Digital Scholarship in the Humanities, 2024, **39**, 215–227
https://doi.org/10.1093/llc/fqad108
Advance access publication 31 January 2024
Full Paper

Normalization of kinship relations to enrich family network analysis: case study on China biographical database

Bin Li [1,2], **Yiguo Yuan** [1,2,*], **Xuehui Lu**[1,2], **Peter K. Bol**[3]

[1]School of Chinese Language and Literature, Nanjing Normal University, No.122, Ninghai Road, Nanjing 210097, China
[2]Center of Language Big Data and Computational Humanities, Nanjing Normal University, No.122, Ninghai Road, Nanjing 210097, China
[3]East Asian Languages and Civilizations, Harvard University, Cambridge, MA 02138, USA

*Corresponding author. Center of Language Big Data and Computational Humanities, Nanjing Normal University, China.
E-mail: lexcliff1023@gmail.com

Abstract

Kinship is an important issue in history studies. The kinship database is the key resource to analyze the structure, succession, and evolution of families. However, one kinship could be expressed by different words, and one kinship word may be vague and ambiguous in natural languages, especially in pre-modern Chinese. As in the well-known China Biographical Database, which contains 484,066 kinship instances, there are more than 400 kinship words. Thus, the relations extracted from history texts cannot be directly used to build family networks. In this article, we put forward a novel method to normalize kinship relations by three basic relations: father–descendant, mother–descendant, and husband–wife, as well as the gender of each person. All types of kinships are normalized to these three basic relations. In this way, we identified 178,390 basic kinship relations to fully describe the original 462,147 unambiguous kinship instances, while finding 3,989 inconsistencies and inferring 5,805 missing persons. Then, we generate 29,423 families by basic kinship relations and analyze the properties of families, such as their sizes, depths, and intermarriages across families. This type of family analysis had been almost impossible prior to normalizing kinship relations. Therefore, this technique enables improved family database construction and deeper quantitative analysis.

Keywords: kinship; family network; normalization; network analysis; pre-modern China.

1. Introduction

Kinship is a basic social relation, playing an important role in history studies. Kinship data are fundamental to reconstruct the families in history. Families exist both for the well-being of their members and for the well-being of society, and they offer predictability, structure, and safety as members mature and participate in the community (Collins et al. 2012). The conventional kinship databases are built on the hand-inputed records. Publicly available online family trees are collected from genealogy enthusiasts and used to partition the genetic architecture of longevity (Kaplanis et al. 2018). At present, there are many commercially developed genealogy packages and websites for individuals to query their family history. These databases are applied to predict political, economic, and genetic activities. Using the data collected from the National Household Targeting System for Poverty Reduction (NHTS-PR), intermarriage networks are reconstructed for more than 15,000 villages of 709 towns in Philippines, which proves the correlation between politicians' family networks and election results (Cruz, Labonne, and Querubin et al. 2017). The application of both database and genealogical programs, such as Genopro and Heredis, for family studies has helped historians to develop their research over the last few decades (Pérez García 2011).

However, a large number of studies obtained family data by various ways. The funerary inscriptions are used to study the nuclear family of ancient Rome (Saller and Shaw 1984) and study the construction of ancient Roman families (Martin 1996). Household register is also an effective source of family data. By using the household registers from Dunhuang and Turfan in China, household forms and related demographic characteristics in a western region of the Tang Empire are examined (Liao 2001). Considering the

quantity of archaeological objects, the scales of the family data involved in these studies are relatively small. With the development of computational technology, database has become an important means for large family data research (Warren et al., 2016; De Nooy et al., 2018). The Integrated Public Use Microdata Series Database is used to explore the origins of African-American family structure (Ruggles 1994). And the Utah Population Database is used to evaluate the influence of family history on longevity (Kerber et al. 2001).

Empirical studies on families rely on well-formed family data, either by hand-input or by gene test. Thus, the collection of such family data is a primary task. However, one kinship could be expressed by different words. For example, the kinship father–descendant is expressed by *father, son, daughter, second son, elder daughter*, etc. On the other hand, one kinship word may be vague and ambiguous. For example, *descendant* and *relative* are not clear enough to determine the exact kinship between two persons. These two issues are much harder to handle in Chinese than in English, because there are many more kinship words. Appendix 1 is a kinship chart of mandarin Chinese, showing the common words used in describing the kinships in a family.

China has a large amount of pre-modern books, recording the kinship genealogy. China Biographical Database (CBDB)[1] is a high-quality resource recording 484,066 relation instances with 255,149 ancient Chinese persons. It collects ancient Chinese historical records, funerary inscriptions, official documents, and other biographical materials systematically. CBDB has been applied to studies of ancient China by many scholars (Liu and Wang, 2017; Shang and Huang 2018).

There are more than 400 types of words expressing kinship relations in CBDB. This huge number of words is manually collected from thousands of books. Each word has to be clarified to a certain type of kinship. At the meantime, there are conflicts in different documents recording the same pair of persons. Thus, error detection and correction are also important in data consistency. Even these two issues can be solved, the family network does not have a good structure. For example, if person A is the son of B, and C is the grandfather of B, then A and C are the adjacent nodes of B in the network. But kinship research needs a tree structure dominating the family relations. A should be C's great-grandson in the family tree.

The motivation of our research is exploring the normalization of kinship relations to basic kinship relations. Many efforts have been made to effectively represent and analyze kinship relations. Family trees are mathematical graph structures that capture two fundamental processes: mating and parenthood (Kaplanis et al. 2018). Kinship networks can also be represented as graphs, and cumbersome genetic graph can be replaced with "p-graph", a more succinct parental graph that has genetic graphs as its dual composed of three partial orders: male ancestral trees, female ancestral trees, and composite forests of trees (White and Jorion 1996). The basic kinship relations used to draw such trees are still in debate, although the kinship relation between two persons can be computed by a third person whose kinship relations to each of the two persons are already known (Read 2001). One theory of the "atom of kinship" (Lévi-Strauss 1969) is based on filiation (father–descendant, brother–sister), affinity (husband–wife), and filiation combined with affinity (wife's brother–sister's son), where father and husband play crucial roles in this theory.

In this article, we introduce our work of normalizing kinship relations from original kinship words to three basic kinship relations, father–descendant, mother–descendant, and husband–wife, as well as the gender of each person, to build family data in CBDB and analyze ancient Chinese families quantitatively. The kinship data is extracted from CBDB, which contains 409 kinship word types that need to be normalized. After normalization using PYTHON, we find that the data consistency and the kinship relations can be enriched by logic inference. For example, if A is the grandfather of B, and A is the father of C, then C is the father of B.

Most of the kinship instances are normalized and the patrilineal trees are built as family trees. The patrilineal tree is chosen from the above three partial orders because males were more frequently recorded in biographies and the number of males is about 4.15 times greater than that of females in CBDB. Finally, when considering marriages across families using the husband's relation, the kinship data forms a network. Thus, the normalization and inferences supply a well-structured database to make a quantitative analysis of ancient Chinese families.

2. Data sources

The CBDB is a freely accessible relational database with biographical information about approximately 484,066 individuals (version 2018), primarily from the 7th through 19th centuries. CBDB contains many kinds of detailed information of ancient Chinese persons, but we only extracted individual ID, name, gender, year of death (see Table 1), and kinship data (see Table 2) in this article.

Table 2 shows the kinship between two persons. The person on the right is the kinsfolk (described by the kinship word) of the person on the left, such as 13059_李淵 is the father(父) of 13060_李世民. We added persons' ID in front of their names to distinguish them. The three kinship instances can be expressed by logic form as shown in Table 3.

The former in parentheses is the kinsfolk (described as the kinship word) of the latter. To be precise, the grandfather in Chinese refers to the paternal grandfather unless it is clearly marked as the maternal grandfather, so does the grandmother, great-grandfather, and uncle. The form of the expressions in Table 3 will also be used to represent relations between two persons in the rest of this article.

We get 484,416 kinship instances and 245,371 persons from CBDB. More than 95 per cent of the data are used shown in Table 4. Some kinship relations like 亲戚 (relatives) and 族兄 (clan sibling) are dropped, as they are not definite kinships, thus cannot correspond to a certain type of kinship.

Within the huge number of kinship instances, there are many redundancies and conflicts. On one hand, a kinship relation can be recorded in several kinship word types and instances. For example, the father–descendant relation between 13059_李淵 and 13060_李世民 can also be recorded as 子 *(son)*, 次子 *(second son)*. On the other hand, a kinship instance may conflict with other kinship instances or persons' information. Thus, we need to detect them automatically before further processing.

3. Normalization

3.1 Basic kinships

Basic kinships are important for the normalization of kinship words of ancient Chinese families. Here, we sort out and draw a schematic diagram of kinship words in modern Chinese families centered around a man who possess both son and husband identities (see Appendix 1). It can be seen that kinship words of modern Chinese families are more complicated than English, due to the marriage system which includes concubines. Basic kinships are the core of kinship relations, with which any kind of kinship can be defined as a group of simplest kinship relations. Father and husband are regarded as the foundation of the atom of kinship (Lévi-Strauss 1969). Considering the marriage system in pre-modern China, there are three basic kinships that are enough to make up a family: father–descendant, mother–descendant, and husband–wife relations, abbreviated as *father*, *mother*, and *husband*.

Table 1. Persons' information.

Person_ID	Person_name	Gender	Year of death
13059	李淵 (Li Yuan)	Male	625 AD
13060	李世民 (Li Shimin)	Male	649 AD
140340	李貞 (Li Zhen)	Male	684 AD
3767	蘇軾 (Su Shi)	Male	1095 AD
27127	釋惟簡 (Shi Weijian)	Male	1071 AD

3.2 Normalization

The instances of basic kinships are the basis for the construction of family data in the next step. In CBDB, there are sixty-four kinship words directly expressing the three basic kinships. For example, the kinship father is expressed by the words like 長子 (the first son), 獨女 (the only daughter), 二女 (the second daughter), 妾之子 (concubine's son), etc. The sixty-four kinship words are listed in Appendix 2. There is an overlapping part between kinship words describing father relation and mother relation, depending on the genders of the parents.

We extract kinship instances by these kinship words and normalize them to remove redundancy. For example, *father*(A, B) is equal to *son*(B, A). As shown in Table 5, we obtained 121,829 basic kinship instances. And after the normalization, we even found some conflicts (see Section 3.3).

3.3 Error detection and correction

As mentioned in Data Sources, a kinship instance may conflict with other kinship instances, such as *father*(A, C) ^ *father*(B, C), or conflict with persons' gender, such as *father*(D, E) ^ *female*(D). Without data correction, family analysis will be an impossible task due to tangled relations. We detected 3,989 conflicting instances by checking rules and then corrected them.

3.3.1 Kinship instances conflicting with persons' information

Kinship instances conflicting with persons' information have many types and different processing modes. Table 6 shows three types of conflicts and errors.

Gender conflict is an ordinary conflict type. Some gender conflicts come from incorrect labeling of the persons' gender which can be corrected automatically.

Table 2. Kinship between persons.

No.	Person1_ID	Person1_name	Kinship word	Person2_ID	Person2_name
1	13060	李世民(Li Shimin)	父(father)	13059	李淵(Li Yuan)
2	140340	李貞(Li Zhen)	祖父(grandfather)	13059	李淵(Li Yuan)
3	3767	蘇軾(Su Shi)	族兄(clan sibling)	27127	釋惟簡(Shi Weijian)

Table 3. Expressions of the above kinship instances.

No.	Expression
1	*father*(13059_李淵, 13060_李世民)
2	*grandfather*(13059_李淵, 140340_李貞)
3	*clan sibling*(27127_釋惟簡, 3767_蘇軾)

Table 4. The adoption of kinship data extracted from CBDB.

Data	No. of kinship word types	No. of kinship instances	No. of person types
Original data	409	484,416	245,371
Data adopted	168	462,147	238,830
Adoption ratio	41.08%	95.40%	97.33%

Table 5. The normalization of basic kinship relations.

Kinship	# of normalized instances	# of kinship word types	Examples of kinship words
father	76,597	52	長子(the first son), 獨女(the only daughter), 二女(the second daughter), 妾之子(concubine's son)
mother	18,839	49	母(mother), 獨女(the only daughter), 次子(the second son)
husband	26,393	10	丈夫(husband), 第一任妻(the first wife), 第二任丈夫(the second husband)

Table 6. Types of conflicts and errors.

Types	No. of instances	Examples (checking rules using predicate logic)
Gender	1,647	*mother*(A, B) ^ *male*(A); husband(A, B) ^ husband(B, A)
Surname	131	*father*(A, B) ^ (*surname*(A) != *surname*(B))
Year	243	*grandfather*(A, B) ^ (*year*(A)—*year*(B) > 100 ˇ *year*(B)—*year*(A) > 150)
Total	2,021	–

And some gender conflicts are totally wrong, which need manual correction.

Surname conflict is another ordinary conflict type, because the surnames of the sons and daughters always are the same as their father's surname in China. When the surnames have conflicts, the instances will be picked out automatically for manual check. For example, 26109_陳愷(Chen Kai) is the father of 385739_鄭次牧(Zheng Cimu) in CBDB. The surname of the father is 陳, while the surname of the son is 鄭. When there is no other evidence to prove the kinship relation, the surname conflicting instances are usually deleted.

The detection of year conflict is the third type. It is a common sense that two persons' year of death cannot be too far away when they have basic kinship relations. So, we set different year-of-death gap thresholds for different kinship relations to detect such year conflicts. For instance, 13501_張行成(Zhang Xingcheng) is the father of 31226_張洺客(Zhang Mingke) in CBDB, but the year of death of 13501_張行成 is 1162 AD, which of 31226_張洺客 is 676 AD. We delete this kind of instances that are obviously wrong.

3.3.2 Conflicting kinship instances

Instances conflicting with other instances often come from paternal relations such as father and grandfather. For example, a person cannot have two fathers, except one father is a stepfather. We sort out 1,968 conflicting instances, from which 987 wrong instances were deleted manually. For example, *father*(32434_顏惟真, 95025_顏真卿) ^ *father*(191913_顏惟貞, 95025_顏真卿) is in the original data. The pronunciation of the names of both fathers is the same, but with different IDs (32434 versus 191913). In order to judge who is the right father, we have to check the persons' other information in CBDB (shown in Table 7).

The most intuitive information is year of death. If these years of death differ significantly, it will be easy to find the correct instance, while in this example, these two persons have so similar year of death that we need to use other information. It can be seen that 32434_顏惟真 have more detailed kinship words with 95025_顏真卿 and more kinship instances with others. So we regard *father*(32434_顏惟真, 95025_顏真卿) as the correct instance, and links other kinship relations of 191913_顏惟貞 to 32434_顏惟真.

3.4 Kinship inference

In Section 3.2, we introduced basic kinship instances extracted directly from the CBDB, which can be used to reconstruct the family trees. But these instances are not enough, because there are many missing kinship relations and persons, which will split a family tree to several trees. Thus, we put forward a set of logic rules to infer basic kinship instances by using other seven relations: grandfather, grandmother, father-in-law, great-grandfather, great-grandmother, uncle, and sibling. The seven relations are expressed by over 100 kinship words in CBDB.

Table 8 gives four rules of how to use other relations to infer basic relations. If A is the grandfather of B, and X is the father of B, then A is the father of X.[2]

Table 7. Persons' information for error correction.

Person_ID	Person_name	Year of death	Kinship words with 95025_顏真卿	No. of instances
32434	顏惟真(Yan Weizhen)	712 AD	父(father), 七子(the seventh son)	22
191913	顏惟貞(Yan Weizhen)	716 AD	父(father), 子(son)	9

Table 8. Formula of kinship words to infer basic relations.

Kinship	Examples of kinship words	Inference by natural language	Rules	Basic relations inferred
Grandfather	祖父(grandfather)	father of father	*grandfather*(A, B) := *father*(A, X) ^ *father*(X, B) ^ *male*(A) ^ *male*(X)	Father
Father-in-law	女婿(son-in-law)	father of wife	*father-in-law*(A, B) := *father*(A, X) ^ *husband*(B, X) ^ *male*(A) ^ *male*(B) ^ *female*(X)	Father
Great-grandmother	曾祖母(great-grandmother)	wife of great-grandfather	*great-grandmother*(A, B) := *husband*(X, A) ^ *great-grandfather* (X, B) ^ *female*(A) ^ *male*(X)	Husband
Sibling	長兄(eldest brother)	children of the same father	*sibling*(A, B) := *father*(X, A) ^ *father* (X, B) ^ *male*(X)	Father

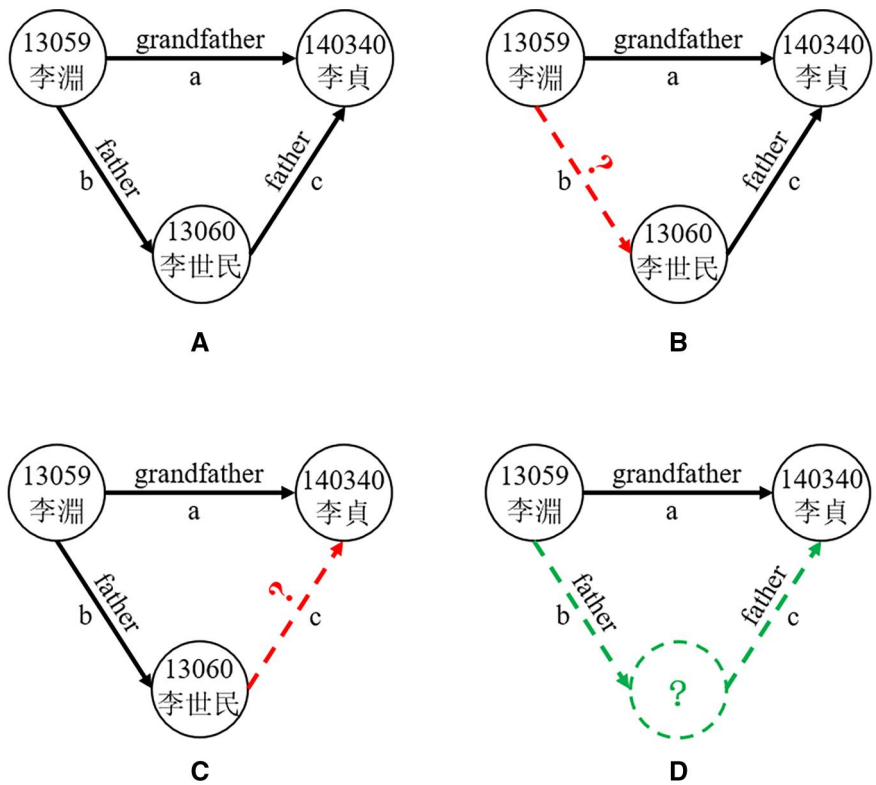


Figure 1. Using grandfather relation to infer father relation.

This method will calculate many missing basic kinship instances. Figure 1 shows an example of four typical situations when a grandfather kinship instance is decomposed into father kinship instances in CBDB. Case *A* is the situation that *grandfather*(13059_李淵, 140340_李貞), *father*(13059_李淵, 13060_李世民) and *father*(13060_李世民, 140340_李貞) are all true. But such cases with all the three relations are seldom. In most cases, there are only two of the three relations, which can be used for inferring the missing relations. Case *B*, *C*, and *D* are the other possible situations. Case *B* and *C* are both missing a certain relation marked with a dashed arrow in Fig. 1. Arrow *B-b* can be inferred as a new father relation instance, while *C-c* may be either a father relation or an uncle relation.

Table 9 gives the number of inferred new (missing) kinship instances by grandfather, great-grandfather, grandmother, great-grandmother, and father-in-law. The first two kinships have a large amount and a very high ratio to infer missing basic kinships.

3.5 Kinship inference with missing persons

When inferring the kinship relations, we found that it is necessary to add some imaginary persons (hereinafter called missing persons) to enrich the family networks. In another word, by the set of kinship logic rules, many missing persons will be inferred. And these missing persons can avoid the split of a family tree. For example, in Fig. 1D, we inferred nothing. But if we add a missing person instead of the question mark (shown in Fig. 2E), the family tree will be more complete. For the missing persons can be automatically inferred from other kinship relations, we have to assign a unique ID to each missing person, and unifying the missing persons which are likely to be the same person. For example, sibling kinship instances are used to minimize the number of missing persons (see Fig. 2F). With the condition *grandfather*(13059_李淵, 140340_李貞) ^ *grandfather* (13059_李淵, 142261_李麗質) ^ *sibling*(140340_李貞, 142261_李麗質), adding a new connective node for the missing person will make this family more natural and complete for further family analysis.

In this way, we applied the method on three relations: grandfather, great-grandfather, and ancestor. As a result, 5,805 missing persons and 10,337 missing father relations are inferred (see Table 10).

Finally, as shown in Table 11, we got 178,390 basic kinship instances by normalization and inference from the 484,066 kinship instances described by more than 400 kinship words in CBDB. Actually, only 462,147 kinship instances are used, because more than 10,000 instances are vague or ambiguous. The kinship relations are clean and clear after normalization. The basic kinship instances support the family tree/network reconstruction and family analysis.

4. Family network analysis

Family tree is a typical structure for family analysis. The tree will be a patrilineal tree when father–

Table 9. The inferred new instances without missing persons.

Kinship	No. of original instances	Inferred kinship	No. of inferred new instances	Inferred ratio (%)
Grandfather	23,236	Father	15,853	68.23
Grandmother	437	Husband	183	41.88
Great-grandfather	18,162	Father	15,218	83.79
Great-grandfather	18,162	Grandfather	96	0.53
Great-grandmother	145	Husband	58	40.00
Father-in-law	7,998	Father	453	5.67

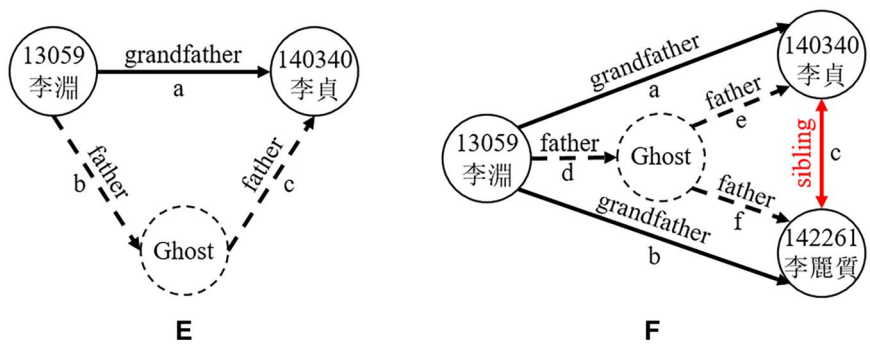


Figure 2. Using grandfather relations to infer father instances with connective nodes.

Table 10. The number of ghost types and inferred kinship instances.

Kinship used	Inferred kinship	No. of inferred ghost types	No. of inferred kinship instances
Grandfather	Father	3,401	6,872
Great-grandfather	Father	1,540	2,347
Ancestor	Father	864	1,118
Total	–	5,805	10,337

Table 11. The basic relations instances after normalization and inference.

Kinship	No. of instances
Father	118,019
Mother	19,006
Husband	41,365
Total	178,390

descendant is the dominate relation. And it will be a matrilineal tree, when mother–descendant is the dominate relation. Thus, when father–descendant, mother–descendant, and husband–wife relations are considered at the same time, the tree structure will not be sufficient for the representation of the family, but the graph structure will. Thus, the family network, made up of vertices (person nodes) and arcs (kinship relations), entails concepts about parentage and marriage (Zanette, 2019).

4.1 Basic statistics

Using the basic kinship relations after normalization and inference, we are able to reconstruct the patrilineal tree, matrilineal tree, and the family network. In this article, we focus on the patrilineal tree, which is the Chinese traditional family structure. Then we can calculate how large could a family be, how many families are recorded in CBDB, how many generations could a family last, and how many intermarriages across families, which were very hard to know before.

Patrilineal tree is a traditional family tree, which can be generated by the relation *father*. Table 12 is an example (△ means male and ○ means female). The *father* kinship relations are on the left, the family tree is illustrated on the right.

By root first traversing algorithm, 29,423 family trees are generated (see Table 13). One-person family means a family with only one person (the number of members does not include affinity, wife mainly). There are 25,762 families when excluding 3,661 one-person families.

The maximum depth reaches fifty generations in a family, and the largest family has 2,112 members. The top five families having the largest number of members are listed in Table 14. Each family is led by its genealogical root person's ID and name.

Table 12. From basic kinship relations to a family tree.

Kinship data	Family tree
	△ ID1678崔憲(Cui Xian)
father(ID1678, ID119432)	|—○ ID119432崔氏(蘇易簡妻) (Cui, wife of Su Yijian)
father(ID1678, ID3883)	|—△ ID3883崔遵度(Cui Zundu)
father(ID3883, ID19603)	|——△ ID19603崔求(Cui Qiu)
father(ID19603, ID3881)	|———△ ID3881崔著(Cui Zhu)
father(ID1678, ID3884)	|—△ ID3884崔遵用 (Cui Zunyong)

Table 13. Basic statistics on family data.

No. of families (excluding one-person families)	No. of families (including one-person families)	Maximum of pedigree depth	Maximum of members
25,762	29,423	50	2,112

Table 14. The top five largest families.

Genealogical root_ID	Genealogical root_name	No. of members
22466	李暠(Li Gao)	2,112
20261	趙敬(Zhao Jing)	2,082
194410	李叡(Li Rui)	1,262
31855	裴遵(Pei Zun)	623
182753	鄭胤伯(Zheng Yinbo)	595

Table 15. The top five families with the largest pedigree depth.

Genealogical root_ID	Genealogical root_name	Pedigree depth
32410	顏回(Yan Hui)	50
20529	陸烈(Lu Lie)	44
3220	賈誼(Jia Yi)	43
437643	蔣樞(Jiang Shu)	40
18740	薛廣德(Zheng Yinbo)	38

These families in Table 14 are all established families in the Tang and Song Dynasties, and the first two are the royal families of the Tang Dynasty and Song Dynasty.

Similarly, the families with the largest pedigree depth are famous families in pre-modern China. The families in Table 15 differ from those in Table 14. The families with the largest generations are usually famous not for power but for knowledge, like literature and Confucianism.

4.2 Enrich families with inference

In this section, we want to argue that inferring missing persons not only increases the instances of persons and

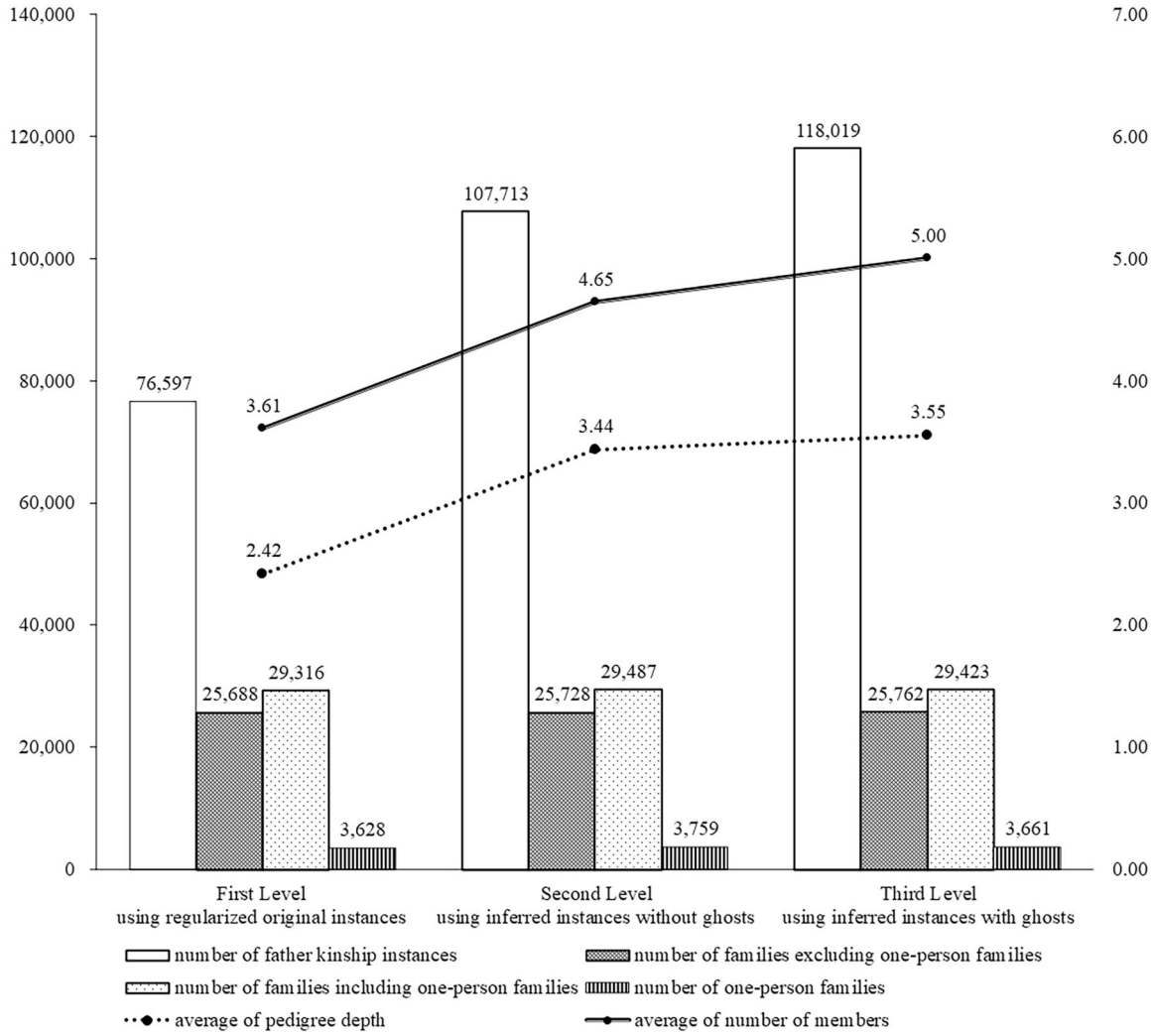


Figure 3. The effect of inference.

kinship relations, but also enriches the families, making the families larger, growing from 3.61 persons per family to 5 persons per family, while the depth grows from 2.42 to 3.55 generations per family.

As shown in Fig. 3, our kinship data can be divided into three levels: normalized original instances, inferred instances without missing persons, and inferred instances with missing persons. The three levels are called as the first level, the second level, and the third level. From the first level to the third level, we inferred more and more family data. And the total number of families is reducing because several families would be merged into one with the missing persons and missing kinship relations.

It can be seen that the number of one-person families decreases with the inference with missing persons. And the increasing trend the average number of members and pedigree depth in a family proves that the inference of kinship data is beneficial to enriching family data.

Figures 4 and 5 show the distribution of pedigree depth and number of members in a family. We count these skewness and kurtosis to describe them. In terms of skewness, the two distributions are both right-skewed distributions and long-tailed distributions, that is, the data are centrally distributed at smaller pedigree depth and small number of members. In terms of kurtosis, the sharpness of the peaks of the two distributions is much higher than that of the normal distribution. Since these are data extracted from pre-modern biographies, which tend to record the most important people of the time, a family here can be understood as a thriving family. According to this logic and the peak of the distribution in Fig. 4, a family generally continued to prosper for four generations in pre-

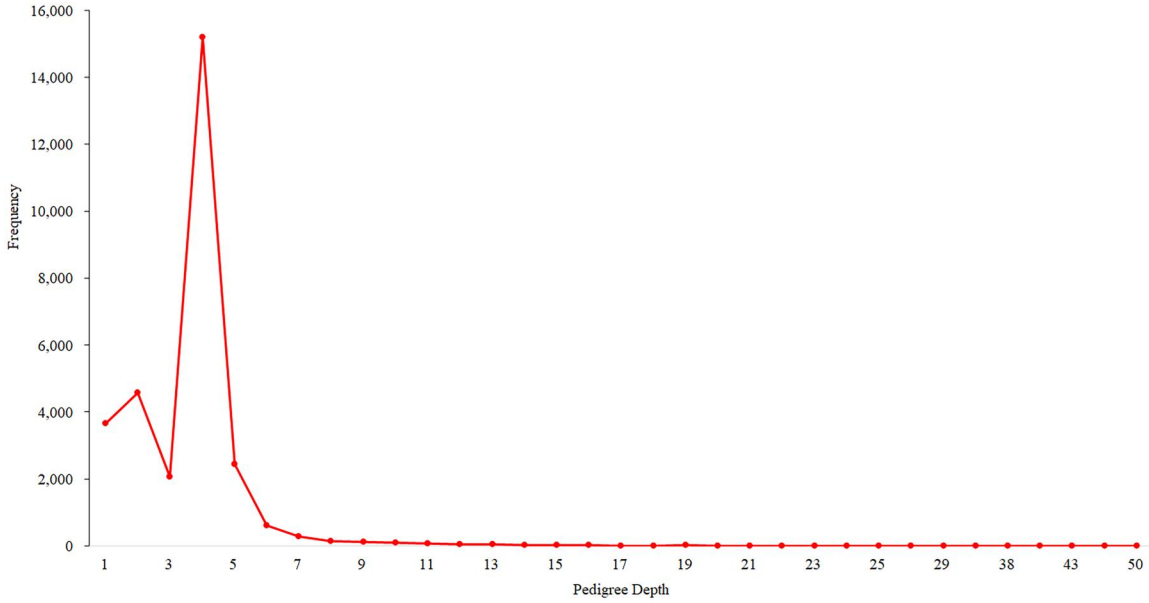


Figure 4. The distribution of pedigree depth.

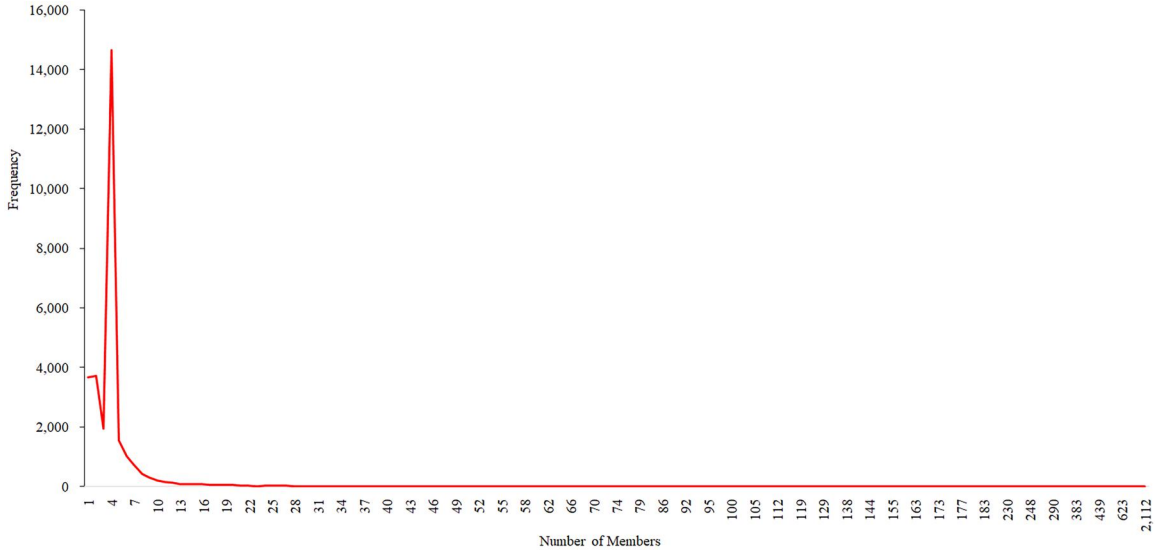


Figure 5. The distribution of number of members.

modern China. It is impossible to draw this conclusion directly from the original kinship data.

4.3 Women's remarriage statistics by using husband instances

Women's remarriage is quite rare in pre-modern China, because most wives have to be a widow even his husband has passed away. We analyze women's remarriages by husband kinship instances, using the condition that if one woman has two or more husbands then she has several marriages. It can be found that there were more remarriages in the Tang Dynasty, but less after the Song Dynasty, as shown in Fig. 6. Years 800–880 AD is the peak, and after the Tang Dynasty, the number of women remarriages dropped sharply because of "Neo Confucianism of Song and Ming Dynasties," which emerged as the dominant ideology during the Song and Ming Dynasties, encouraged women to observe widowhood (Bol 2008).

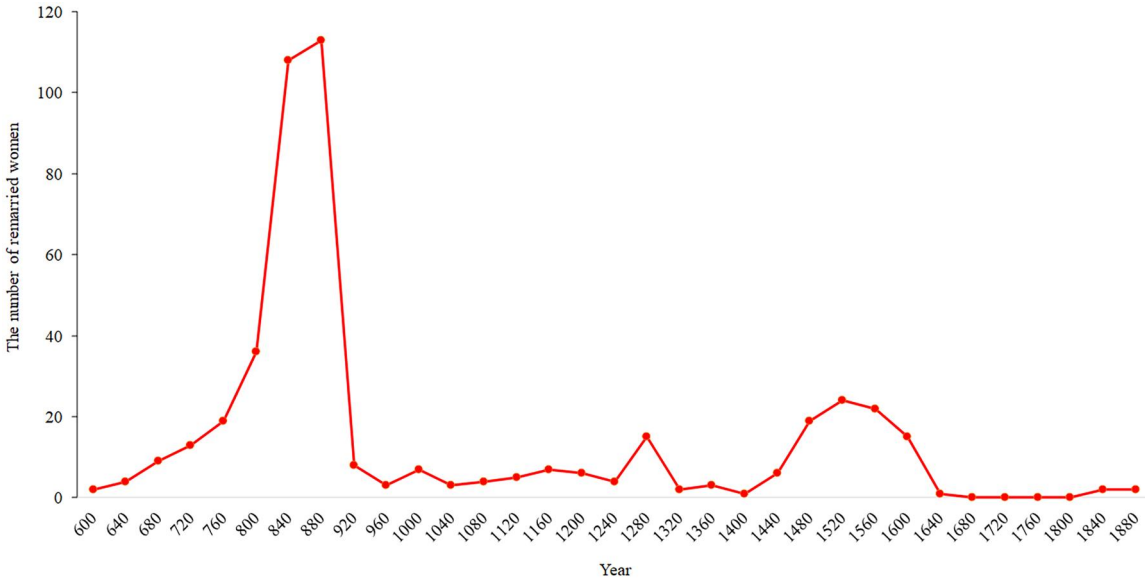


Figure 6. The number of women's remarriage.

Table 16. The top five families with the highest frequency of intermarriage.

Genealogical origin_ID	Genealogical origin_name	Intermarriage frequency
20261	趙敬(Zhao Jing)	351
22466	李暠(Li Gao)	230
44035	耶律雅里(涅里、泥禮)(Yelv Yali)	72
31521	盧植(Lu Zhi)	54
157668	盧玄(Lu Xuan)	36

4.4 Marriages across families

With the family data, it is possible to calculate the intermarriages across families, which is a rather hard and important work in traditional researches. The husband–wife relation across families could be a good way to measure the relations between large and famous families. From the marriage data collected in CBDB (from 7th century to 19th century), we find that there are up to eight intermarriages between two families. One is royal family of the Tang Dynasty called 陇西李氏(Li Family in Longxi) and the other is 太原郭氏(Guo Family in Taiyuan), whose genealogical origin is 19236_郭進(Guo Jin). Here we show the top five families with the highest frequency of intermarriage in Table 16. The first three families are all royal families. And the last two are both established families, they can be named 范阳卢氏 (Lu Family in Fanyang).

Figure 7 is the visualization of intermarriages of the family of 31521_盧植. The dotted line represents less than three intermarriages between two families while the solid line represents more than three intermarriages. We can infer from the figure that, the family of 31521_盧植 have many marriage connections with many other families, but most important families are 鄭稚, 李寶 and 鄭簡.

There is no node of the royal families in the Tang Dynasty in Fig. 7, which reflects the contradiction between the royal family and the aristocratic families, which is also the contradiction between the new imperial examination system and the old ninth-class official system in the Tang Dynasty.

The results also show that in the recorded data, the number of intermarriages in paternal families is 0 while within matrilineal families is 318, which is relatively common. It is not hard to explain why this happens. The pre-modern Chinese system of extramarital marriage was based on the patriarchal system. It is conventional that "the same race does not marry," or more strictly, "the same surname does not marry." Lévi-Strauss (1969) also pointed out that marriage exogamy has theoretical reasons, including inbreeding avoidance, incest taboos, and the formation of

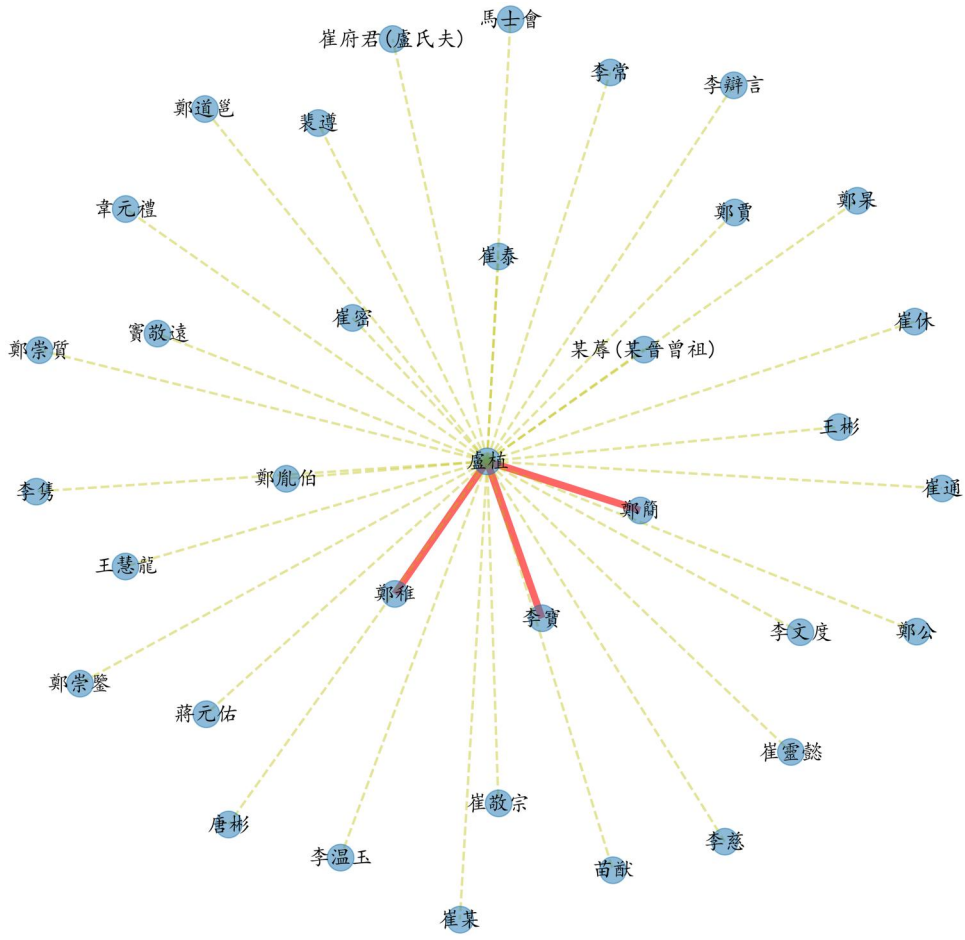


Figure 7. Intermarriage of the family of 31521_盧植.

marriage alliances. Our application of kinship verifies the historical research from the perspective of family data analysis.

5. Conclusions

As natural languages have various and complex kinship words to name different types of kinships, it is difficult to use them to build family networks. So, we normalize the kinship relations by three basic relations: father, mother, and husband, and conduct family network analysis. In this article, the kinship data of CBDB is used to generate the family network in ancient China. We make kinship data available by normalizing kinships into three basic kinships. We infer missing kinship instances and missing persons, and we drop or correct 3,989 conflicting instances to ensure that kinship data can generate correct family data. On the basis of kinship data with 178,390 basic kinship instances, we construct family networks by more complete family trees, which help us analyze the size of families and intermarriages across families. The total number of family trees is 29,423, the maximum depth of a family could reach fifty generations, and the largest family has 2,112 members. Statistics show that between 800 and 880 AD, the number of women's remarriages was the largest. We find 41,314 intermarriages between families and there are up to 8 intermarriages between two families. Without normalization, it is difficult to construct relatively complete family trees, while it is much less possible to quantitatively analyze families and intermarriages.

In the future, with new data from CBDB, we will conduct a deeper analysis on the family networks. Second, matrilineal trees will be analyzed and compared to patrilineal trees and other biographical databases. Third, applying the proposed normalization and reference techniques to other kinship databases. Of

course, the kinship logic rules have to be adapted to the languages' kinship lexicon. In the case of modern Chinese, most rules can be used directly, with the transform from traditional Chinese to simplified Chinese characters. And some rules need to be rewritten or revised.

Finally, we hope our normalization method can be a benefit to the research on family data collection and family network analysis.

Acknowledgements

This research was supported by National Language Commission Project (YB145-41) and National Social Science Foundation of China major project (21&ZD331, 22&ZD262). We are grateful to the reviewers for comments which helped us to improve the article. Special thanks go to Hongsu Wang and Edith Enright for their advices and data management. Thank Ruoting Wang for her design of the figures.

Author contributions

Bin Li (Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Supervision, Validation, Writing—original draft, Writing—review & editing), Yiguo Yuan (Data curation, Formal analysis, Investigation, Resources, Software, Validation, Visualization, Writing—original draft), Xuehui Lu (Data curation, Formal analysis, Resources, Software, Validation, Visualization), and Peter K. Bol (Conceptualization, Investigation, Resources, Supervision, Writing—review & editing)

Notes

1. Harvard University, Academia Sinica, and Peking University, *China Biographical Database* (April 24, 2018) https://projects.iq.harvard.edu/cbdb.
2. But if A is the grandfather of B, and A is the father of X, then X is not surely the father of B, may be B's uncle.

References

Bol, P. K. (2008) *Neo-Confucianism in History*. Harvard East Asian Monographs, number 307. Cambridge, MA: Harvard University Asia Center.

Collins, D., Jordan, C., and Coleman, H. (2012) *Brooks/Cole Empowerment Series: An Introduction to Family Social Work*. Boston: Cengage Learning.

Cruz, C., Labonne, J., and Querubin, P. (2017) 'Politician Family Networks and Electoral Outcomes: Evidence from the Philippines', *American Economic Review*, **107**: 3006–37.

De Nooy, W., Mrvar, A., and Batagelj, V. (2018) *Exploratory Social Network Analysis with Pajek: Revised and Expanded Edition for Updated Software* (Vol. 46). Cambridge: Cambridge University Press.

Kaplanis, J. *et al.* (2018) 'Quantitative Analysis of Population-Scale Family Trees with Millions of Relatives', *Science*, **360**: 171–5.

Kerber, R. A. *et al.* (2001) 'Familial Excess Longevity in Utah Genealogies', *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, **56**: B130–B139.

Lévi-Strauss, C. (1969) *The Elementary Structures of Kinship* (No. 340). Boston, MA: Beacon Press.

Liao, T. F. (2001) 'Were Past Chinese Families Complex? Household Structures During the Tang Dynasty, 618–907 AD', *Continuity and Change*, **16**: 331–55.

Liu C., and Wang H. (2017) 'Matrix and Graph Operations for Relationship Inference: An Illustration with the Kinship Inference in the China Biographical Database', *Proceedings of the 2017 Annual Meeting of the Japanese Association for Digital Humanities (JADH 2017)*, pp. 94–96. Kyoto, Japan.

Martin, D. B. (1996) 'The Construction of the Ancient Family: Methodological Considerations', *The Journal of Roman Studies*, **86**: 40–60.

Pérez García, M. (2011) 'New Technologies Applied to Family History: A Particular Case of Southern Europe in the Eighteenth Century', *Journal of Family History*, **36**: 248–62.

Read, D. W. (2001) 'Formal Analysis of Kinship Terminologies and its Relationship to What Constitutes Kinship', *Anthropological Theory*, **1**: 239–67.

Ruggles, S. (1994) 'The Origins of African-American Family Structure', *American Sociological Review*, **59**, 136–51.

Saller, R. P., and Shaw, B. D. (1984) 'Tombstones and Roman Family Relations in the Principate: Civilians, Soldiers and Slaves', *The Journal of Roman Studies*, **74**: 124–56.

Shang, W., and Huang, W. (2018) 'Investigating the Relationships between Scholars and Politicians in Ancient China: Taking the Yuanyou Era as an Example', *Journal of the Japanese Association for Digital Humanities*, **3**: 33–48.

Wang, K. (2017) 'The Digitalization of Huizhou Genealogy—Focusing on the Information of Characters and Geography', *Library Tribune*, **37**: 10–17.

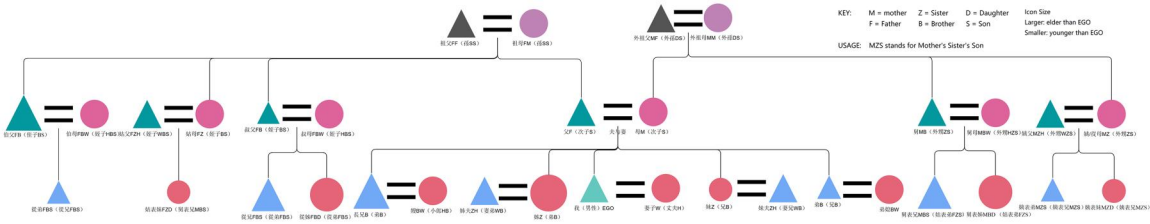
Warren, C. *et al.* (2016) 'Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks', *Digital Humanities Quarterly*, **10**(3): 1–16.

White, D. R., and Jorion, P. (1996) 'Kinship Networks and Discrete Structure Theory: Applications and Implications', *Social Networks*, **18**: 267–314.

Zanette, D. H. (2019) 'Statistical Properties of Model Kinship Networks', *Journal of Statistical Mechanics: Theory and Experiment*, **2019**: 094008.

Appendix 1

The kinship chart of modern Chinese.



Appendix 2

The kinship words for the three basic kinship relations.

Kinship words	English interpretation	Kinship words	English interpretation
父	Father	十子	The tenth son
母	Mother	十一子	The eleventh son
女兒	Daughter	十二子	The twelfth son
長女	The first daughter	十三子	The thirteenth son
二女	The second daughter	十四子	The fourteenth son
三女	The third daughter	十五子	The fifteenth son
四女	The fourth daughter	十六子	The sixteenth son
五女	The fifth daughter	十七子	The seventeenth son
六女	The sixth daughter	十八子	The eighteenth son
七女	The seventh daughter	十九子	The nineteenth son
八女	The eighth daughter	二十子	The twentieth son
九女	The ninth daughter	二十一子	The twenty-first son
十女	The tenth daughter	二十二子	The twenty-second son
十一女	The eleventh daughter	二十三子	The twenty-third son
十二女	The twelfth daughter	二十四子	The twenty-fourth son
十三女	The thirteenth daughter	二十五子	The twenty-fifth son
十四女	The fourteenth daughter	季子	The youngest son
獨女	The only daughter	私生子	The illegitimate son
妾之女	The concubine's daughter	私生子之本生父	The natural father of the illegitimate son
子	Son	私生子女之本生母	The natural mother of illegitimate children
獨子	The only son	妾之子	The concubine's son
唯一幸存的兒子	The survived son	庶子	The son of a concubine
長子; 第一子	The first son	丈夫	Husband
幸存的長子	The survived son	妻子	Wife
次子	The second son	第一任妻	The first wife
三子	The third son	第二任妻	The second wife
四子	The fourth son	第三任妻	The third wife
五子	The fifth son	第四任妻	The fourth wife
六子	The sixth son	第一任丈夫	The first husband
七子	The seventh son	第二任丈夫	The second husband
八子	The eighth son	第三任丈夫	The third husband
九子	The ninth son	妾	Concubine

Digital Scholarship in the Humanities, 2024, **39**, 215–227
https://doi.org/10.1093/llc/fqad108
Advance access publication 31 January 2024
Full Paper