APPENDIX A: FEATURE GENERATION

This appendix provides details of feature generation, step (b) in our process, as described in Section 3.2. In demonstrating this, we extracted the examples from the Isle of Skye dataset [1].

Node attributes

The Isle of Skye(IoS) dataset [1] contains attributes recording information related to the births of children, including child's first and last name, date of birth, gender, birth address, parish in which the birth was recorded, parent's names, date and place of marriage, their occupation, the informant of the birth and the relationship with the informant.

We ignored attributes with a high number of missing values due to the impracticality of imputing because missing values have a meaning in this domain. For example, in a birth certificate an empty first name resembles a baby who has died soon after birth. So even though the birth is registered, the baby is not named. Similarly, an empty parent's marriage date represents unmarried parents. While each birth record forms a node in the graph, the features constructed utilizing the attributes act as the characteristics of those nodes. Table 1 shows the attributes we considered to generate the IoS node features. Please note that all values provided in this table are anonymized for confidentiality purpose.

Table 1. Attributes of the Isle of Skye dataset with values for three made-up birth certificates.

Attribute list	Birth record 1	Birth record 2	Birth record 3
Child's first name	Alan	Isabella	Margaret
Child's last name	Chisholm	Cameron	Chisholm
Source parish	Bracadale	Duirinish	Bracadale
Gender	Male	Female	Female
Date of birth	03/07/1867	4/05/1866	15/03/1869
Address	Balgown	Coishletter	Struan
Father's occupation	Shepherd	Tailor	Farmer
Mother's first name	Katie	Elisa	Katie
Mother's maiden name	Mcleod	Maccaskill	Mcleod
Mother's last name	Chisholm	Cameron	Chisholm
Father's first name	John	Angus	John
Father's last name	Chisholm	Cameron	Chisholm
Parent's marriage date	24/04/1863	30/09/1865	24/04/1863

Node Features

From each attribute shown in Table 1, one or more binary features can be generated. Our approach is capable of generating features from attributes containing textual, numerical, or categorical data. Following are the types of binary features we generated for nodes.

- **Presence check features:** these features check for the presence or absence of the attribute value. For example, for the attribute gender, its corresponding presence check binary feature is *GenderPresent*. The feature value is set to $True$ if a gender is present, whereas it is set to $False$ if the value is absent.
- **Frequency based features:** for each attribute value, the count of occurrence of a feature value is obtained and compared against a threshold such as the mean or median value of all values of this attribute to determine if the feature value under consideration is rare or common. For example, for the attribute child's first name, its corresponding frequency based

feature is *CommonFnamePopulation*. The feature value is set to $True$ if the first name is common within the population, and it is set to $False$ if the value is uncommon.

- **Multi attribute features:** these are features generated by considering two different attributes. For example, the two attributes child's date of birth and parent's date of marriage are compared to generate the feature *PrenupPregnancy*. If the time difference between the values of the two attributes is not at least 7 months apart, then this indicates a prenuptial pregnancy. Hence where there is prenuptial pregnancy, the feature value is set to $True$.

Edge Features

Edge features are generated using the attributes of the participating nodes of an edge. In vital records where a node represents a birth record, two nodes are connected if they are identified as siblings, such as for example between birth record 1 and 3 in Table 1. Thus created edge features will belong to either one of the following types.

- **Equality features:** these features are generated by comparing matching attributes of the two participating nodes. For example, the two last names of birth records identified as siblings can be matched to form the feature *SameLname*. If the names are a match, the feature value is set to $True$. If the last names are mismatching, the feature value is set to $False$.
- **Spatial features:** these features are generated by considering geographical factors such as parish, postcode and so on. For example, features determining the parish and village the birth of the siblings have taken place. This provides information about the migration of families possibly due to change of occupations of the parents, which is further confirmed by the change of occupation across siblings.
- **Temporal features:** temporal constraints consider date/time factors to generate features. For example, compare the number of days or months between two births to validate an edge between two siblings. If the difference between two dates of birth is is 0 or 1, the two siblings are considered as twins. If the births are not at least 7 months apart, then either the two babies are not siblings, there is a data entry error, or there is a historical reasoning behind the two births [1].

REFERENCES

[1] Alice Reid, Ros Davies, and Eilidh Garrett. 2006. Nineteenth-Century Scottish Demography From Linked Censuses and Civil Registers: A 'Sets of Related Individuals' Approach. *History and Computing* 14, 1–2 (2006), 61–86.