Nominative Linkage of Records of Officials in the China Government Employee Dataset-Qing (CGED-Q)

By Cameron Campbell and Bijia Chen

To cite this article: Campbell, C., & Chen, B. (2022). Nominative Linkage of Records of Officials in the China Government Employee Dataset-Qing (CGED-Q). *Historical Life Course Studies, 12*, 233–259. https://doi.org/10.51964/hlcs11902

HISTORICAL LIFE COURSE STUDIES

Content, Design and Structure of Major Databases with Historical Longitudinal Population Data

VOLUME 12, SPECIAL ISSUE 5, 2020

GUEST EDITORS
George Alter
Kees Mandemakers
Hélène Vézina

EHPS NETWORK

MISSION STATEMENT

HISTORICAL LIFE COURSE STUDIES

Historical Life Course Studies is the electronic journal of the *European Historical Population Samples Network* (EHPS-Net). The journal is the primary publishing outlet for research involved in the conversion of existing European and non-European large historical demographic databases into a common format, the Intermediate Data Structure, and for studies based on these databases. The journal publishes both methodological and substantive research articles.

Methodological Articles
This section includes methodological articles that describe all forms of data handling involving large historical databases, including extensive descriptions of new or existing databases, syntax, algorithms and extraction programs. Authors are encouraged to share their syntaxes, applications and other forms of software presented in their article, if pertinent, on the EHPS-Net website.

Research articles
This section includes substantive articles reporting the results of comparative longitudinal studies that are demographic and historical in nature, and that are based on micro-data from large historical databases.

Historical Life Course Studies is a no-fee double-blind, peer-reviewed open-access journal supported by the European Science Foundation (ESF, http://www.esf.org), the Scientific Research Network of Historical Demography (FWO Flanders, http://www.historicaldemography.be) and the International Institute of Social History Amsterdam (IISH, http://socialhistory.org/). Manuscripts are reviewed by the editors, members of the editorial and scientific boards, and by external reviewers. All journal content is freely available on the internet at hlcs.nl.

Co-Editors-In-Chief:
Joana Maria Pujadas-Mora (Open University of Catalonia & Center for Demographic Studies, Autonomous University of Barcelona) &
Paul Puschmann (Radboud University)

Associate Editor:
Eva van der Heijden (Utrecht University)

hislives@kuleuven.be

The European Science Foundation (ESF) provides a platform for its Member Organisations to advance science and explore new directions for research at the European level. Established in 1974 as an independent non-governmental organisation, the ESF currently serves 78 Member Organisations across 30 countries. EHPS-Net is an ESF Research Networking Programme.

The European Historical Population Samples Network (EHPS-net) brings together scholars to create a common format for databases containing non-aggregated information on persons, families and households. The aim is to form an integrated and joint interface between many European and non-European databases to stimulate comparative research on the micro-level.
Visit: http://www.ehps-net.eu.

HISTORICAL LIFE COURSE STUDIES
VOLUME 12 (2022), published 08-09-2022

Nominative Linkage of Records of Officials in the China Government Employee Dataset-Qing (CGED-Q)

Cameron Campbell

The Hong Kong University of Science and Technology & Central China Normal University & 2022–23 Fellow, Center for Advance Study in the Behavioral Sciences, Stanford University

Bijia Chen

Renmin University

ABSTRACT

We introduce our approach to the nominative linkage of records of Qing officials who were included in the China Government Employee Datasets-Qing (CGED-Q) Jinshenlu (JSL) and Examination Records (ER). We constructed these datasets by transcription of quarterly rosters of civil and military officials produced by the government and by commercial presses, and records of examination degree holders. We assess each of the primary attributes available in the original sources in terms of their usefulness for disambiguation, focusing on their diversity and potential for inconsistent recording. For officials who were not affiliated with the Eight Banners, these primary attributes include surname, given name, and province and county of origin. For the small subset of officials who were affiliated with the Bannermen, we assess the available data separately. We also assess secondary attributes available in the data that may be useful for adjudicating candidate matches. We then describe the approach that we developed that addresses the issues we identified with the primary and secondary attributes. The issues we have identified and the approach that we have developed will be of interest to researchers engaged in similar efforts to construct and link datasets based on elite males in historical China.

Keywords: China, Nominative linkage, Elites, Careers

e-ISSN: 2352-6343
DOI article: https://doi.org/10.51964/hlcs11902

© 2022, Campbell, Chen
This open-access work is licensed under a Creative Commons Attribution 4.0 International License, which permits use, reproduction & distribution in any medium for non-commercial purposes, provided the original author(s) and source are given credit. See http://creativecommons.org/licenses/.

1 INTRODUCTION

We describe our approach to the large-scale nominative linkage of records of elite males in two Qing dynasty (1644–1911) historical datasets that we have constructed: the China Government Employee Dataset-Qing Jinshenlu (CGED-Q JSL) and Examination Records (CGED-Q ER). By transcribing records of Qing civil and military officials in quarterly personnel rosters from the period between 1762 and 1911 to produce the CGED-Q JSL and then linking those records over time, we have reconstructed the career histories of officials. By linking officials in the CGED-Q JSL to their records in the CGED-Q ER, we have also attached information about their year of birth, exam performance, ancestry, and other attributes to their career records. This allows us to examine a major topic in the sociological study of stratification: the roles of family background and 'ability' (as measured by exam performance) in the appointment, promotion, and exit from work of officials. Attaching information on year of birth to career histories allows for the study of the age structure of officialdom and the age dynamics of appointment, promotion, and exit from work.

We arrived at the approach we describe here iteratively, building on experience analyzing career histories in the CGED-Q JSL in a series of publications on appointment, promotion, and exit of Qing officials (Campbell, 2020; Chen, Campbell, & Lee, 2018; Hu, Chen, & Campbell, 2020; Hu, Hu, Chen, & Campbell, 2021; Xue & Campbell, 2022), a visualization platform (Wang et al., 2021), an introduction to the CGED-Q JSL (Chen, Campbell, Ren, & Lee, 2020) and a dissertation (Chen, 2019). Each analysis brought to light issues with the sources, the transcription process, and linkage procedures that had not arisen previously and required adjustments. As our dataset expanded, meanwhile, we adjusted our code and obtained substantial improvements in speed. In the end, as described below, we used probabilistic linkage as implemented in the STATA package dtalink (Kranker, 2018).

The most important contribution of the paper is the thorough documentation of the many problems that arise in the recording of names, place of origin, and other attributes in Qing administrative sources, the implications of these problems for nominative linkage, and our solutions to them. We hope that our experience will be useful to researchers carrying out large-scale nominative linkage in other Chinese sources and to users of the CGED-Q JSL public releases that we have made available for download (Campbell, Chen, Ren, & Lee, 2019).[1] The problems that we identify and our solutions to them should be general to historical Chinese sources. Common problems include the replacement of characters in surnames and given names with variant forms, homonyms, and similar-looking characters, and the inconsistency in the recording of locations because of changes in administrative boundaries. To facilitate work by others who are carrying out nominative linkage with historical Chinese sources, we have also made the complete tabulations that are the basis of most of our tables available for download.[2]

We organize our paper as follows. Section 2 presents a brief review of the literature on nominative linkage in historical and Chinese language sources. Sections 3 and 4 introduce the two datasets that we link: the China Government Employee Dataset-Qing Jinshenlu (CGED-Q JSL) and the Examination Records (CGED-Q ER). We describe the attributes of officials recorded in the data that may be used for linkage, distinguishing between primary attributes available for all officials in both sources, and secondary attributes only available in the CGED-Q JSL. We identify issues that arise in the primary attributes that need to be addressed when carrying out linkage. Section 5 describes our current approaches to linkage in the CGED-Q JSL and CGED-Q ER. Section 6 concludes with discussion of implications of these results and prospects for the future.

1 We have released CGED-Q JSL data for the years 1850–1864 and 1900–1912. The data and documentation may be downloaded at the Lee-Campbell Group page at the HKUST Dataspace (https://doi.org/10.14711/dataset/E9GKRS) and the Lee-Campbell Group page at the Harvard Dataverse (https://doi.org/10.7910/DVN/GMQWVZ). We will release additional tranches of CGED-Q JSL data every two years until we have made it all available, and then release the CGED-Q ER.

2 As a resource for other researchers carrying out nominative linkage with historical Chinese sources, we have made the complete tabulations that were the basis of tables 2 through 8 available for download at the Lee-Campbell Group dataverses at the HKUST Dataspace (https://doi.org/10.14711/dataset/M8HQEA) and at the Harvard Dataverse (https://doi.org/10.7910/DVN/4OSP8V). These should help researchers who need to address issues related to inconsistency in the recording of names or places of origin develop approaches to handle such problems.

2 BACKGROUND

Large-scale automated nominative linkage of records of individuals from archival sources is a key tool for production of longitudinal 'big data' for ongoing studies of European and North American population, social and economic history. Common applications are the linkage of census records for the same individual at different points in time, and linkage of individuals across birth, death, and marriage records. Linkage may also include other, more specialized sources including tax records, health records, and retirement and pension records that supplement information routinely available in census records and vital registration. The resulting linked data not only provide life histories of individuals, but in some cases, histories of families across multiple generations. As a result of this activity, methods for large-scale nominative linkage of individuals in sources written in English and other languages that use phonetic scripts are relatively mature. A large literature discusses challenges associated with nominative linkage and offers various solutions, and relevant software packages are readily available.[3]

The literature on large-scale nominative linkage of records of individuals with names written in English and other phonetic scripts in historical sources is already large because efforts to construct massive longitudinal databases for social and economic history by linkage of censuses, vital records and other administrative data have been underway in the United States, Canada, and a variety of European countries for at least two decades.[4] An early example was the initiative by the Minneapolis Population Centre to create a statistically representative sample of records in the 1860, 1870, 1900 and 1910 censuses linked to the complete-count 1880 census described in Ruggles (2002). Since then, methodology has advanced substantially, with explorations of machine learning to fully automate record linkage (Abramitzky, Mill, & Pérez, 2020) and the leveraging of information on residence and relationships to increase linkage rates (Akgün et al., 2020; Helgertz et al., 2020).

Key issues that arise in the linkage of names written in English and other languages with phonetic alphabets include misspellings, name changes, the use of variant spellings, and inconsistencies in the recording of other attributes like age or date of birth, all of which could create false negatives, and overall low diversity of surnames and given names, which could lead to false positives. By 'false negatives', we refer to situations where two records that should have been linked together were not. By 'false positives', we refer to situations where records that should not have been linked together, were. Misspellings occurred because people were inconsistent in the way they wrote their own name, or the way census takers or other officials wrote their name in official records. International migrants might have new names assigned to them by immigration officers who transliterated their original names in ad hoc fashion or might adapt new names on their own. Women typically adopted their husband's surname on marriage. People might use contractions of their name or nicknames in some situations but not in others, for example, writing Bill in some situations and William in others. In many communities in Europe, diversity of surnames and given names was low, making it difficult to distinguish whether records of the same name referred to the same or different people.

The issues that arise with names written in Chinese are very different. Surnames are not diverse. In 2020, the top 5 surnames in China accounted for 30.8% of the population, and the top 100 surnames accounted for 85.8% of the population.[5] Given names are potentially more diverse since they are typically two characters, and for each of those two characters there are thousands to choose from. The actual diversity of given names depended on naming practices in different periods and social classes. While names of elite males during the Qing and the first half of the 20th century should have been very diverse because well-off families could showcase their erudition by including rare characters with literary, historical or philosophical connotations in the names of their sons, names for people born between the 1960s and 1980s were much less diverse than for those born before or after because

3 See, for example, the Linkage Library at https://www.icpsr.umich.edu/web/pages/about/linkage-library.html.

4 See *Historical Methods* Special Issues 51(2) and 53(4) on historical record linkage for introductions to relevant projects (Sylvester & Hacker, 2020).

5 See the 2019 and 2020 *Nian Quanguo Xingming Baogao* [National Surname and Given Name Report] published by the Public Security Bureau of the People's Republic of China, retrieved from http://www.gov.cn/xinwen/2021-02/08/content_5585906.htm.

single character names with political or patriotic implications became more popular (Cai, Xi, Yi, Liu, & Jing, 2018; Bao, Cai, Jing, & Wang, 2021).[6]

Developing procedures for record linkage is important because there are numerous efforts ongoing to create biographical databases of historical Chinese individuals. Prominent examples include the China Biographical Database (Chen & Wang, 2022; Fuller, 2021; Tsui & Wang, 2020), the Modern China Historical Database (Armand, Guo, Henriot, Hu, & Van den Bosch, 2022), and the various projects of the Lee-Campbell Group (Campbell & Lee, 2020). Such databases are the basis of prosopographical studies of social groups (Stone, 1971), especially elites, in historical China. The creators of these databases carry out what they refer to as 'disambiguation' to assess whether the same name and other attributes appearing in two or more sources refers to the same person or different people, and then attach unique identifiers to each appearance of a person in the dataset. The underlying task is similar to the record linkage that we carry out in the CGED-Q, but somewhat broader in that it may also involve individuals named in unstructured texts like newspaper articles, dynastic histories, or gazetteers.[7] Pronunciation-based approaches developed for linkage of individuals with names written in phonetic scripts are not immediately useful for these Chinese language sources because the prevalence of homonyms in Chinese means that names with identical pronunciations can be completely different. Meanwhile, characters that look similar and may be mistakenly replaced with each other during the production process can be pronounced differently and have different meanings.

The studies of Chinese language nominative linkage and disambiguation that we have located focus on names in contemporary unstructured Chinese language texts (for example, web pages) not on structured records like in the CGED-Q. We mention them here because they could eventually help with the linkage of the officials in the CGED-Q to mentions of them in unstructured texts. Chen and Huang (2010) assessed issues that arise in the disambiguation of the names of individuals in Chinese language texts. They report that single character given names are more challenging than two character given names. Combinations of surname and single-character name that are also commonly used words are especially difficult to disambiguate. For example, the combination *Gaofeng* (高峰) could be the surname Gao followed by the given name Feng but could also be the word for 'peak'.[8] Han, Zu and Zhao (2011) and Fan and Li (2021) describe approaches based on clustering in which the same names appearing in different documents are disambiguated by reference to other words appearing with them in the text. The problems these papers address is different to the one we face in our own linkage of names in tabular datasets where the surname and given name are clearly specified in fields of their own, but relevant for efforts by others to extract and disambiguate names in unstructured historical texts like newspaper articles, books, and essays.

Several studies discuss the disambiguation of Chinese names of authors of texts. Han et al. (2017) focus on the specific case of disambiguating the names of authors of Chinese language publications, and introduce a method based on the names of the co-authors, the author's institution, and 'semantic fingerprints'. Kim, Kim and Kim (2021) shows that disambiguation of the names of Chinese authors of English language publications is easier if their name in Chinese characters is available alongside their phoneticized names. Yin, Motohashi and Dang (2020) presents the results of an effort to disambiguate the names of inventors listed on Chinese patents between 1985 and 2016. They use supervised learning approach that begins with hand-labelled data for training.

Another line of studies offers potentially useful approaches for measuring similarity in the sound and the appearance of Chinese characters and then using this to assess the similarity of strings of Chinese characters. Liu, Rus, Liao and Liu (2017) offer a method for encoding Chinese characters in terms of their sound,

6 See Chua (2021) for an overview of contemporary naming practices in China and descriptive results on the popularity of different kinds of names during the 20th century. The analysis was based on the Chinese Name Database (1930–2008) created by Han-Wu-Shang (Bruce) Bao and shared at https://github.com/psychbruce/ChineseNames.

7 Campbell and Lee (2020) and Chen and Campbell (2023) include brief, non-technical overviews of linkage in the CGED-Q as part of their overviews of the methods used in the project. We describe linkage procedures for two of our other publicly released datasets, the China Multigenerational Panel Datasets (CMGPD) Liaoning (LN) and Shuangcheng (SC), in Appendix A of Lee and Campbell (1997), Lee, Campbell and Chen (2010) and Wang et al. (2013). According to personal communication with the leaders of the China Biographical Database and Modern China Historical Database projects, they do not yet have any publications describing their procedures for linkage and disambiguation.

8 Segmentation of text is also important because in the absence of spaces between words, there are instances where the last character of one word and the first character of the word that immediately follows might be mistaken for a name.

appearance, and meaning, and then ranking pairs of characters according to their similarity. Chen et al. (2018) proposes a "SoundShape Code" for Chinese characters that reflects their pronunciation and appearance, and which may be used as a basis of measuring similarity between two characters in a pair. Xu, Zheng and Li (2020) combine the SoundShape Code for individual Chinese characters with the Dice similarity measure for strings of potentially different lengths. Such methods address a challenge that we describe below: because of errors in the original source or errors during our transcription, the names of the same individual may appear with slightly different characters in different records in our dataset. Characters may be replaced by a homonym that looks different, or with a visually similar character that is pronounced very differently.

3 CHINA GOVERNMENT EMPLOYEE DATASET-QING JINSHENLU (CGED-Q JSL)

We constructed the China Government Employee Dataset-Qing Jinshenlu (CGED-Q JSL) from *Jinshenlu* (縉紳錄) and *Zhongshubeilan* (中樞備覽) rosters of Qing civil and military officials respectively that were produced every three months. We have described the CGED-Q JSL and the sources from which it was constructed in detail elsewhere (Chen et al., 2020; Ren, Chen, Hao, Campbell, & Lee, 2016, 2019) and only provide key details here. Official editions of the *Jinshenlu and Zhongshubeilan* were produced by the Qing Ministries of Personnel and War, respectively.[9] The government used the official editions to keep track of posts and the officials who held them. In the 19th century, commercial publishers produced and sold editions that supplemented information on officials from the official editions with additional information collected by the publishers.[10] Purchasers of commercial editions used them for a variety of purposes, including searching for vacant positions and locating kin, classmates, or other connections who they knew were officials.

At the time of writing, the CGED-Q JSL contains 4,433,600 records from 275 *Jinshenlu* editions and 75 *Zhongshubeilan* editions. Each *Jinshenlu* roster lists 13,000 to 15,000 posts in the civil service and identifies the officials who held them. *Zhongshubeilan* rosters each list approximately 8,000 military posts and the officers who held them. The editions in the CGED-Q JSL are from the period 1762 to 1912. Coverage is sparse before 1830, but very complete after that year. From 1830 to 1911, the CGED-Q JSL includes at least one *Jinshenlu* edition from nearly every year. In many years, it includes all four quarterly editions. *Zhongshubeilan* are sparser and the gaps between them are longer.

78.9% of officials were ordinary citizens (*minren*, 民人) and almost all the remainder were Bannermen (*qiren*, 旗人). The vast majority of *minren* were what we would now refer to as Han Chinese.[11] Bannermen were hereditary affiliates of the Eight Banners, originally the army used to conquer China and establish the Qing in 1644, and in the 18th and 19th centuries, an organization used by the Qing state to maintain political and military control. Most officials who were Bannermen were Manchu or Mongol, but 16.4% were Han Chinese. The latter were referred to as Han Martial Bannermen (*hanjun qiren*, 漢軍旗人). They were the descendants of Han Chinese who had been incorporated into the Eight Banners. Bannermen had a privileged position in the Qing government, with their own pathways to appointment and promotion, and quotas for certain positions. Thus, even though Bannermen accounted for only 2%–4% of the population of the Qing (Elliott, Campbell, & Lee, 2016), they accounted for one-fifth of civil officials overall, two-thirds of civil officials serving in the capital Jingshi (now Beijing) and 90% of officials in the secondary capital Shengjing (now Shenyang) (Chen et al., 2020, p. 454).

To produce career histories by longitudinal linkage of CGED-Q JSL records of officials, we distinguish between what we refer to as the primary and secondary attributes recorded for officials. We define primary

9 We refer to the dataset constructed from *Jinshenlu* and *Zhongshubeilan* rosters as CGED-Q Jinshenlu (JSL) because when *Zhongshubeilan* are available, it is usually as part of a set with a *Jinshenlu* edition for the same season. The resulting sets are typically catalogued by libraries and archives as a *Jinshenlu* edition. We only have a small number of freestanding *Zhongshubeilan* editions that are not part of a set.

10 See Chen et al. (2020) for a detailed discussion of the differences in the contents of the official and commercial editions.

11 What we refer to as *minren* likely also included members of what since the 1950s have been officially designated as minority ethnic groups, but the *Jinshenlu* does not record any information that would allow us to distinguish them.

attributes as basic and stable information about an official that are available in all or nearly all records and should be available in almost any other source that we might wish to link to. The most important of these are the names. We define secondary attributes as characteristics that are specific to the CGED-Q JSL and may not be available in other sources or recorded in every edition of a *Jinshenlu* or *Zhongshubeilan*. They may also be attributes that vary over time, for example, the official's current position. These may be used to adjudicating candidate links made based on the primary attributes, but on their own are not sufficient for linkage within the CGED-Q JSL or between the CGED-Q JSL and CGED-Q ER.

For linkage, we separate officials according to whether they had a surname recorded because the primary attributes available for officials with surnames differed from those available for those without surnames. Officials with surnames accounted for 80.2% of records. These included all the *minren* and one-third of the Han Martial Bannermen.[12] Basic information recorded for them included not only their surname (*xing*, 姓) and given name (*ming*, 名) but also their place of origin. The latter was usually the province and prefecture or county of origin, though there are complications that we discuss below. Officials without surnames included all Manchu (*Manzhou*, 滿洲) and Mongol (*Menggu*, 蒙古) Bannermen and two-thirds of Han Martial Bannermen.[13] The only attributes recorded for officials without surnames that were in principle stable were given name and Banner affiliation (*qifen*, 旗分). We use these as the primary attributes for Bannermen.

The primary attributes for officials with and without surnames differ in terms of their ability to uniquely identify officials within an edition. For officials with a surname, the combination of surname, given name, and province and county of origin was usually unique within an edition. If these were all recorded reliably and consistently across every edition, they would in principle be sufficient for linkage. Table 1 summarizes the number of repetitions of combinations of primary attributes within each quarterly *Jinshenlu* edition. For officials with a surname, 95.0% of the combinations of surname and given name were unique within their edition. In other words, for 95.0% of records, there was no other record in the same edition with the same surname and given name. For 4.4% of records, there was only one other record in the same edition with the same surname and given name. 98.1% of records of officials with a surname were unique within their edition in terms of the combination of surname, given name, and place of origin. Our investigations have revealed that for these officials, most repetitions within the same edition all refer to the same official. If an official held more than one post, there was a separate record for each of them.

Table 1 *Uniqueness of primary attributes of officials within each quarterly edition of the* Jinshenlu*, 1760–1912*

	Officials with a surname		**Officials without a surname**	
	Surname and Given name	**+ Place of origin**	**Given name**	**+ Banner**
Repetitions within an edition[a]	%	%	%	%
1	95.0	98.1	64.0	88.0
2	4.4	1.7	19.9	9.9
3	0.5	0.2	8.2	1.4
4	0.1	0.0	4.0	0.4
5 or more	0.01	0.0	3.9	0.4
Total	100	100	100	100
Records	2,817,156	2,817,156	784,502	784,502

[a] *Repetitions refers to the total number of records in the same quarterly edition with the specified combination primary attributes.*

For officials without surnames, given name by itself is not sufficient for linkage. Only two-third of records recorded a given name that was unique within the quarterly edition. One-third of records had a name that appeared in one or more other records. When Banner affiliation was added, 88% of records became unique within their quarterly edition in terms of the primary attributes. 12% of records had a given name and Banner affiliation that appeared in at least one other record in the same

12 Exactly one-third of the officials recorded as Han Marital Bannermen had a surname recorded. The remainder did not, presumably because they had taken Manchu names. See Campbell, Lee and Elliott (2002) for a discussion of the adaptation of Manchu names by Han Chinese in northeast China.

13 Manchu and Mongol Bannermen accounted 71.4% and 12.2% of Bannermen, respectively.

quarterly education. Based on our investigations, these reflect some cases where the same official held more than one office, as well as cases where two different officials had the same name.

These results highlight that the approaches to linkage must differ according to whether a surname was available. For officials with a surname, as discussed above, the combination of surname, given name, and province and county of origin all written in Chinese characters is likely to be unique, and false positives in which records of different officials are mistakenly linked together should be rare. The main task for linkage of officials with a surname is avoiding false negatives in which an inconsistency in the recording of the name or some other attribute prevents a link from being made. For officials without a surname, the risk of false positives is high because surnames and place of origin are not available, and there are enough officials who share the same combination of given name and Banner affiliation to raise concerns that two records with identical name and Banner affiliation may refer to different officials.

Below we introduce the primary and secondary attributes in detail and assess their usefulness for linkage, with a focus on their homogeneity or heterogeneity. We divide our discussion of attributes between those available in records of officials with surnames and those available in records of officials without surnames.

3.1 ATTRIBUTES AVAILABLE FOR OFFICIALS WITH SURNAMES

3.1.1 SURNAMES

Because a small number of surnames accounted for a large share of the records of officials, surnames are of limited utility as a primary attribute for linkage. According to Table 2, which presents the cumulative percentages of records accounted for by the 100 most common surnames in the CGED-Q JSL, the five most common surnames appeared in one-quarter of the records. These were Wang (王) , Zhang (張), Li (李), Chen (陳) and Liu (劉). The top 10 surnames accounted for 38.3% of the records of officials with surnames. The top 20 surnames accounted for approximately one-half of the records, and the top 200 accounted for 95.1%. There were a total of 1626 distinct surnames recorded, though the actual number was lower because in this tabulation a surname may have more than one entry if the character appears in more than one form.

Table 2 *Cumulative percentages of the top 100 most common surnames in the CGED-Q JSL, 1760–1912*

	1–20		21–40		41–60		61–80		81–100	
	Surname	%	Surname	%	Surname	%	Surname	%	Surname	%
1	王	6.6	林	51.6	蔡	65.4	魏	75.0	薛	81.5
2	張	12.7	謝	52.4	韓	65.9	戴	75.4	廖	81.8
3	李	18.7	郭	53.3	唐	66.5	盧	75.7	白	82.0
4	陳	23.5	高	54.1	鄧	67.1	田	76.1	嚴	82.3
5	劉	27.7	許	54.9	蔣	67.6	崔	76.5	萬	82.6
6	楊	30.6	馮	55.6	方	68.2	夏	76.8	施	82.8
7	周	32.8	吳	56.4	孔	68.7	熊	77.2	賈	83.1
8	吳	34.7	羅	57.1	蕭	69.3	陶	77.5	洪	83.3
9	徐	36.5	梁	57.8	袁	69.8	秦	77.8	雷	83.6
10	趙	38.3	姚	58.5	曾	70.3	俞	78.2	邱	83.8
11	朱	40.0	葉	59.2	董	70.8	江	78.5	姜	84.1
12	孫	41.5	程	59.9	章	71.3	譚	78.8	孟	84.3
13	胡	42.9	余	60.5	傅	71.7	鄒	79.2	賀	84.5
14	馬	44.2	宋	61.1	錢	72.2	史	79.5	毛	84.8
15	沈	45.4	潘	61.7	顧	72.6	于	79.8	侯	85.0
16	黃	46.6	丁	62.4	范	73.0	鍾	80.1	尹	85.2
17	何	47.8	彭	63.0	杜	73.4	龔	80.4	武	85.4
18	鄭	48.8	陸	63.6	蘇	73.8	邵	80.7	郝	85.6
19	黃	49.7	曹	64.2	任	74.2	石	80.9	葛	85.8
20	汪	50.7	金	64.8	呂	74.6	湯	81.2	倪	85.9

Note: Based on authors' calculations on 3,244,484 CGED-Q JSL records with a legible surname.

One issue that arises with linkage based on surnames is that a character may be replaced with one that looks similar in an adjacent edition. Of the 1,559,380 pairs of records in editions which were no more than one year apart and almost certainly referred to the same official because they recorded the same two-character given name, province and county of origin, and position and broad category of degree qualification, 20,055 pairs (1.3%) differed on the character written for the surname.[14] Table 3 presents the cumulative frequencies of discordant pairs of surnames. The most common discordant pair (黃 黃) accounted for 22.4% of discordant pairs overall, the top 20 accounted for nearly two-thirds (63.1%) and the top 100 accounted for 79.2%.

Inspection of the results in Table 3 reveals two common issues that may generate false negatives, in which records of the same official are not properly linked. The first issue is that some pairs are the same character written in variant forms (*Yitizi*, 異體字). The four most common pairs in Table 3 are examples: 黃 and 黃, 吳 and 吳, 高 and 高, and 呂 and 呂 are different ways of writing the surnames Huang, Wu, Gao, and Lu respectively. In the Unicode standard these are recognized as different representations of the same character, and as we describe below, this is straightforward to address. The second and more challenging issue is that sometimes between editions a character for a surname is replaced by one that looks similar but is a completely different character. Examples in Table 3 include the fifth entry (叚, Xia and 段, Duan), the seventh entry (宋, Song and 朱, Zhu), the 10th entry (汪, Wāng and 王, Wáng), and the 15th entry (馬, Ma and 馮, Feng). These issues reflect either inconsistencies in the production process across different editions or transcription errors by coders. There are also examples of discordant pairs in Table 3 that consist of characters that are clearly different, for example the 24th entry 張 章 (Zhang and Zhang) and the 28th entry 程 陳 (Cheng and Chen). In most of these cases, one or both characters are relatively common surnames. While there is some possibility that these could be from records of different people, they may also be transcription errors that occurred during data entry.

Table 3 *Cumulative percentages of the top 100 most common discordant pairs of surnames in adjacent editions in the CGED-Q JSL*

	1–20		21–40		41–60		61–80		81–100	
	Pair	%	Pair	%	Pair	%	Pair	%	Pair	%
1	黃 黃	22.4	衛 衛	63.7	鄧 鄭	70.7	盧 虞	74.6	蔣 薛	77.2
2	吳 吳	35.7	關 關	64.2	曹 曾	71.0	丁 于	74.7	翰 韓	77.3
3	高 高	41.4	孫 馮	64.7	章 童	71.2	関 關	74.9	向 尚	77.4
4	呂 呂	44.8	張 章	65.1	杜 林	71.5	馮 馮	75.0	俞 喻	77.5
5	叚 段	47.2	柳 柳	65.6	余 徐	71.7	葉 蔡	75.1	褚 諸	77.6
6	錢 錢	49.3	劉 陳	66.0	徐 涂	71.9	曾 魯	75.3	徐 許	77.7
7	宋 朱	51.3	寗 甯	66.4	全 金	72.1	張 陳	75.4	朿 束	77.8
8	閆 閻	52.8	程 陳	66.8	鄔 鄢	72.4	董 黃	75.5	冠 寇	78.0
9	汪 王	54.1	楊 陽	67.1	董 黃	72.6	刑 邢	75.7	樂 欒	78.1
10	凌 淩	55.3	余 金	67.5	員 貟	72.8	宋 宗	75.8	張 楊	78.2
11	賴 賴	56.5	楊 湯	67.8	于 王	72.9	萬 黃	75.9	苑 范	78.3
12	余 俞	57.5	毛 王	68.1	李 陳	73.1	強 强	76.1	郭 鄧	78.4
13	龐 龐	58.4	佘 余	68.4	吳 呂	73.3	王 黃	76.2	娄 婁	78.5
14	温 溫	59.2	竇 竇	68.8	晉 晋	73.5	曹 曺	76.3	柏 栢	78.6
15	馬 馮	59.9	季 李	69.1	曹 賈	73.6	潘 王	76.4	丁 李	78.7
16	冷 涂	60.6	嵇 稽	69.4	童 董	73.8	湛 諶	76.6	褚 褚	78.8
17	顏 顏	61.2	龍 龔	69.6	劉 鄧	74.0	杜 樊	76.7	範 范	78.9
18	関 關	61.9	侯 候	69.9	邊 邊	74.1	唐 康	76.8	廉 廉	79.0
19	江 汪	62.5	朱 李	70.2	瞿 翟	74.3	寇 寇	76.9	呂 吳	79.1
20	鍾 鐘	63.1	陳 陸	70.5	宮 宮	74.4	荆 荊	77.0	孫 張	79.2

Note: Of the 1,559,380 pairs of records in adjacent editions no more than one year apart that were identical on given name, province and county of origin, broad category of degree qualification, and position, 20,055 (1.3%) were discordant.

14 Degree qualification refers to the examination or purchased degree that qualified a *minren* official for appointment to office. This is a secondary attribute that we discuss below.

3.1.2 GIVEN NAMES

Given names (Table 4) were the most diverse of the primary attributes available for officials with surnames, and therefore the most useful for record linkage. We distinguish between records of officials with two- and one-character names. The former accounted for 85% of the records and the latter accounted for the remainder. A total of 102,648 distinct given names appeared in our data, 98,745 of which were two-character names, with the remaining 3,903 being one-character names. According to Table 4, two-character names were very diverse. The top 100 accounted for only 5.7% of records, the top 200 accounted for 9% of records, the top 1,000 accounted for 23% of records, and the top 10,000 accounted for only 61% of records. The diversity of two-character names reflects the large number of characters available to choose from: we found that at least 5,764 different characters made at least one appearance in a two-character given name in the CGED-Q JSL.[15]

Table 4 *Cumulative percentages of the top 100 most common two-character given names of officials with surnames in the CGED-Q JSL*

	1–20		21–40		41–60		61–80		81–100	
	Given name	%	Given name	%	Given name	%	Given name	%	Given name	%
1	汝霖	0.1	樹棠	1.7	瑞麟	2.9	祖培	3.9	錫麟	4.9
2	文炳	0.2	炳文	1.8	桂芳	3.0	繼昌	4.0	登雲	4.9
3	得勝	0.3	雲龍	1.8	殿元	3.0	沛霖	4.0	文彬	4.9
4	占魁	0.4	桂林	1.9	玉麟	3.1	祖蔭	4.1	安邦	5.0
5	兆麟	0.5	占鰲	2.0	國泰	3.1	鴻鈞	4.1	錫疇	5.0
6	作霖	0.6	逢春	2.0	維藩	3.2	其昌	4.2	建勳	5.1
7	廷棟	0.7	廷桂	2.1	恩培	3.2	鵬飛	4.2	鴻恩	5.1
8	秉鈞	0.8	鳳翔	2.2	紹曾	3.3	炳章	4.3	毓麟	5.2
9	承恩	0.9	步雲	2.2	文蔚	3.3	炳南	4.3	玉堂	5.2
10	慶雲	1.0	國楨	2.3	殿魁	3.4	國祥	4.4	樹森	5.2
11	世昌	1.0	煥章	2.3	桂森	3.4	長庚	4.4	念祖	5.3
12	步瀛	1.1	文藻	2.4	國華	3.5	定邦	4.4	桂芬	5.3
13	兆熊	1.2	長春	2.5	光祖	3.5	振邦	4.5	學海	5.4
14	培元	1.2	登瀛	2.5	國瑞	3.6	萬春	4.5	連陞	5.4
15	文光	1.3	慶元	2.6	廷珍	3.6	慶恩	4.6	家駒	5.4
16	維翰	1.4	維城	2.6	世榮	3.7	永清	4.6	錫祺	5.5
17	樹勳	1.4	恩榮	2.7	恩溥	3.7	永清	4.7	文治	5.5
18	文煥	1.5	錫齡	2.7	維新	3.8	廷杰	4.7	濟川	5.6
19	錫恩	1.6	國棟	2.8	春華	3.8	榮光	4.8	占春	5.6
20	振聲	1.6	壽昌	2.8	遇春	3.9	廷楨	4.8	鶴年	5.7

Note: Based on authors' calculations on 2,718,433 CGED-Q JSL records with a legible surname and a legible two-character given name.

Like surnames, characters in given names may also be inconsistent across different quarterly editions. If not addressed, this may also lead to false negatives. Table 5 repeats the exercise for surnames carried out in Table 3 for the characters in two-character given names.[16] It presents the cumulative percentages of discordant pairs, defined as characters in given names that differ between records in editions that are no more than one year apart, and where the surname, one of the two characters in the given name, place of origin, position, and degree qualification are all identical. Out of 1,539,198 such pairs of records, 4.34% (66,994) differed on one character in the given name. Discordant pairs of characters in given names were much more diverse than was the case for surnames. The most common

15 We have included the complete tabulation of characters making at least one appearance in a two-character given name as one of the files available for download at the Harvard and HKUST Dataverses for this paper.

16 Restricting to a two-character name and then including the requirement that at least character in the name matches substantially increases the likelihood that two records that match on everything else refer to the same person.

discordant pair (淸 and 清) accounted for only 3.7% of discordant pairs. The top 20 accounted for one-fifth (20.3%) of discordant pairs, and the top 100 accounted for 39.2%.

Once again, the most common issue is that between one edition and the next, a character was replaced with a variant, of which the seven most frequent pairs are all examples. 淸 and 清, for example, are both ways of writing the same character (Qing). However, there are also cases where a character is replaced by one that is different but looks similar. The 12th, 14th, 22nd and 39th entries are examples: 傅 (Fu) and 傳 (Chuan), 思 (Si) and 恩 (En), 增 (Zeng) and 曾 (Ceng), and 先 (Xian) and 光 (Guang), respectively. Again, this likely reflects a problem during the production of the source, or during the transcription.

Single-character names were less diverse. According to Table 6, the top 10 most common single-character names accounted for 6.6% of records with single-character names and the top 100 accounted for 37%. According to separate tabulations, the top 200 accounted for 54% and the top 500 accounted for 78%. According to a separate tabulation like the ones in Tables 3 and 5 but not shown here, the patterns in discordant pairs are like those in Table 5. Most discordant pairs consisted of the same characters written differently or similar looking characters that could be mistaken for each other. There were examples, however, of characters that were clearly different, at least raising the possibility that they were men from the same county with the same surname and post who should not be linked. Accordingly, we link records of officials with single-character names separately, with more stringent criteria for match on other attributes when assessing candidate links.

Table 5 *Cumulative frequencies of the 100 most common discordant pairs of characters in two-character given names in records of officials with surnames in adjacent editions in the CGED-Q JSL*

	1–20		21–40		41–60		61–80		81–100	
	Pair	%	Pair	%	Pair	%	Pair	%	Pair	%
1	淸 清	3.7	嗚 鳴	20.8	覲 觀	27.8	元 光	32.7	得 德	36.4
2	勳 勳	5.7	增 曾	21.2	之 芝	28.1	堯 尧	32.9	台 臺	36.6
3	齡 齡	7.0	曾 會	21.6	穀 穀	28.4	峯 峰	33.1	宜 宣	36.7
4	鳳 鳳	8.3	遠 遠	22.0	春 椿	28.6	榮 荣	33.3	城 成	36.9
5	壽 壽	9.5	延 廷	22.4	壁 璧	28.9	世 士	33.5	掟 捷	37.0
6	寶 寶	10.7	廉 廉	22.8	緒 緒	29.2	寬 寬	33.7	嘉 家	37.2
7	晉 晋	11.7	懷 懷	23.2	燮 燮	29.4	顯 顯	33.9	彝 彝	37.3
8	煥 煥	12.7	燿 耀	23.6	凌 淩	29.7	為 爲	34.1	日 曰	37.5
9	賓 賨	13.6	愼 慎	24.0	瀚 翰	29.9	惟 維	34.3	如 汝	37.6
10	彥 彦	14.4	濂 濂	24.3	繩 繩	30.1	甲 申	34.5	連 運	37.8
11	恆 恒	15.2	熙 熙	24.7	保 葆	30.4	煇 輝	34.7	宗 崇	37.9
12	傅 傳	15.9	献 獻	25.0	崧 松	30.6	昭 照	34.8	誠 誡	38.1
13	靑 青	16.7	瀾 瀾	25.4	均 鈞	30.9	恩 榮	35.0	弼 弼	38.2
14	思 恩	17.3	棻 芬	25.7	柱 桂	31.1	瑞 端	35.2	煙 燮	38.4
15	鍾 鐘	17.9	蕃 藩	26.0	聯 聯	31.3	祿 禄	35.4	柏 栢	38.5
16	鎭 鎮	18.5	高 高	26.3	豊 豐	31.6	繩 繩	35.6	彝 彝	38.6
17	庭 廷	19.0	啓 啟	26.7	方 芳	31.8	丙 炳	35.7	讓 讓	38.8
18	熙 熙	19.4	樹 澍	27.0	廸 迪	32.0	璋 章	35.9	鰲 鼇	38.9
19	達 達	19.9	先 光	27.3	祐 祐	32.3	堂 棠	36.1	萼 萼	39.1
20	聯 聯	20.3	聯 聯	27.6	擧 舉	32.5	淸 靑	36.2	達 達	39.2

Note: In the 1,539,198 pairs of records with legible surname and two-character given names in adjacent editions no more than one year apart that were identical on surname, one character of the given name, province and county of origin, broad category of degree qualification, and position, there were 66,994 discordant pairs.

Table 6 *Cumulative percentages of the top 100 most common one-character given names of officials with surnames in the CGED-Q JSL*

	1–20		21–40		41–60		61–80		81–100	
	Character	%	Character	%	Character	%	Character	%	Character	%
1	鈞	0.9	芳	12.0	煜	20.2	璋	26.9	溶	32.5
2	榮	1.7	煦	12.4	勳	20.6	煥	27.2	琦	32.7
3	鑑	2.4	淦	12.9	潤	21.0	桐	27.5	瑛	33.0
4	炳	3.0	源	13.4	濤	21.3	鎔	27.8	坦	33.2
5	鏞	3.7	澐	13.8	鵬	21.7	玉	28.1	超	33.5
6	鈺	4.3	浩	14.3	鴻	22.0	筠	28.4	鎬	33.7
7	瀛	4.9	培	14.7	沅	22.4	治	28.7	貴	34.0
8	湘	5.5	棠	15.1	椿	22.7	傑	29.0	鐸	34.2
9	楷	6.0	謙	15.6	釗	23.0	榕	29.3	翰	34.5
10	堃	6.6	溥	16.0	均	23.4	坤	29.5	芬	34.7
11	震	7.2	泰	16.4	琳	23.7	增	29.8	藻	34.9
12	杰	7.7	燦	16.8	雲	24.0	濬	30.1	模	35.2
13	銘	8.2	斌	17.2	華	24.3	燾	30.4	炘	35.4
14	彬	8.6	澍	17.6	瑞	24.7	煌	30.6	棟	35.7
15	霖	9.1	熙	18.0	鉞	25.0	琛	30.9	寅	35.9
16	森	9.6	英	18.4	林	25.3	焜	31.2	濟	36.1
17	俊	10.1	煒	18.7	元	25.6	桂	31.4	淳	36.3
18	楨	10.6	瀚	19.1	灝	25.9	蘭	31.7	濂	36.6
19	鼎	11.0	照	19.5	珍	26.3	銑	32.0	塏	36.8
20	銓	11.5	錦	19.9	璜	26.6	麟	32.2	錕	37.0

Note: Based on authors' calculations on 514,417 CGED-Q JSL records with a legible surname and a legible one-character given name.

The given names recorded in the CGED-Q JSL should otherwise be stable and in our experience are the ones recorded for officials in their family genealogies and other sources like the CGED-Q ER, not their courtesy name (*biaozi*, 表字) or style name (*hao*, 號). We have shared data with researchers who have constructed datasets from lineage genealogies, and they report success linking men in the genealogies to officials in the CGED-Q JSL based on the names in the genealogies. Users of our CGED-Q JSL search page also report success locating ancestors or other figures based on names recorded in genealogies or other sources.[17] As for the stability of names, while we have not explicitly searched for cases where an official appeared to change their given name, we are not aware of any cases where someone appeared with two different given names except as the result of problems with the sources or transcription process that we discuss below.[18]

3.1.3 PLACE OF ORIGIN

For place of origin, the available level of detail differed between the civil officials recorded in the *Jinshenlu* and the military officials in the *Zhongshubeilan*. The place of origin was where an official had first sat for an exam. In most cases this was where their family lived, but as we will discuss below, there were exceptions. 95% of the records of civil officials with surnames in the *Jinshenlu* specified county of origin and either specified province of origin or allowed for it to be inferred from the province in which the official was currently serving.[19] Of the records of military officers with surnames in the *Zhongshubeilan*, 13% had both province and county of origin, 84% only had province of origin and 3% had county of origin.

17 The search page is located at http://vis.cse.ust.hk/searchjsl/.

18 If evidence emerges that officials did change their name, we will have to revisit our procedures for linkage within the CGED-Q JSL to produce career histories, as well as our procedures for linkage to other sources like the CGED-Q ER.

19 Province of origin could be imputed from province of current post because the *Jinshenlu* typically omitted province of origin for officials serving in their home province.

For civil officials, the place of origin was diverse, though not as diverse as the given name. Table 7 presents the cumulative percentages for the 100 most common places of origin as recorded for officials with surnames in the *Jinshenlu*. In most cases this is the province and county or prefecture where an official earned the *shengyuan* (生員) degree that made them eligible to sit for further exams or purchase the degrees that would qualify them for office. Since usually the county was recorded, not the prefecture, below we will only refer to county. A total of 10,156 distinct combinations of province and county or prefecture appeared in the CGED-Q JSL. For reasons that we discuss below, this is larger than the actual number of counties and prefectures at any given time.

The top 10 places of origin accounted for 16.7% of records and the top 100 locations accounted for 45.4% of records. The two most common locations, Daxing (大興) and Wanping (宛平) in Shuntian (順天) require explanation. These were the locations where the sons and possibly other male kin of officials serving in the capital originally sat for their exams. In these and a small number of other cases, the official's family's place of origin was somewhere else. [20] As long as province and county of origin of these officials were recorded consistently in every edition in which they appeared as the prefecture in Shuntian where they took the linkage, there is no problem for linkage. Many of the other top counties of origin were in Zhejiang, traditionally an important source of exam passers, degree purchasers, and officials. Other top counties of origin included Changsha (長沙) in Hunan (湖南) in fifth place, Tianjin (天津) in Zhili (直隸) in seventh place, and Chengdu (成都) in Sichuan (四川) in ninth place.

Table 7 *Cumulative percentages of the 100 most common places of origin of civil officials with surnames*

	1–20		21–40		41–60		61–80		81–100	
	Province and County	%	Province and County	%	Province and County	%	Province and County	%	Province and County	%
1	順天大興	5.2	湖北漢陽	23.4	四川華陽	31.6	山西平定	37.5	江西建昌	41.9
2	順天宛平	7.5	江蘇吳縣	23.9	廣東順德	31.9	四川重慶	37.7	山東歷城	42.1
3	浙江山陰	9.6	湖南善化	24.3	陝西同州	32.2	江西新建	38.0	浙江餘姚	42.3
4	浙江會稽	11.1	貴州貴陽	24.8	直隸河間	32.5	山西介休	38.2	江西南豐	42.5
5	湖南長沙	12.2	山東濟南	25.2	廣東嘉應	32.9	浙江慈谿	38.5	湖南湘潭	42.7
6	浙江仁和	13.2	福建閩縣	25.7	江蘇元和	33.2	安徽合肥	38.7	直隸清苑	42.9
7	直隸天津	14.2	河南開封	26.1	安徽涇縣	33.5	廣東番禺	38.9	陝西長安	43.1
8	浙江錢塘	15.1	陝西西安	26.5	雲南昆明	33.8	直隸永平	39.2	河南固始	43.3
9	四川成都	15.9	廣西桂林	26.9	廣東肇慶	34.1	江蘇金匱	39.4	安徽太平	43.5
10	浙江山陰	16.7	浙江紹興	27.3	安徽歙縣	34.4	安徽甯國	39.6	安徽懷甯	43.7
11	廣東廣州	17.4	浙江蕭山	27.8	山西汾州	34.7	江蘇無錫	39.8	江蘇常州	43.9
12	浙江歸安	18.1	福建侯官	28.2	江蘇長洲	35.0	直隸保定	40.1	江蘇蘇州	44.0
13	安徽桐城	18.8	河南祥符	28.6	河南光州	35.3	江蘇常熟	40.3	浙江秀水	44.2
14	江西南昌	19.5	廣西臨桂	29.0	浙江烏程	35.6	江西南城	40.5	山東武定	44.4
15	福建福州	20.1	浙江杭州	29.4	山西太原	35.9	安徽婺源	40.7	廣東香山	44.6
16	江蘇上元	20.8	浙江嘉興	29.8	廣東南海	36.1	山東諸城	40.9	湖北黃州	44.7
17	江蘇陽湖	21.3	湖北武昌	30.1	江蘇吳縣	36.4	浙江上虞	41.1	河南南陽	44.9
18	江蘇武進	21.9	貴州貴筑	30.5	山東萊州	36.7	江西吉安	41.3	江蘇儀徵	45.1
19	順天通州	22.5	江蘇江甯	30.9	雲南臨安	37.0	直隸順天	41.5	江西新城	45.3
20	湖北江夏	23.0	江蘇丹徒	31.2	山東登州	37.2	貴州遵義	41.7	河南衛輝	45.4

Note: Based on 2,615,955 records of officials with surnames in the Jinshenlu with both a province and county of origin recorded. We exclude military officials in the Zhongshubeilan because they rarely had county of origin included.

There were two main reasons that the number of combinations of province and county appearing in the data was larger than the number of counties at any given time, and these require attention during

20 There were too few secondary places of origin included to be of much use in linkage. Of the records that included a province and county of origin, only 13,533 (0.39%) listed an additional place of origin.

linkage. First, the province of origin listed for an official could change between editions even when the county of origin did not.[21] Out of 1,789,985 pairs of records in adjacent editions with identical surname and given name, position, and degree qualification, there were still 0.1% (1941) in which the province changed. This could occur because a provincial boundary was redrawn, but in other cases it was likely the result of a mistake during the production of the edition or the transcription by coders. Several sets of adjacent provinces stood out for the frequency with which one was replaced by the other across two records of the same official with the same county of origin listed: 1) Guangdong and Guangxi, 2) Zhejiang, Jiangsu, Jiangxi, and Anhui, 3) Hubei and Hunan, 4) Shandong and Shanxi, 5) Shuntian and Zhili, and 5) Shaanxi and Gansu.[22]

Second, the characters used to write the name of a county could differ across editions. Out of 1,581,616 pairs of records in adjacent editions that had an identical surname, given name, province of origin, degree qualification, and position recorded, there were 3.6% (57,066) in which the county differed. Almost all of these were situations where a character within a county name was replaced with a variant form of the same character, as happened above with the surnames and characters that were part of given names. For example, the third and 10th most common counties (山陰 and 山陰) in Zhejiang (浙江) are the same county (Shanyin) but the second character of the county's name appears in the original source in two different forms. Similarly, the 22nd and 57th most common counties (吳縣 and 吳縣) in Jiangsu (江蘇) are the same county (Wuxian), but the first character appears in the original in two different forms. Other examples include Qiantang (錢塘 and 錢塘) in Zhejiang and Qingyuan (淸苑 and 淸苑) in Zhili.[23]

The cumulative implication of the discrepancies for surname, given name, and location for nominative linkage across the career of all the records of an official across their career is serious. By combining the discrepancy rates for the primary attributes, we can produce estimates that for two records of the same official in two adjacent editions, at least one of the four primary attributes differs. Assuming independence between the probabilities of each of the four primary attributes differing, we have 1-(1-0.035)(1-0.001)(1-0.0434)(1-0.0128) = 0.0896, or 8.96%. Assuming a typical career length of five years, or 20 quarterly editions, the probability of a discrepancy in at least one pair of records is 83.2% (1-(1-0.0896)[19]). In other words, assuming independence of these probabilities, it is almost certain that for any official whose career lasted for more than a few years that at least one of their records will not match exactly, and in the absence of measures to accommodate discrepancies, the records of many if not most officials with careers of more than just a few years of service will be split incorrectly into two or more officials. Below, we will present tabulations from career histories of officials produced by our linkage to show that such discrepancies were indeed common.

3.1.4 SECONDARY ATTRIBUTES

Secondary attributes help adjudicate in situations where the primary attributes in a pair of records of officials with a surname are close but not an exact match. As we discuss below, they may be useful to confirm a candidate match, but by themselves they are rarely adequate to rule one out because they are not recorded completely, may not be recorded in a consistent fashion, or may change. For example, commercial editions tended to recorded more details that could be used as secondary attributes than

21 The *Zhongshubeilan* editions had additional complications. In the *Zhongshubeilan* rosters of military officials, Huguang (湖廣) appeared as a province of origin in some late 18th century and early 19th century editions. This was a combination of Hunan and Guangdong. We assigned the four counties that were associated with Huguang to Hunan. These were 慈利 (Cili), 祁陽 (Qiyang), 衡陽 (Hengyang), and 道州 (Daozhou). Similarly, in the *Zhongshubeilan* and sometimes in the *Jinshenlu*, counties in Jiangsu, Zhejiang, Anhui and sometimes Jiangxi were listed as being in Jiangnan (江南).

22 When we compared records in adjacent editions that were less than three years apart and which were identical on the surname, given name, county of origin, degree qualification and position, there were 36 cases where an official from Lingui (臨桂) county was listed as being from Guangdong in one record and Guangxi in another, 25 cases where someone was listed with Changping (昌平) as county of origin and were listed as being from Shuntian province in one record and Zhili province in the other, 21 cases where an official from Dantu (丹徒) county was listed as being from Jiangsu in one record and Jiangxi in the other, and 19 cases where an official from Hanyang (漢陽) was listed as being from Hubei in one record and Hunan in the other. Counties that switched between Shuntian and Zhili in more than 10 cases included Baoding (保定), Wuqing (武淸), Ninghe (寧河) and Wanping (宛平).

23 We have made a list of pairs of discordant counties available at the same website as the other tables.

official editions (Chen et al., 2020). Available secondary attributes for officials with surnames include the exam or purchased degree that qualified an official for appointment, the official position, courtesy or style name, and title.

The most important of these are degree qualifications. 84.2% of the records of officials with surnames included the examination or purchased degree that qualified them for appointment (*chushen*, 出身). For some officials who held a *jinshi* or *juren* examination degree, the name of the degree wasn't included in the record, but the year (*ganzhi*,干支) in which they earned their degree was included. Since the provincial and metropolitan exams were the basis of the *juren* and *jinshi* were held in different years, whether an official held a *juren* or *jinshi* could be inferred from the exam year. When *jinshi* or *juren* inferred from exam year are included, 93.2%of records of officials with surnames specified a degree qualification. Hundreds of different degrees were recorded in the original, but for 89.3% of them, the degree fell into one of the following five broad categories: 1) *Jinshi* (進士) degrees for graduates of the Metropolitan Exam, 2) *Juren* (舉人) degrees for graduates of the provincial exam, 3) Regular *gongsheng* (正途貢生) degrees earned by examination, 4) Irregular *gongsheng* (異途貢生) degree acquired by purchase, or 5) Purchased *jiansheng* (監生) degree.[24] Of 1,405,138 pairs of records in adjacent editions that matched on surname, given name, place of origin, and post and which had a degree qualification recorded in the original source, only 7.5% (106,007) changed their degree between two editions. Nearly all these changes were within the broad categories above and represented different ways of writing the same degree. Actual transitions between broad categories were rare.[25]

Official post is useful for confirmation of candidate matches. Relevant information includes an official's job title (*guanzhi*, 官職). For officials in the capital, their ministry and department were recorded. For officials outside the capital, their province, prefecture, and county were recorded. According to our calculations based on record pairs in adjacent editions that were identical on all primary attributes, 7.3% of job titles changed between editions, either because the official changed jobs, or because the title was written differently. If we consider the entire post, including the geographic location or ministry and department, 12.6% changed between editions. Again, this reflected not only actual changes, but inconsistencies across editions in recording. The recorded post had high specificity: for 85% of the records of officials with a surname, the combination of geographic location or ministry and department and job title was unique within the quarterly edition. We have also mapped posts to the numeric bureaucratic ranks used in the civil service (*pinji*, 品级) and then categorized these numeric ranks as high, middle, low, and unranked. Below, this helps us assess whether two records with the same name belong to the same or different officials.[26]

Some other attributes were recorded only for a few officials, but when they were recorded, could be useful for helping to confirm a match. One of these was the official's courtesy name (*biaozi*, 表字) or style name (*hao*, 號). 11.7% of the records of officials with a surname included a courtesy or style name alongside the given name. Whether or not these names were recorded also varied across editions: In 74 of 275 *Jinshenlu* editions, no courtesy or style names were recorded at all. They are also not systematically available in the CGED-Q ER, limiting their usefulness for linkage to that dataset. Titles (*juewei*, 爵位) were recorded consistently, but only 0.5% of civil officials with a surname had one. Year of appointment to the current post and related information could be useful but they are only available for 60.2% of records of officials with a surname in the CGED-Q JSL, and not available at all in the CGED-Q ER. 57 *Jinshenlu* editions do not record year of appointment to the current post.

24 A small number of civil officials in the *Jinshenlu* and many military officials in the *Zhongshubeilan* had military exam (武舉) degrees. A small number of officials were yinsheng (蔭生), that is holders of a hereditary honorary status. See Chen et al. (2020) for a detailed discussion of these degrees, including tabulations and trends over time.

25 When transitions between broad categories did occur, they were upward, occurring when an official passed a higher exam while serving. The most common was from *jiansheng* to *juren*, of which there were 1022 cases. There were 633 transitions from *juren* to *jinshi*.

26 When officials held two or more posts at the same time, they tended to be within the same rank category or in adjacent categories. Similarly, when officials changed post between editions, it was usually between posts in the same or adjacent rank categories. Transitions from high to low or high to unranked were extremely rare.

3.2 ATTRIBUTES AVAILABLE FOR OFFICIALS WITHOUT SURNAMES

3.2.1 GIVEN NAMES

26,727 distinct given names appeared for officials without surnames in the data. In principle, all or almost all of these officials should have been Bannermen, mostly Manchu but in some cases Mongol. 84.1% of the given names consisted of only two characters, 11.2% three characters and less than 1% four or more characters. According to Table 8, the top 100 names accounted for 8.6% of records. This was only slightly higher than the 6.6% accounted for by the top 100 given names of officials with a surname. The main difference is that the distribution of given names of officials without surnames has a shorter tail: separate calculations reveal that the top 200 account for 13%, the top 1,000 account for 36%, and the top 10,000 account for 92%. By contrast, the top 10,000 names accounted for only 64% of the records of officials with surnames. While the smaller number of officials without surnames may have accounted for the overall smaller number of distinct given names, it should not have affected the shape of the distribution.

Table 8 *Cumulative percentages of the top 100 most common given names of officials without surnames in the CGED-Q JSL*

	1–20		21–40		41–60		61–80		81–100	
	Given name	%	Given name	%	Given name	%	Given name	%	Given name	%
1	文光	0.2	錫麟	2.4	恩壽	4.1	恒安	5.7	明安	7.0
2	祥麟	0.3	松林	2.5	德興	4.2	恒昌	5.7	慶昌	7.1
3	玉山	0.4	桂森	2.6	祥安	4.3	慶雲	5.8	崇勳	7.1
4	英俊	0.6	瑞麟	2.7	文海	4.4	玉崑	5.9	文溥	7.2
5	文英	0.7	松齡	2.8	延齡	4.5	奎文	5.9	桂斌	7.3
6	文明	0.8	文治	2.9	吉昌	4.5	恩慶	6.0	恩承	7.3
7	長春	0.9	恩光	3.0	崇福	4.6	祥瑞	6.1	定保	7.4
8	慶安	1.0	鍾秀	3.0	恩榮	4.7	祥泰	6.2	清安	7.4
9	慶福	1.2	榮慶	3.1	玉衡	4.8	榮桂	6.2	長慶	7.5
10	毓秀	1.3	常明	3.2	松壽	4.8	文成	6.3	文斌	7.6
11	奎英	1.4	松秀	3.3	文桂	4.9	文惠	6.4	桂昌	7.6
12	恩霖	1.5	文貴	3.4	榮昌	5.0	雙福	6.4	全福	7.7
13	扎拉芬	1.6	慶恩	3.5	榮恩	5.1	佛爾國春	6.5	英奎	7.7
14	英秀	1.7	榮安	3.6	景福	5.1	德馨	6.6	慶祥	7.8
15	慶麟	1.8	崇禧	3.6	景昌	5.2	春慶	6.6	托克托布	7.9
16	德祿	1.9	文瑞	3.7	吉順	5.3	恩明	6.7	英麟	7.9
17	慶瑞	2.0	興奎	3.8	恩隆	5.4	麟祥	6.7	文敬	8.0
18	崇恩	2.1	文麟	3.9	德麟	5.4	桂芬	6.8	常興	8.0
19	桂林	2.2	文秀	4.0	榮光	5.5	德克精額	6.9	松年	8.1
20	文興	2.3	桂芳	4.1	恩綸	5.6	文俊	6.9	全順	8.2

Note: Based on 811,580 records of officials without surnames in the CGED-Q JSL.

The given names recorded for officials without surnames were transliterations into Chinese of originally Manchu or Mongol names. Bannerman officials had different combinations of characters to choose from for the transliteration of their name. For example, the most common name in term of toneless pronunciation, Qing'an, appeared variously as 慶安, 清安, and 淸安. In the latter two, 清 and 淸 are variants of the same Chinese character. The next most common name in terms of toneless pronunciation, Xilin, appeared as 錫麟, 錫霖, 熙麟 and 西林. These are all different characters. As a result, our tabulations of the romanized names without tones reveals that they were less diverse than names written as Chinese characters. There were 14,560 distinct names if we only consider the pronunciations without tones. The top 100 accounted for 11.8% of records, the top 200 accounted for 19.4% of records, the top 1000 accounted for half of records and the top 10,000 accounted for 99.0% of records.

In the CGED-Q JSL, changes in the transliterations of the same Manchu or Mongol name across different editions appear to have been rare. While officials who had the same Manchu or Mongol name may

have had different transliterations to choose from at the beginning of their career, once they chose one they do not seem to have changed it later. Of 560,559 pairs of records of officials without surnames in editions no more than one year apart that were identical in terms of the toneless Mandarin pronunciation of the name, Banner affiliation, and post, the Chinese characters used to write the name changed in only 2.3% of pairs (13,128). Our further inspection revealed that many of these apparent changes were the result of replacement of one character in the name with a variant form of the same character.

3.2.2 BANNER AFFILIATION

Banner affiliation was stable enough to help confirm candidate links, but there were enough changes to suggest caution against reliance on it to exclude possible links. Every Bannermen were associated with one of eight banners defined by a combination of either Plain or Bordered and one of four colours: Yellow, White, Red, and Blue.[27] When we examined 488,734 pairs of records of Manchu and Mongol Bannermen in adjacent editions with identical names in Chinese characters, identical location or ministry and department, and identical job title, 4.4% (21,634) changed banner. More than one-quarter of these were between Plain and Bordered Banners of the same colours. Most of the changes are among officials with the same three job titles as above: clerk (*bitieshi*, 筆帖式), *yuanwailang* (員外郎) or *zhushi* (主事). At present we are unclear of the process by which officials changed Banners, and we will need to conduct further inquiries with the help of Qing historians.

3.2.3 SECONDARY ATTRIBUTES OF BANNERMEN

The posts recorded for officials without surnames within a quarterly edition were not unique. Table 9 presents the tabulation of the concatenation of job title and administrative unit for officials without surnames. For those serving in the capital, the administrative unit was their ministry and department. For those serving outside the capital, it was the province and possibly prefecture and county where they were assigned. Only 16.7% of job titles (*guanzhi*, 官職) were unique within an edition. More than three-quarter appeared five or more times within an edition. The most common were clerks (*bitieshi*, 筆帖式), *yuanwailang* (員外郎) and *zhushi* (主事). Even when we consider the combination of location or ministry and department and job title, less than one-third of positions were unique. For more than half of positions, there were 5 or more records in the same edition with an identical position. Most of the repeated positions were clerks who were in pools assigned to the central government ministries.

Table 9 *Uniqueness of given names and posts for officials without surnames in the* Jinshenlu

	Job title (Guanzhi, 官職)	**+ Location or Ministry and Department**
Repetitions within edition	%	%
1	16.7	31.1
2	2.5	8.5
3	1.2	4.0
4	1.4	3.8
5	78.2	52.7
Total	100	100
Records	784,502	784,502

Officials without surnames had other details recorded that are potentially useful as secondary attributes, but which are only available for small numbers of records. Those who were members of the main line (*zongshi*, 宗室) or collateral line (*jueluo*, 覺羅) of the Imperial Lineage were recorded as such and accounted for 7.4% of the civil officials who had no surname and 1.7% of civil officials overall. Over the entire course of the Qing and into the Republican era, the Imperial Lineage only had 83,656 male members total, thus its members were heavily overrepresented among officials. One-third (35.8%) of civil officials who were Bannermen had an examination or purchased degree recorded. This tended to be more common later in the 19th century. 11.6% of Bannermen had a courtesy or style name recorded. Year of appointment is only recorded in 7.5% of the records of Bannermen.

27 The upper three were Bordered Yellow (鑲黃旗), Plain Yellow (正藍) and Plain White (正白旗). The lower five were Plain Red (正紅旗), Bordered White (鑲白旗), Bordered Red (鑲紅旗), Plain Blue (正藍旗) and Bordered Blue (鑲藍旗).

4 CHINA GOVERNMENT EMPLOYEE DATASET-QING EXAMINATION RECORDS (CGED-Q ER)

The China Government Employee Dataset-Qing Examination Records (CGED-Q ER) consists of records of examination degree holders transcribed from originally separate lists of exam passers from different sittings of the exam. The most important sources are lists in books self-published by the exam degree holders who had passed at the same sitting of an exam and thought of themselves as classmates. Most of these were titled *Tongnianchilu* (同年齒綠), though some appeared with other titles. Hereafter we refer to them as Classmate Books. Each one listed the surname, given name, and province and county of origin for exam passers at a single sitting along with their current post, if any, and names and degrees held for their father and paternal grandfather and great-grandfather. In most cases they also provide age at passing the exam. They also list other kin, but such information is less systematic. Most of the Classmate Books we have transcribed are for *jinshi* (進士) degree holders who passed the Metropolitan Exam (*Huishi*, 會試) held every three years in the capital and *juren* (舉人) degree holders who passed the Provincial Exam (*Xiangshi*, 鄉試) that qualified them to sit for the Metropolitan Exam. We have also entered similar books for holders of the *Gongsheng* (貢生) degree. For Classmate Books, at present we have entered 5,724 *jinshi* records, 26,870 *juren* records, and 11,990 other records.

We also have less detailed official records of the passers of the Provincial and Metropolitan Exams. For the Provincial Exams, *Xiangshilu* (鄉試錄) rosters record the surname and given name, county of origin, exam rank, and ages of passers of a single sitting. Province of origin is inferred from the location of the exam. Some of these are for sittings of provincial exams for which we also have Classmate Books and are therefore redundant. For all passers of the Metropolitan Exam, the *Jinshi Timinglu* (進士題名錄) lists exam year, surname and given name, province, and county of origin, and ranks in the Metropolitan Exam and the follow-up Court Exam (*Dianshi*, 殿試). For many 19th century sittings of the Metropolitan Exam, we already have Classmate Books that provide more detailed information, thus the *Jinshi Timinglu* is useful mainly for its records of *jinshi* not covered by Classmate Books.

For the CGED-Q ER, we have two linkage tasks. The first is to link records of the same degree holder across Classmate Books and the official records *Xiangshilu* and *Jinshi Timinglu*. This facilitates deduplication of records in situations where we have multiple records of the same exam. This could occur if we have *Xiangshilu* and Classmate Books for the same sitting of a Provincial Exam, or if a sitting of the Provincial Exam is covered by a Classmate Book specific to that sitting and a separately published Classmate Book from the same year that compiles results from multiple provinces. Within the CGED-Q ER, we can also link between the different levels, connecting the records of *juren* to their records as *jinshi*. This allows us to examine how characteristics of a *juren* influenced their chances of going on to earn the *jinshi*. The second task is to link the information about degree holders in the CGED-Q ER to their career records in the CGED-Q JSL. This allows us to examine how the characteristics of degree holders including their family background and their exam performance affected their chances of being appointed subsequently being promoted.

For these linkage tasks we make use of surname, given name, province and county of origin, the year in which the degree was earned, and the type of degree recorded in the CGED-Q JSL and ER. Issues related to the use of surname, given name, and province and county of origin are similar to those in the CGED-Q JSL. The combination of surname, given name, and province and county of origin is almost always unique for degree holders with surnames who earned their degrees at the same time, thus we do not repeat the detailed analysis for officials in the CGED-Q JSL from above. There is also the possibility that across different sources, characters may be replaced by variants. The approach we describe below for dealing with this in the CGED-Q JSL will also work for linkage of exam records. Exam year is useful because it allows us to constrain matching to exclude situations where someone appears to earn the *jinshi* before the *juren*, or else earns it more than a decade after the *juren*.

5 LINKAGE

We carry out linkage in four stages. First, as we describe in 5.1, we prepare for linkage by constructing standardized versions of key attributes. Second, as described in 5.2, we carry out simple deterministic linkage to form groups of records that match exactly on a variety of primary and secondary attributes

and therefore are unambiguously the same official. We then extract the first record in each group to produce the dataset that will be used in the later stages. This substantially reduces the number of records to be considered in the later stages. Third, as described in 5.3, we make use of the capability in the STATA probabilistic linkage package dtalink (Kranker, 2018) to specify attributes to be used for 'blocking', according to which pairs of records are selected for scoring in the probabilistic linkage only if they have an exact match on those attributes. By excluding large numbers of record pairs that are clearly not matches, for example ones in which records differ on both surname and given name, it yields another order of magnitude reduction in the time required for linkage. In the fourth stage (5.4), we carry out probabilistic linkage, again with dtalink. Candidate pairs of records left over after the formation of record groups and application of blocking are scored and then based on these scores, linked together by assignment of a unique identifier to all records that have been associated with a specific official.

5.1 PREPARATION

We prepare the datasets for linkage by producing standardized versions of the primary and secondary attributes. To reduce the chances that inconsistencies in the recording of a given name for the same person across different editions will produce false negatives during linkage, we create transformed versions of the surname and given name. We begin by consolidating the characters in surnames and given names recognized in the Unicode standard as different versions of the same character.[28] Examples in Table 6 include 淸 and 清 (Qing) and 勳 and 勛 (Xun). We refer to these as the CV versions of the names, for 'Consolidated Variants'. We then carry out a second round of consolidation on the CV versions which we group sets of characters in given names that are not recognized as variants in the Unicode standard but look like each other.[29] Examples include the ones mentioned in the discussion of Table 6: 傅 (Fu) and 傳 (Chuan), 思 (Si) and 恩 (En), 增 (Zeng) and 曾 (Ceng), and 先 (Xian) and 光 (Guang). We refer to these as the SC versions, for 'Similar Characters'. At the end of the process, each record contains the given name as originally entered, and fields for the CV and the SC version.

We also produce standardized versions of the surnames. We first consolidate variant forms of characters based on the Unicode standard to produce CV versions. We then consolidate similar looking CV characters to produce SC versions. To do this, we manually reviewed the results of the tabulation that produced Table 3 to identify the most common discordant pairs that were not variant forms that would be accounted for by consolidation on the Unicode standard. As we noted in our discussion of Table 3, there were pairs of characters that were different enough that we concluded that they may have been for different people who were otherwise similar on the attributes we matched on. After excluding these, for the time being we have settled on 12 sets of characters that were especially like to appear in place of each other, and which we thought were similar enough that they could be swapped by mistake between editions, either during the production of the editions, or during transcription by our coders.[30] This is a more conservative approach than we took with the characters in given names because surnames are less diverse than the characters in given names, and accordingly the risk is higher that two people who are the same on other attributes but differ on the their surname really are different people. We may adjust our approach later.

We produce standardized versions of the province and county of origin. We create two versions of the county name romanized by Hanyu pinyin to account for the possibility that characters in the name of a county were replaced with homonyms by mistake. These are listed in Table 10. The first version (PY) includes tone marks, and the second version (PY TL) excludes them. Finally, to address inconsistency in the association of counties with provinces, we create a version of province of origin in which Anhui, Jiangsu, Jiangxi and Zhejiang are all combined into Jiangnan, and Hunan, Guangdong, and Guangxi

28 This includes converting characters mistakenly typed in simplified form into traditional form. See https://unicode.org/reports/tr38/ for a report on the latest version of the Unicode Han Database. We downloaded the Unicode database for Han Chinese characters from https://www.unicode.org/Public/UCD/latest/ucd/Unihan.zip

29 We did this by carrying out a tabulation like the one that produced Table 6 but which only used the CV versions of the characters to produce a list of pairs of characters that are commonly swapped. We manually assessed each of the resulting pairs to flag those that were visually similar enough that it is plausible that they could be switched. We use the resulting pairs to map sets of similar characters to a single character.

30 These were 1) 宋, 朱,宗, 2) 段, 段, 3) 王, 汪,江, 4) 馬, 馮, 馮, 5) 柳, 栁, 6) 季, 李, 7) 龍, 龔, 8) 余, 徐, 涂, 9) 湛, 諶, 10) 寇, 寇, 11) 樂, 欒, and 12) 褚, 諸.

are all combined into Huguang. We refer to this as the C version of province. In the very small number of records in which a second province and county of origin were listed in the original source, we used that instead of the first listed province and county of origin.

5.2 DETERMINISTIC LINKAGE

We group records that match exactly on a large number of primary and secondary attributes and are in editions less than one year apart and create an extract of the data that only includes the first record in each of these groups. We make the criteria for inclusion of a record in one of these groups so exacting as to rule out false positives in which records of different officials are accidentally linked.[31] The creation of these record groups by deterministic linkage is straightforward and we do not discuss it further. Because the number of record groups that need to be linked is an order of magnitude less than the original number of records to be linked, the time required for the second and third stages is substantially reduced.

5.3 BLOCKING

We divide blocking for the CGED-Q JSL and CGED-Q ER linkage into six types based on the attributes available in the records involved and the risks of false positives or negatives. For linkage within the CGED-Q JSL to produce career histories, we distinguish three types: 1) officials with a surname who had a single character given name, 2) officials with a surname who had two character given names, and 3) officials without surnames. We link officials with surnames and one-character given names separately because comparison of Tables 4 and 5 suggests that the risk of a false positive is higher, compared with the ones with a two-character given name. This requires stricter criteria for matching on other attributes. Because the combination of surname and two-character given name is more likely to unique, for linkage of officials with given names who had two-character given names we can be more forgiving for other attributes. Officials without surnames have only the given name and Banner affiliation as primary attributes, which combination is less likely to be unique, thus we must put more weight on secondary attributes. Linkage within the CGED-Q ER forms the fourth type. Here, we treat all the records the same. The total number of records is small enough that false positives for degree-holders with a surname and only a one-character given name are unlikely. For linkage between the CGED-Q JSL and CGED-Q ER, we distinguish the fifth and sixth types according to whether men with surnames have one- or two-character given names.

Table 10 summarizes the attributes used for blocking for each of the six types of linkage. In each case, we balance the risk of false negatives associated with use of overly strict criteria against the increased linkage time associated with the use of loose criteria. In general, we make the blocking criteria as loose as possible while seeking to prevent clearly impossible pairs through to be scored. Thus, for example, we typically block on SC versions of names rather than CV versions of names, and then use scoring on other attributes to assess pairs that match on the SC but not CV versions. For blocking within the CGED-Q JSL, we apply different criteria for each linkage type. For the first type, officials with surnames who had two-character names, we block on the SC and pinyin versions of the surname and given name. That is, if two records have the same SC or pinyin version of the surname and given name, they are a candidate match and go on to be scored on the other attributes, including the CV versions of the names. We do not use the CV version of the names for blocking because it would be too strict, and would preclude making matches based on the looser criteria associated with use of the SC versions. For the second type, officials with surnames and only one-character given names, we only allow pairs of records with the same SC versions of the names. Our experiments with allowing for matches on the pinyin version of the surname and given name yielded too many false positives. For the third type, officials with no surname, we block on the SC version of the given name and Banner affiliation, or on the combination of the pinyin version of the name, the Banner affiliation, Imperial Lineage affiliation, title, and complete post. In other words, a pair in which the SC version of the name doesn't match but the pinyin version does match can still be treated as a candidate pair and scored if there is an exact match on a variety of other characteristics. We allow candidate pairs that match on the pinyin name

31 For officials with surnames, records in a group must have the same CV version of the surname and given name, the same C version of the province of origin, and the same pinyin for the county of origin. For Bannermen, records in a group must have the same CV version of the given name, the same Banner affiliation, and the same government post, which is the concatenation of the administrative unit and job title. We require records in Bannermen sets to match on post as well because as Table 1 showed, the combination of given name and Banner affiliation was not unique within an edition.

only when several additional secondary attributes also match because allowing candidate pairs based on pinyin given name alone would substantially expand the number of pairs to be considered. For the fourth type, linkage within the CGED-ER, the SC versions of the surname and given name are sufficient for blocking. Rather than have a separate approach to blocking in the CGED-Q ER for men with a surname and a single character given name, as we describe below, we apply tighter criteria for scoring candidate pairs involving such records.[32] For the fifth type, linkage between the CGED-Q ER and CGED-Q JSL of men with a surname and a two-character given name, we allow for candidates pairs that match on the SC or toneless pinyin versions of the surname and name.[33] For the sixth type, linkage between the CGED-Q ER and CGED-Q JSL of men with single-character given names between, we only allow candidate pairs that match on the SC versions of the names.

Table 10 *Attributes used for blocking in linkage of the CGED-Q JSL and CGED-Q ER*

Linkage Type		Blocking
Within the CGED-Q JSL		
1	Officials with surnames and two-character given names	Surname SC + Given name SC OR Surname PY+ Given name PY
2	Officials with surnames and one-character given names	Surname SC+ Given name SC
3	Officials without surnames	Given name SC + Banner affiliation OR Given name PY + Banner affiliation + Imperial Lineage status + Noble title + Post
Within the CGED-Q ER		
4	All records	Surname SC + Given name SC
Between the CGED-Q ER and CGED-Q JSL		
5	Men with surnames and two-character given names, and men without surnames	Surname SC + Given name SC OR Surname PY NT + Given name PY NT
6	Men with surnames and one-character given names	Surname SC + Given name SC

5.4 PROBABILISTIC LINKAGE

Since probabilistic linkage is already widely used and described in detail elsewhere, here we only provide a summary of the basic concept. Probabilistic matching considers every possible pair of records in a dataset left over after blocking and then scores each pair for similarity according to criteria specified by the user. For the scoring, the user specifies the attributes to compare, and the amount to be added to or subtracted from the score if they match or differ. Calipers may also be specified according to which some amount may be added to the score for a pair if two numeric attributes are within some range of each other, and some other amount may be deducted if they are not. A match is made by comparing the scores of candidate pairs and selecting the ones with the highest score that also meet a cutoff score set by the user.

We scored the candidate pairs of record groups left over after blocking according to their concordance or discordance on specified primary and secondary attributes. Tables 11 and 12 summarize our current rewards and penalties for concordance or discordance on each primary or secondary attribute for our six types of linkage. The rewards ("+" in the tables) are added to the score for a candidate pair if the condition specified in the row heading is satisfied. The penalties ("-" in the tables) are subtracted from the score if the condition is not satisfied. Tables 11 and 12 also include the cutoffs that a score had to be greater than or equal to in order for a match to be made. For each of the six linkage tasks, we choose the amounts to be added to or subtracted for a match or mismatch on a specified attribute to balance the risks of false negatives and false positives. We apply more stringent criteria when there are larger numbers of records to be linked, most notably within the CGED-Q JSL, and therefore a higher chance that separate individuals will have the same primary attributes. We apply looser criteria when the chances of a false positive are lower, usually because there are fewer records to be linked. Linkage within the CGED-Q ER is one example.

32 We do not have separate blocking for Bannermen when linking between the CGED-Q JSL and CGED-ER because there are too few of them (1.2% of records overall) in the CGED-Q ER to warrant special handling.

33 We include Bannermen with officials with surnames because only a small proportion (1.23%) of exam degree holders in the Classmate Books we have coded were Bannermen, and the chances of different individuals having the same name were small.

Table 11 *Rewards and penalties for concordance or discordance on attributes in candidate pairs for linkage within the CGED- Q JSL*

	CGED-Q JSL					
	Type 1 Surname and two-character given name		**Type 2 Surname and one-character given name**		**Type 3 No surname**	
	+	-	+	-	+	-
Primary attribute						
Surname (CV) + Given name (CV) + County (PY)						
Surname (SC) + Given name (SC) + County (SC)						
Surname (SC) + Given name (SC) + Province (C)						
Surname (CV) + Given name (CV)	100	0	100	0		
Surname (CV) + Given name (SC)						
Surname (SC) + Given name (PY)						
Given name (CV)					50	0
Given name (SC)					50	0
Province (C)	100	-400	100	-400		
County 1 (Original)						
County 1 (PY)	200	-100	200	-100		
County 1 (PY) in Jiangnan, Huguang	0	-200	0	-200		
Banner affiliation					50	-100
Secondary attribute						
Courtesy or Style name	300	0	300	0	200	0
Imperial Lineage status					100	0
Title					100	0
Post						
Province	25	0	25	0		
Ministry, Agency, or Prefecture	25	0	25	0		
Department or County	25	0	25	0		
Job title	25	0	25	0		
Complete	100	0	100	0		
Rank (Pinji) category						
Same					0	-25
Differ by less than 2					0	-50
Differ by less than 3					0	-400
Degree						
Original	50	0	50	0	50	0
Broad category	0	-100	0	-100	0	-100
Broad category in Jiangnan or Huguang	0	-100	0	-100		
Year						
Same					50	0
< 5 years apart					0	-50
< 10 years apart		-100		-100	0	-100
< 20 years apart		-200		-200		
< 30 years apart						
< 40 years apart		-500		-500	0	-400
Surname and Given name of record above	50	0	50	0	50	0
Surname and Given name of record below	50	0	50	0	50	0
Cutoff		100		100		150

Table 12 *Rewards and penalties for concordance or discordance on attributes in candidate pairs for linkage within the CGED- Q ER and between the CGED-Q ER and CGED-Q JSL*

	CGED-Q ER Type 4 All		CGED-Q JSL to CGED-Q ER Type 5 Surname and two- character given name, OR no surname		Type 6 Surname and one- character given name	
	+	-	+	-	+	-
Primary attribute						
Surname (CV) + Given name (CV) + County (PY)			500	0	500	0
Surname (SC) + Given name (SC) + County (SC)	300	0	200	0	200	0
Surname (SC) + Given name (SC) + Province (C)	100	0	200	0		
Surname (CV) + Given name (CV)			150	0	200	0
Surname (CV) + Given name (SC)			100	0	150	0
Surname (SC) + Given name (PY)			50	0		
Given name (CV)						
Given name (SC)						
Province (C)		-200	0	-200		
County 1 (Original)			100	0		
County 1 (PY)	0	-200	100	0		
Year						
Same						
< 5 years apart						
< 10 years apart	0	-100	100	0	100	0
< 20 years apart	0	-300				
< 30 years apart						
< 40 years apart			0	-500	0	-500
Cutoff		100		200		200

We arrived at the rewards, penalties, and cutoffs in Tables 11 and 12 iteratively. We inspected the results every time we ran the linkage. We located false negatives by searching the data for groups of records that matched exactly on secondary attributes such as position and degree and most but not all of the primary attributes, and which were not associated with a single official. We examined these groups to assess whether the records in the group should all have been assigned to the same official. This helped clarify how often characters were replaced with ones that looked similar and inspired our effort not only to create the CV and SC versions of names. It led us to increase the rewards for exact matches on such secondary attributes as courtesy name and complete post that were highly unlikely to match by chance. It also led to our discovery of inconsistencies in the recording of province.

We searched for false positives by identifying groups of records that had all been assigned to the same official, but which differed on at least one primary attribute, for example, surname, or one character in a two-character given name. This led to our realization that we needed to apply more stringent criteria for individuals with single character given names and led us also to increase penalties for mismatches on attributes such as province of origin or broad category of purchased or examination degree that should be stable. Users working with extracts of the data to study topics of their own, most commonly the appointment and promotion of specific categories of officials, also reported problems that they noticed, and our investigations revealed.[34]

34 For example, the analyses that underpinned Chen et al. (2018), Hu et al. (2020), Hu, Hu, Chen and Campbell (2021) and Xue and Campbell (2022) all led to discovery of problems that were addressed by refinements to linkage procedures.

For linkage within the CGED-Q JSL (Types 1 through 3), we assigned the largest rewards to concordance on attributes like given name, post, or county that are the most diverse and therefore the least likely to match by chance. Even though blocking differed for one- and two-character names, scoring was the same. We gave large rewards to matches on secondary attributes like courtesy or style name and complete post. This helped counter the effects of inconsistencies in the recording of province and county of origin that were not addressed by the transformations described above. Since posts were listed in the same order from one edition to the next, we also rewarded concordance on the name of the official in the record above or below. Rewards are smaller for concordance on attributes like province or broad category of examination or purchase degree that are less diverse and more likely to match by chance.

We apply the largest penalties for discordance on attributes like province or county of origin that should have been stable and were less diverse. A mismatch on a less diverse attribute like the C version of the province, Banner affiliation, or broad category of degree qualification will lead to a large penalty. We apply a penalty for a mismatch on county, with an additional penalty if the province in which the counties are located are part of Huguang or Jiangnan.[35] We also penalize matches of records that are further apart in time, and in the case of records so far apart that it is implausible for them to be the same person, we apply a penalty so large that it will preclude a match from being made. For officials without surnames, we also penalize candidate matches if the categories of bureaucratic rank (*pinji*, 品级) are too far apart. This helps reduce the chances that a record of a high official will be linked to those of another officials with the same given name who is a low-ranking clerk. We apply a smaller penalty for mismatches on attributes that are more prone to inconsistent recording, like detailed examination or purchase degree. Courtesy and style names were diverse, often missing, and sometimes seem to have changed, thus we do not apply a penalty for a mismatch on them. Similarly, because complete positions and the components that made up the position were expected to change when an official was promoted or reassigned, and because different editions could record positions differently even when the official was not promoted or reassigned, we do not apply a penalty for a mismatch on position.

For linkage within the CGED-Q ER (Type 4), we began with surname, name, province and county of origin, and exam year. We created CV and then SC versions of the surname and name. We blocked on the SC version of the surname and name. We rewarded matches on the combination of SC surname, SC name, and county or province, and heavily penalized discordance on the pinyin (PY) version of the county or C version of the province. We used the SC version of the name rather than the CV version because the overall number of men to be linked was much smaller than in the CGED-Q JSL and the risk of a false positive accordingly smaller. We applied only a mild penalty for a gap between exam years because we wanted to allow for links between records of *juren* and *jinshi* degrees earned in different years but applied a much larger penalty if the exam years were so far apart that the rules would not have allowed a *juren* to sit for the Metropolitan exam in the specified year.

For linkage between the CGED-Q JSL and CGED-Q ER (Types 5 and 6), we relied on surname, given name, province and county of origin, CGED-Q ER exam year, and CGED-Q JSL edition year. We blocked on the SC version of the surname and given name. We gave very large rewards for matches on the CV version of the surname and name and smaller rewards for matches on the SC versions. We allowed for matches not only on the province and county of origin in the CGED-Q JSL, but also on the province and county of origin (籍貫) listed in the CGED-Q JSL for officials who sat for the exam someplace other than their actual place of origin, usually Shuntian. We allowed up to 30 years for the time between earning a degree and being appointed for the first time.

5.5 RESULTS

To illustrate how the approach describe above reduces false positives and false negatives, while also reducing the amount of time required, we present the results of linkage within the CGED-Q JSL, that is Types 1, 2 and 3. We focus on linkage with the CGED-Q JSL because it was the most challenging and complex, and made use not only of primary attributes, but a wide range of secondary attributes. Linkage of the 4,108,586 records in the CGED-Q JSL with a name and other information required by the approach described in the sections above yielded 326,315 sets of linked records, each a career history of a single official. For each of the three types of linkage within the CGED-Q JSL, Table

35 Because place of origin could change because of the redefinition of administrative units, we set the penalty for mismatch on county so that it can still be offset by rewards for matches on other attributes. A mismatch on county, in other words, doesn't preclude a match if other attributes are in correspondence.

13 presents the original number of records to be linked, the number of groups remaining after the deterministic linkage described in 5.2, the number of candidate pairs left after the blocking described in 5.3, and the final number of officials produced by the probabilistic linkage described in 5.4. According to Table 13, grouping records with deterministic linkage on the primary and some secondary attributes substantially reduces the number of items to be linked. In the case of Type 1 linkage, the number of items to be linked is reduced by 88.6% percent, from 2,676,108 to 315,015. The resulting number of candidate pairs to be scored is modest. For Type 1 linkage, the number of candidate pairs is lower than the number of groups because many groups are isolates: blocking left them without any other groups to be paired with and scored, and they went straight to being recognized as an official. The number of candidate pairs for Type 3 linkage is much larger because only the given name and Banner affiliation are available for blocking, and these are less diverse than the surname, given name and province and county of origin of officials who have surnames.

Table 13 *Results for Type 1, 2, and 3 linkage, CGED-Q JSL dataset*

	Type 1 Surname and two-character given name	**Type 2 Surname and one-character given name**	**Type 3 No surname**
Records for linkage	2,767,108	527,570	813,908
Number of groups after deterministic linkage	315,015	76,885	171,449
Candidate pairs after blocking	199,263	46,231	398,353
Career histories after linkage	218,946	45,965	64,940

Probabilistic linkage on standardized primary attributes that compensates for discrepancies when there are matches on secondary attributes reduces the number of false negatives. Had we required exact matches on the primary attributes as originally recorded, each distinct combination within one of the histories produced by our linkage would have been associated with a separate official. Table 14 tabulates the career histories according to the numbers of distinct combinations of surname, name and province and county of origin or Banner affiliation within them in the original data. In 28% (100-28) of the career histories of officials with a one-character given name, more than one surname, given name, or place of origin appeared. The corresponding figure for officials with two-character given names was 29.9% (100-70.1). In the career histories of officials of without surnames produced by linkage, 13.9% (100-86.1) had more than one name or Banner affiliation appeared. According to our calculations, linkage by requiring exact matching on the original primary attributes and not using probabilistic linkage with the standardized versions of the names consolidated CV or SC versions of names to allow for discrepancies would have led to the creation of 453,375 career histories. Career histories that in our probabilistic linkage were attributed to a single official would have been separated. The total number of career histories, in other words, would have been inflated by 38%. The gains associated with applying probabilistic linkage within the CGED-Q ER and between the CGED-Q JSL and CGED-Q ER are similar: the number of *juren* degree holders who are linked to *jinshi* records increases substantially, as do the numbers of *juren* and *jinshi* linked to the CGED-Q JSL.

Table 14 *Combinations of surname, given name, and province and county of origin or banner affiliation in original data within sets of records for officials produced by linkage in the CGED-Q JSL*

Distinct combinations of primary attributes observed within career histories produced by linkage	**Type 1 Surname and two-character name** %	**Type 2 Surnames and one-character name** %	**Type 3 No surname** %	**Total** %
1	70.1	72.0	86.1	73.5
2	20.9	20.0	11.7	19.0
3	5.7	5.0	1.7	4.8
4	2.1	1.8	0.4	1.7
5	1.3	1.4	0.1	1.1
Total	100	100	100	100
Number of officials	218,946	45,965	64,940	329,851

6 CONCLUSIONS

This is unlikely to be the final word, especially for the linkage of officials without surnames. Based on manual examination of the resulting data we are confident that our linkage of officials with surnames is close to optimal in terms of its balance between avoiding false positives and false negatives, and that any further accommodation of additional discrepancies we have noticed would open the door to false positives in which the records of clearly different officials would be combined. Any further adjustments to the linkage of officials with surnames are likely to consist of small refinements to the lists of similar characters, and adjustments to the handling of problems with provinces and counties. For Bannermen, however, we suspect that the lack of diversity in the combination of names and Banner affiliation means that we still have too many false positives.

Our experiences, and our descriptive results about patterns in names, should be useful to other teams that are carrying out large-scale record linkage in datasets constructed from historical Chinese sources. The issues we discuss here and our approach to linkage are most relevant for the linkage of highly structured data transcribed from rosters and related records, the descriptive results on the consistency and potential for overlap in the recording of names may be of interest to those conducting disambiguation in unstructured data like newspaper articles. Particular attention needs to be paid to the possibility that across difference sources, the characters in the names of individuals to be linked may be replaced with variant forms of the same character, or entirely different characters that are superficially similar.

We now have ongoing projects to construct, link, and analyze datasets of individuals during the Republican era (1911–1949). Our efforts to create datasets from university student records are the furthest along (Ren et al., 2020), but we have other projects to create datasets of Republican officials, professionals, and other elites. While we expect some of our experiences with Qing records to be relevant, we also anticipate that there will be other issues specific to the Republican data. Naming patterns may have changed. Consistency in the usage of genealogical given names as opposed to courtesy or style names may have changed as well. Customs for the recording of place of origin may also have evolved.

ACKNOWLEDGMENTS

This research was supported by Hong Kong Research Grants Council General Research Fund 16602621 (Campbell PI). We are grateful to members of the Lee-Campbell Group, especially Hao Dong, Lawrence Zhang, James Lee, and Matthew Noellert, for their feedback and suggestions. We are also grateful to Loretta Kim for sharing her knowledge of Manchu naming practices. We are also grateful to Xue Qin, Chen Jun, and other users of the data who brought issues they discovered to our attention, leading directly or indirectly to adjustments in our linkage procedures.

REFERENCES

Abramitzky, R., Mill, R., & Pérez, S. (2020). Linking individuals across historical sources: A fully automated approach. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 53*(2), 94–111. doi: 10.1080/01615440.2018.1543034

Akgün, Ö., Dearle, A., Kirby, G., Garrett, E., Dalton, T., Christen, P., Dibben C., & Williamson, L. (2020). Linking Scottish vital event records using family groups. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 53*(2), 130–146. doi: 10.1080/01615440.2019.1571466

Armand, C., Guo, W., Henriot, C., Hu, Y., & Van den Bosch, N. (2022). *Modern China Biographical Database (MCBD). User manual*. Aix en Provence: ENP-China, Aix-Marseille University. Retrieved from https://bookdown.enpchina.eu/mcbd_usermanual/

Bao, H., Cai, H., Jing, Y., & Wang, J. (2021). Novel evidence for the increasing prevalence of unique names in China: A reply to Ogihara. *Frontiers in Psychology, 12*(731244), 1–6. doi: 10.3389/fpsyg.2021.731244

Cai, H., Xi, Z., Yi, F., Liu, Y. & Jing, Y. (2018). Increasing need for uniqueness in contemporary China: Empirical evidence. *Frontiers in Psychology, 9*(554), 1–7. doi: 10.3389/fpsyg.2018.00554

Campbell, C. D. (2020). Qingmo keju tingfei dui shiren wenguan qunti de yingxiang-jiyu weiguan dashuju de hongguan xin shijiao [The influence of the abolition of the examinations at the end of the Qing on the holders of exam degrees]. *Shehui kexue jikan [Social Science Journal], 4*(249), 156–166.

Campbell, C. D., Chen, B., Ren, Y., & Lee, J. Z. (2019). China Government Employee Database-Qing (CGED-Q) Jinshenlu public release [Database]. DataSpace@HKUST, V14. doi: 10.14711/dataset/E9GKRS

Campbell, C. D., & Lee, J. Z. (2020). Historical Chinese microdata. 40 years of dataset construction by the Lee-Campbell research group. *Historical Life Course Studies, 9*, 130–157. doi: 10.51964/hlcs9303

Campbell, C. D., Lee, J. Z., & Elliott, M. (2002). Identity construction and reconstruction: Naming and Manchu ethnicity in Northeast China, 1749–1909. *Historical Methods, 35*(3),101–116. doi: 10.1080/01615440209601201

Chen, B. (2019). *Origins and career patterns of the Qing government officials (1850–1912): Evidence from the China Government Employee Dataset-Qing (CGED-Q)* (PhD dissertation). Hong Kong University of Science and Technology Division of Social Science, China.

Chen, B., & Campbell, C. D. (2023). Cong yizhong dao duozhong shiliao: Lijie Qingdai guanyuan shitu de xin fangfa [From one source to many sources: New methods for understanding the careers of Qing officials]. *Shixue Yuekan [History Monthly]*. Forthcoming publication.

Chen, B., Campbell, C. D., & Lee, J. Z. (2018). Qingmo xinzheng qianhou qiren yu zongshi guanyuan de guanzhi bianhua chutan-yi jinshenlu shujuku wei cailiao de fenxi [The transition of banner and imperial lineage officials during the late Qing reform period: Evidence from the Qing Jinshenlu Database]. *Qingshi Yanjiu [Studies in Qing History], 2018*(4), 10–20. Retrieved from http://qsyj.iqh.net.cn/CN/abstract/abstract2384.shtml

Chen, B., Campbell, C. D., Ren, Y., & Lee, J. Z. (2020). Big data for the study of Qing officialdom: The China Government Employee Database-Qing (CGED-Q). *The Journal of Chinese History, 4*(2), 431–460. doi: 10.1017/jch.2020.15

Chen, M., Du, Q., Shao, Y., & Long, H. (2018). Jiyu yinxingma de hanzi xiangsidu bidui suanfa [Chinese characters similarity comparison algorithm based on phonetic code and shape code]. *Xinxi Jishu [Information Technology], 11*, 73–75. doi: 10.13274/j.cnki.hdzj.2018.11.016

Chen, S., & Wang, H. (2022). China Biographical Database (CBDB): A relational database for prosopographical research of pre-modern China. *Journal of Open Humanities Data, 8*(4). doi: 10.5334/johd.68

Chen, Y., & Huang, C. (2010). Exploring personal name disambiguation from name understanding. *2010 4th International Universal Communication Symposium*, 345–349, doi: 10.1109/IUCS.2010.5666185

Chua, I. (2021). What can we tell from the evolution of Han Chinese names? *Kontinentalist.* Retrieved from https://kontinentalist.com/stories/a-cultural-history-of-han-chinese-names-for-girls-and-boys-in-china

Elliott, M. C., Campbell, C. D., & Lee, J. Z. (2016). A demographic estimate of the population of the Qing banners. *Études Chinoises, 35*(1), 9–40.

Fan, C., & Li, Y. (2021). Chinese personal name disambiguation based on clustering. *Wireless Communications and Mobile Computing, 3790176*. doi: 10.1155/2021/3790176

Fuller, M. A. (2021). *The China Biographical Database user's guide. Revised version 3.3.* Retrieved from https://projects.iq.harvard.edu/cbdb/supporting-documents

Han, H., Yao, C., Fu, Y., Yu, Y., Zhang, Y., & Xu, S. (2017). Semantic fingerprints-based author name disambiguation in Chinese documents. *Scientometrics, 111*, 1879–1896. doi: 10.1007/s11192-017-2338-6

Han, W., Xu, X., & Zhao, T. (2011). Study on Chinese person name disambiguation based on multi-stage strategy. *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, 1177–1181. doi: 10.1109/FSKD.2011.6019646

Helgertz, J., Price, J., Wellington, J., Thompson, K. J., Ruggles, S., & Fitch, C. A. (2022). A new strategy for linking U.S. historical censuses: A case study for the IPUMS multigenerational longitudinal panel. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 55*(1), 12–29. doi: 10.1080/01615440.2021.1985027

Hu, C., Hu, H., Chen, B., & Campbell, C. D. (2021). Qingdai zhou de zhengqu fendeng yu zhizhou xuanren de lianghua fenxi [Quantitative analysis on the local government administrative categorization system and the appointment of department prefects during the Qing]. *Shuzi Renwen Yanjiu [Digital Humanities Research], 1*(1), 34–47.

Hu, H., Chen, C., & Campbell, C. D. (2020). Qingdai zhifu xuanren de kongjian yu lianghua fenxi-yi zhengqu fendeng, <jinshenlu> shujuku wei zhongxin [The appointment of prefects during the Qing: A spatial and quantitative analysis focusing on the system of administrative division and using the CGED-Q]. Xinya Xuebao *[New Asia Journal], 37*, 339–398.

Kim J., Kim, J., & Kim, J. (2021). Effect of Chinese characters on machine learning for Chinese author name disambiguation: A counterfactual evaluation. *Journal of Information Science, OnlineFirst.* doi: 10.1177%2F01655515211018171

Kranker, K. (2018). DTALINK: Stata module to implement probabilistic record linkage. Statistical Software Components S458504, Boston College Department of Economics, revised 16 Feb. 2019. Retrieved from https://ideas.repec.org/c/boc/bocode/s458504.html

Lee, J. Z., & Campbell, C. D. (1997). *Fate and fortune in rural China. Social organization and population behaviour in Liaoning, 1774–1873*. Cambridge: Cambridge University Press.

Lee, J. Z., Campbell, C. D., & Chen, S. (2010). *China Multi-Generational Panel Dataset, Liaoning (CMGPD-LN) 1749–1909. User guide.* Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Liu, M., Rus, V., Liao, Q., & Liu, L. (2017). Encoding and ranking similar Chinese characters. *Journal of Information Science and Engineering, 33*(5), 1195–1211. doi: 10.6688%2fJISE.2017.33.5.6

Ren, B., Chen, L., & Lee, J. Z. (2020). Meritocracy and the making of the Chinese academe, 1912–1952. *The China Quarterly, 244*, 942–968. doi: 10.1017/S0305741020001289

Ren, Y., Chen, B., Hao, X., Campbell, C. D., & Lee, J. Z. (2016). Qingdai jinshenlu lianghua shujuku yu guangliao qunti yanjiu [The Qing Jinshenlu database: A new source for the study of Qing officials]. *Qingshi Yanjiu [Qing History Research], 2016*(4), 61–77.

Ren, Y., Chen, B., Hao, X., Campbell, C. D., & Lee, J. Z. (2019). *Zhongguo lishi guanyuan lianghua shujuku-qingdai jinshenlu (1900–1912). Shiduan gongkaiban yonghu zhinan. [The China Government Employee Database-Qing (CGED-Q) Jinshenlu (JSL) 1900–1912. Public release user guide]*. doi: 10.14711/dataset/E9GKRS

Ruggles, S. (2002). Linking historical censuses: A new approach. *History and Computing, 14*(1–2), 213–224. doi: 10.3366/hac.2002.14.1-2.213

Stone, L. (1971). Prosopography. *Daedalus, 100*, 46–79.

Sylvester, K., & Hacker, J. D. (2020). Introduction to special issues on historical record linking. *Historical Methods: A Journal of Quantitative and Interdisciplinary History, 53*(2), 77–79. doi: 10.1080/01615440.2020.1707445

Tsui, L. H., & Wang, H. (2020). Harvesting big biographical data for Chinese history: The China Biographical Database (CBDB). *Journal of Chinese History, 4(*2), 505–511. doi: 10.1017/jch.2020.21

Wang, H., Chen, S., Dong, H., Noellert, M., Campbell, C. D., & Lee, J. Z. (2013). *China multi-generational panel dataset, Shuangcheng (CMGPD-SC) 1866–1914. User guide.* Ann Arbor, MI: Inter-university Consortium for Political and Social Research. doi: 10.3886/ICPSR35292.v9

Wang, Y., Liang, H., Shu, X., Wang, J., Xu, K., Deng, Z., Campbell, C. D., Chen, B., Wu, Y., & Qu, H. (2021). Interactive visual exploration of longitudinal historical career mobility data. *IEEE Transactions on Visualization and Computer Graphics, 28*(10), 3441–3455. doi: 10.1109/TVCG.2021.3067200

Xu, S., Zheng, M. & Li, X. (2020). String comparators for Chinese-characters-based record linkages. *IEEE Access, 9*, 3735–3743. doi: 10.1109/ACCESS.2020.3047927

Xue, Q., & Campbell, C. D. (2022). Qingji gaige shiyu xia libu guanyuan qunti de renshi dishan yu jiegou bianqian (1898–1911) — Yi "jinshenlu" shujuku wei zhongxin [Change and constancy: The personnel evolution and structural change of the ministry of personnel during the reform in Qing dynasty — Based on China Government Employee Database-Qing (CGED-Q)]. *Shehui Kexue Yanjiu [Social Science Research], 2(259),173–182.*

Yin, D., Motohashi, K., & Dang, J. (2020). Large-scale name disambiguation of Chinese patent inventors (1985–2016). *Scientometrics, 122*, 765–790. doi: 10.1007/s11192-019-03310-w