user if the two input graphs are only partially overlapping. In Publication V we present a variation of Equation (4.1), where term $f|U(\mathcal{X})|$ has been omitted from the objective but the following new constraint has been added to the optimization problem: $|U(\mathcal{X})| \leq N_e$, where N_e is the maximum number of entities. This variant is useful in the case that we have prior knowledge about the number of entities, since it allows the method to find an accurate alignment robustly without having to tune parameter f.

would be to estimate β based on already-queried nodes, updating the estimate as new queries are made. Moreover, the already-queried nodes could also be used to tune the network-aligner parameters. Designing new active query strategies that do not try to merely identify the most uncertain nodes given a fixed set of parameters but also try to query nodes such that the uncertainty about the parameters would be reduced seems like a promising future direction.



Node	Marginal distribution	Certainty	Query
A	$\{A_1 : 40\%, \underline{A_2 : 60\%}\}$	60%	
B	$\{\underline{B_1 : 80\%}, B_2 : 20\%\}$	80%	
C	$\{\underline{C_1 : 40\%}, C_2 : 40\%, C_3 : 20\%\}$	40%	✓

Figure 4.5. A toy-example illustration of the proposed four-step approach to the problem of active network alignment. The goal is to decide which node to query: A, B, or C. The original network alignment problem is first transformed into a bipartite matching problem (Step 1), then we sample a set of matchings (Step 2), quantify the certainty of each node based on the matchings (Step 3), and finally query the most uncertain node (Step 4). In this example, the algorithm ends up querying node C.

5. Analyzing Inferred Genealogical Networks

Both the vital-record-linking methods discussed in Chapter 3 and the network-alignment methods discussed in Chapter 4 can be used to infer large-scale genealogical networks. Such networks are naturally useful for genealogists and historians but they also provide a valuable resource for other fields. To demonstrate the wide applicability of the inferred networks, we present a longitudinal analysis on *assortative mating* using a computationally inferred genealogical network. We also discuss the opportunities these networks can offer specifically for the field of *computational social science*.

5.1 Assortative Mating

Humans have a tendency to marry someone with a similar socioeconomic status. A recent study shows that this phenomenon—known as *assortative mating* or *social homogamy*—has a big societal impact as it contributes to income inequality (Greenwood et al., 2014). The study also finds that assortative mating has been on the rise between 1960 and 2005 in the United States (Greenwood et al., 2014), whereas an earlier work on assortative mating in Norway, 1750–1900, finds that assortative mating was declining in the mid-1700s after which it stayed fairly constant.

We leverage an inferred genealogical network to address two questions concerning this phenomenon:

(Q1) Can assortative mating be detected in Finland during the years 1735–1885?

(Q2) How has the intensity of assortative mating evolved during this time period?

To operationalize assortative mating, we compare the socioeconomic statuses of spouses inferred by linking vital records. The status of a person is estimated based on his or her father's occupation. Comparing father occupations is more straightforward and more robust than comparing spouse occupations directly, since occupations were strongly gendered in the 18th and 19th centuries, which are the focus of this study.

In addition to using occupations directly, we also map them to the historical international classification of occupations (HISCO) (Van Leeuwen et al., 2002) and divide the HISCO classes into four main categories: (1) upper and middle class, (2) peasants (who own land), (3) crofters (who rent land), and (4) laborers (who live at another person's house). The occupational category of a person's father is labeled CLASS4. The HISCO classes can also be mapped into an occupational stratification scale called HISCAM (Lambert et al., 2013), which assigns a real-valued number from 0 to 100 to each occupation. HISCAM measures the social interaction distance of people based on their occupations (Lambert et al., 2013).

A high percentage p of spouses with the same father occupation is indicative of assortative mating but this percentage might also be affected by external factors such as the number of available occupations, which depends on the context where the spouses lived. To control for such external factors, we propose a null model, which randomly assigns each person to a new spouse from the same city and the same birth year ±10 years. Then we compute the percentage p_n of matching occupations under the null model. The strength of assortative mating is defined as the ratio between the two percentages p/p_n. This ratio measures how much more likely people marry someone with a similar social status compared to a null model where the marriages are randomized. Thus a ratio larger than 1 is a sign of assortative mating.

Similarly, we can quantify assortative mating based on the occupational categories (CLASS4) by computing the percentages of matching categories q and q_n and their ratio q/q_n. For the real-valued HISCAM scores, we compute the mean absolute difference of the scores of the spouses (δ) and of the scores of the spouses under the null model (δ_n). In this case, assortative mating is quantified by the inverse ratio δ_n/δ, so that again, a ratio larger than 1 indicates assortative mating.

Data. In Chapter 3, Section 3.4, we applied the linking method BINCLASS to link a dataset of 5 million historical birth records from Finland and showed that the estimated link probabilities correlate with the true probability of a link to be correct. In this analysis, we consider the same set of inferred links but we only use the links that have a probability of 90% or more. Spouses are defined as persons who have had a child together. The analysis is limited to years 1735–1885, since outside this time period the number of reliably inferred spouses with known father occupations is small. After filtering out the spouses with unknown occupations, we are left with 6 402 spouse pairs. The CLASS4 categories and the HISCAM scores can be assigned to 6 248 of these pairs.

Results. The similarity of spouse occupations is plotted on the left-hand side of Figure 5.1, using the occupations directly (top), using the CLASS4 categories (middle), and using the HISCAM scores (bottom). The corresponding ratios, which quantify assortative mating, are shown on the right-hand side. To high-

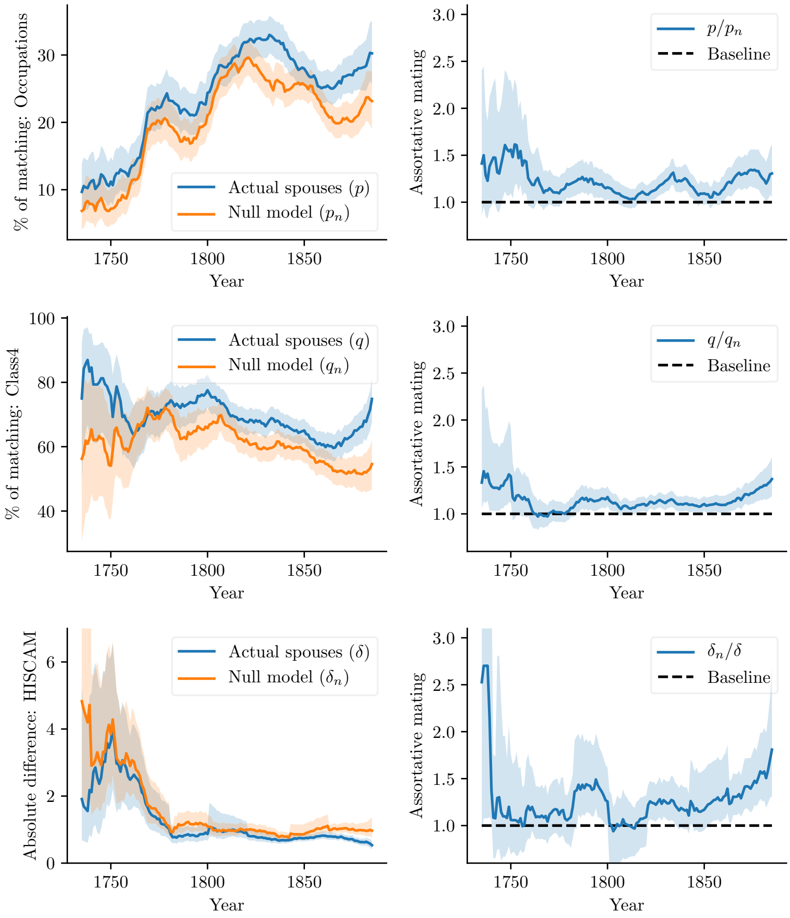


Figure 5.1. Assortative mating is detected in the inferred genealogical network for Finland (1735–1885), but the phenomenon is not monotonously decreasing or increasing.

light long-term trends, the curves show moving averages where a data point at year y uses the inferred spouses from years $[y-10, y+10]$. Confidence intervals are computed using 95% bootstrap confidence intervals. Even though the shapes of the comparison curves on the left vary significantly across the three measures, the shapes of the ratio curves on the right are fairly consistent across the measures. This highlights the importance of using a null model.

The ratios are mostly between 1 and 1.5 with all three methods, suggesting that assortative mating did occur in Finland (Q1). Up to 1770 there is a slight decreasing trend and after 1850 a slight increasing trend. Overall, however, we do not find evidence for a monotonically decreasing or increasing trend across the whole time period (Q2). These results suggest that even though one might assume that in the long run our societies become less stratified, this does not necessarily happen by itself.

5.2 Implications to Computational Social Science

Computational social science (CSS) is a recently emerged field, which studies social phenomena by quantitatively analyzing digital traces of humans (Lazer et al., 2009). This field is driven by the surge in the popularity of online social networking services, such as Twitter and Facebook, which record granular information about people's behavior and interaction with other people—mainly in the online but also in the offline world through smartphone sensors such as GPS trackers. These data sources are available to researchers in companies and, to a varying extent, to researchers in academic institutions via public APIs.

The new online data sources enable numerous opportunities to analyze different aspects of human behavior, including migration and mobility (Gonzalez et al., 2008; Malmi et al., 2012, 2013; Song et al., 2010; Zagheni et al., 2014), social influence (Bond et al., 2012; Onnela and Reed-Tsochas, 2010), the structural properties of social networks (Newman et al., 2011; Onnela et al., 2007), and sociolinguistics (Cheng et al., 2015; Danescu-Niculescu-Mizil et al., 2013), to name a few. Compared to traditional survey-based methods, the new data sources make it possible to significantly scale up sample sizes in order to capture more nuanced behavioral differences (see Kramer et al., 2014) and to collect more granular data without having to resort to self-report studies which may suffer from various biases such as the social desirability bias. On the other hand, in the recent years the CSS research community has also started to acknowledge the biases present in online behavioral datasets, such as sampling biases, causing the datasets to be non-representative (Baeza-Yates, 2016; Malmi and Weber, 2016; Wagner et al., 2015).

Large-scale genealogical networks, which can be inferred with the methods proposed in this thesis, and which are also becoming available through online crowd-sourcing sites, such as *Geni.com* and *WikiTree*, offer typically less granular behavioral data than online-social-network datasets. However, the key advantage of genealogical networks is the long temporal scale that they offer; for instance, Twitter and Facebook are both less than 15 years old, whereas genealogical networks can cover multiple centuries. This makes it possible to conduct *longitudinal* studies on phenomena that take multiple generations to occur or evolve, such as assortative mating, which was studied earlier in this chapter.

Other concrete research areas that could be tackled with automatically inferred genealogical networks include: intergenerational social mobility (Mare, 2011; Zijdeman, 2009), the heritability of longevity (Fire and Elovici, 2015; Kaplanis et al., 2017), family effects on longevity (Lahdenperä et al., 2004), large-scale migration patters (Schich et al., 2014), epidemics and mortality (death records often mention the cause of death), and the effects of wars and other external events on society. Furthermore, combining genealogical data with genetic data (Gauvin et al., 2015; Kaplanis et al., 2017) seems a promising future direction.

6. Conclusions

Entity resolution, or *record linkage*, has been an active research area for the past 50 years, but more recently, an increasing number of works have been published in the area of *collective* entity resolution. The idea of collective entity resolution is to leverage *relational information* about the entities or the records to be linked. In this thesis, we presented several methods for collective entity resolution. From the methodological perspective, our main focus was in the problem of *network alignment*, which is an instance of collective entity resolution, although it has prominent applications also outside entity resolution. For this problem, we proposed methods for *multiple* network alignment and for *active* network alignment.

From the application perspective, an overarching theme in the thesis has been the problem of *inferring genealogical networks*. We proposed methods for this task based on aligning existing genealogical networks and by linking vital records. The latter methods were employed to infer a genealogical network consisting of millions of individuals. Although the accuracy of the fully-automatic methods is not yet comparable to a careful human genealogist who has access to vital records not yet digitized, the probabilities of the inferred links can be used to distinguish the high-confidence links reliably. Furthermore, automatically linking 5 million birth records took less than an hour, which shows that the automatic approach is much more scalable than a manual approach. Moreover, the inferred link probabilities can be used to guide the manual process. In addition to supporting genealogical research, large-scale genealogical networks provide a valuable resource to the field of *computational social science* because of the long temporal span of the networks. This was demonstrated by analyzing *assortative mating* in Finland over a period of 150 years.

6.1 Future Directions

Some open problems related to DAG aggregation and network alignment were discussed earlier in Section 4.4. At the end of the previous chapter, we listed potential future directions regarding the analysis of automatically inferred genealogical networks. An important thing to bear in mind when conducting such an analysis is that if an analyzed feature is used as input to the network inference, any bias in the training data or in the inference method will potentially bias also the analysis results. For example, if we want to study human migration patterns through birth and burial records, similar to the work of Schich et al. (2014), and the genealogist who has linked the death records to the birth records has only been studying records from a single parish to make the manual linking task more feasible, the inference method might learn that people never move to another city during their lifetime. One approach to avoid such bias would be to simply ignore the location features when training the genealogical network inference method. However, this could decrease the accuracy of the inferred links considerably since people's mobility was much more restricted in the past centuries, so that two records geolocated far from each other are unlikely to refer to the same person even though it is possible.

Moreover, even ignoring the analyzed features when learning the model does not guarantee bias-free analysis results. In the analysis of assortative mating presented in the previous chapter, the analysis is based on spouse *occupations*. The occupations are not used by the inference method, but in theory they might still bias the analysis results; if occupations were biasing the genealogist who linked the spouses in the training dataset and the occupations correlate with some of the features used by the inference method, then the occupational bias of the genealogist could still manifest itself in the assortative mating analysis results. Interestingly, the emerging field of *algorithmic bias* (Baeza-Yates, 2016) is set to combat such scenarios. Therefore, the methods developed for reducing algorithmic bias could be helpful also when conducting computational social science studies based on computationally inferred genealogical networks.

It would also be useful to test the generalizability of the vital-record linking methods by applying them to records from other countries. In Finland, the indexed records are particularly extensive since all parishes were mandated by law to keep the records and a significant fraction of the records have been scanned and transcribed by volunteers in a project called "HisKi", which started already in 1980s.[1] Nevertheless, similar vital records were kept in many other countries as well and we can expect them to become amenable to computational analyses in large volumes in the future, through projects, such as READ,[2] which develop computational methods for handwritten text recognition.

The proposed vital-record linking methods first link the non-ambiguous burial records into birth records, and then link the birth records into their parents' birth records. Ideally, the linking of all vital records, including birth, burial, marriage, and migration records, should be formulated as a single optimization task. One idea for such an integrated method would be to consider vital-record linking as a multiple network alignment problem, where each vital record corresponds to an input network with one to three nodes, corresponding to the people mentioned

[1] For more information, see `http://hiski.genealogia.fi/hiskitalkoot/` (only in Finnish).

[2] `https://read.transkribus.eu/`

in the record. However, the multiple network alignment formulation presented in Section 4.2 would need to be adapted somehow to capture constraints such as: “burial node b and mother node m should not be aligned to the same entity node if the child of m has a birth date after the death date of b.” A more promising approach for an integrated vital-record linking method would be to adapt the PSL approach proposed by Kouki et al. (2017). Furthermore, to estimate a probability distribution over parent candidates, their approach could be used together with the BINCLASS method discussed in Section 3.1 by replacing the XGBoost classifier with the PSL method.

Another interesting future direction would be to incorporate DNA data to the genealogical network inference problem.[3] DNA tests can provide an estimate for the family relationship between two tested individuals and these relationships could be used as (soft) constraints in the genealogical network inference task. This direction could benefit from the large body of literature on *phylogenetic inference* (Huelsenbeck and Ronquist, 2001).

[3] This idea has been proposed at: `https://stats.stackexchange.com/questions/295801/principled-way-to-formalize-unknown-parents-in-a-genealogy-tree-optimization-p`

References

L. Backstrom and J. Kleinberg. Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 831–841. ACM, 2014.

L. Backstrom, E. Sun, and C. Marlow. Find me if you can: Improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th International Conference on World Wide Web*, pages 61–70. ACM, 2010.

R. Baeza-Yates. Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science (extended abstract)*. ACM, 2016.

M. Bayati, D. F. Gleich, A. Saberi, and Y. Wang. Message-passing algorithms for sparse network alignment. *ACM Transactions on Knowledge Discovery from Data*, 7(1):3, 2013.

J. L. Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 18(9):509–517, 1975.

I. Bhattacharya and L. Getoor. A latent Dirichlet model for unsupervised entity resolution. In *Proceedings of the Sixth IEEE International Conference on Data Mining*, pages 47–58. IEEE, 2006.

I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data*, 1(1):5, 2007.

M. Bilgic and L. Getoor. Active inference for collective classification. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, 2010.

G. Bloothooft, P. Christen, K. Mandemakers, and M. Schraagen, editors. *Population Reconstruction*. Springer, 2015.

R. M. Bond, C. J. Fariss, J. J. Jones, A. D. Kramer, C. Marlow, J. E. Settle, and J. H. Fowler. A 61-million-person experiment in social influence and political mobilization. *Nature*, 489(7415):295–298, 2012.

J.-C. Borda. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*, 1781.

T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

J. Cheng, C. Danescu-Niculescu-Mizil, and J. Leskovec. Antisocial behavior in online discussion communities. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, pages 61–70, 2015.

L. Chindelevitch, C.-S. Liao, and B. Berger. Local optimization for global alignment of protein interaction networks. In *Proceedings of the Pacific Symposium on Biocomputing*, volume 15, pages 123–132, 2010.

P. Christen. Automatic record linkage using seeded nearest neighbour and support vector machine classification. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data mining*, pages 151–159. ACM, 2008.

P. Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, 2012.

P. Christen. Application of advanced record linkage techniques for complex population reconstruction. *ArXiv e-prints*, Dec. 2016.

P. Christen, D. Vatsalan, and Z. Fu. Advanced record linkage methods and privacy aspects for population reconstruction—a survey and case studies. In *Population Reconstruction*, pages 87–110. Springer, 2015.

C. Clark and J. Kalita. A comparison of algorithms for the pairwise alignment of biological networks. *Bioinformatics*, 30(16), 2014.

M. Cochinwala, V. Kurien, G. Lalk, and D. Shasha. Efficient data reconciliation. *Information Sciences*, 137(1):1–15, 2001.

C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

K. Cranmer, J. Pavez, and G. Louppe. Approximating likelihood ratios with calibrated discriminative classifiers. *ArXiv e-prints*, Mar. 2016.

C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 307–318. ACM, 2013.

X. Dong, A. Halevy, and J. Madhavan. Reference reconciliation in complex information spaces. In *Proceedings of the 2005 ACM SIGMOD International Conference on Management of Data*, pages 85–96. ACM, 2005.

H. L. Dunn. Record linkage. *American Journal of Public Health and the Nations Health*, 36(12):1412–1416, 1946.

C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web*, pages 613–622. ACM, 2001.

J. Efremova, B. Ranjbar-Sahraei, H. Rahmani, F. A. Oliehoek, T. Calders, K. Tuyls, and G. Weiss. Multi-source entity resolution for genealogical data. In *Population Reconstruction*, pages 129–154. Springer, 2015.

M. El-Kebir, J. Heringa, and G. W. Klau. Natalie 2.0: Sparse global network alignment as a special case of quadratic assignment. *Algorithms*, 8(4), 2015.

A. Elmsallati, C. Clark, and J. Kalita. Global alignment of protein-protein interaction networks: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(4):689–705, 2016.

R. Fagin, R. Kumar, M. Mahdian, D. Sivakumar, and E. Vee. Comparing partial rankings. *SIAM Journal on Discrete Mathematics*, 20(3), 2006.

I. P. Fellegi and A. B. Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

M. Fire and Y. Elovici. Data mining of online genealogy datasets for revealing lifespan patterns in human population. *ACM Transactions on Intelligent Systems and Technology*, 6(2):28, 2015.

D. Firmani, B. Saha, and D. Srivastava. Online entity resolution using an oracle. *Proceedings of the VLDB Endowment*, 9(5):384–395, 2016.

J. Fisher, P. Christen, and Q. Wang. Active learning based entity resolution using markov logic. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 338–349. Springer, 2016.

M. L. Fisher. The Lagrangian relaxation method for solving integer programming problems. *Management science*, 27, 1981.

J. Flannick, A. Novak, B. S. Srinivasan, H. H. McAdams, and S. Batzoglou. Græmlin: general and robust alignment of multiple large interaction networks. *Genome research*, 16(9), 2006.

J. Flannick, A. Novak, C. B. Do, B. S. Srinivasan, and S. Batzoglou. Automatic parameter learning for multiple local network alignment. *Journal of Computational Biology*, 16 (8), 2009.

H. Gauvin, J.-F. Lefebvre, C. Moreau, E.-M. Lavoie, D. Labuda, H. Vézina, and M.-H. Roy-Gagnon. GENLIB: an R package for the analysis of genealogical data. *BMC Bioinformatics*, 16(1):160, 2015.

O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. P. Gummadi. On the reliability of profile matching across large online social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1799–1808. ACM, 2015.

M. C. Gonzalez, C. A. Hidalgo, and A.-L. Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, 2008.

J. Greenwood, N. Guner, G. Kocharkov, and C. Santos. Marry your like: Assortative mating and income inequality. *The American Economic Review*, 104(5):348–353, 2014.

P. H. Guzzi and T. Milenković. Survey of local and global biological network alignment: the need to reconcile the two sides of the same coin. *Briefings in bioinformatics*, 2017.

S. Hashemifar and J. Xu. HubAlign: an accurate and efficient method for global alignment of protein–protein interaction networks. *Bioinformatics*, 30(17):i438–i444, 2014.

T. N. Herzog, F. J. Scheuren, and W. E. Winkler. *Data Quality and Record Linkage Techniques*. Springer Science & Business Media, 2007.

J. Hu, B. Kehr, and K. Reinert. NetCoffee: a fast and accurate global alignment approach to identify functionally conserved proteins in multiple networks. *Bioinformatics*, 30 (4):540–548, 2013.

J. P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001.

K. Jain and V. V. Vazirani. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and lagrangian relaxation. *Journal of the ACM*, 48(2):274–296, 2001.

E. Jiménez-Ruiz, B. C. Grau, Y. Zhou, and I. Horrocks. Large-scale interactive ontology matching: Algorithms and implementation. In *Proceedings of the 20th European Conference on Artificial Intelligence*, pages 444–449. Ios Press, 2012.

D. Jurgens. That's what friends are for: Inferring location in online social media platforms based on social relationships. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, pages 273–282, 2013.

J. Kaplanis, A. Gordon, M. Wahl, M. Gershovits, B. Markus, M. Sheikh, M. Gymrek, G. Bhatia, D. G. MacArthur, A. Price, and Y. Erlich. Quantitative analysis of population-scale family trees using millions of relatives. *bioRxiv*, 2017.

M. Kendall. A new measure of rank correlation. *Biometrika*, 30, 1938.

S. Kim, N. Kini, J. Pujara, E. Koh, and L. Getoor. Probabilistic visitor stitching on cross-device web logs. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1581–1589, 2017.

A. Kimmig, S. Bach, M. Broecheler, B. Huang, and L. Getoor. A short introduction to probabilistic soft logic. In *Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications*, pages 1–4, 2012.

G. W. Klau. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, 10(Suppl 1), 2009.

P. Kouki, C. Marcum, L. Koehly, and L. Getoor. Entity resolution in familial networks. In *Proceedings of the 12th Workshop on Mining and Learning with Graphs*, 2016.

P. Kouki, J. Pujara, C. Marcum, L. Koehly, and L. Getoor. Collective entity resolution in familial networks. In *Proceedings of the 2017 IEEE International Conference on Data Mining*, 2017.

A. D. Kramer, J. E. Guillory, and J. T. Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proceedings of the National Academy of Sciences*, 111(24):8788–8790, 2014.

O. Kuchaiev and N. Pržulj. Integrative network alignment reveals large regions of global network similarity in yeast and human. *Bioinformatics*, 27(10):1390–1396, 2011.

O. Kuchaiev, T. Milenković, V. Memišević, W. Hayes, and N. Pržulj. Topological network alignment uncovers biological function and phylogeny. *Journal of the Royal Society Interface*, 2010.

H. W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics*, 2(1-2):83–97, 1955.

M. Lahdenperä, V. Lummaa, S. Helle, M. Tremblay, and A. F. Russell. Fitness benefits of prolonged post-reproductive lifespan in women. *Nature*, 428(6979):178–181, 2004.

P. S. Lambert, R. L. Zijdeman, M. H. Van Leeuwen, I. Maas, and K. Prandy. The construction of HISCAM: A stratification scale based on social interactions for historical comparative research. *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 46(2):77–89, 2013.

D. Laming. *Human judgment: the eye of the beholder*. Cengage Learning EMEA, 2003.

D. Lazer, A. S. Pentland, L. Adamic, S. Aral, A. L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al. Life in the network: the coming age of computational social science. *Science*, 323(5915):721–723, 2009.

J. Lian and X. Xie. Cross-device user matching based on massive browse logs: The runner-up solution for the 2016 CIKM Cup. In *Proceedings of the CIKM Cup Workshop*, 2016.

C.-S. Liao, K. Lu, M. Baym, R. Singh, and B. Berger. IsoRankN: spectral methods for global alignment of multiple protein networks. *Bioinformatics*, 25(12):i253–i258, 2009.

E. Malmi and I. Weber. You are what apps you use: Demographic prediction based on user's apps. In *Proceedings of the 10th International AAAI Conference on Web and Social Media*, pages 635–638, 2016.

E. Malmi, T. M. T. Do, and D. Gatica-Perez. Checking in or checked in: comparing large-scale manual and automatic location disclosure patterns. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia*, page 26. ACM, 2012.

E. Malmi, T. M. T. Do, and D. Gatica-Perez. From Foursquare to my square: Learning check-in behavior from multiple sources. In *Proceedings of the 7th International AAAI Conference on Weblogs and Social Media*, pages 701–704, 2013.

E. Malmi, N. Tatti, and A. Gionis. Beyond rankings: comparing directed acyclic graphs. *Data Mining and Knowledge Discovery*, 29(5):1233–1257, 2015.

E. Malmi, S. Chawla, and A. Gionis. Lagrangian relaxations for multiple network alignment. *Data Mining and Knowledge Discovery*, pages 1–28, 2017a.

E. Malmi, E. Terzi, and A. Gionis. Active network alignment: A matching-based approach. In *Proceedings of the 26th ACM International Conference on Information and Knowledge Management*, pages 1687–1696. ACM, 2017b.

N. Malod-Dognin and N. Pržulj. L-GRAAL: Lagrangian graphlet-based network aligner. *Bioinformatics*, pages 2182–2189, 2015.

R. D. Mare. A multigenerational view of inequality. *Demography*, 48(1):1–23, 2011.

J. McGee, J. Caverlee, and Z. Cheng. Location prediction in social media based on tie strength. In *Proceedings of the 22nd ACM International Conference on Information and Knowledge Management*, pages 459–468. ACM, 2013.

L. Meng, A. Striegel, and T. Milenković. Local versus global biological network alignment. *Bioinformatics*, 32(20):3155–3164, 2016.

S. Mohammadi, D. F. Gleich, T. G. Kolda, and A. Grama. Triangular alignment (tame): A tensor-based approach for higher-order network alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(6):1446–1458, 2017.

H. Nassar and D. F. Gleich. Multimodal network alignment. In *Proceedings of the 2017 SIAM International Conference on Data Mining*, pages 615–623. SIAM, 2017.

M. Newman, A.-L. Barabasi, and D. J. Watts. *The structure and dynamics of networks*. Princeton University Press, 2011.

N. F. Noy, M. A. Musen, et al. Algorithm and tool for automated ontology merging and alignment. In *Proceedings of the 17th National Conference on Artificial Intelligence*, 2000.

J.-P. Onnela and F. Reed-Tsochas. Spontaneous emergence of social influence in online systems. *Proceedings of the National Academy of Sciences*, 107(43):18375–18380, 2010.

J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A.-L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, 2007.

H. Paulheim, S. Hertling, and D. Ritze. Towards evaluating interactive ontology matching tools. In *Extended Semantic Web Conference*, pages 31–45. Springer, 2013.

B. Ranjbar-Sahraei, J. Efremova, H. Rahmani, T. Calders, K. Tuyls, and G. Weiss. HiDER: Query-driven entity resolution for historical data. In *Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 281–284. Springer, 2015.

M. Richardson and P. Domingos. Markov logic networks. *Machine learning*, 62(1): 107–136, 2006.

S. B. Roy, M. De Cock, V. Mandava, S. Savanna, B. Dalessandro, C. Perlich, W. Cukierski, and B. Hamner. The Microsoft Academic Search dataset and KDD cup 2013. In *Proceedings of the 2013 KDD Cup Workshop*. ACM, 2013.

S. M. E. Sahraeian and B.-J. Yoon. SMETANA: Accurate and scalable algorithm for probabilistic alignment of large-scale biological networks. *PLOS ONE*, 8(7), 2013.

C. Sarasua, E. Simperl, and N. F. Noy. Crowdmap: Crowdsourcing ontology alignment with microtasks. In *International Semantic Web Conference*, pages 525–541. Springer, 2012.

S. Sarawagi and A. Bhamidipaty. Interactive deduplication using active learning. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 269–278. ACM, 2002.

M. Schich, C. Song, Y.-Y. Ahn, A. Mirsky, M. Martino, A.-L. Barabási, and D. Helbing. A network framework of cultural history. *Science*, 345(6196):558–562, 2014.

V. Sehgal, L. Getoor, and P. D. Viechnicki. Entity resolution in geospatial data integration. In *Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems*, pages 83–90. ACM, 2006.

B. Settles. Active learning literature survey. Technical report, University of Wisconsin–Madison, 2010.

F. Shi, J. Li, J. Tang, G. Xie, and H. Li. Actively learning ontology matching via user interaction. In *International Semantic Web Conference*, pages 585–600. Springer, 2009.

P. Shvaiko and J. Euzenat. Ontology matching: state of the art and future challenges. *IEEE Transactions on Knowledge and Data Engineering*, 25(1):158–176, 2013.

R. Singh, J. Xu, and B. Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 105(35):12763–12768, 2008.

P. Singla and P. Domingos. Entity resolution with Markov logic. In *Proceedings of the Sixth IEEE International Conference on Data Mining*, pages 572–582. IEEE, 2006.

C. Song, Z. Qu, N. Blumm, and A.-L. Barabási. Limits of predictability in human mobility. *Science,* 327(5968):1018–1021, 2010.

Y. Tay, C.-M. Phan, and T.-A. N. Pham. Cross device matching for online advertising with neural feature ensembles: First place solution at CIKM Cup 2016. In *Proceedings of the CIKM Cup Workshop*, 2016.

M. H. Van Leeuwen, I. Maas, and A. Miles. *HISCO: Historical international standard classification of occupations*. Leuven University Press, 2002.

V. Verroios, H. Garcia-Molina, and Y. Papakonstantinou. Waldo: An adaptive human interface for crowd entity resolution. In *Proceedings of the 2017 ACM International Conference on Management of Data*, pages 1133–1148. ACM, 2017.

V. Vijayan and T. Milenković. Multiple network alignment via multiMAGNA++. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, PP(99), 2017.

M. Volkovs and R. S. Zemel. Efficient sampling for bipartite matching problems. In *Advances in Neural Information Processing Systems*, pages 1313–1321, 2012.

C. Wagner, D. Garcia, M. Jadidi, and M. Strohmaier. It's a man's Wikipedia? Assessing gender inequality in an online encyclopedia. In *Proceedings of the 9th International AAAI Conference on Web and Social Media*, pages 454–463, 2015.

W. E. Winkler. String comparator metrics and enhanced decision rules in the Fellegi–Sunter model of record linkage. In *Proceedings of the Section on Survey Research Methods*, pages 354–359. American Statistical Assn., 1990.

E. Zagheni, V. R. K. Garimella, I. Weber, and B. State. Inferring international and internal migration patterns from Twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 439–444. ACM, 2014.

J. Zhang and P. S. Yu. Multiple anonymized social networks alignment. In *Proceedings of the IEEE International Conference on Data Mining*, pages 599–608. IEEE, 2015.

R. L. Zijdeman. Like my father before me: intergenerational occupational status transfer during industrialization (Zeeland, 1811–1915). *Continuity and Change*, 24 (3):455–486, 2009.

Aalto-DD 73/2018

9 789526 079493

ISBN 978-952-60-7949-3 (printed)
ISBN 978-952-60-7950-9 (pdf)
ISSN 1799-4934 (printed)
ISSN 1799-4942 (pdf)

Aalto University
School of Science
Department of Computer Science
www.aalto.fi

BUSINESS +
ECONOMY

ART +
DESIGN +
ARCHITECTURE

SCIENCE +
TECHNOLOGY

CROSSOVER

DOCTORAL
DISSERTATIONS