The Journal of
HISTORICAL
NETWORK
RESEARCH

5 | 2021
special issue

Beyond *Guanxi*: Chinese Historical Networks

HENRIKE RUDOLPH | SONG CHEN

C²DH
LUXEMBOURG CENTRE FOR
CONTEMPORARY AND DIGITAL HISTORY

uni.lu
UNIVERSITY OF
LUXEMBOURG

DOAJ DIRECTORY OF OPEN ACCESS JOURNALS

Imprint

Université du Luxembourg 2021

Luxembourg Centre for Contemporary and Digital History (C²DH)
Université du Luxembourg
Belval Campus
Maison des Sciences Humaines
11, Porte des Sciences
L-4366 Esch-sur-Alzette

The publication of this special issue was in part supported by the Max Weber Foundation and the Fritz Thyssen Foundation.

ISSN 2535-8863

Contact
Principal Contact
JHNR-editors@historicalnetworkresearch.org
Support Contact
Dr. Marten Düring (Université du Luxembourg)
JHNR-support@historicalnetworkresearch.org

Typesetting
text plus form, Dresden, Germany
Cover image
Chinese star chart, British Library, Or.8210/S.3326 recto,
https://www.bl.uk/collection-items/chinese-star-chart
Copyediting
Andy Redwood, Barcelona, Spain

Published online at
https://doi.org/10.25517/jhnr.v5i1

The Journal of HISTORICAL NETWORK RESEARCH

MICHAEL FULLER/HONGSU WANG

Structuring, Recording, and Analyzing Historical Networks in the China Biographical Database

Journal of Historical Network Research 5 (2021) 248–270

Keywords China Biographical Database project, social network analysis, prosopography, relational database

Abstract The China Biographical Database (CBDB) is a relational database of over 470,000 individuals from pre-modern Chinese history. CBDB is distinctive as a prosopographical database in that it allows users to generate kinship and social networks for individuals – and groups of individuals – in the database. At the beginning of the project, to develop CBDB, we sought models among other digital prosopography projects but realized that, with CBDB's focus on analytic procedures and extracting data from the vast resources of the historical Chinese textual archive, we were developing a highly different model for digital prosopography. This paper presents an overview of the China Biographical Database, its capacities for exploring networks, and how we are extending those capacities through the ever-broadening extraction of data from the corpora of historical sources.

C²DH
LUXEMBOURG CENTRE FOR CONTEMPORARY AND DIGITAL HISTORY

UNIVERSITY OF LUXEMBOURG

I. Introduction*

Robert Hartwell's "Chinese Historical Studies" database system was the precursor to the current China Biographical Database (CBDB). Hartwell tracked both kinship and social relations among the Song dynasty officials he was studying, but he did not take the next step of exploring the networks defined by his data. When Michael Fuller rewrote Hartwell's database, he added recursive searches, starting with kinship data, to build social and kinship networks.

Once we realized the power of adding recursive searches to CBDB's structure, we understood the importance of collecting the data on kinship and social associations in carefully structured ways. Because the coders collecting the data from the texts wanted to stay as close as possible to the wording of the textual sources, their categories for coding associations expanded quickly, and we needed systematic ways to control the addition of new categories. Similarly, the variations in types of kinship relations that coders identified in the sources led to a large number of kinship categories, which complicated the analytic routines.

We solved the problem of the thicket of association codes through the aggregation of those codes through layers of higher-order classifications, so that scholars could decide how broadly or narrowly they wanted to construct the network of associations they sought to explore. We also added a halting-condition, restricting the node distance – the number of edges between the starting nodes and the final nodes – to be included in the network.

For kinship relations, we defined all relations in terms of generational (ancestor and descendent generations), lateral (sibling), and affinal metrics (in which a father's brother's wife shared the same metrics with a brother's wife's father). All combinations of kinship terms can be described within these metrics, and scholars can specify limits to their searches by setting upper bounds to these metrics.

Once scholars have results from the searches that build networks, they can export them to standard SNA software for further analysis. Because the CBDB is a relational database with significant additional information on the people in networks, scholars can look for additional patterns within their resultant networks. The first, immediate additional dimension is GIS data for the individuals, as the CBDB also outputs data on networks to standard GIS software.

---

* **Acknowledgements:** We would like to thank Chen Song for the thorough discussion and editorial assistance.
Corresponding author: Hongsu Wang, hongsuwang@fas.harvard.edu

Because the CBDB is a relational database, scholars can constrain the groups of individuals whose networks they wish to explore: doctors in the Song dynasty, then Ming, then Qing, or degree-holders from Sichuan over time, etc. With these variations, the CBDB has developed into a powerful tool for using flexible criteria to generate networks from data on over 470,000 individuals in pre-modern Chinese history and to then explore them through many dimensions.

1.1 In the Beginning

Robert M. Hartwell (1932–1996) explained the goals of his ambitious Chinese Historical Studies (CHS) database project in a 1991 essay "A Computer-Based Comprehensive Analysis of Medieval Chinese Social and Economic History:"

> Although research on Medieval China expanded at a rapid rate during the past two decades, these investigations suffer from the absence of a comprehensive understanding of the complex interrelationships among the social, economic, political and intellectual variables that, in the aggregate, constituted the structure or nature of Chinese Society. Such investigations as the many disparate and skillfully crafted monographs on separate localities, specific families or lineages, incumbents in one or another bureau of government, the functioning of one or another governmental institution, the growth of single industries and the nature of separate fiscal policies suggest many propositions about China from Han through early Ming. However, the collective information contained in these studies (Chinese, Japanese and Western) does not (and never will), provide the quantitative data necessary to test these propositions through investigating interregional and inter-institutional similarities and differences and changes in them over specified periods of time. The goal of the project described in this essay was to create the requisite database and use it to construct a framework for future research on Traditional China comparable to the ones available to students of European and American history.[1]

Upon Hartwell's death in 1996, he left the *Chinese Historical Studies* database to the Harvard-Yenching Institute, and the database – revised in its structure and moved to a new software platform – became the *China Biographical Database*. Peter K. Bol, with the crucial support of Chen Song and his successor project managers, organized a collaboration with an expanding board of partners in China, Taiwan, and Japan to develop the CBDB.

Along the way, we broadened CBDB's capacity to produce data on historical networks, a feature largely made possible by the accident of the language in which the database was written. Hartwell had created CHS in the dBase program-

1 Robert M. Hartwell, "A Computer-Based Comprehensive Analysis of Medieval Chinese Social and Economic History," in *Characters and Computers,* ed. Victor H. Mair and Yongquan Liu (Amsterdam: IOS Press, 1991), 89.

ming language, and we ported the database to MS Access, where we used Visual Basic to create the database's analytic functions. We realized that we could use Visual Basic to loop through the records for kinship and associations to build networks. Once we understood the power of linking networks with the host of other historical "entities" in the CBDB (e.g., entry into office, office holding, and place of origin), we focused a great deal of effort both on extending the capacity to generate and contextualize social networks, and on expanding the data needed for it.

2. The China Biographical Database as a Model of Relations within Pre-Modern Chinese Society

The design of the CBDB as a relational database derives from the ongoing effort, started by Hartwell, to model the core elements that shaped the lives of individuals in pre-modern Chinese society. Although the CBDB is based on the components of Hartwell's Chinese Historical Studies database, we expanded his initial design by applying an entity-relationship model to the historical sources. We identified nine basic components around which we built the CBDB. The current CBDB entity-relationship model thus includes these nine basic entities and the relations between them:

1) **People** (the central entity to which all other entities relate)
2) **Place** (not only people, but most other entities potentially have place attributes)
3) **Kinship Structure**
4) **Social Relations** in the society
5) The **Official Bureaucracy**
6) The **Modes of Entering** the official bureaucracy
7) The **Social Institutions** in which people participate (academies, temples, hospitals, etc.)
8) The **Fields of Social Distinction** through which individuals acquire status
9) **Texts** (texts often participate in the delineation of social networks, e.g., the exchange of writings and writings for specific occasions are expressions of social relationships)

The relationships between people and the other entities are one-to-many.

The CBDB database converts this model into digital form by transforming each of the nine entities into code tables and transforming the relationships between entities into data tables linking the entities.

The power of a relational database is that one can easily explore the interaction of different factors. Consider, for example, the question of whether the families of officials who entered office through the "Presented Scholar" (*jinshi* 進士) examination made local marriage alliances and, more specifically, whether the pat-

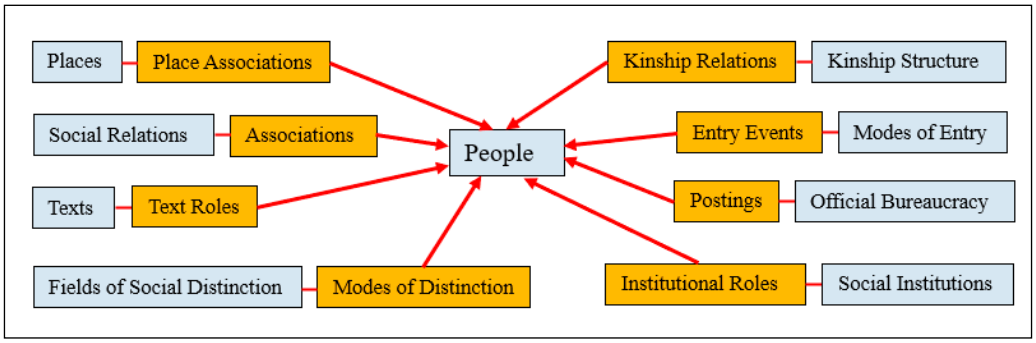

Fig. 1 The Basic Entities and Relationships in the CBDB.

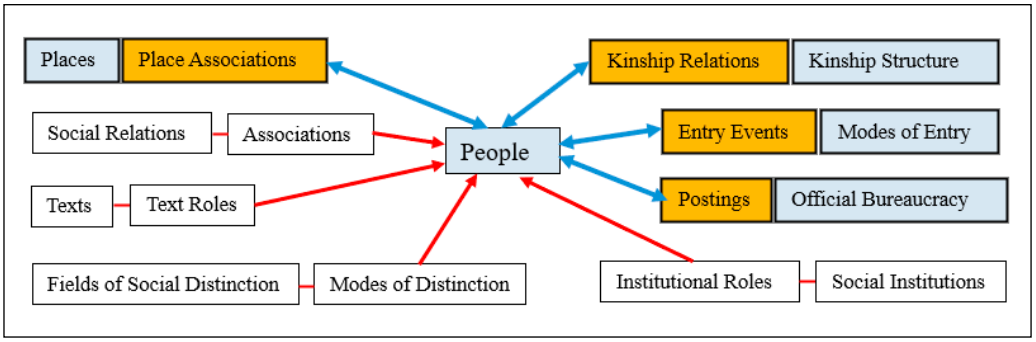

Fig. 2 The Interaction of Entities in Examinations, Office-holding and Social Networks.

tern differed from region to region and over time. One can trace the data required to answer the question by looking at the correlation of Place, Entry, Office, and Kinship.

Historical networks in pre-modern China interacted with – and were informed by – a host of structural factors that shaped social experience. One of the goals of the CBDB is to allow researchers to tease out these interactions.

3. Querying the China Biographical Database

The basic strength of the CBDB lies in its structure, as well as the ever-increasing data included in the database. Because the CBDB is complex, however, Michael Fuller has developed a series of forms to help users explore the contents of the database. Once users become familiar with the structure of the CBDB, they can then use the MS Access Query Builder to create customized SQL queries.

3.1 The Forms "Look at Entry" and "Look at Office Holding"

Two of the forms, "Looking at Entry" and "Looking at Office" report groups of individuals whose networks can then be explored. Qualifying for government service was one of the major life events for the sons of elite households in imperial China. From the Song dynasty (960–1279) onward, the most prestigious mode of qualifying was to pass the "Presented scholar" examination, which often took years or decades of preparation and many attempts. Passing the examination created networks with the cohort of successful examinees, with the officials overseeing the examination, and with the families of other examination graduates in one's native community. However, there were other routes to qualifying for government service, such as "yin privilege", which allowed officials to recommend relatives for admittance to the bureaucracy. Thus a key historical question is the differential patterns in marriage and social networks in which officials who followed different modes of entry participated. The forms "Look at Entry" and "Look at Office Holding" identify the men for these analyses of entry, office-holding, and participation in networks.

"Look at Entry" is perhaps the simplest CBDB form. One simply specifies the target mode of entry or a more general category, a range of years (for either the date of the entry-event or the index year of the person), and a location (either the index place for the person or the place of the entry-event).

Two of these data for individuals – index year and index place – merit additional comment. Because the CBDB is a historical database, it needs to be able to link the people in it to dates (years) that are as specific as possible. We have done this linking through an approach Hartwell started with CHS; since Hartwell was particularly interested in senior officials in the Song dynasty, he most wanted to know when someone turned 60. We adopted this practice and developed an algorithm to deal with the fact that we often do not know the year of birth or death. When possible, we calculate the index year based on a hierarchy of other information to estimate the index year. In our most recent release, however, we have replaced "age 60" with the birth year, either known or calculated. Because geographic information is also centrally important in Chinese prosopography, we treat "index place" in a similar manner to index year. We then use a descending hierarchy of place information to assign the index place.[2]

2 Since "place" is an administrative unit in the imperial order – usually a county or a prefecture – and because these designations change over time, the CBDB has developed two ways to handle both the hierarchy of administrative units and the changes across time. "Look at Entry" implements these methods. There are check-boxes on the form to allow the user to include all the subordinate units under the jurisdiction of a target unit (the counties in a selected prefecture, etc.) and to use the coordinates of the administrative seat of the target unit (and its subordinate units) to identify other units in the same location across the specified time period.

Once one has selected the mode of entry and specified the additional parameters, one runs the query. The form in Figure 3 shows the 1402 men who passed "regular examinations" between 1130 and 1160. The attribute data provided by this form, such as each man's place of origin and the exact year of his examination success, are potentially useful for constructing a two-mode network. Or one can examine the geographic distribution of these men on a map by exporting this result as a GIS file. To look at the other forms of networks in which these people participated, one can also save the IDs for reuse in other queries in the CBDB, such as queries on kinship and social networks, by simply clicking a button.

"Look at Office" is very similar to "Look at Entry." Since the official bureaucracy is complex, the CBDB allows one to specify any level within that bureaucracy and choose either all the offices attached to the level or select particular offices. One can filter by index year, dynasty, index place, and the location of the office.

The example in Figure 4 is a search for all the people who served as magistrates at the county level in the Southern Song dynasty (1127–1279). The attribute data reported here for the magistrates, like those provided in the query output in "Look at Entry," may be used to construct a two-mode network, but most of the interest will be in the other types of networks in which the office-holders participated. For instance, since the northern elite had fled to the south after the Jurchen conquest of the north that ended the Northern Song (960–1127), one question is

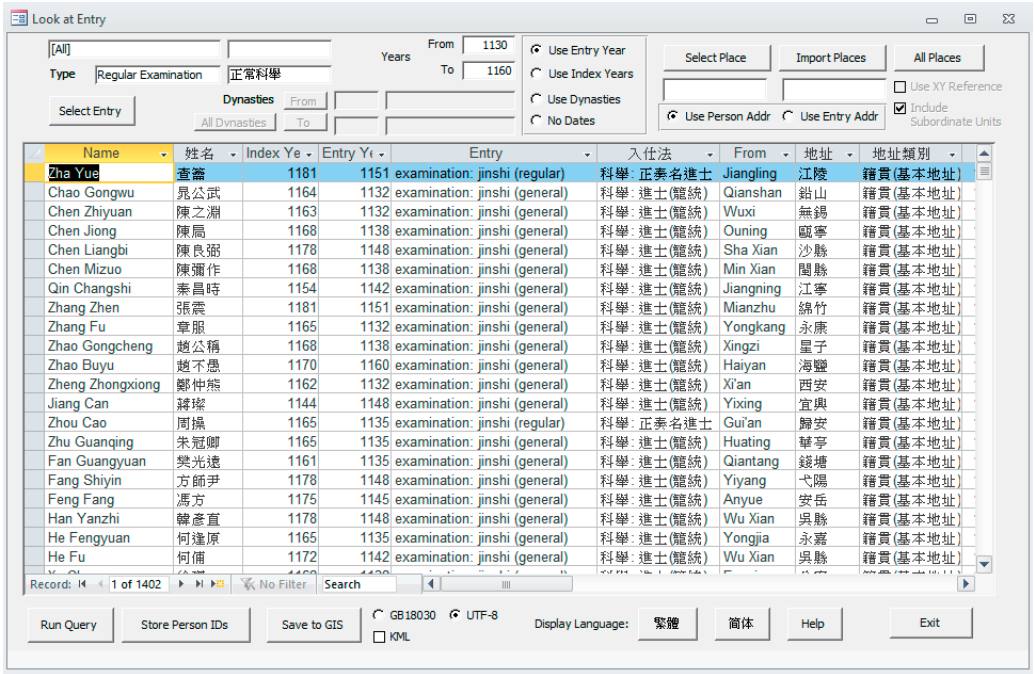

Fig. 3 Early Southern Song Entry by Examination in "Look at Entry."

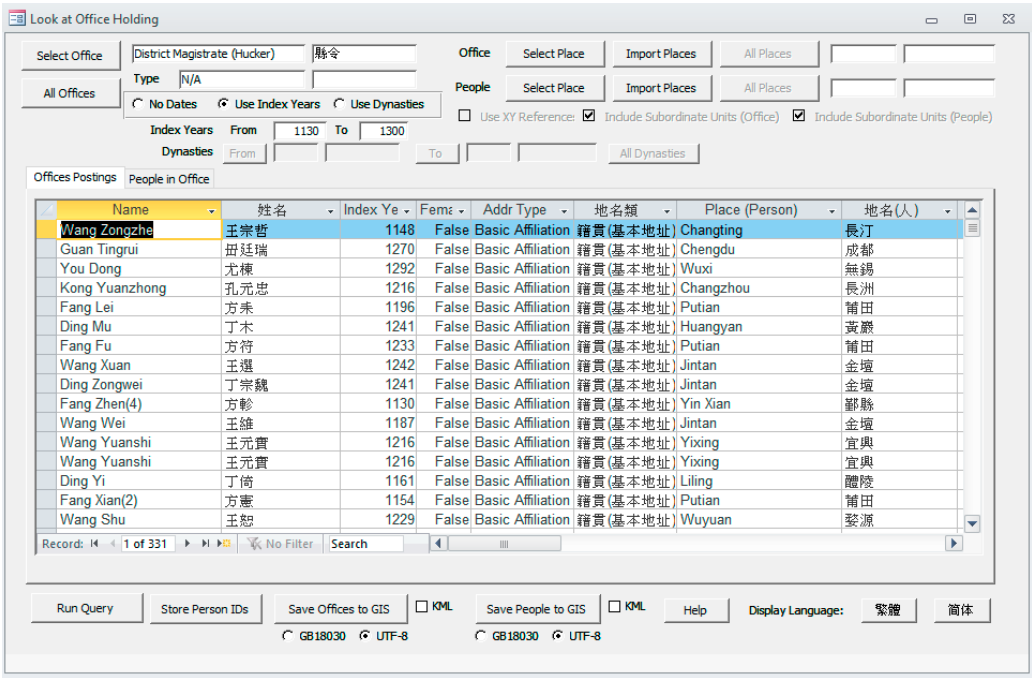

Fig. 4 Magistrates in the Southern Song Dynasty in "Look at Office Holding."

how the refugee elite reestablished itself in the south. Did the displaced northern magistrates in the early Southern Song, for example, establish kinship and social networks with the local elite families in their jurisdictions? "Look at Office" can provide a list of local officials in the late Northern Song which one can explore from the GIS perspective, and whose networks one can also explore through "Looking at Kinship" and "Looking at Social Networks."

3.2 The Form "Look at Associations"

As of May 2020, the CBDB has records for 149,611 instances of associations for the 472,090 individuals in the database. The CBDB's form for exploring the basic grouping of people through categories of associations is "Look at Associations."

As Figure 5 shows, the CBDB records 2,824 associations between people defined through some form of teacher-student relationship. From these data, one can ask whether there were, for example, regional networks of teacher-student relations or whether the networks were empire-wide. When do the networks start appearing in the texts, and do they change over time in size and geographic distribution? The CBDB form allows one to look at large categories of associations or more narrowly at subcategories or specific forms of associations. One can also constrain the search to specific places and time periods. The results can be saved as spreadsheets or as files for further SNA and GIS analysis. Equally significantly,

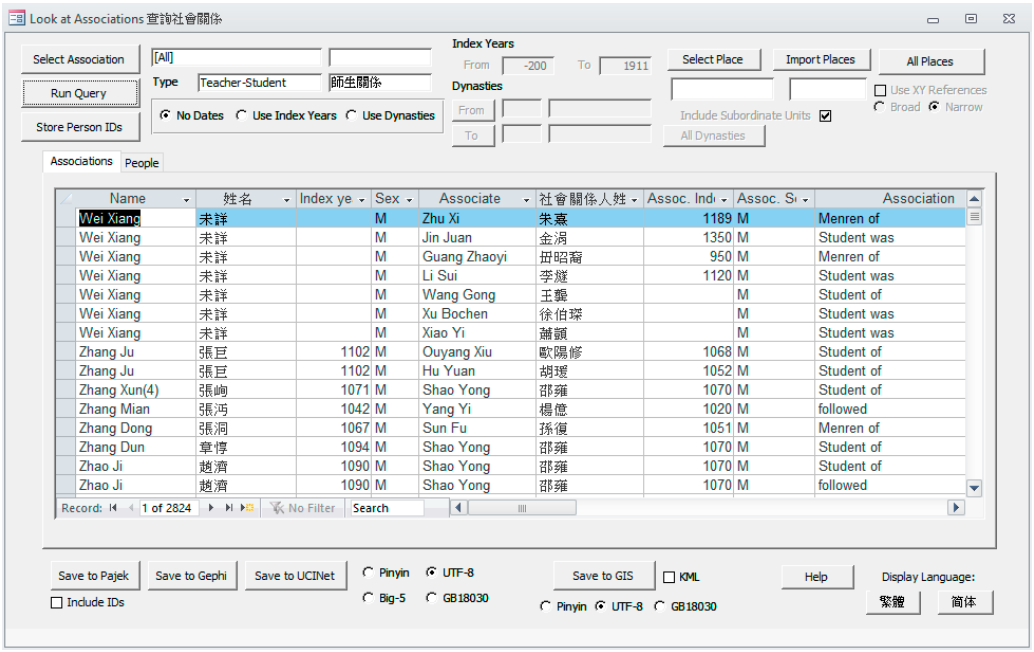

Fig. 5 Teacher Student Associations in Pre-Modern China ("Look at Associations").

one can save the IDs of the people participating in the associations for other forms of analysis. For example, what did it mean to be a student of someone and to be a member of a cohort of students? Did these students, teachers, and their families form other sorts of kinship and social alliances? Thus, for example, one can save the IDs of the people in teacher-student relations and explore them in the form "Looking at Kinship."

3.3 The Form "Looking at Kinship"

At present, the CBDB has data on 482,979 instances of kinship relations for the 472,090 people in the database. Given the interconnectedness of elites in pre-modern China, the challenge has been how to control the scope of the kinship networks that "Looking at Kinship" dynamically generates from the data.

Because of both the structure of kinship terms and the level of detail recorded for kinship relations, the CBDB has accumulated 479 codes for kinship. To control the size of the kinship networks defined through these many terms, in "Looking at Kinship" we adopted a simple four-value metric for inclusion in networks. While scholars who edit the results of searches for their own use can apply more specific criteria for kinship distance (for which our metric is not a historically meaningful substitute), the metric we apply simply tells the routine when to stop searching. The routine counts the accumulation of ancestor generations

(F [Father], M [Mother]), descendent generations (S [Son], D [Daughter]), sibling links (B [Brother], Z [Sister]), and affinal relations (H [Husband], W [Wife]) as it adds nodes to the network. The texts that preserve kinship relations for pre-modern Chinese individuals use many terms, some of which have no unambiguous translation into English. A *biaodi* 表弟, for example, can be either the younger son of one's father's or mother's sister or the younger son of one's mother's brother. Where possible, the CBDB uses the available data to distinguish between these three options. Still, it turns out that all three (FZS-, MZS-, and MBS-) share the same metric: ancestral generation = 1, descendant generation = 1, sibling = 1, marriage = 0. In constructing kinship networks, CBDB treats adopted siblings, bastard siblings, and half-siblings alike, even though it preserves the particularity of the edge relationship and therefore allows the user to winnow the data as necessary. Disambiguating kinship terms in natural language with a string of basic kinship symbols that can be concatenated in iterative searches lets us determine the four-value metric for kinship distance and produce kinship networks from the CBDB data.

Figure 6 shows the results for Huang Tingjian 黃庭堅 (1045–1105), a prominent Northern Song dynasty cultural figure, when the query metrics are set to 3 ancestral generations, 3 descendent generations, 1 sibling link, and 1 affinal link. If one looks at Huang's kinship data in the basic browser for individuals, the CBDB has only 32 records for instances of kinship for him, while "Looking at Kinship" identifies 295 that meet the defined parameters. Thus iteratively searching through the CBDB's kinship data allows users to discover kinship relations that would otherwise be difficult to discover. A user can quickly experiment with the search results by changing the parameters. We assume that it is better to get too large a network and cull, than to create a network that misses important connections.

The CBDB uses an automatic routine to simplify eight relations produced by the concatenation of edge relations in the construction of the network: BZ > Z, BB > B, ZZ > Z, ZB > B; SB > S, SZ > D, DB > S, DZ > D. That is, in the first iteration the CBDB identifies Li Bu 李布 as Huang Tingjian's maternal uncle (MB, i.e., mother's brother). A search of Li Bu's kinship data then reveals that Li Chang 李常 was Li Bu's older brother (B+) and therefore, with the concatenation, Huang's mother's brother's older brother (MBB+), but this relationship can be reduced to simply another of Huang's mother's brothers (MB). Because each of these simplifications decreases the sibling-link count by 1, it broadens who meets the sibling-link search metric, and the additional nodes will then generate additional nodes and edges to add to the network. Consider the impact of this simplification on the results for Huang Tingjian:

Lady Li (MBB+Z– > MZ), or Li Shi as she is called in Chinese, is Huang Tingjian's mother's sister, but the network arrived at Lady Li through Li Bu (MB) and Li Bu's older brother Li Chang (MBB+ > MB). The routine then added Hong Dan 洪亶 as Lady Li's husband (MBB+Z–H > MZH). We believe that further improving

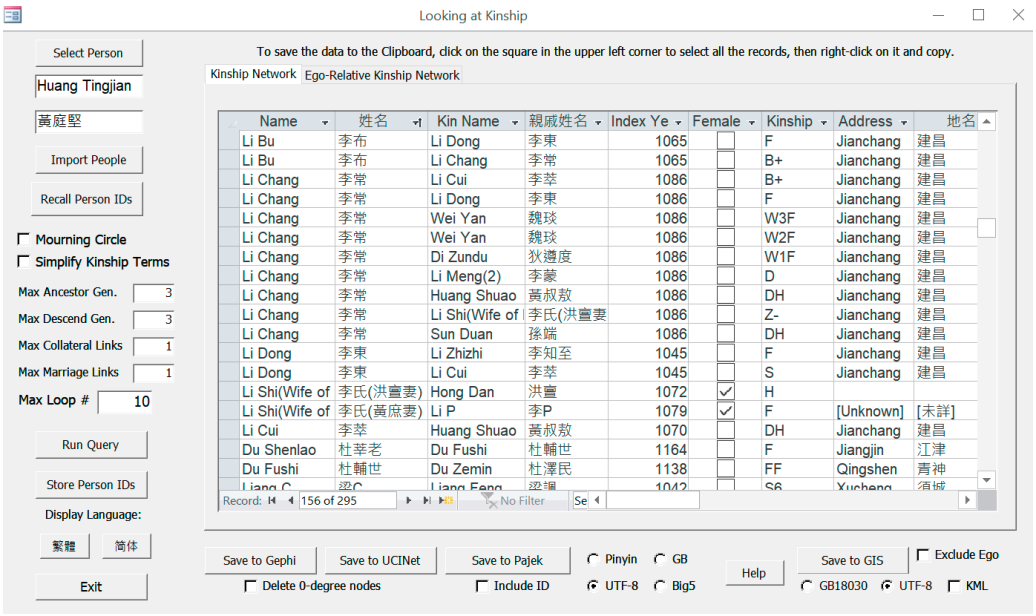

Fig. 6 Huang Tingjian's Kinship Network ("Looking at Kinship").

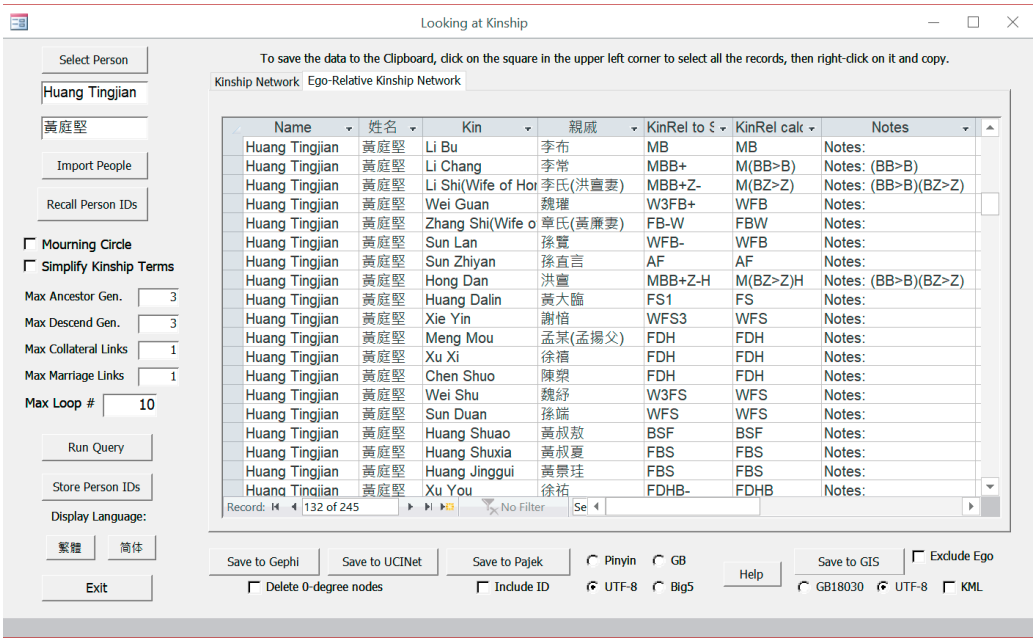

Fig. 7 Huang Tingjian's Ego-Relative Kinship Relations ("Looking at Kinship").

the CBDB's procedures for dynamically handling the simplification of kinship relations that extend beyond sibling relations is one place where future collaboration may prove very fruitful.

3.4 The Form "Looking at Social Networks"

"Looking at Social Networks" is the main form for exploring historical networks in the CBDB. As Figure 8 shows, in appearance, the form is far more complex than "Looking at Kinship" because it must account for place and time constraints in constructing networks. The form also allows users to delimit the categories of association around which the network is to be built, and also allows the user to intermix social and kinship networks. Yet the algorithm for finding a stopping point for the search is far simpler than for kinship: one merely specifies the maximum allowed node-distance from the initial target individual(s).

The search for networks produces three types of data: the individual edges, the individual nodes, shown in the tab "People in the Social Network", and a version of the relationship that merges parallel edges, shown in the tab "Aggregated Social Networks." The GIS export function saves the place information for the individuals in the network, while exporting to UCINet and Gephi uses the individual edges in "Social Network Relationship," and files for Pajek use the aggregated edges.

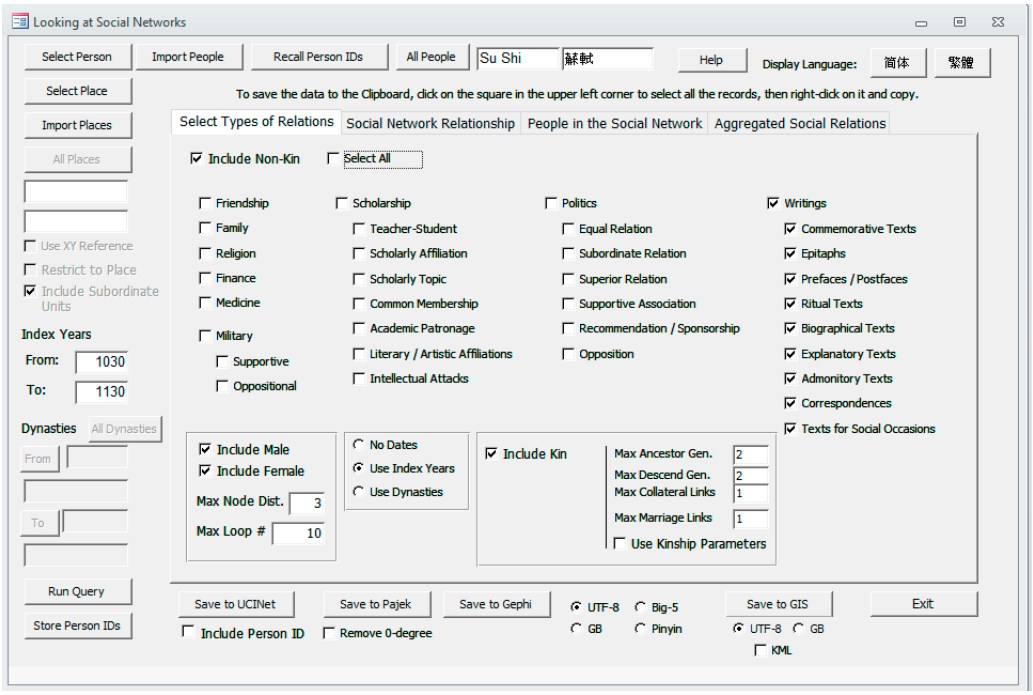

Fig. 8 "Looking at Social Networks," the Form for Querying Networks.

To make "Looking at Social Networks" easier and more convenient to use, we added additional features. As discussed above, while records for many individuals lack adequate data to determine an index year, we at least know the dynasty, as one can now filter by dynasty. Similarly, because administrative units change name and size in ways that most users lack the expertise to master, we have added the option of using the coordinates for a location (or group of locations) to identify all relevant place codes when searching over long time periods. When including kinship relations in the construction of networks, we optionally use the approach to kinship distance developed to deal with the complexity of kinship relations in "Looking at Kinship." And users can recall the IDs of people produced in other forms of searches (status, office-holding, mode of entry, place association, etc.) to explore the social networks among those people and, conversely, one can store the IDs of people participating in a network to make the list available for other queries.

3.5 The Forms "Look at Pair-Wise Associations" and "Look at Place"

The CBDB offers two additional forms to allow users to explore networks that are otherwise difficult to approach within a large dataset. The first is "Look at Pairwise Associations." This form allows one to enter two people (or a list of people) and see the network of associations linking them either directly or through intermediaries. The example in Figure 9 shows the intermediaries between the great Northern Song dynasty literary figure Su Shi and his contemporary, the Neo-Confucian philosopher Cheng Yi, with whom he had a contentious relationship. This search is restricted to older and younger contemporaries and shows 114 instances of relationships among 15 people. That is to say, a substantial community provided lines of communication between these two cultural rivals.

The form allows one to search for links mediated by both one and two persons. The form has wide application for exploring interactions between groups. For example, one can identify Southern Song dynasty Buddhist monks and the Neo-Confucian advocates of the period – who, in theory, were strongly opposed to Buddhism – and see the extent of their intersecting social networks.

The final form, "Look at Place," provides the user with a list of people who participated in any forms of connection to a specified place: people who served in office there, taught in or were students in an academy there, had family or associates from there, or passed an examination there. Looking at the results, the user may discover overlapping networks that brought people together in a place in unanticipated ways that would otherwise be difficult to sort out from the raw data.

Figure 10 provides the 3,353 records in the CBDB that connect people to Jinhua county in the late Southern Song dynasty. These records are raw data for further analysis, but they conveniently draw all the data together.

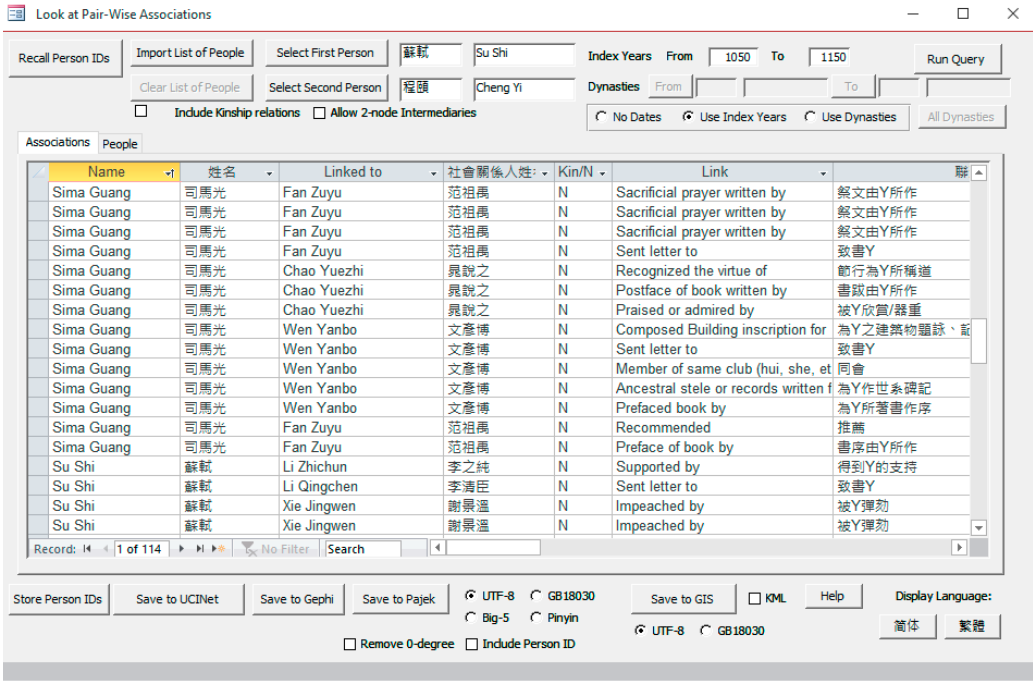

Fig. 9 "Look at Pair-Wise Associations."

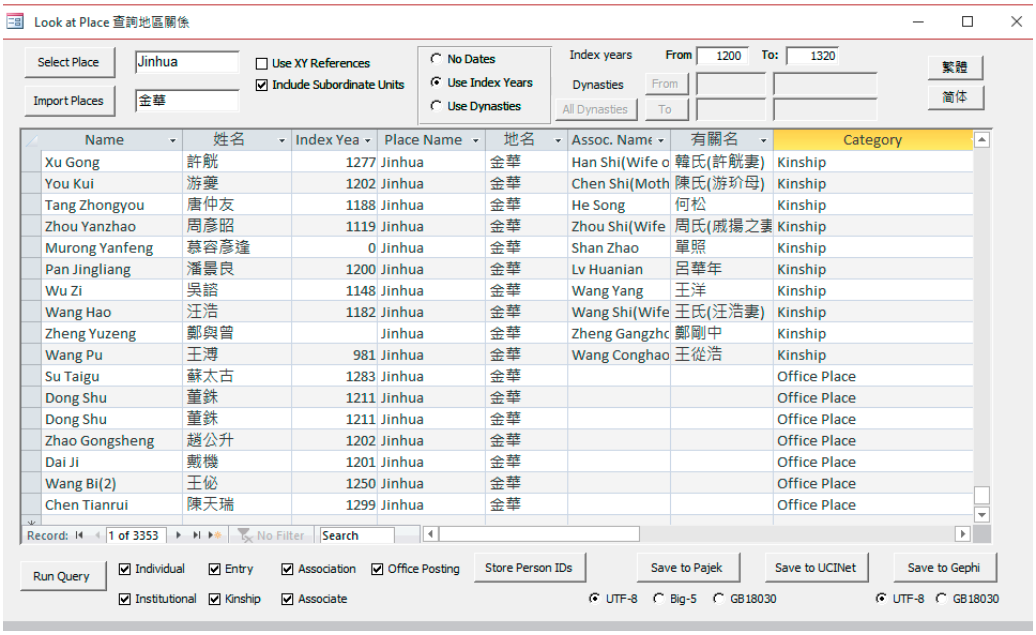

Fig. 10 People Associated with Jinhua County in the Southern Song Dynasty.

4. A Summary of Network Data in the China Biographical Database

4.1 Data on Non-Kin Social Associations

To support prosopographical research, the China Biographical Database had collected 482,979 instances of kinship relations and another 149,610 instances of non-kin social associations by May 2020. Most of our data concern the period between 600 and 1912 CE (Figure 11).

Reflecting the history of the CBDB, whose earliest developers, including Robert Hartwell, are predominantly scholars specializing in the Tang (618–907) and Song (960–1279) dynasties, until recently more than 75% of its data on social association pertained to this period. It is our plan, however, for the next few years to significantly expand its coverage by collecting large amounts of data on social associations between the fourteenth and nineteenth centuries.

To capture the nuances of different types of relationships found in the historical record, but also provide convenience to general users, the CBDB adopts a two-

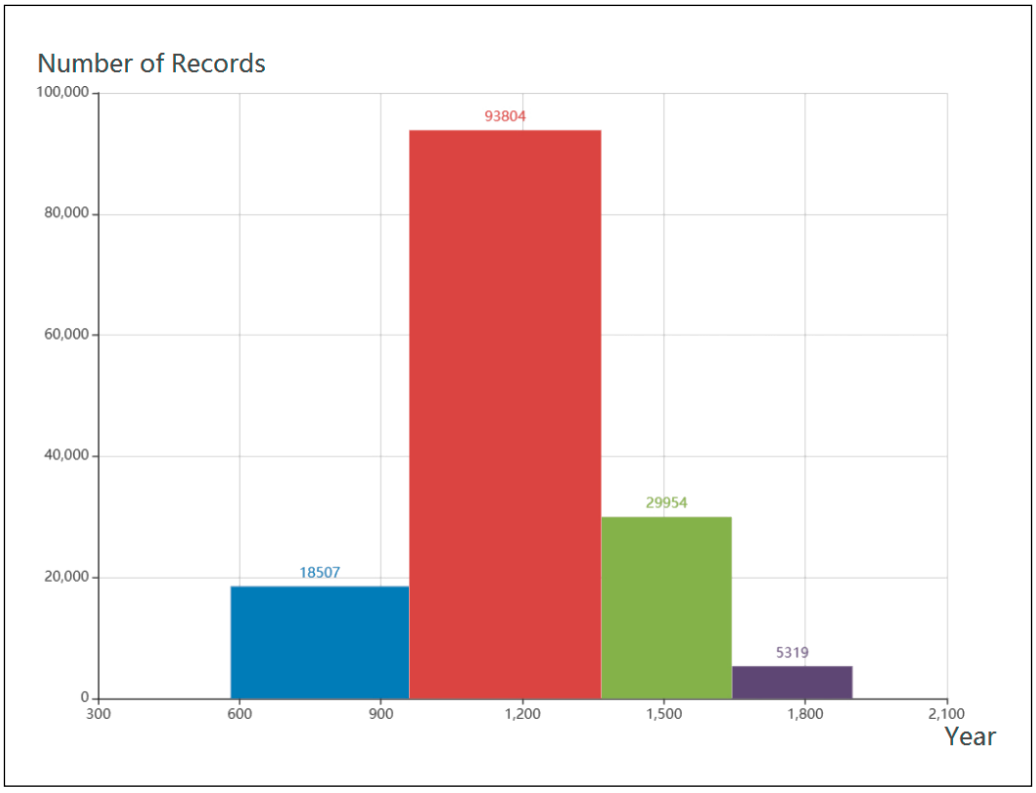

Fig. 11 The temporal distribution of CBDB data on social associations.

tier approach in recording social connections. First, we define 480 types of social association, every two of which form a pair, such as "epitaph written by" and "epitaph written for," "purged" and "purged by," etc. We then classify these different types of associations into ten broad categories, among which writings, politics, and scholarship have the most records (Figure 12). This is a result of the nature of the CBDB's data sources. In sharp contrast to Europe, where church records and bank records abound, the Chinese historical record before the twentieth century is rich in political, scholarly, and literary interactions. The CBDB systematically harvests data from biographical indexes, literary collections, and local gazetteers, which document these social interactions using formulaic expressions conducive to semi-automated data extraction.

A visualization of all social association data in the CBDB generates a massive network with 33,433 persons (nodes) and 149,610 connections (edges). Where multiple instances of associations are documented for a given pair of nodes, duplicate edges are counted and removed and the count is added to the remaining edge as its weight. This reduces the total number of edges in the network to 55,799. Analyzed as an undirected network, it has an average degree of 3.35, and a network diameter of 19. The average path length in the network is 5.509, suggesting that on average a node is 5 to 6 steps away from other nodes in the network.

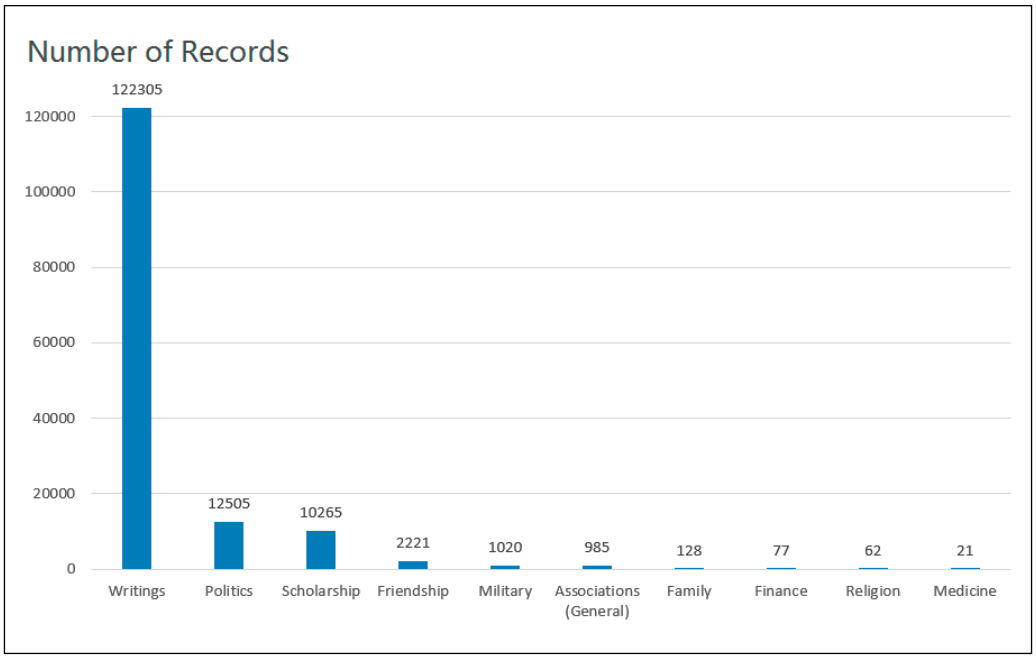

Fig. 12 The number of records for different social association categories in the CBDB.

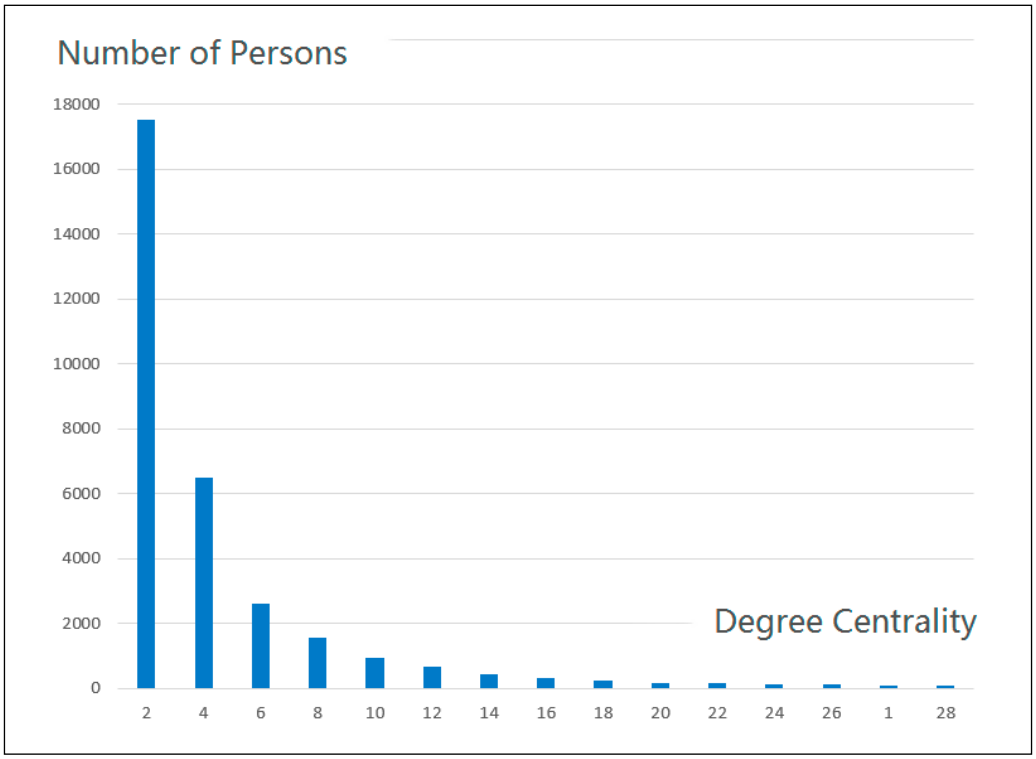

Fig. 13 The distribution of weighted degree centrality values in the CBDB social association network (number of persons ≥ 100).

The distribution of the weighted degree centrality of each node follows a power-law curve: 90.20% of persons have a degree centrality of only four or less, whereas a tiny percentage (3.69%) of persons have a degree centrality above 10 (Figure 13).

Likewise, the network's Gini coefficient (0.63), computed on the basis of different degree ranks, indicates a high degree of inequality in how social connections are distributed among the persons in the CBDB (Figure 14).

One may also examine the network's clustering tendency by comparing it to the Erdős-Rényi random graph, where n = 33,433 (nodes), m = 55,799 (edges),[3] there are on average 119 3-cliques but only $1.91*10^{-5}$ 4-cliques. In other words, in a random graph, the possibility of having a k-clique (i.e., a subgroup composed of *k* nodes and where each node in the subgroup has a tie to every other node) in the network where the k value is greater than or equal to 4 is very low. In contrast, the observed network in CBDB reports a total of 313 3-cliques, 154 4-cliques,

3 The value p(prob.) of Erdős-Rényi model for CBDB data is $p = m/\binom{n}{k} = 0.000267$.

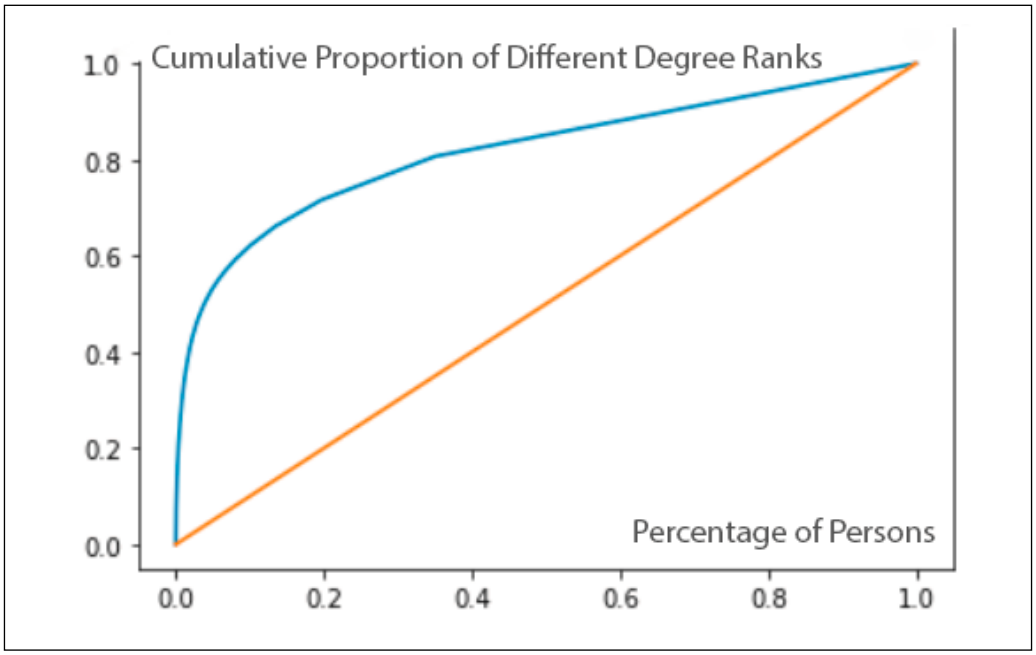

Fig. 14 The Lorenz curve of the degree centrality ranks for CBDB social association data.

54 5-cliques, and 27 6-cliques. This implies a strong tendency towards local clustering in the CBDB that deserves the attention of humanities scholars. It is also noteworthy that while the observed network in the CBDB includes 1099 components, 92% of the persons are members of the giant component.

4.2 Data on Kinship Ties

The CBDB has recorded 481,476 kinship ties that link together 244,658 persons. These ties are classified into 479 types, of which thirteen have more than 5,000 records: son (S), father (F), elder brother (B+), younger brother (B–), husband (H), grandfather (FF), wife (W), grandson (SS), great-grandfather (FFF), great-grandson (SSS), mother (M), wife's father (WF) and daughter's husband (DH) (Figure 15).

Analyzed as undirected and unweighted, this kinship network has an average degree of 1.97, an average path length of 24.235, and a diameter of 79.[4] This network

4 The CBDB records only kinship ties explicitly mentioned in our data sources. As illustrated by the example of Huang Tingjian in the preceding section, for example, one of a person's maternal uncles may be directly registered under that person, while his other

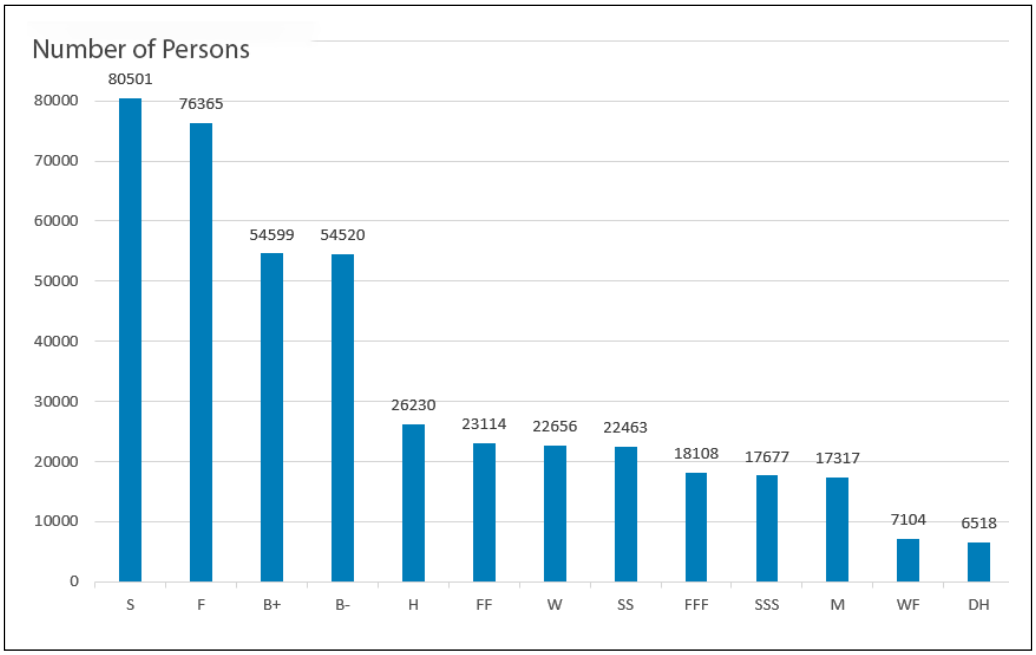

Fig. 15 The number of records for different types of kinship ties in CBDB (number of records > 5000).

is more fragmented than the non-kin social network. It contains 23,532 components, but only 51,738 persons (i.e., 21.15% of the total) are part of the giant component, and even the second largest component has only 357 nodes.

15,573 persons in the kinship network are also found in the non-kin social network. Therefore,

$$\frac{\text{Number of overlapping nodes between CBDB kinship and social networks}}{\text{Number of nodes in the CBDB social network}}$$

$$= \frac{15{,}573}{33{,}433} = 46.58\%$$

maternal uncles and aunts may not, despite their presumably equally close ties to that person. Therefore, some network metrics – such as network density, diameter, average degree, and average path length – are not accurate measures of the structural properties of a network constructed from unprocessed CBDB kinship data. It is more meaningful to interpret these metrics as indicating how kinship data are registered in the CBDB and its data sources.

On the other side,

$$\frac{\text{Number of overlapping nodes between CBDB kinship and social networks}}{\text{Number of nodes in the CBDB kinship network}}$$

$$= \frac{15{,}573}{244{,}658} = 6.37\%$$

These numbers suggest that a person's relatives are more likely to enter the historical record if that person's social connections are recorded in the historical sources. This implies that family members of a socially prominent individual tended to have a greater chance of being influential.

5. Extracting Data from Texts

5.1 Identifying Suitable Historical Sources

The CBDB prioritizes the collection of data from digitized texts that contain large amounts of reliable biographical information and present them in a clear and consistent format, which makes it possible for us to extract data from these texts in a systematic manner using computational methods. The systematic harvesting of data has not only increased the efficiency of expanding our data coverage, but it also ensures that research using CBDB data is statistically meaningful. For example, *Quan Song wen,* a collection of all preserved prose style literature from the Song dynatsy (960–1279), is an ideal source for data collection. Its table of contents, in essence a nearly exhaustive list of more than 178,000 unique pieces of writing from the three centuries, includes letters, funerary biographies, elegies, commemorative inscriptions for private and public constructions, etc. that signify different types of social relationships.

5.2 Acquiring and Preparing Texts for Data Mining

Whenever possible, the CBDB prefers to work with texts for which searchable digitized editions already exist, and that the copyright owner agrees to share these editions with us for data mining. In some cases, where working with non-digitized texts is inevitable, we have to assess whether it is more efficient to create a digital edition by ourselves or simply have someone type the data directly into the computer. We have developed an open-source online inputting system for data entry. But for some texts, we have editors first enter data into spreadsheets, then code and upload them to the database in batches. We find simple spreadsheets with a limited number of value fields most effective for this purpose.

When the need arises for us to create a digital edition by ourselves, we scan the documents and use Abbyy, an optical character recognition (OCR) package,

to transform them into searchable text files. The precision of OCR depends on a number of factors, such as the layout and quality of the scanned pages. We typically start off by checking the precision rate of a few sample pages and proceed only if it exceeds 95%.

5.3 Data Mining

Patterns of language use vary from text to text. To efficiently extract data, we have employed different data mining techniques for different texts. The most common technique relies on regular expressions. The basic idea here is string searching based on patterns of language use. Therefore, it works best only on texts that use a finite number of formulaic expressions, which has the benefit of not requiring any training data. Take teacher-disciple relations, for example. A common expression of this relationship in ancient Chinese texts describes the disciple as "following" (*cong* 從) the teacher in his "travels" (*you* 遊). One typically encounters the teacher's name sandwiched between these two Chinese characters meaning "follow" *(cong)* and "travel" *(you)* respectively, and therefore to find the teachers named in a person's biography, we can ask the computer to scan through the document for all text strings that have this pattern.

For texts where patterns of language use are less obvious and more variegated, regular expressions have to give way to more sophisticated methods. We have tried several methods of machine learning for different genres of texts, with varying degrees of success.

Random forest is a supervised learning algorithm that we have been experimenting with to extract social association data from the biographies. First, we collect from this corpus all the sentences that contain two person names and use a small number of these sentences as the training data. Next, we use a feature vector to represent each sentence in the training data based on the frequency of each word in the sentence. We then read each sentence in the training data and assign it to a specific type of social association that CBDB has coded. This allows us to gain knowledge of how the mathematical properties of each feature vector, which represents a sentence in the training data, correlate with the type of social association it expresses. In the final step, we apply this knowledge to those sentences that are not yet in the training set, trying to predict what type of social association they express based on their mathematical properties. Thus, we in effect resolve the problem of detecting and extracting network data in a corpus by transforming it into a problem of identifying sentences with name collocations and then classifying them into a finite group of social association types.

Unlike random forests that handle the extraction of named entities as a classification problem, BiLSTM-CRF approaches it as a sequential tagging problem. Compared to random forests, we find the combination of BERT and BiLSTM-

CRF particularly effective. BERT is an unsupervised language representation technique that transforms texts into vectors, but BERT surpasses all previous techniques for being deeply bidirectional and taking into account the context for each occurrence of a given word (here, a given Chinese character). After vectorizing our texts with BERT, we apply the state-of-art sequence tagging model, called BiLSTM-CRF (Bidirectional-Long Short Term Memory-Conditional Random Field), to predict whether the occurrence of a given character in our texts is part of a person's name, official title, place name, kinship term, or something else. This prediction is based on both the immediate context of each occurrence of the character (the Bidirectional LSTM model), but also on more general knowledge of how the character is used in the entire training set (the CRF model). This method is particularly useful for handling Chinese texts, where the names of entities (e.g., persons, offices, and places) have no obvious stylistic features, such as the use of uppercase letters, and are not separated from other words in the sentence by white spaces. This proved successful when we applied it to local gazetteers, from which we extracted biographical information, including the officials' degrees, careers, and kinship relations.

5.4 Data Standardization

Our data mining algorithms typically export the results in the format of a collection of tagged XML documents or a large spreadsheet that includes information on a person's relationship to many different entities (e.g, offices held, social associations, kinship ties). Since the CBDB stores each entity in a separate code table and the relationship between each entity and a person in a separate data table, we convert these tagged texts and large spreadsheets into a group of linked tables that fit the CBDB data model.

The principal challenge in this step is disambiguation. The CBDB uses a unique ID for each person, as it does for other entities, too. Yet a person may appear in historical sources under different names, and two persons may have also had exactly the same name. Before uploading newly harvested data to the CBDB, we disambiguate each occurrence of a person's name within the new dataset and against existing data in the CBDB. Although we occasionally seek advice from historians specializing in the relevant topic or period, the sheer size of our data requires that we automate this process as much as possible. To disambiguate person names, we draw upon our knowledge of the data source, attribute data to each person, and also his or her networks. We assume, for example, a name that repeatedly appears in different chapters of the same book most likely refers to the same person. We assume that persons who had the same name but hailed from different places, lived in different centuries, or passed the civil service examination in different years were merely homonymic by chance. These biographical details, however, are not always available. It is worth noting here that we have also found network data harvested by the CBDB useful for disambiguation purposes. In recent years, we have had success in developing disambiguation algorithms on the assumption

that two "persons" with the same name whose kinship and non-kin social networks overlapped significantly were most likely the same person.

6. Appendix: Publications Using CBDB Social Association Data

Bol, Peter. "Changing Literati Networks: Kinship and Collegiality, 1100–1400." *Journal of Historical Network Research* 5 (2021): 87–113.

Chen, Song 陳松. "Governing a Multicentered Empire: Prefects and Their Networks in the 1040s and 1210s." In *State Power in China, 900–325,* edited by Patricia Ebrey and Paul J. Smith, 101–152. Seattle: University of Washington Press, 2016.

De Weerdt, Hilde, Brent Ho, Allon Wagner, Qiao Jiyan, and Chu Mingkin. "Is There a Faction in This List?" *Journal of Chinese History* 4, no. 2 (2020): 347–389.

Hsu, Ya-hwei 許雅惠. "Bei Song wanqi jinshi shoucang de shehui wangluo fenxi" 北宋晚期金石收藏的社會網絡分析 [The Social Networks of Antiquities Collectors in the Late Northern Song]. *Xinshixue* 新史學 29, no. 4 (2018): 71–124.

Liu, Feiyan 劉飛燕, and Gao Jianbo 高劍波. "Sui Tang zhi Song shiqi jingying shehui wangluo donglixue de yanhua yanjiu" 隋唐至宋時期精英社會網絡動力學的演化研究 [Dynamical Evolution of Social Networks of Elites from Sui-Tang to Song Dynasty]. *Shuzi renwen* 數字人文 1 (2020): 118–127.

Tackett, Nicolas. "The Evolution of the Tang Political Elite and its Marriage Network." *Journal of Chinese History* 4, no. 2 (2020): 277–304.

Yan, Chengxi 嚴承希, and Wang Jun 王軍. "Shuzi renwen shijiao: jiyu fuhao fenxifa de Songdai zhengzhi wangluo keshihua yanjiu" 數字人文視角：基於符號分析法的宋代政治網絡可視化研究 [Digital Humanistic Perspective: A Study on the Visualization of Political Network in Song Dynasty Based on Symbolic Analysis]. *Zhongguo tushuguan xuebao* 中國圖書館學報 44, no. 5 (2018): 87–103.