Author(s): Eric Malmi, Aristides Gionis, and Arno Solin

Title: Computationally Inferred Genealogical Networks Uncover Long-Term Trends in Assortative Mating

Year: 2018

Version: Final published version

Computationally Inferred Genealogical Networks Uncover Long-Term Trends in Assortative Mating

Eric Malmi
Aalto University
Espoo, Finland
eric.malmi@aalto.fi

Aristides Gionis
Aalto University
Espoo, Finland
aristides.gionis@aalto.fi

Arno Solin
Aalto University
Espoo, Finland
arno.solin@aalto.fi

ABSTRACT

Genealogical networks, also known as family trees or population pedigrees, are commonly studied by genealogists wanting to know about their ancestry, but they also provide a valuable resource for disciplines such as digital demography, genetics, and computational social science. These networks are typically constructed by hand through a very time-consuming process, which requires comparing large numbers of historical records manually. We develop computational methods for automatically inferring large-scale genealogical networks. A comparison with human-constructed networks attests to the accuracy of the proposed methods. To demonstrate the applicability of the inferred large-scale genealogical networks, we present a longitudinal analysis on the mating patterns observed in a network. This analysis shows a consistent tendency of people choosing a spouse with a similar socioeconomic status, a phenomenon known as *assortative mating*. Interestingly, we do not observe this tendency to consistently decrease (nor increase) over our study period of 150 years.

KEYWORDS

genealogy; family tree; pedigree; population reconstruction; probabilistic record linkage; assortative mating; social stratification; homogamy

Figure 1: A subgraph of a genealogical network automatically inferred by linking birth records. The subgraph spans 13 generations.

Legend:
★ 1. Upper and middle class ◆ 3. Crofters ● Unknown
▼ 2. Peasants ■ 4. Labourers

1 INTRODUCTION

Where do we come from? What are we? Where are we going? These questions, posed by the famous painter Paul Gauguin, resonate with many people. An example of the allure of the first question can be attested in the popularity of genealogy, the study of family history. Genealogical research is typically conducted by studying a large number of historical vital records, such as birth and marriage records, and trying to link records referring to the same person. This process is nowadays facilitated by numerous popular online services, such as *Ancestry.com*, *MyHeritage*, and *Geni*, and increasingly with genetic anal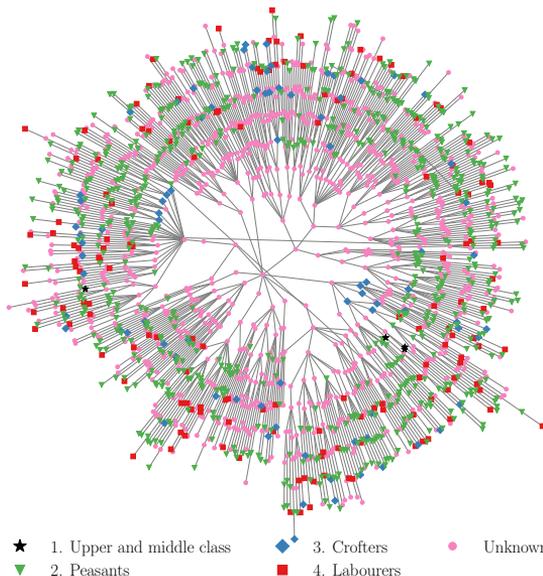ysis. Nevertheless, constructing a genealogical network (also known as family tree or population pedigree) is still a very time consuming process, entailing lots of manual work.

In recent years, lots of efforts have been focused on *indexing* historical vital records available in physical archives as well as in online repositories in the format of scanned images. Crowdsourcing projects organized, for example, by online genealogy services aim to do this kind of research by attracting people capable of interpreting old handwriting. Furthermore, there are also recent efforts at developing new optical character recognition (OCR) techniques to automate the process [13]. The availability of indexed records makes genealogical research amenable to new computer-supported and, to some extent, even fully-automatic approaches.

Our goal in this paper is to develop novel computational methods for inferring large-scale genealogical networks by linking vital records. We propose two supervised inference methods, BINCLASS and COLLECTIVE, which we train and test on Finnish data from the mid 17th to the late 19th century. More specifically, we apply these methods to link a large collection of indexed vital records

from Finland, constructing a genealogical network whose largest component contains 2.6 million individuals. A small subgraph of this component is visualized in Figure 1, where people are colored by their social class which is based on their father's occupation. The class information is used later when analyzing assortative mating. The construction of the full network takes only about one hour, which shows that it is vastly more scalable than the traditional manual approach which would probably require at least dozens of man-years for the same task. However, at its current stage, the accuracy of the automatic approach is not comparable to a careful human genealogist, but it can still support the work of the genealogist by providing the most probable parents for each individual, narrowing the search space.

From the methodological point of view, the main idea behind the proposed approaches is to cast the network-inference task into multiple binary classification tasks. Furthermore, our second approach, Collective, aims at capturing the observation that people tend to have children with the same partner, so if, for example, a father has multiple children, the mothers of the children should not be inferred independently. Incorporating this notion to our optimization problem, interestingly, leads to the well-known facility-location problem as well as an increase in the accuracy of the links.

In addition to family history, genealogical networks can be applied to several other domains, such as digital demography [30], genetics, human mobility, epidemiology, and computational social science. In this work, we demonstrate the applicability of the inferred network by analyzing *assortative mating*, that is, a general tendency of people to choose a spouse with a similar socioeconomic background. Mating choices have been shown to be an important driver for income inequality [14].

From the perspective of computational social science [21], genealogical networks offer particularly interesting analysis opportunities because of the long time window these networks cover. Compared to social-media data, which is typically used for computational social-science studies, genealogical networks contain less granular data about the people in the network, but they allow us to observe phenomena that occur over multiple generations. This aspect of genealogical networks enables us to quantify long-term trends in our society, as done for assortative mating in this work. Therefore, genealogical networks can provide answers not only to the first question asked in the beginning of this section but also to the third one: *Where are we going?*

Our main contributions in this paper are summarized as follows:

- We propose a principled probabilistic method, BinClass, for inferring large-scale genealogical networks by linking vital records.
- We perform an experimental evaluation, which shows that 61.6% of the links inferred by BinClass are correct. The accuracy obtained by BinClass surpasses the accuracy (56.9%) of a recently proposed NaiveBayes method [24]. Furthermore, we show that the link probability estimates provided by BinClass can be used to reliably control the precision–recall trade-off of the inferred network.
- We present a novel inference method, called Collective, which aims to improve disambiguation of linked entities, and does so by considering the genealogical-network inference

task as a global optimization task, instead of inferring family relationships independently. Collective further improves the overall link accuracy to 65.1%. However, contrasted to BinClass, Collective does not provide link probabilities.
- Finally, we demonstrate the relevance and applicability of automatically inferred genealogical networks by performing an analysis on assortative mating. The analysis suggests that assortative mating existed in Finland between 1735 and 1885, but it did not consistently decrease or increase during this period.

2 DATA

Our genealogical network inference method is based on linking vital records. The method is trained and evaluated using a human-constructed network. Next, we present these two data sources.

2.1 Population Records

The Swedish Church Law 1686 obliged the parishes in Sweden (and in Finland, which used to be part of Sweden) to keep records of births, marriages, and burials, across all classes of the society. The 'HisKi' project, an effort started in the 1980s, aims to digitally index the hand-written Finnish parish registers. The digitized data contains about 5 million records of births and a total of 5 million records of deaths, marriages and migration. The HisKi dataset is publicly available at http://hiski.genealogia.fi/hiski?en, except for the last 100 years due to current legislation.

Each birth record typically contains the name, birth place, and birth date of the child in addition to the names and occupations of the parents. The goal of the genealogical network inference problem is to link the birth records to the birth records of the parents, creating a family tree with up to millions of individuals.

Currently, we only have access to data from Finland, but similar indexed datasets can be expected to become available for other countries through projects such as READ[1] which develop optical character recognition (OCR) methods for historical hand-written documents.

2.2 Ground Truth

We have obtained a genealogical network consisting of 116 640 individuals constructed by an individual genealogist over a long period of time. To use this network as a ground truth, we first match these individuals to the birth records in the HisKi dataset. An individual is considered matched if we find exactly one birth record with the same normalized first name and last name and the same birth date. Then, we find parent–child edges where both individuals are matched to a birth record, yielding 18 731 ground-truth links.

Finally, the ground-truth links are split into a training set (70%) and a test set (30%). This is done by computing the connected components of the network, sorting the components by size in descending order, and assigning the nodes into two buckets in a round-robin fashion: First, assign the largest component to the training set. Second, assign the second largest component to the test set. Continue alternating between the buckets, however, skipping a bucket if its target size has been reached. Compared to directly

---
[1]https://read.transkribus.eu/

splitting people into the two buckets, this approach ensures that there are no edges going across the two buckets that would thus be lost. In the end, we get 5 631 test links out of which 42% are between mother and a child.

3 GENEALOGICAL NETWORK INFERENCE

Genealogical networks are typically constructed manually by linking vital records, such as birth, marriage, and death certificates. The main challenges in the linking process are posed by duplicate names, spelling variations and missing records.

3.1 Problem Definition

A *genealogical network* is defined as a directed graph where the nodes correspond to people and the edges correspond to family relationships between them. We consider only two type of edges: *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 they 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 $F_a \in V$. In addition to the birth records, the inference method gets as input a set of mother candidates $C^m \subseteq V$ and a set of father candidates $C^f \subseteq V$ for each child. In each method studied in this paper, the candidate sets are defined as the people who were born between 10 and 70 years before the child and whose normalized first and last name match to the parent name mentioned in the child's record.[2]

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

3.2 Naive Bayes Baseline

This baseline method [24] first constructs an attribute similarity vector γ for each (child, candidate parent) pair. The vector consists of the following five features: 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.

Assuming that the family links are independent, the probability of each mother candidate m (and similarly of each father candidate) for person a can be written as follows, using the Bayes' rule

$$p(M_a = m \mid \gamma) = \frac{p(M_a = m) \frac{p(\gamma^{a,m}|M_a=m)}{p(\gamma^{a,m}|M_a \neq m)}}{\sum_{m' \in C_a^m \cup \emptyset} p(M_a = m') \frac{p(\gamma^{a,m'}|M_a=m')}{p(\gamma^{a,m'}|M_a \neq m')}}, \quad (1)$$

where $p(M_a = m)$ is the prior probability of m and $p(\gamma^{a,m} \mid M_a = m)$ denotes the likelihood of observing the attribute similarities $\gamma^{a,m}$. The derivation of Equation (1) is given in the earlier work of Malmi et al. [24], and we adopt it in this paper.

Since the links are assumed to be independent, the log-likelihood function over all links is given by the sum of the link log-probabilities.

Let $x_{a,m} \in \{0, 1\}$ be a random variable denoting whether person a is linked to parent m. This allows us to write the likelihood function as

$$\max_x \left[\sum_{a,m} \log p(M_a = m \mid \gamma) x_{a,m} \right.$$
$$\left. + \sum_{a',f} \log p(F_{a'} = f \mid \gamma) x_{a',f} \right] \quad (2)$$

$$\text{such that} \quad \sum_m x_{a,m} = 1, \quad \text{for all } a, \quad (3)$$

$$\sum_f x_{a,f} = 1, \quad \text{for all } a, \quad (4)$$

$$x_{a,m}, x_{a,f} \in \{0, 1\}, \quad \text{for all } a, m, f. \quad (5)$$

The maximum likelihood solution for this non-collective genealogical network inference problem can be obtained simply by optimizing the parent links independently. The method assumes that the components of $\gamma^{a,m}$, corresponding to different attribute similarities, are also independent. Therefore, we call this method NAIVE-BAYES.

Next we present the two methods proposed in this paper, which improve over NAIVEBAYES.

3.3 Binary-Classification Approach

The key observation behind this approach is that likelihood ratios can be approximated with probabilistic discriminative classifiers [9]. This means that instead of estimating the component-wise probability distributions $p(\gamma_i^{a,m} \mid M_a = m)$, as done in NAIVEBAYES, we can approximate the two likelihood ratios in (1) by training a probabilistic binary classifier to separate attribute similarity vectors $\gamma^{a,m}$ corresponding to matching and non-matching (child, candidate parent) pairs. The most straightforward approximation is given by

$$\frac{p(\gamma^{a,m} \mid M_a = m)}{p(\gamma^{a,m} \mid M_a \neq m)} \approx \frac{s(\gamma^{a,m})}{1 - s(\gamma^{a,m})},$$

where $s(\gamma^{a,m}) \in [0, 1]$ is the output of a probabilistic binary classifier trained to separate matching $\gamma^{a,m}$ vectors from non-matching ones. In some cases, the probabilities predicted by the classifier can be distorted, which can be countered by calibrating the probabilities [26].

We choose XGBoost [6] as the classifier s since it has been successfully employed for a record linkage task [22, 28] as well as other classification tasks previously. In our case, the probabilities predicted by the XGBoost classifier are fairly accurate (calibration does not improve the Brier score [4] of the classifier) so calibration is not used. This approach is called BINCLASS.

The key advantages that BINCLASS offers over NAIVEBAYES are: (*i*) we do not have to make the independence assumption for attribute similarities, (*ii*) we can use any existing classifier that provides probabilities as output, and (*iii*) it is very easy to add new features to the attribute similarity vector and retrain the model, whereas in NAIVEBAYES we have to estimate two new likelihood distributions for each new feature, which requires manually selecting a suitable distribution based on the type of the new feature.

In total, we use 20 attribute similarity features γ, which can be grouped into the following categories:

(1) **Candidate age** : The age of the candidate parent at the time of the child's birth and the difference to the reported parent age if available (for mothers, an approximated age is reported in 40% of the birth records).

(2) **Geographical distance** : The distance between the birth places of the child and the candidate parent.

(3) **Names** : The similarity of the *first* names, *middle* names (if any), *last* names, and *patronyms* reported in the child's record for the parent and in the candidate's record for the child.

(4) **NaiveBayes** : The probability estimated by the NAIVEBAYES method is used as a feature.

(5) **Gender** : A binary variable indicating whether we are matching a father or a mother.

(6) **Location** : The coordinates of the child's birth location.

(7) **Birth year** : The birth year of the child.

(8) **Candidate death** : A binary variable indicating whether there is a non-ambiguous death record, which indicates that the candidate had died before the birth of the child, and another variable indicating how long before the birth the death occurred (0 if it is not known to occur before the birth).
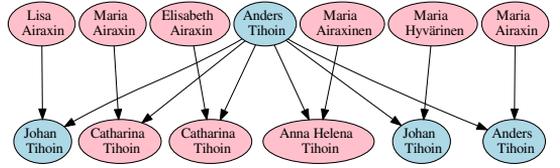
Feature groups (6)–(8) as well as middle name and patronym similarities are not used by the NAIVEBAYES model. Feature groups (6) and (7) are not useful on their own, since they are the same for each candidate parent of a child, but they might be informative together with other features. For instance, name similarity patterns might be time- or location-dependent.

To compute the features regarding the death year of the candidate (feature group (8)), we first match the death records to birth records by finding the record pairs that (*i*) have the same normalized first name, last name, and patronym, (*ii*) for which the reported age at death matches to the birth year ±1 year, and (*iii*) the birth and the death location are at most 60 kilometers apart. Then we link a death record to a birth record only if the death record has only one feasible match. This approach allows us to infer the death time for 12% of the birth records.[3]

3.4 Collective Approach

Both NAIVEBAYES and BINCLASS assume that the family links are independent, which leads to some unlikely outcomes. In particular, the number of spouses per person becomes unrealistically high which is illustrated in Figure 2 which shows a subgraph inferred by NAIVEBAYES. The algorithm has inferred a person called *Anders Tihoin* to have fathered six children—each with a different mother. While this is 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's name for these children is almost the same *(Maria Airaxin(en))*.

To address this problem, we propose to minimize the number mother–father pairs in addition to maximizing the probability of the inferred links. Let $y_{m,f} \in \{0, 1\}$ indicate whether any child has been assigned to mother–father pair (m, f). Now we can write the

[3]Based on the ground-truth data (Section 2.2), 96.4% of the obtained death record matches are correct, but we are able to match only 18% of the 3.3 million death records. Ideally, the death record matching should be done in a probabilistic way, jointly with the birth record linking.

Figure 2: A sample subnetwork inferred by NAIVEBAYES **with colors corresponding to gender. If each child is matched to the most probable father–mother pair independently, the number of spouses per person can be unrealistically high.**

collective genealogical network inference problem as

$$\max_{x,y} \quad \left[-\lambda \sum_{m,f} y_{m,f} + \sum_{a,m} \log p\left(M_a = m \mid \boldsymbol{\gamma}\right) x_{a,m} \right.$$
$$\left. + \sum_{a',f} \log p\left(F_{a'} = f \mid \boldsymbol{\gamma}\right) x_{a',f} \right] \qquad (6)$$

$$\text{such that} \quad \sum_{a} x_{a,m} x_{a,f} \leq y_{m,f}, \quad \text{for all } m, f, \qquad (7)$$

$$\sum_{m} x_{a,m} = 1, \quad \sum_{f} x_{a,f} = 1, \quad \text{for all } a, \qquad (8)$$

$$x_{a,m}, x_{a,f}, y_{m,f} \in \{0, 1\}, \quad \text{for all } a, m, f, \qquad (9)$$

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

This optimization problem is an instance of the *uncapacitated facility-location problem*, where parent pairs correspond to facilities, child nodes to demand sites, and the parameter λ to the facility opening cost. The uncapacitated facility-location problem is **NP**-hard for general graphs so we adopt a greedy approach presented in Algorithm 1. Assuming that each person has c candidate mothers and fathers, the time complexity of this algorithm is $O\left(c^2|V| + |V| \log|V|\right)$, which shows that it scales well in the number of people to be linked. The input probabilities p are computed with BINCLASS. Note that the facility assignment costs, that are based on the probabilities, are not necessarily a metric, so the approximation guarantees for methods, such as [16], do not necessarily hold in our problem setting.

This method is called COLLECTIVE. It outputs a genealogical network, where some of the links inferred by BINCLASS have been rewired in order to reduce the number of spouses. A limitation of COLLECTIVE is that it does not recompute the link probabilities (the marginal distributions of the random variables x, indicating the parents). One approach for computing the marginal distributions would be to adopt a Markov chain Monte Carlo (MCMC) method, which samples genealogical networks by proposing swaps to the parent assignments. However, to obtain a meaningful level of precision for the link probability estimates we need to sample a very large number genealogical networks, which renders such an MCMC approach non-scalable. Thus, in this work, we are not experimenting with link-probability estimation for COLLECTIVE.

Algorithm 1 A greedy method for collective genealogical network inference.

Input: Child nodes V, parent candidates C^m, C^f and their probabilities p, parameter λ.
Output: Inferred parent–child links R.
1: $R = \emptyset$
2: $Q = \emptyset$ ▷ Set of used mother–father pairs.
3: Sort children $a \in V$ by $\max_{m,f} \log p\,(M_a = m \mid \boldsymbol{\gamma})\ +$
  $\log p\,(F_a = f \mid \boldsymbol{\gamma})$ in a descending order.
4: **for** $a \in V$ **do**
5:   $p_{\max} = -1$ ▷ Maximum probability.
6:   $m_{\max} = -1$ ▷ Best mother.
7:   $f_{\max} = -1$ ▷ Best father.
8:   **for** $(m, f) \in C_a^m \times C_a^f$ **do**
9:     $p_{m,f} = \log p\,(M_a = m \mid \boldsymbol{\gamma}) + \log p\,(F_a = f \mid \boldsymbol{\gamma})$
10:     **if** $(m, f) \notin Q$ **then**
11:       $p_{m,f} = p_{m,f} - \lambda$
12:     **if** $p_{m,f} > p_{\max}$ **then**
13:       $p_{\max} = p_{m,f}$
14:       $m_{\max} = m$
15:       $f_{\max} = f$
16:   $R = R \cup (m_{\max}, a) \cup (f_{\max}, a)$
17:   $Q = Q \cup (m_{\max}, f_{\max})$
18: **return R**

Table 1: Accuracy of the links inferred with different methods.

Method	RandomCand	NaiveBayes	BinClass	Collective
Accuracy	12.5%	56.9%	61.6%	**65.1%**

4 EXPERIMENTAL EVALUATION

Next, we compare the proposed methods, BinClass and Collective, to two baseline methods, NaiveBayes [24] and RandomCand, the latter of which randomly assigns the parents among the set of candidates. The methods are evaluated by computing their *link accuracy* which is the fraction of ground-truth child–parent links correctly inferred by the method. To use Collective, we first need to optimize λ—the penalty induced by each extra parent pair. The optimization is done using the training ground-truth data and the results are shown in Figure 3a. The training accuracy is maximized when $\lambda = 1.6$.

The results for all methods are presented in Table 1. BinClass clearly outperforms the two baseline methods, and 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 probabilities and compute accuracy within each bin. The results presented in Figure 3b show that the estimated link probabilities are somewhat pessimistic, but overall, they are well in line with the link accuracies. Therefore, we can use the link probabilities to filter out links below a desired certainty level when analyzing the inferred network, as done in the next section. This filtering improves the precision of the inferred network but naturally also decreases the recall as illustrated in Figure 3c. In total,

BinClass finds 1.8 million child–mother links and 2.5 million child–father links. If we require a minimum link probability of 90%, as done in the next section, we are still left with 253 814 child–mother links and 341 010 child–father links.

5 CASE STUDY: ASSORTATIVE MATING

Assortative mating, also known as *social homogamy*, refers to the phenomenon that people tend to marry spouses with a similar socioeconomic status and it is one instance of *social stratification*. A recent study shows that assortative mating contributes to income inequality and it has been on the rise between 1960 and 2005 [14]. In this section, we leverage the inferred genealogical network to address two questions:

 (*i*) Can we detect assortative mating in historical Finland?
 (*ii*) How has the intensity of assortative mating evolved in historical Finland?

5.1 Estimating Social Status

To detect assortative mating, we compare the socioeconomic status of the spouses inferred by our linking method. We use *occupation* as a proxy for status, and rather than comparing the occupations of the spouses directly, we compare the occupations of the spouses' fathers. The father occupations are more comparable since occupations were strongly gendered in the 18th and 19th centuries, which are the focus of this study. Furthermore, there is a separate field for father occupation in the birth records so we do not need to separately infer the fathers but only the spouses.

The numerous variants and alternative abbreviations of the same occupation pose a challenge when comparing occupations. For instance, *farmer*, which is the most common occupation in the dataset, goes by the following (non-comprehensive) list of titles: *bonden, bd., b., bd:, b:, b:n, bdn, b:den, talollinen, talonpoika, tal., talp., tl., tln., talonp*. To address this challenge, we normalize the occupation titles by first removing special characters and then comparing the title to lists of abbreviations available online [1, 15]. We have also manually normalized the most common abbreviations not found in the online lists to increase the coverage of the normalization. In total, we are able to normalize 77.5% of all non-empty father occupation titles found in the birth records.

In addition to comparing the normalized occupations directly, we also map them to the historical international classification of occupations (HISCO) [29]. Then we divide the HISCO classes into four main classes: (1) upper and middle class, (2) peasants (who own land), (3) crofters (who rent land), and (4) labourers (who live at another person's house). The class to which the father of a person belongs is called Class4. We also map the HISCO classes into an occupational stratification scale HISCAM [20].[4] HISCAM is a real-valued number between 0 and 100 which measures the social interaction distance of people based on their occupations. The Class4 and HISCAM codes are obtained for 75.5% of the people.

5.2 Measuring Assortative Mating

A high percentage of matching spouse father occupations (p) is a signal of assortative mating but this percentage might also be

---
[4]More specifically, we use the "U2: Male only, 1800–1938" scale. For more information on the different scales, see: http://www.camsis.stir.ac.uk/hiscam/
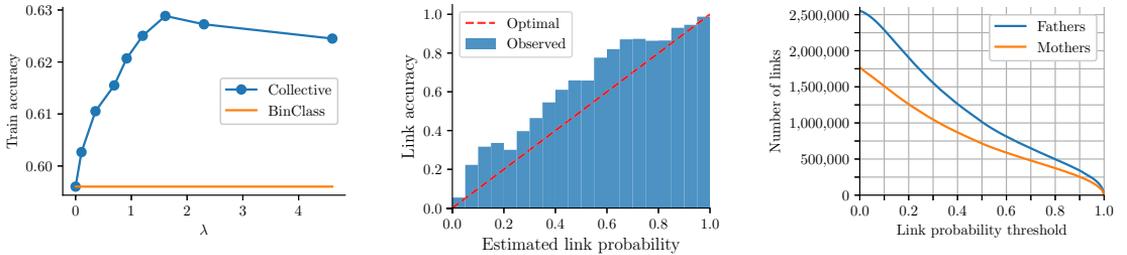
(a) Optimizing parameter λ which controls the penalty induced by each extra parent pair in the Collective linking method.

(b) The link-probability estimates by BinClass correlate strongly with the accuracy of the links binned by their probability.

(c) The number of inferred family links with the estimated link probability above a given threshold.

Figure 3: Experimental results on genealogical network inference.

affected by external factors such as the number of distinct occupations people had at a given time in a given city or the availability of data from different cities at different time periods. To control for these external factors, we introduce a null model, which shuffles the spouses within a city and a time window of 20 years and then computes the percentage of matching occupations (p_n). Then we measure assortative mating as the ratio of 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 one is a sign of assortative mating.

The higher-level occupation categories Class4 can also be used to compute match percentages (q and q_n) and their ratio q/q_n. With the HISCAM scores, we compute the mean absolute difference of the scores for the spouses (δ) and for the null model (δ_n), instead of the percentages. The assortative mating measure, in this case, is defined as δ_n/δ, so that again, a ratio larger than one indicates assortative mating.

5.3 Data

As the set of spouses to be analyzed, we use all pairs of people who have been inferred to be the mother and the father of the same child. Both parent links are required to have at least a 90% probability (this threshold is varied in Appendix A). We limit the analysis to the period with the most records from 1735 to 1885. Requiring both parents to have a sufficiently high link probability limits the number of spouses to 14 542 pairs. Out of these, the father occupation is known for both spouses in 6 402 pairs and Class4 and HISCAM in 3 128 pairs. Although these filtering steps significantly reduce the number of spouses, we are still left with a sufficiently large dataset to perform a longitudinal analysis on assortative mating.

5.4 Results

The percentages of matching spouse father occupations over time and the corresponding assortative mating curve are shown on the top row of Figure 4. In order to highlight long-term trends, the curves show moving averages where a data point at year y uses the inferred spouses from years $[y-10, y+10]$ (the effect of varying

the delta value of 10 years is studied in Appendix A). The figure also shows the 95% bootstrap confidence intervals.

The middle and the bottom rows show the corresponding curves for the occupations grouped into four main classes (Class4) and for the numerical HISCAM scores of the occupations, respectively.

All three measures of assortative mating suggest that assortative mating did occur in Finland between years 1735 and 1885 since the assortative mating curves are fairly consistently above the baseline ratio of 1. The intensity of the phenomenon varies mainly between 1 and 1.5 but interestingly, there is no monotonically decreasing or increasing trend.

Finally, we observe that the spouse similarity curves p, q and δ have very different shapes, whereas the three assortative mating curves are clearly correlated (the correlation coefficients are 0.69, 0.23, and 0.62 for pairs (p/p_n vs. q/q_n), (p/p_n vs. δ_n/δ), and (q/q_n vs. δ_n/δ), respectively). This suggests that the proposed measures of assortative mating, which account for a null model, measure the phenomenon robustly.

6 RELATED WORK

Genealogical network inference, also known as *population reconstruction* [3], is one application of *record linkage*, which has been an active research area for many decades and was mentioned already in 1946 by Halbert L. Dunn [10]—interestingly, in the context of linking birth, marriage, and other vital records. Two decades later, Fellegi and Sunter published a widely cited paper on probabilistic record linkage [12]. This approach considers a vector of attribute similarities between two records to be matched and then computes the optimal decision rule (*link* vs. *possible link* vs. *non-link*), assuming independent attributes. The BinClass method, proposed in this work, adopts a similar approach, however, employing a supervised classifier, which has the advantage of capturing dependencies between variables.

Record linkage has been mostly studied for other applications, but recently, with the rise in the number of indexed genealogical datasets, several studies have applied record linkage techniques for
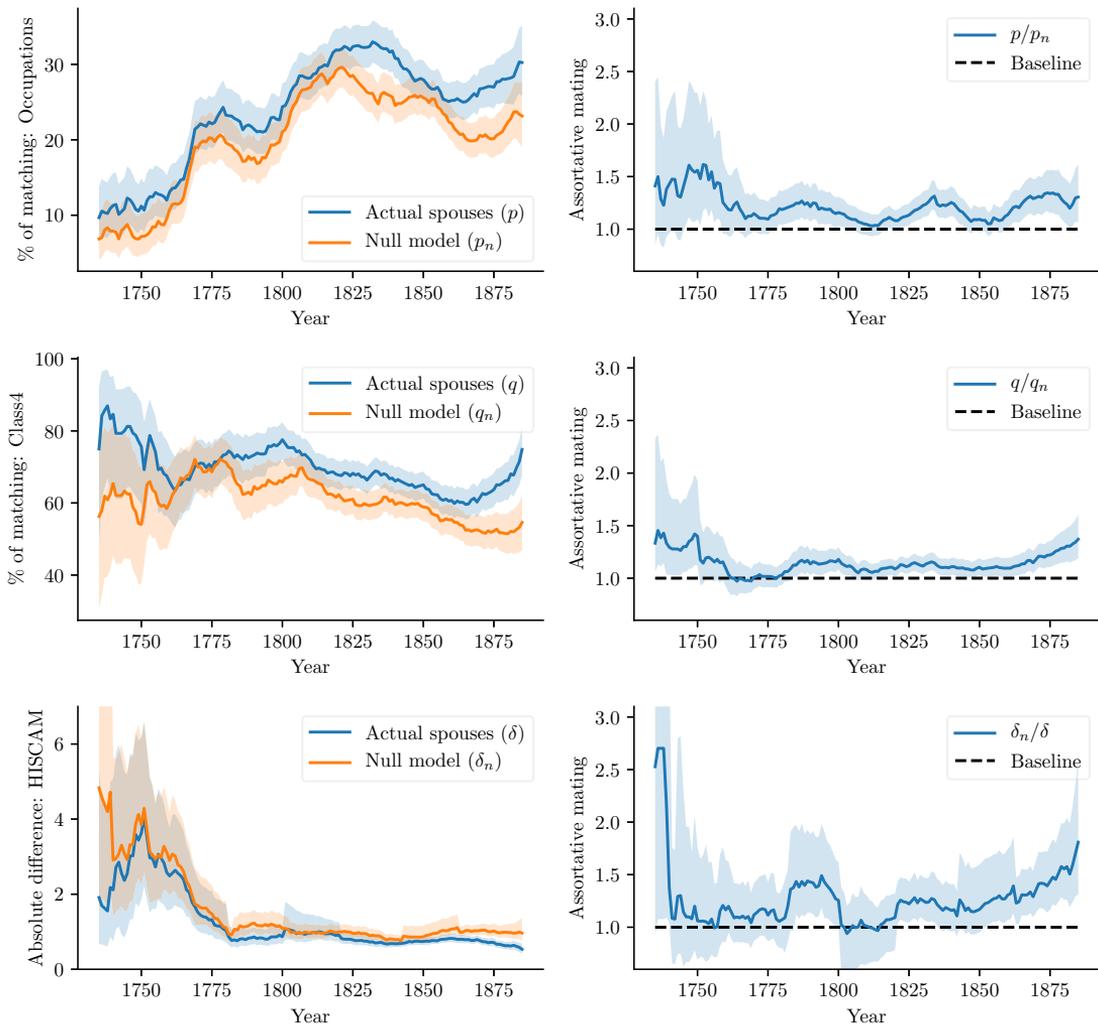
Figure 4: Assortative mating is detected in the inferred genealogical networks for Finland (1735–1885), but the phenomenon is not monotonously decreasing or increasing. Shaded areas correspond to the 95% bootstrap confidence intervals.

genealogical data [7, 8, 11, 18, 19, 23, 25, 27]. The most closely related to our work are the papers by Efremova et al. [11], Christen [7], and Kouki et al. [19].

Efremova et al. [11] consider the problem of linking records from multiple genealogical datasets. They cast the linking problem into supervised binary classification tasks, similar to this work, and find name popularity, geographical distance, and co-reference information to be important features. 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 probability of the candidate.

Christen [7] and Kouki et al. [19] propose collective methods for linking vital records. The former method is evaluated on historical Scottish data and it is concluded that due many intrinsically difficult linking cases, the results are inferior to a linkage constructed by a domain expert. The latter method is evaluated on a more recent dataset collected by the National Institutes of Health and it

yields a high linking F-measure of 0.946. This method is based on probabilistic soft logic (PSL), which makes it possible to add new relational rules to the model without having to update the inference method. In comparison to these works, we evaluate our methods on a dataset that is two orders of magnitude larger, and we show that the proposed BinClass method not only provides accurate matches but also reliably quantifies the certainty of the matches—an important feature of a practical entity-resolution system.

Assortative mating,[5] also known as *social homogamy*, has been studied widely in the sociology literature. The phenomenon has a significant societal impact since it has been shown to be connected to income inequality [14]. Greenwood et. al. [14] also find that assortative mating has been on the rise 1960–2005. Bull [5] studies assortative mating in Norway, 1750–1900, focusing on farmers and farm workers. They find that assortative mating was declining in the mid-1700s but after that it stayed fairly constant. This finding is largely in line with our results for Finland from the same time period. To control for the variations in the sizes of the groups, Bull computes odds ratios [5, 17], whereas we use a null model, which randomizes spouses.

7 DISCUSSION AND CONCLUSIONS

We presented a principled probabilistic machine-learning approach, BinClass, for inferring genealogical networks. This approach was applied to a dataset of 5.0 million birth records and 3.3 million death records based on which it inferred a network, containing 13 generations and a connected component of 2.6 million individuals, taking about one hour when parallelized over 50 machines. A comparison against a large human-compiled network yielded a link accuracy of 61.6%, outperforming a naive Bayes baseline method. We showed that the accuracy can be further improved to 65.1% by a collective approach called Collective.

A key feature of BinClass is that it outputs probabilities for the inferred links. This allowed us to separate links that have a sufficiently high probability and to perform an analysis on the mating patterns observed in the network. The main findings of the analysis are that: (i) assortative mating—the tendency to select a spouse with a similar socioeconomic status—did occur in Finland, 1735–1885, and (ii) assortative mating did not monotonically decrease nor increase during this time period.

Limitations and future work. There is always uncertainty involved when analyzing records that are several centuries old and even an experienced domain expert can make incorrect linking decisions. Since our method relies on human-generated training data, any mistakes or biases in the training dataset affect the resulting model and potentially the downstream analyses. Although the automatic approach enables data generation for various analyses at an unprecedented scale, one has to be careful when drawing conclusions. For instance, an interesting phenomenon that could be studied with the inferred networks would be long-term human mobility patterns. However, since the model uses features that are based on location, any potential biases learned by the model would directly affect the analysis of the resulting mobility patterns.

Many exciting future directions are left unexplored. First, it would be useful to extend the problem formulation so that it tries to jointly link all available record types, as now we only link birth records, using some features based on death records linked to the births separately. Second, it would be interesting to incorporate additional collective terms to the optimization problem, capturing patterns such as namesaking (a practice to name a child after an ancestor) or constraints on the family relations between individuals, which could be obtained from DNA tests. Third, there are many options for extending the assortative mating analysis presented in this work. We could study the differences in the strength of this phenomenon between social classes and between, for example, the first and the latter children. We could also extend the analysis to a related phenomenon of *social mobility*, which considers the status differences between parents and children instead of spouses. Fourth, with the advent of improved handwritten text recognition methods, similar historical birth record datasets can be expected to become available for many other countries. This will enable studying the generalizability of the proposed methods and the assortative mating analysis beyond Finland. Finally, the presented inference and analysis methods could be applied to other types of genealogical networks such as the genealogical networks of Web content [2].

To facilitate future research on these and other related research problems, the data and the code used in this paper have been made available at: https://github.com/ekQ/genealogy.

A ASSORTATIVE MATING SENSITIVITY ANALYSIS

In this section, we study the effect of varying the minimum link probability threshold (p_{th}) of 90% and the moving average delta (Δt) of 10 years used in the analysis of assortative mating in Section 5. In the interest of space, we consider only the first assortative mating measure p/pN, which is based on the normalized occupations. The results are shown in Figure 5.

First, we note that the larger the threshold p_{th} or the smaller the smoothing parameter Δt, the less data points we have per year and thus larger the confidence intervals. Second, varying the threshold does not seem to affect the p/pN curves significantly. Third, decreasing the smoothing parameter to $\Delta t = 3$ makes the curves considerably less smooth, which suggests that a higher smoothing parameter might highlight the long-term trends more effectively.

We conclude that these results support the findings that assortative mating did occur in Finland, 1735–1885, but it did not monotonically decrease nor increase during this time period.

---

[5]Assortative mating can also refer to *genetic* assortative mating, but here we use it exclusively to refer to sociological assortative mating.
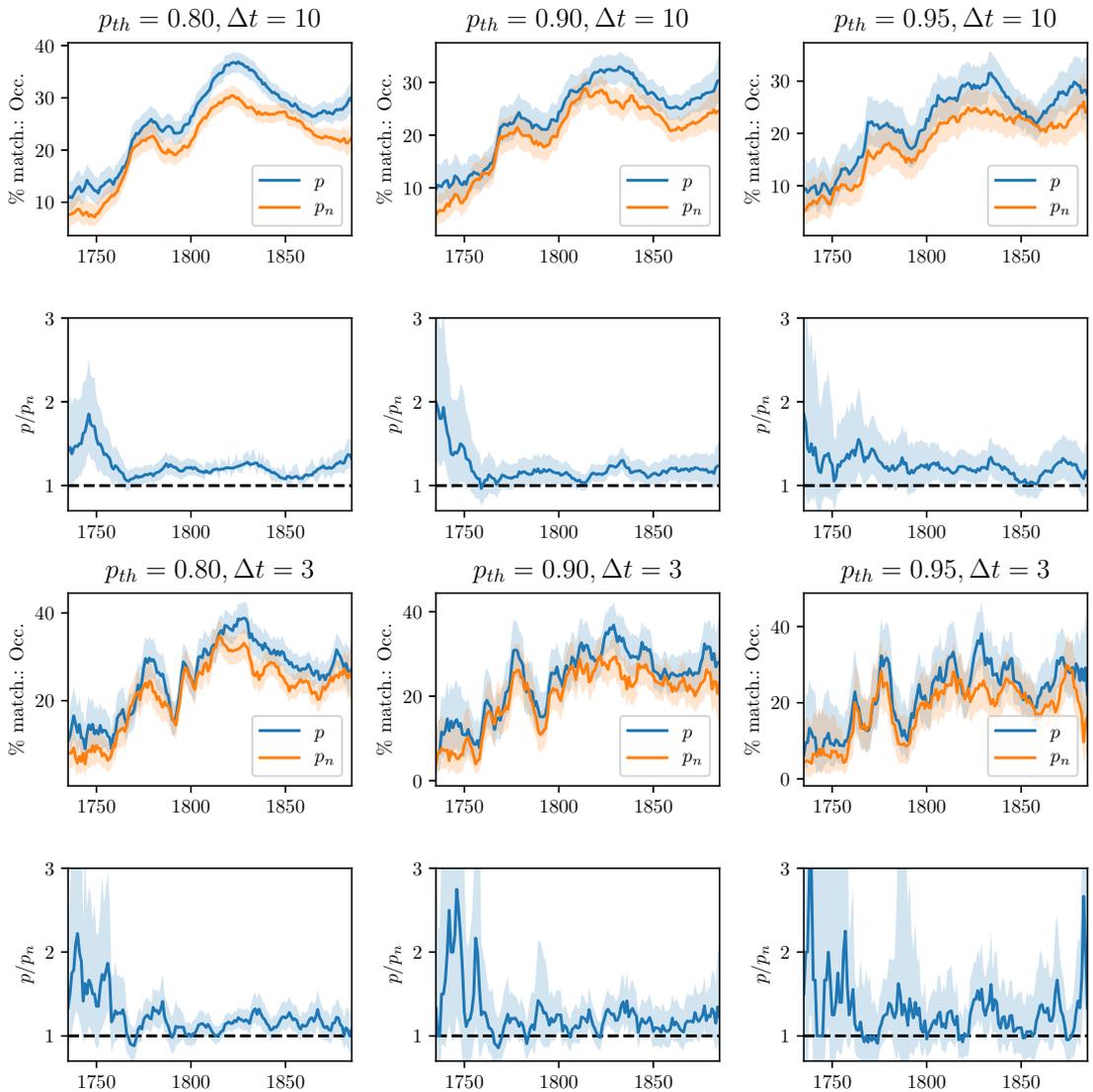
Figure 5: A sensitivity analysis showing the effect of varying the minimum link probability threshold (p_{th}) and the moving average delta (Δt) on the assortative mating curves from Figure 4 (top row).

REFERENCES
[1] 2016. Lyhenteitä (Abbreviations). http://www.saunalahti.fi/hirvela/historismi_sivut/termitnew.html. (2016). Accessed: 2017-10-25.
[2] Ricardo Baeza-Yates, Álvaro Pereira, and Nivio Ziviani. 2008. Genealogical trees on the Web: a search engine user perspective. In *Proc. WWW*.
[3] Gerrit Bloothooft, Peter Christen, Kees Mandemakers, and Marijn Schraagen. 2015. *Population Reconstruction*. Springer.

[4] Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthey Weather Review* 78, 1 (1950), 1–3.
[5] Hans Henrik Bull. 2005. Deciding whom to marry in a rural two-class society: Social homogamy and constraints in the marriage market in Rendalen, Norway, 1750–1900. *International review of social History* 50, S13 (2005), 43–63.
[6] Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proc. KDD*.
[7] P. Christen. 2016. Application of advanced record linkage techniques for complex population reconstruction. *ArXiv e-prints* (Dec. 2016). arXiv:cs.DB/1612.04286

[8] Peter Christen, Dinusha Vatsalan, and Zhichun Fu. 2015. Advanced record linkage methods and privacy aspects for population reconstruction—A survey and case studies. In *Population Reconstruction*. Springer, 87–110.

[9] Kyle Cranmer, Juan Pavez, and Gilles Louppe. 2016. Approximating likelihood ratios with calibrated discriminative classifiers. *ArXiv e-prints* (March 2016). arXiv:stat.AP/1506.02169

[10] Halbert L Dunn. 1946. Record linkage. *American Journal of Public Health and the Nations Health* 36, 12 (1946), 1412–1416.

[11] Julia Efremova, Bijan Ranjbar-Sahraei, Hossein Rahmani, Frans A Oliehoek, Toon Calders, Karl Tuyls, and Gerhard Weiss. 2015. Multi-source entity resolution for genealogical data. In *Population Reconstruction*. Springer, 129–154.

[12] Ivan P Fellegi and Alan B Sunter. 1969. A theory for record linkage. *AmerStatAssoc* 64, 328 (1969), 1183–1210.

[13] Angelos P Giotis, Giorgos Sfikas, Basilis Gatos, and Christophoros Nikou. 2017. A survey of document image word spotting techniques. *Pattern Recognition* 68 (2017), 310–332.

[14] Jeremy Greenwood, Nezih Guner, Georgi Kocharkov, and Cezar Santos. 2014. Marry your like: Assortative mating and income inequality. *The American Economic Review* 104, 5 (2014), 348–353.

[15] Aija Greus and Harri Hirvelä. 2016. Ammatit (Occupations). http://www.saunalahti.fi/hirvela/historismi_sivut/ammatit%20a-k.html. (2016). Accessed: 2017-10-25.

[16] Kamal Jain and Vijay V Vazirani. 2001. Approximation algorithms for metric facility location and k-median problems using the primal-dual schema and Lagrangian relaxation. *Journal of the ACM (JACM)* 48, 2 (2001), 274–296.

[17] Matthijs Kalmijn. 1998. Intermarriage and homogamy: Causes, patterns, trends. *Annual review of sociology* 24, 1 (1998), 395–421.

[18] Pigi Kouki, Christopher Marcum, Laura Koehly, and Lise Getoor. 2016. Entity resolution in familial networks. In *Proc. MLG*.

[19] Pigi Kouki, Jay Pujara, Christopher Marcum, Laura Koehly, and Lise Getoor. 2017. Collective Entity Resolution in Familial Networks. In *Proc. ICDM*.

[20] Paul S Lambert, Richard L Zijdeman, Marco HD Van Leeuwen, Ineke Maas, and Kenneth Prandy. 2013. 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 (2013), 77–89.

[21] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, et al. 2009. Life in the network: the coming age of computational social science. *Science* 323, 5915 (2009), 721–723.

[22] Jianxun Lian and Xing Xie. 2016. Cross-device user matching based on massive browse logs: The runner-up solution for the 2016 CIKM Cup.

[23] Eric Malmi, Sanjay Chawla, and Aristides Gionis. 2017. Lagrangian relaxations for multiple network alignment. *Data Mining and Knowledge Discovery* (2017), 1–28.

[24] Eric Malmi, Marko Rasa, and Aristides Gionis. 2017. AncestryAI: A tool for exploring computationally inferred family trees. In *Proc. WWW Companion*.

[25] Eric Malmi, Evimaria Terzi, and Aristides Gionis. 2017. Active Network Alignment: A Matching-Based Approach. In *Proc. CIKM*.

[26] Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Proc. ICML*.

[27] Bijan Ranjbar-Sahraei, Julia Efremova, Hossein Rahmani, Toon Calders, Karl Tuyls, and Gerhard Weiss. 2015. HiDER: Query-driven entity resolution for historical data. In *Proc. ECML PKDD*.

[28] Yi Tay, Cong-Minh Phan, and Tuan-Anh Nguyen Pham. 2016. Cross device matching for online advertising with neural feature ensembles: First place solution at CIKM Cup 2016.

[29] Marco HD Van Leeuwen, Ineke Maas, and Andrew Miles. 2002. *HISCO: Historical international standard classification of occupations*. Leuven University Press.

[30] Ingmar Weber and Bogdan State. 2017. Digital Demography. In *Proc. WWW Companion*.