From Textual to Historical Networks: Social Relations in the Bio-graphical Dictionary of Republican China

Cécile Armand, Christian Henriot

JHNR

ARMAND, CÉCILE
HENRIOT, CHRISTIAN

From Textual to Historical Networks: Social Relations in the Biographical Dictionary of Republican China

Journal of Historical Network Research x (202x) xx-xx

Keywords

Biography, China, Cooccurrence; elites, NLP

C²DH
LUXEMBOURG CENTRE FOR
CONTEMPORARY AND DIGITAL HISTORY

uni.lu
UNIVERSITÉ DU
LUXEMBOURG

Abstract

In this paper, we combine natural language processing (NLP) techniques and network analysis to do a systematic mapping of the individuals mentioned in the *Biographical Dictionary of Republican China*, in order to make its underlying structure explicit. We depart from previous studies in the distinction we make between the subject of a biography (bionode) and the individuals mentioned in a biography (object-node). We examine whether the bionodes form sociocentric networks based on shared attributes (provincial origin, education, etc.). Our major contribution consists in annotating the links between individuals in order to (1) question the assumption that word cooccurrences equate with actual relations; (2) define a more accurate classification of relationships among elites in republican China. We demonstrate that political and professional relations in this population outweigh the types of social ties commonly accepted in the scholarship on modern China. We eventually develop a method that can be applied to similar corpora in a critical and comparative perspective.

1 Introduction*

A biographical dictionary is by definition a work centered on individuals whose lives provide the backbone of distinct biographical narratives. The amount and scope of information in such works fall short of the breadth and depth of full biographical works in which the life of an individual is minutely described and usually closely intermeshed with the social, political, economic events of the times. Even with the best of efforts and intention to offer a macro-reading of historical events — this was an explicit goal of the editors of the *Biographical Dictionary of Republican China* (BDRC) — the format of more or less short biographical notes by necessity curtailed this ambition.[1] This holds especially true for the social relations and contacts that an individual had in the course of

* **Acknowledgements**: We would like to acknowledge the contribution of Pierre Magistry (Aix-Marseille University) and Baptiste Blouin (Aix-Marseille University) who prepared the data used in this paper, advised on the methods, and organized the annotation workflow. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 788476).

Corresponding author: Cécile Armand, Aix-Marseille University, IrAsia, cecile.armand@gmail.com

[1] Howard L. Boorman and Richard C Howard, *Biographical Dictionary of Republican China* (New York: Columbia University Press, 1967), I, vii.

his/her life. In the condensed biographical notes that make up a dictionary, all the related historical actors are reduced to brief and often unique mentions in the body of the text. Moreover, due to the involvement of many contributors — vs. a single author in a biographical work — such mentions are unsystematic with no apparent rationale as to the selection of the people included beyond the subjects of the biography.

In this paper, we propose a systematic mapping of all the individuals whose names appear in the biographical notes in order to make the networks underlying the BDRC explicit. We argue that the links between individuals in the biographical texts create an interlinked reference network of the biographical texts.[2] This network can in turn be used to examine to what degree the cooccurrence of names is constitutive of relations between individuals and whether these relations can be further qualified. We follow in the steps of previous experiments on the relevance of network analysis in exploring a world of word cooccurrences (named individuals) in biographical texts and establishing the existence of actual social networks based on these named entities.[3] Our approach, however, differs from previous studies in the distinction that we make between the two different kinds of population in the dictionary: those who were the subject of a biography and those who were just mentioned in a biography. There is a considerable imbalance in the wealth of information on each group. The latter is reduced simply to a name, except when they belonged to the group of biographed individuals. We argue that this distinction is necessary when analyzing the data in network analysis.

This paper is structured as follows. Section 1 describes how we build a reference network from robustly recognized person-to-person cooccurrences at the highest possible accuracy. In our case, we built the network as a directed

[2] Christopher N. Warren et al., "Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks," *Digital Humanities Quarterly* 010, no. 3 (July 12, 2016).

[3] Matje van de Camp and Antal van den Bosch, "The Socialist Network," *Decision Support Systems* 53, no. 4 (November 2012): 761–69; Matje van de Camp and Antal van den Bosch, "A Link to the Past: Constructing Historical Social Networks," in *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11 (Stroudsburg, PA, USA: Association for Computational Linguistics, 2011), 61–69; Minna Tamper, Eero Hyvönen, and Petri Leskinen, "Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research," EasyChair Preprints, EasyChair Preprints (EasyChair, April 8, 2019); Pablo Aragon et al., "Biographical Social Networks on Wikipedia: A Cross-Cultural Study of Links That Made History," in *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration - WikiSym '12* (Eighth Annual International Symposium, Linz, Austria: ACM Press, 2012), 1.

network in which the nodes are individuals, and when individual B is mentioned in the biography of A, we added a directed edge from A to B. We examine this network of cooccurrences (textual links) in the first section of the paper to study the underlying structure of the BDRC and propose an alternative reading of the dictionary and its population at a global scale. In section 2, we explore whether sociocentric subnetworks form on the basis of the specific attributes (provincial origins, education, etc.) that we extracted from the biographies. In section 3, we shift the point of observation from the study of networks of cooccurrences to that of social networks. To this end, we enriched the network of cooccurrences with annotations that qualified the nature of relations, in order to (1) distinguish mere cooccurrences from actual social relationships, and (2) build subnetworks based on the nature of relations, which we can compare with the corresponding subnetworks of attributes.

2 The BDRC as a network of cooccurrences

In this section, we proceed in two steps. First, we describe the workflow for extracting named entities from the BDRC and building the reference network of cooccurrences. Second, we experiment with various methods (global and local metrics, pruning tables, clustering) in order to analyze its structure.

The BDRC consists of four volumes published between 1967 and 1971[4] and has served generations of China historians. It was produced under the editorship of Howard L Boorman, with contributions from about 100 different authors. Even if the biographies eventually went through the hands of a small group of editors, the first thing that our digital forensics has revealed is inconsistencies in the vocabulary used to describe individuals, positions, and institutions. The four volumes contain 589 individual biographies of unequal length — from 576 to 18,000 tokens — that feature “eminent Chinese” of the Republican period (1912–1949). This constitutes a very small sample of the Chinese Republican elites, by any standard, and the criteria for selecting this group of historical figures have proved debatable. Yet, a great number of people (3,178) are mentioned in these 589 biographies, which come under three main categories: family members, authors, and other individuals.

As a rule, the biographers provided information on the family of the biographed individuals, usually starting the biographical notes with the genitors or those who raised them, if known. The latter may not be the same as

[4] Howard L Boorman and Richard C Howard, *Biographical Dictionary of Republican China* (New York: Columbia University Press, 1967-1971).

the genitors due to death or adoption. Each biography thus starts with birth and childhood, with a discussion of the family background. In most cases, the biographers cited only the name of the father, almost never that of the mother, even in the case of prominent families. The father and mother of people of humble origin were simply not named. Generally, at the end of each biography, there is also often, but not always, a list of the biographed character's direct family members, namely wife/wives and children. Most of the time, the information is sketchy, especially for the wives, except when they were themselves prominent figures, socially, intellectually, or politically (the Song sisters, Ding Ling, etc.).

Under the category of authors, we grouped all the individuals who wrote about the biographed character and whose works are cited in the biography. Some of them are people who were effectively in contact with the biographed person in the course of his/her life. This is the case of former students who compiled and edited the writings of their former mentor, or sometimes next-of-kin (son-in-law, nephew, etc.). The majority of such works, however, were produced *ex post facto* by individuals who were unrelated to the person. Finally, the group of authors also includes historians and professional biographers who wrote extensive monographs or papers on the major figures in the BDRC.

To index the content and to identify all the named entities in the text, we processed the 589 biographies with Stanford CoreNLP.[5] Data extraction produced a raw file of 3,178 persons that listed all the biographed persons and all the individuals mentioned in their biographies. The high number of cooccurrences is indicative of the wealth of data available in the BDRC beyond the 589 biographed persons, something that one may perceive through conventional reading but will fail to embrace to its full extent. The number of individuals mentioned in each biography varies greatly, from 124 for Jiang Jieshi (蔣介石) to just one for Li Yizhi and Ma Buqing. Based on this list, we built a directed network linking each biographed person (thereafter "bionode") to the individuals mentioned in his/her biography (thereafter "object-nodes").[6] The direction of arrows indicates whether an individual mentions (outgoing edges)

[5] https://stanfordnlp.github.io/CoreNLP

[6] We borrowed this distinction between bionodes and nodes from Henrike Rudolph, "Structures of Empowerment: A Network Exploration of the Collective Biographies of Women Activists in Twentieth- Century China," *Elites, Knowledge, and Power in Modern China: The Formation and Transformation of Elites in Modern China* (Aix en Provence, 2019), 25.

or is mentioned by (incoming edges) another individual.[7] We counted each pair of individuals only once, even if an individual is mentioned several times in a biography. Given the high number of cooccurrences and the nature of the BDRC — a collection of individual biographies — what is the relational structure of the dictionary? Is it merely an aggregate of multiple ego-networks or does it form a interconnected global network?

The network of cooccurrences generated from the extracted data comprises 3,254 nodes and 9,524 edges. It is made up of a total of 11 components, with one giant component (3,177 nodes with 9,377 edges) and ten disconnected components. All of the latter, except one, are in fact isolated ego-networks built around one single bionode. The exception is a small component consisting of two small ego-networks that centered on — Kang Cheng (Ida Kahn) and Shi Meiyu (Mary Stone) respectively. In fact, these two figures were the first Chinese women physicians trained in the United States at the end of the 19th century. Both received the help and support of the same woman missionary (Gertrude Howe), through whom their ego-networks are interconnected and form a small component.[8] The other ego-networks revolve around individuals with very specific profiles: some had careers that unfolded mostly before the Republican era (Ye Changchi, Wang Ganchang) or in religious organizations (Wei Zhuomin, Zheng Hefu), others were scientists who either spent a lot of time abroad or even made most of their career outside of China (physicists Wu Jianxiong, Qian Xuesen) or whose profile fully diverged from the "mainstream population" in the BDRC (Li Yizhi, Pei Wenzhong). It cannot be said that these individuals were poorly connected since what we catch here are just mentions of names in their biographies. What can be said is that neither they nor the individuals named in their biographies were related to any of the nodes in the main component. Why were these individuals selected if they seemed quite off the mark? The probable answer lies between the editors' decision to have "representatives" of different sectors of society and the availability of source materials.

Within the main component, how and to what extent are the biographies interconnected? How far do they rely on object-nodes to be interconnected? Do the latter contribute to the connectedness of the global network? Previous studies of biographical dictionaries have generally focused only on the

[7] The extraction process was done in R-studio. The data and the script are available on GitLab (https://gitlab.com/enpchina/brelations).

[8] Connie Anne Shemo, *The Chinese Medical Ministries of Kang Cheng and Shi Meiyu, 1872-1937: On a Cross-Cultural Frontier of Gender, Race, and Nation* (Bethlehem: Lehigh University Press, 2011).

biographed individuals and their relations.[9] In this paper, we move a step further and compare the entire network of cooccurrences with the network consisting only of bionodes. Once the object-nodes are excluded, the number of disconnected components increases to 16, with 15 isolated individuals. The network of bionodes, however, remains highly connected. If we compare the global metrics of the two networks, the density increases tenfold (entire network = 0.002 / bionode network = 0.028) and the clustering coefficient multiplies by a factor of four (entire network = 0.083 / bionode network = 0.337).[10] In both networks, there is still a high degree of connectedness given the considerable number of nodes in each.

Beyond global metrics, we use various centrality measures in order to examine the relative position of nodes and bionodes: edge count (number of neighbors), indegree (number of incoming edges), outdegree (number of outgoing edges), and betweenness centrality. The edge count displays a long-tailed distribution, with a minimum of one tie and a maximum of 367 (Jiang Jieshi).[11] We defined six thresholds as shown in Table 1:

Table 1. Edge count

It can be observed from Table 1 that 2,154 individuals are mentioned only once in the BDRC. These nodes are each linked only to a single biography and are strictly related to the life of this very character. Although the vast majority becomes part of the main component of the BDRC network by virtue of their association with at least one bionode, their presence makes sense only in relation to this particular individual. These individuals all came from the object-node category, except for Li Yizhi (engineer, 1882-1938), who happened to have the lowest number of edges among the bionodes. The group of 496 individuals with 2 to 5 ties includes 459 object-nodes and 37 bionodes. The latter represents a group of individuals of lesser historical importance, more peripheral individuals (scientists), or individuals with a short lifespan. Within the network

[9] Aragon et al., "Biographical Social Networks on Wikipedia"; van de Camp and van den Bosch, "The Socialist Network"; Tamper, Hyvönen, and Leskinen, "Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research."

[10] Technically, network density shows how densely the network is populated with edges. It is a value between 0 and 1. A network which contains no edges and solely isolated nodes has a density of 0. In contrast, the density of a clique is 1. The clustering coefficient measures the ratio of the number of edges between neighbors and the maximum number of edges that could possibly exist in the network.

[11] We do not include isolated nodes (13) in this table.

of only bionodes, a significant percentage (19 percent) of bionodes also have only this range of ties.

In the next group of 396 individuals with between 6 and 24 ties, we find mostly bionodes (345) who directly connected to one another. Yet, it also includes 51 object-nodes who appear in a good number of biographies. This category comprises a wide range of profiles and includes both Chinese and foreign elites.[12] Among the foreigners, we can distinguish two groups of people that are highly connected, although they are from different social circles - American philosophers or military advisers and Soviet/Comintern agents. In both cases, this is about foreign experts involved in Chinese politics. The Chinese in this group include political figures from the late imperial period (emperor Guangxu, Li Hongzhang, Zeng Guofan) and intellectual figures from the Republican period (Mao Dun, Zhang Junmai, Yan Huiqing, Ding Wenjiang), some with political connections. Zhang Zhidong appears in the biographies of individuals of a similar generation with whom he had direct (Sheng Xuanhuai) or indirect contact (Shen Jiaben, Yan Fu), but also he was a figure of inspiration to younger people (Guo Bingwen).

Outside bionodes, the two individuals with the highest number of edges include Joseph Stalin (30) and Michael Borodin (27), both due to their direct or indirect role in Chinese politics. Stalin is mentioned mostly in the biographies of members of the Chinese Communist Party (CCP) (17), but also in relation with major political figures of the Guomindang (Jiang Jieshi, Jiang Jingguo, Song Ziwen, etc.) and left-wing personalities (Song Qingling, Liao Chengzhi, Deng Yanda, etc.). Borodin, in contrast, is hardly mentioned in the biographies of CCP figures despite his close interaction with communist leaders in China as the main Comintern representative. His network includes most of the major Guomindang figures, from Sun Zhongshan to Jiang Jieshi, which reflects his activity and direct relations with these individuals in the mid-1920s in Guangzhou.

Finally, the top three categories with more than 50 edges include exclusively bionodes. This group consists of the main figures of the Guomindang (Sun Zhongshan, Jiang Jieshi, Wang Jingwei, Hu Hanmin), two communist leaders (Mao Zedong, Zhou Enlai), all the main warlords (Duan Qirui, Zhang Zuolin,

[12] By order of importance: Zhang Zhidong (23), Li Hongzhang (18), general George Marshall (17), Zhang Junmai (Carson Chang), Mao Dun (16 each), John Dewey and Yan Huiqing (W.W. Yen) (15 each), Wu Chaoshu (C. C. Wu) and Zeng Guofan (14), Emperor Guangxu (13), Komintern agent Gregory Voitinsky, warlord Lu Yongxiang, and the Indian writer Rabindranath Tagore (12 each).

Feng Yuxiang, Wu Peifu, Zhang Xueliang), the transitional figure of Yuan Shikai, and two intellectuals (Liang Qichao, Hu Shi).

In order to better understand the underlying structure of the network of bionodes, we apply Marilyn Levine's method of dissecting the "hairball" by using pruning tables.[13] We take the edge count as the criterion for pruning the network. As shown on the pruning tables and the pruned graphs below, no significant change occurs until we remove the bionodes with 25 ties or more. From this point, the number of ties and nodes in the network is more than halved at each step. In the final step (Graph F), there only remains the four pivotal figures in the BDRC — Yuan Shikai, Sun Zhongshan, Jiang Jieshi, and Mao Zedong — i.e. the four major leaders that shaped the conventional narrative of the Republican period. This is just a preliminary exploration. More systematic utilization of the pruning method will help penetrate more deeply the structure of such complex networks.

Table 2. Pruning table with the range of edge counts, the ratio between remaining ties and nodes, and the remaining bionodes in the BDRC network at each threshold.

Figure 1. Prunings of network graph. A. 574 bionodes with >= 2 ties. B. 360 bionodes with >=15 ties. C. 205 bionodes with >= 20 ties. D. 118 bionodes with >= 25 ties. E. 54 bionodes with >= 35 ties. F. 4 bionodes with >= 200 ties (Mao Zedong [Mao Tse-tung], Yuan Shikai [Yuan Shih-k'ai], Sun Zhongshan [Sun Yat-sen], and Jiang Jieshi [Chiang Kai-shek]). Color and size of node proportionate to their edge count.

The hierarchies based solely on edge count, however, do not take into account the directed nature of the network. In order to refine our analysis, we need to distinguish between incoming and outgoing edges. To this end, we selected the individuals with an outdegree above 20 and calculated the ratio between indegree and edge count. Table 3 presents the results for the top 25 bionodes ranked by edge count (i.e., bionodes with more than 70 mentions in the BDRC). The ratio between indegree and edge count (last column) serves as a general indicator of how often an individual was mentioned in other biographies. It reinforces the impression that individuals with a very high rate of indegree are those who play an important role in the biographies of a large number of other individuals. Based on this ratio, we identified three major profiles: (1) individuals with a relative balance between incoming and outgoing edges (ratio ≈ 50%); (2) "source" figures with a greater number of outgoing

[13] Marilyn Levine, "Post WWI Chinese Revolutionary Leaders in Europe," *Journal of Historical Network Research* xx, no. xx (n.d.): xx–xx.

edges (ratio <50%), which include mostly political or military leaders who controlled the chain of command at the top of institutions and were often the source of action or decision; (3) "referential" figures with a greater number of incoming edges (ratio >50%).

The referential figures are those who are mentioned far more frequently in other people's biographies than other people get mentioned in their biography. For instance, Duan Qirui (400%), Wang Jingwei (371%), Yan Xishan (321%), Li Yuanhong (300%), and Liu Bocheng (300%) were each mentioned three to four times more than they mentioned other individuals. In absolute number, the most frequently mentioned individuals include, in descending order, Jiang Jieshi, Sun Zhongshan, Yuan Shikai, Wang Jingwei, Feng Yuxiang, Mao Zedong, and Duan Qirui. As we elaborate later, these individuals were either solicited for advice or mentioned as contextual references (that is, they are not mentioned as part of an actual relationship, but as an element of historical context). This is a central question that we discuss in the third section. It points to a major shortcoming of previous studies that often assume that cooccurrences are the expression of social relationships. In the last section, we challenge this assumption through a close analysis of the nature of relationships, based on the computer-assisted annotations of biographies.

Table 3. The 25 bionodes with an edge count above 70
Note: Bionodes are ranked by edge count in descending order, with their respective indegree (number of incoming edges), outdegree (outgoing edges), and ratio indegree/edge count.

Edge count, however, fails to convey the importance of certain individuals, whose significant position in the network does not rely solely on the number of ties. Betweenness centrality offers an alternative way of measuring the importance of object-nodes, not just bionodes, who hold a central position in the network.[14] Although they have a relatively low number of neighbors, certain individuals play a structuring role as mediators between different parts of the global network. For example, Paul Pelliot, the French archaeologist, with only 4 ties, holds a connecting position that joins two peripheral branches to the main component. If we remove him, the main component loses 95 nodes and 1,488 links. Similarly, the Qing official Zeng Guofan presents another intriguing case of a object-node with just 14 edges, who nevertheless connects 544 nodes and

[14] Betweenness centrality scores are particularly informative because they highlight the individuals who served as essential bridges ("brokers") between individuals and communities. Technically, betweenness centrality measures the number of shortest paths that travel through a node.

7,795 links. This is due to his connection to major bionodes such as Jiang Jieshi and Mao Zedong.

The list and range of bionodes with the highest scores of betweenness centrality remain very much aligned with the hierarchy defined by edge count. This group comprises 19 political, military, and intellectual figures who appear highly connected among themselves (123 edges). When their direct neighbors are included, these 19 individuals form a network of 870 nodes (including both the bionodes and the object-nodes) (26.7 percent of all nodes) and 6,226 edges (65.4% of all edges). The distinctive feature in the ranking by betweenness centrality is the emergence of a few prominent intellectuals such as Hu Shi, Liang Qichao, and Guo Moruo, who are placed 4th, 6th, and 8th respectively, well ahead of the major political and military figures who, by edge count, rank far above other nodes. It can be argued that the more versatile profiles of these three intellectuals who had a foot in various circles explain their position as eminent "brokers" in the BDRC.

The last method we apply to the bionode-only network is clustering. We seek to identify sub-communities of more densely connected bionodes. The algorithm (GLay) detected 23 communities with great variations in their size.[15] The largest cluster (cluster 4) comprises 208 bionodes and 2,279 ties, but the 14 smallest "clusters" each consist of just one bionode. The eight largest clusters are listed in Table 4. For each cluster, we report the number of nodes and ties it contains, list the most representative individuals, and give it a label that best describes its composition based on the names of its members. Some of them are clearly centered on major historical figures or clearly identified social groups. Cluster 4, for instance, revolves around famous military and political leaders (Jiang Jieshe, Sun Zhongshan, Yuan Shikai). Cluster 1 includes major intellectuals such as Hu Shi, Cai Yuanpei, and Liang Qichao. Cluster 3 groups together major Communist leaders (Mao Zedong, Zhou Enlai). Cluster 6 represents the business circle, whereas cluster 9 connects several prominent scientists (physician Wu Liande, biochemist Wu Xian). Other clusters, however, exhibit more complex patterns with less obvious rationale for their grouping (such as clusters 7 and 16). Little can be said about these communities relying

[15] We used Cytoscape Glay algorithm (Fast-greedy), which relies on the greedy optimization of modularity score, with different corrections on edge density and cluster size. Previous studies have demonstrated its dramatic performance advantage in handling large networks. Gang Su et al., "GLay: Community Structure Analysis of Biological Networks," *Bioinformatics* 26, no. 24 (December 15, 2010): 3135–37.

solely upon the names of their members. In the next section, we address the question of whether they coalesce on the basis of specific attributes.

Table 4. The eight largest clusters of bionodes detected by the GLay algorithm.
Note: For each cluster, the table reports the number of nodes and ties, the most representative individuals, and the label that best describes its composition, based on the names of its members.

Figure 2. Clustered network of bionodes
Note: Size of node is proportionate to degree centrality.

To conclude the first section, our analysis points to the dual structure of the BDRC as a network of cooccurrences, featuring, on the one hand, several small ego-networks isolated from the main component, and on the other hand, a polycentric network (the main component) that represent a relatively well-connected group of biographies polarized by a limited number of prominent figures. In brief, because of the considerable number of individuals mentioned in the BDRC, the cooccurrences of names eventually constitute a relatively well-connected network, with a giant component of interconnected individuals. The pruning method based on the edge count and an analysis of betweenness centrality have helped define distinct groups of individuals who are important to varying degrees and play different roles in maintaining the interconnectedness of the global structure. Moreover, beneath this massive "hair ball," clustering analysis has also revealed subgroups of more densely connected individuals. Do the various subgroups that contribute to the global structure of the BDRC coalesce sociocentric networks based on specific attributes? This is the core question we address in the next section.

3 Networks of attributes

After we identified the persons in the BDRC, we used NLP techniques to retrieve a wide range of information related to these persons, such as institutions, positions, locations, events, etc. We propose to use this data as attributes to enrich the network of cooccurrences. Our analysis focuses on bionodes only, because the BDRC provides information on the provincial origin, education, and other details of all the biographed individuals, but such information is entirely missing for those mentioned in their biographies. Retrieving such information at this stage would require a huge amount of time, especially because for many Chinese people mentioned in the biographies of others, the BDRC gives only the initials of their given names.

In this section, we examine whether individuals who share common attributes tend to group together so as to form "sociocentric" networks in the

BDRC.[16] We focus on five major attributes: provincial origin (well-studied by historians and often recognized as central in Chinese society[17]), military background (well represented in the BDRC[18]), education abroad (also a frequent feature in the BDRC population[19]), CCP affiliation (a self-contained, easily identifiable, and well-studied group[20]), and gender (women[21]). For each attribute, we built the network in two steps. First, we identified all the bionodes with the selected attribute, and second, we built an extended network that includes these bionodes and their first neighbors. It is these extended networks that we study below.

The hypothesis that provincial origin could provide the basis for specific networks did not pan out in general. For example, natives of Zhejiang are well represented in the BDRC, with 79 biographed individuals (13.9 percent of all the bionodes) who are connected to 434 other individuals in the extended network. Two isolated ego-networks around two scientists (Qian Xuesen and Wang Ganchang) from Zhejiang have no direct or indirect connection with the other Zhejiang natives. Furthermore, almost all the provinces of China are represented in the main component (1,040 nodes and 6,084 edges) of the Zhejiang network, with Jiangsu (57) and Hunan (56) as the most represented provinces, followed

[16] A “sociocentric network” is a network based on shared social attributes. This notion is borrowed from Tamper, Hyvönen, and Leskinen, “Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research.”

[17] William T Rowe, *Hankow: Commerce and Society in a Chinese City, 1796-1889* (Stanford, Calif.: Stanford University Press, 1984); Bryna Goodman, *Native Place, City, and Nation: Regional Networks and Identities in Shanghai, 1853-1937* (Berkeley: University of California Press, 1995); Richard Belsky, *Localities at the Center: Native Place, Space, and Power in Late Imperial Beijing* (Cambridge, Mass.: Harvard University Asia Center : Distributed by Harvard University Press, 2005).

[18] Diana Lary, *Region and Nation: The Kwangsi Clique in Chinese Politics, 1925-1937* (London; New York: Cambridge University Press, 1974); Jerome Ch’ên, *The Military-Gentry Coalition: China under the Warlords* (Toronto: University of Toronto-York University Joint Centre on Modern East Asia, 1979); Edward Allen McCord, *The Power of the Gun: The Emergence of Modern Chinese Warlordism* (Taipei: SMC Pub., 1997).

[19] Y.C. Wang, *Chinese Intellectuals and the West, 1872-1949* (Chapel Hill, University of North Carolina Press, 1966).

[20] Marilyn Levine, *The Found Generation: Chinese Communists in Europe during the Twenties.* (Seattle, Wash.: University of Washington, 1993); Steve Smith and Taylor & Francis, *A Road Is Made: Communism in Shanghai 1920-1927*, 2018.

[21] Gail Hershatter, Berkeley University of California, and Area Global and International Archive, *Women in China’s Long Twentieth Century* (Berkeley: Global, Area, and International Archive : University of California Press, 2007); Barbara Mittler, Michael Hockx, and Joan Judge, eds., *A Space of Their Own: Women and the Periodical Press in China’s Long Twentieth Century* (Cambridge: Cambridge University Press, forthcoming).

by Guangdong (48). Due to their importance in the Guomindang elites, Guangdong natives apparently offered the prospect of a tighter-knit group. Their network includes 376 bionodes and 4,050 edges, with 67 Cantonese nodes that possess a total of 239 edges. As in the case of the Zhejiang natives, however, a wide range of 23 provinces are represented, and the Cantonese account for only 18 percent of the total. We also observed that the network was actually made up of seven separate components, with six ego-networks built around specific personalities who had no link with each other.[22] In other words, there is little evidence of homophily by provincial origin among the Zhejiang and Guangdong natives in the BDRC.

Only one group based on the same provincial origin stands out in the BDRC: the natives of Hunan province form a large network of six components with 722 nodes and 4,908 edges. The presence of such prominent figures as Mao Zedong may have introduced a bias in terms of both provincial origin and political affiliation (CCP). Yet, even after removing Mao Zedong, the main component of the Hunan network still reveals a group of densely connected Hunanese whose pillars are also CCP members (Cai Hesen, Liu Shaoqi, Peng Dehuai, Peng Shuzhi, etc.). In this network, 72 of the bionodes are Hunanese, who have links to Zhejiang (47), Jiangsu (40), and Cantonese (37) natives.[23] Altogether 21 provinces are represented. Yet the main feature is the considerable number of CCP members in this network (112 bionodes), of which the Hunan natives claim a substantial share (37). One can argue that the Hunanese clearly form a more homophilic network that intersects with political affiliation.

The BDRC includes a high number of military figures. Individuals with any kind of military positions in their careers form a population of 466 bionodes with 4,035 edges. The size of this network is an indication of the high level of cooccurrences in the biographies of these individuals. Of these individuals, however, only a much smaller number (43 bionodes with 94 edges) received a military education or held military positions continuously. Taking only these forty-three individuals into account, this smaller network displays a high density (0.322 vs. 0.074 in the larger network), although five individuals constitute isolated components with few links each (Xie Bingying, Huang Kecheng, Sun Lanfeng, Sun Liren, He Zhonghan). The major broker in the main

[22] The six individuals were Wei Zhuomin (Religious leader), Hu Die (actress), Xu Guangping (women writer), Li Fanggui (linguist), Dai Ailian (woman writer), and Luo Dengxian (labor activist).

[23] On the place of Hunanese in CCP networks, see Levine, "Post WWI Chinese Revolutionary Leaders in Europe."

component is Li Zongren, who has the highest degree and betweenness centrality. At the next level and almost at par, three lesser-known figures emerge: Liu Zhi, Xue Yue, and Yang Sen, who each form their own clusters. Only Xue Yue and Yang Sen are directly connected. In terms of indegree, Li Zongren still ranks first, followed by Duan Qirui and Bai Chongxi.

There is no clear evidence that graduation from the same academic institution was a strong connecting factor, except for the military officers who graduated from Japanese military academies. Graduates from Shinbun Gakko and Shikan Gakko, for instance, show a strong propensity to group together, but it needs to be established whether this is the sign of actual social relationships or an artifact of contextual mentions. The generational factor may reinforce the impact of cooccurrences. There is a clear overrepresentation of individuals born during the decade 1886-1896 who graduated mostly from the new military schools and academies established by the Qing or from the Japanese military academies between 1906 and 1920. This generational group includes 25 bionodes (58%) and 60 edges (67%) and accounts for all the main brokers in the military-only network.

The Communists form a very identifiable and self-contained set of individuals who are included in the BDRC exclusively because of their affiliation with the party (CCP). The network of CCP members includes 865 nodes and 5,391 edges. It is a very large network that reflects the share of CCP members in the BDRC (121 bionodes) and the high number of individuals tied to them (760 edges). The network of CCP members alone is made up of nine components, with eight isolated individuals, and it has only four women in it. The outdegree distribution highlights 20 individuals with more than 30 neighbors. We can delineate four groups. The first group with an outdegree above 35 includes six distinct individuals (Mao Zedong, Zhou Enlai, Lin Biao, Zhu De, He Long, Li Dazhao). These six individuals actually connect almost every CCP member (93/99). At the next level (outdegree between 25 and 34), we find a second group of seven tightly knit individuals (Liu Shaoqi, Ye Ting, Li Lisan, Zhang Guotao, Li Jishen, Chen Duxiu, Guo Moruo). Taken together, these thirteen individuals (Guo Moruo as an outlier) form the backbone of the CCP network in the BDRC. Betweenness centrality reveals a limited number of mediators (10), yet with large discrepancies between them. The whole network is clearly centered around Mao Zedong, who serves as the main broker (0.3), followed by Zhou Enlai (0.12), Zhou Yang (0.06), Ye Ting (0.05), and Lin Biao (0.04). Within the CCP network, as discussed above, the Hunanese lead the pack with 65 individuals, followed by natives of Zhejiang (52), Guangdong (44), Jiangsu (44), Hebei (27), Hubei (21), and Jiangxi (18). The CCP network includes

13 military officers, but these CCP military figures do not form a specific community.

Most of the biographed characters in the BDRC received a high level of education. 519 (88 percent) received a college degree. Among them, many had the opportunity to study abroad, which place them in the particular category of "returned students." These returned students are commonly grouped according to where they studied (country, university). In the BDRC, they form a population of 200 individuals who attended and graduated from a total of 343 different academic programs. Since the returned students established alumni associations or held events in China that brought together those who had studied in the same country or region, one can hypothesize that networks may have been built on this basis.[24]

The United States ranks first in the BDRC with 70 returned students, followed closely by Japan with 67 individuals. Europe received the next largest batch, but the returned students from Europe were distributed across several countries: United Kingdom (24), Germany (14), France (15), Soviet Union (8), and a host of other countries (6). We built networks based on the country of study to examine to what extent the returned students from a given country had a propensity to connect with each other. Each such network includes bionodes who were not educated in the country of reference. For example, for the American-trained Chinese, 43 percent of their neighbors were not trained in the United States. Previous works have demonstrated their propensity to mingle with diverse communities, which corroborates our observation.25 We found the same ratio among Europe returnees. It is only among the Japan-returned students that the ratio was slightly lower, indicating possibly a greater homophily. Yet, as we discuss below, other factors may explain this higher level of homogeneity.

The American-trained students form one of the most interesting networks. In fact, it can be read as a miniature of the BDRC global network. On the one hand, a fair number of individuals (14) are not mentioned in any of the

[24] Stacey Bieler, *"Patriots" or "Traitors"?: A History of American-Educated Chinese Students* (New York: Routledge, 2003); Liu Xiaoqin, "Minguo Liumei Shetuan Yu Liumei Sheng de Shehui Wangluo -- Yi Chengzhihui Zhang Boling Fenxi Wei Zhongxin (Social networks and student associations in the United States in the republican era: A study of Zhang Boling and the Chengzhihui)," *Huaqiao Huaren Lishi Yanjiu*, no. 4 (2019): 88-95.

[25] Cécile Armand, "Foreign Clubs with Chinese Flavor: The Rotary Club of Shanghai and the Politics of Language," in *Knowledge, Power, and Networks: Elites in Transition in Modern China*, ed. Cécile Armand, Christian Henriot, and Huei-min Sun (Leiden: Brill, 2021).

biographies of their peers. Their networks do not intersect with the individuals in the main component, nor with the dyads formed by another five individuals. The 51 bionodes in the main component, however, tend to exhibit a higher level of connectedness. Hu Shi, one of China's leading intellectuals, serves as the main broker in this network (betweenness centrality = 0.15). He is connected to 21 peers through a total of 37 edges. His peers include mostly intellectuals but very few political figures. He is not linked to the second most important broker, Kong Xiangxi, an eminent figure in the dual world of business and politics. Kong is connected essentially to political figures, including his family relations (Song Ziwen, the Song sisters, and their father). The only intellectuals in his network are scholars with a foot in administration (Guo Bingwen, Ma Yinchu). Two other figures also play an important role in the network of American returnees, each in a different register: Song Ziwen, with a profile quite similar to his brother-in-law, Kong Xiangxi, and Jiang Menglin, a multi-faceted intellectual bridging the worlds of education, culture, and politics. While an exclusive pattern of homophily based on the country of education cannot be established among the American-trained students, their network suggests that they shared a common cultural, educational and linguistic background that may have served for establishing professional and political relationships in the course of their life.

Figure 3A. Network of American-returned students (main component)
Note: Size of nodes is proportional to betweenness centrality.

The network of Japan-trained students presents a very different structure. A striking but obvious feature is the centrality of Jiang Jieshi (by degree and betweenness measures). He connects almost all the individuals in the Japan network, and more importantly, in this network, those who received military training. If we remove Jiang from the network, two other figures emerge: Li Liejun and Wang Jingwei, each at the center of a substantially different network. Wang is connected mostly to political figures, with very few military leaders. Li, by contrast, reaches out to all the military leaders. If we enlarge the scope of observation to include all the bionodes connected to the Japan-returned students, we find a network made up largely of military figures trained in China or elsewhere. In other words, the single most important factor in the Japan-trained individuals is less the country where they studied than the military education that they received there, which put them on a career path that connected them to a wider circle of military figures. If we compare it with the American returnees, one could say that the degree of heterophily based on the country of education is higher among the Japan returnees than their American counterparts.

Figure 3B. Network of Japan-returned students (main component)
Note: Size of nodes is proportionate to betweenness centrality.

Very few women were selected for a biography in the BDRC. Altogether, they account for only 25 of the 589 biographies. What compelled the editors to select these women? Was it for their own profile and intrinsic importance, or because they were related to prominent men?[26] How do these women fit into the network structure of the BDRC? How do they contribute to male-dominated specific communities? In other words, can we identify a women's network in the BDRC?

The network of women includes 189 nodes and 990 edges, with four components: one main component and three ego or bi-ego networks. The bi-ego networks are those identified previously in the global analysis of the BDRC network of cooccurrences, namely the first two women physicians trained in the United States who graduated in 1896. Their network is composed exclusively of foreigners — which reflects the fact that Boorman focused on their period of education and training before they returned to China. The other small component revolves around Wu Jianxiong, a woman physicist who made most of her career outside China, also with a large number of foreigners in her network.

Women as such do not form any cohesive network. Their limited number may be part of the explanation for the lack of more obvious networking. Five of them stand alone, and four women form two distinct pairs. The main component is made up of a core of a very small number of highly connected women — Song Qingling, Ding Ling (writer), and Song Meiling. Except for Ding Ling, who stands apart, Song Qingling and Song Meiling - two sisters from an influential family and among the most prominent and politically active women of Republican China - are interconnected at the same level. Their marriage with men of prominence (Sun Zhongshan and Jiang Jieshi) placed them within a larger network that included many main figures of the Republican period. Yet, even after removing the three main male figures (Mao, Jiang, and Sun), the centrality of the three women remains the same. On the other hand, they fail to connect directly with any significant number of women, and they even do not connect with each other. In the case of Song Meiling, it is through He Xiangning, the wife/widow of Liao Zhongkai (d. 1925), that she makes the connection with

[26] Henrike Rudolph, "Structures of Empowerment: A Network Exploration of Women Activists' Collective Biographies in Twentieth-Century China," in *Knowledge, Power, and Networks: Elites in Transition in Modern China*, ed. Cécile Armand, Christian Henriot, and Huei-min Sun (Leiden: Brill, 2021).

Chen Bijun (wife of Wang Jingwei) and Deng Yingchao (wife of Zhou Enlai). The remaining group of five women are even more tenuously connected.

Ding Ling's network branches out in two main directions: CCP members, including the men of letters in the party (Zhou Yang, Hu Feng) and literary figures (two other women writers, [Xie Wanying and Su Xuelin] and three male writers [Ye Shengcao, Lu Xun, Cao Yu]). Song Meiling's network is made up of powerful men that include all her direct and indirect next-of-kin (father, husband, brother-in-law, etc.), as well as military and political figures who served her husband or her more directly (Yu Hongjun, Wu Guozhen). There is no CCP figure in her network. Song Qingling, the older sister, presents a similar profile in terms of next-of-kin relations, but her network branches out to both Guomindang and CCP figures, which quite accurately reflects her positioning in Republican politics and in the People's Republic of China when she became the willing pawn of the CCP. In brief, the analysis of the main component reveals four main profiles of women that may have served as a guide for including them in the BDRC: scientists, artists, writers, and political women. Political women appear in the BDRC for their own sake but also due to their marriage with important political figures in Republican China, whereas intellectuals such as Ding Ling appear only due to their own merit.

To conclude this section, the study of attribute-based networks reveals that neither provincial origin, education, nor any single attribute alone is sufficient to constitute significant sub-communities in the BDRC. It is only the combination of multiple attributes that can account for the most densely connected clusters of biographies in the BDRC. For instance, this paper has reasserted, in line with previous scholarship, the effect of Hunanese origin combined with CCP affiliation, that of study-abroad experience in Japan combined with military training, and the conjunction of marriage, political influence, but also professional skills in the case of women. The patterns delineated in the study of attribute-based networks tend to support the hypothesis that there is more to the cooccurrence of names than the aggregated mentions of individuals in the various biographies. The repeated mentions suggest the existence of actual relationships. The network of cooccurrences, however, does not permit us to fully ascertain this. A more in-depth examination of the nature of the relationships is needed. This is the purpose of the next section.

4 Cooccurrences or relations?

Our analysis in the previous sections suggests the existence of two major categories of mentions associated with two dominant types of links:

textual/contextual references and actual social contacts. But it has proved difficult, if not impossible, to firmly establish the difference as long as we remain focused on the network of cooccurrences. In the first section, we made the hypothesis that some individuals may be mentioned as elements of historical context or were a source of inspiration for the biographed characters. We also pointed to historical characters from the past, such as Adam Smith or Zhuangzi, who could not actually have met with the individuals biographed in the BDRC. In the particular case of "referential" figures with high indegree centrality, we highlighted that they were most often sought for advice or served as contextual references. The close reading of their biographies further reinforces this impression. These observations led us to question the more general assumption that the cooccurrence of names in a biography could be systematically considered as the expression of a genuine social relationship.

In this section, we move a step further and explore in greater depth the nature of the links between individuals in a selected sample of 36 biographies that we annotated manually (see Appendix). Our approach is to examine the actual ground for the mention of a name in a given biography and qualify the relation between the named individuals. Our ultimate goal is to build networks from these annotations that better reflect historical relationships, and not just textual cooccurrences. This section follows three steps. First, we present the method for annotating the relations in the biographies. Second, we analyze the results statistically. Third, we build and analyze various networks based on the (most significant) extracted annotations, which we compare with the attribute-based networks discussed in section 2.

The first challenge was to constitute a "representative" sample to annotate. We relied on two criteria. First, we selected the individuals on the basis of the edge count. A high edge count was a sign that these individuals were involved in the widest and richest range of possible relations. Second, we refined the sample to include the greatest possible variety of individual profiles in order to correct the bias produced by relying solely on the edge count. The selected biographies represent only 6 percent of the total number of biographies, but 18 percent of the total number of words in the BDRC.

In each biography, we focused only on the relations involving the biographed individual. For instance, in the biography of A, we annotated the relations between A and B and between A and C, but we discarded any potential relation between B and C. The selection of qualifying terms was based on the terms identified through close reading. We ensured that the manual annotations reflect only the language and the terms used in the text without adding any layer of interpretation or external knowledge, because the model for automatic annotations would ultimately rely only on the text itself and the particular combinations of words through which different relations are expressed. As shown in Table 5, we classified the relationships into twelve categories: acquaintance, protégé, friendship, liaison, kinship, *tongxiang* (same native place), education (master/disciple, co-disciple), professional, political, military (military conflict), indirect (no direct relation: context, third-party reference, etc.), and neutral (uncategorized mentions). We were aware that our categories represented a wide range of choices, but previous experiments had convinced us that a narrow range of terms might produce results too broad for analysis.

For instance, a previous study based on a similar corpus — a biographical dictionary of Dutch socialists — had also attempted to qualify the relations between individuals through manual and automatic annotations. All the cases of cooccurrences, however, were considered as meaningful relations. Relying on sentiment analysis, the authors chose to classify the relations into three main types: positive, antagonistic, and neutral. They eventually found that the trilogy was too reductive, especially the neutral one that regrouped too many cases to make it a significant marker.[27] This categorization may have been relevant for relationships within a coherent and like-minded group, but the BDRC presented us with a wider array of very distinct profiles. In our case, we were interested in qualifying the relationships along a richer set of terms based on the words used in the text itself to describe the relationships. Some categories did not really pan out due to the limited number of such relationships (liaison, protégé, *tongxiang*). Yet the process provided a preliminary schema in the form of pre-determined categories to be used eventually in the model for automatic annotations.

Table 5. Types and definitions of annotations in the BDRC.

For the annotation workflow, we relied on InCeption, a machine-assisted interactive annotation platform.[28] Each biography was annotated manually by a pair of annotators who worked independently, and then we curated together their respective biographies.[29] There were significant variations in the

[27] Matje van de Camp and Antal van den Bosch, "A Link to the Past: Constructing Historical Social Networks," in *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '11 (Stroudsburg, PA, USA: Association for Computational Linguistics, 2011), 61–69.

[28] Jan-Christoph Klie et al., "The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation," n.d., 5. See also: https://inception-project.github.io/

[29] The annotators included Cécile Armand (ENP-China, Aix-Marseille University), Guo Weiting (ENP-China, Aix-Marseille University), Christian Henriot (Aix-Marseille University), Jiang Jie (Shanghai Normal University), David Serfass (Inalco), Sun Huei-min (IMH, Academia Sinica).

annotations, mostly due to the inevitable propensity to "interpret" based on prior knowledge of the individuals mentioned in the biography and the difficulty to disentangle multifaceted relationships. Eventually, all the biographies went through the hands of a single curator who homogenized the annotations. These manual annotations were used to train a model for expanding the annotations automatically to the whole corpus in the future.

The annotation workflow produced a total of 3,227 annotated relations. As shown in table 6, the three most frequent types of relations represent 77 percent of the total. They include political relations (34 percent), indirect relations (26 percent), and professional relations (17 percent). Although military relations and kinship relations garner a good number of annotations, they represent only 7 percent and 6 percent of the total, respectively. All the other categories, including that of *tongxiang* (same native place), failed to produce a significant number of annotations. They were not used to train our model. The high percentage of the top three types of relations indicates that the contributors who wrote the biographies centered primarily on the work and public life of the individuals, especially their professional activity. More importantly, although a significant share (one quarter) of annotations points to textual cooccurrences (indirect relations), they do not prevail in the BDRC. The remaining 75% are the expressions of actual social relationships between historical actors.

Table 6. Distribution of annotated relations in a sample of 36 biographies from the BDRC.

We ran a Principal Component Analysis (PCA) on the data extracted from the annotated biographies. Through PCA, we sought to construct *relational profiles* based on the most frequent correlations of specific relationships.[30] For this PCA, we retained as active variables only the most frequent categories of relations: political (1094, 34%), indirect (823, 26%), professional (563, 17%), and military (239, 7%) and we considered the minor relations as supplementary variables. The PCA graph plots all 35 biographed individuals and their relations in a two-dimensional space.[31] We selected the first two dimensions, as they best explain the variance among individuals and capture more than 70% of the information (46% and 24% on each dimension). The graph clearly contrasts individuals with many indirect and political relations on the right and those with few such relations on the left. The second dimension, which is primarily

[30] For conducting our PCA, we relied on the package "Factominer" in R Studio: http://factominer.free.fr/factomethods/hierarchical-clustering-on-principal-components.html

[31] We chose to remove Guo Moruo whose profile was too specific, too different from the rest of the sample.

determined by military and professional relations, separates individuals with strong military and professional relations above the *x* axis from those with few such relations below it.

Figure 4. PCA analysis of the 36 annotated biographies: A. Graph of variables. *Note: Black arrows: active variables; blue arrows: supplementary variables.* B. Hierarchical clustering of individuals.

Based on the PCA, we performed hierarchical clustering on all four dimensions. We observed that professional relations contribute the most to the partition at large (0.693), followed by indirect (0.664), military (0.633), and political (0.506) relations. The algorithm grouped individuals into four clusters based on their relational characteristics. In figure 3 below, each individual is color-coded by cluster. Cluster 4 on the right of the graph isolates two "big names" — Jiang Jieshi and Yuan Shikai — from all the others. What sets them apart is the wealth of professional relations (4.52), the weight of indirect relations (2.88), and finally, the range of political relations (2.41). What dominates cluster 2 is the conjunction of strong professional relations (individuals who held a variety of positions in the army through which they came into contact) and military confrontations, as allies or as enemies. This cluster brings together almost all the military leaders, both nationalists and communists, except for the three warlords (Zhang Xueliang, Wu Peifu, and Zhang Zuolin). Individuals in cluster 3 are characterized by few military and professional relations and a stronger weight of political and indirect relations. This cluster includes Mao Zedong, Hu Hanmin, Zhou Enlai, and Sun Zhongshan. Wu Peifu and Zhang Zuolin, two major warlords of the post-Yuan Shikai era, also belong to this cluster, which suggests their respective biographies define them much more by their political relations than by their professional or military ones.

In the last step, we built a series of networks based on annotated relations. We constructed one network for each of the most significant types of relationships. We expected annotated relations to determine with greater confidence the reality of a relationship and to delineate more precisely the nature of the links between the biographies and between the individuals. The annotated relations may not alter the overall structure of the co-occurrence networks fundamentally, nor the dominant position of high-profile individuals such as Jiang Jieshi or Yuan Shikai — we could only compare them with the co-occurrence network of 36 annotated biographies — but they substantiated the hypothesis that a relationship could, in fact, reveal different configurations of proximities and interactions.

The category of “indirect relations” lumps together different types of relations (contextual mention, source of inspiration, connection through a third party, etc.). In future analyses, we will refine this category to separate the purely contextual relations from the other types of relations, and we will adjust the annotation workflow accordingly. Still, the network of indirect relations confirms the weight of Sun Zhongshan, Mao Zedong, and Jiang Jieshi (in descending order of betweenness centrality) in purely contextual mentions. In this function, they have no direct relationship with the person in whose biography they appear. Most such mentions come under formulas such as “When A came to power,” “At the time of A’s death,” “Under A’s regime,” etc. Moreover, the clustering of this network produced very interesting subgroups centered on specific individuals who shared common characteristics. One of these clusters is dominated by Jiang Jieshi and Sun Zhongshan, a second by the northern military leaders *cum* warlords, another one by CCP leaders, whereas two less densely connected groups revolved around intellectual figures (Hu Shi, Li Shizeng, Li Dazhao) and late Qing revolutionary activists (Huang Xing, Liang Qichao, Zhang Binlin). These indirect relations, even if they do not denote an actual social relationship, provide a sort of “index” of relevance in terms of contextual mentions.

Political relations are the most prominent feature in the connection between individuals. They delineate neatly two separate worlds: CCP leaders on one side and all the other main historical figures on the other side. They highlight the centrality of Jiang Jieshi and Sun Zhongshan (betweenness centrality) in the whole network, while Mao Zedong and Zhou Enlai are significant only within the CCP sub-network. None of them reach out very much into non-communist circles, except Zhou Enlai. Yet this is mainly due to Zhou’s later career as premier of the People’s Republic that considerably extended his contacts internationally. In contrast, Li Dazhao, more of a secondary figure who was executed in 1927 at age 38, had a broad network of relations across political lines although he was executed in 1927 at age 38. Yuan Shikai also built an extensive and very diverse network of political relations with political and mostly non-military figures, including a good number of Qing officials, but also major opponents such as Huang Xing, Song Jiaoren, or allies-turned opponents like Liang Qichao and, of course, Sun Zhongshan.

The exploration of professional relations redraws the previous configurations and leads to a very different network structure. Jiang Jieshi and Yuan Shikai are the two most central figures in this network. By placing emphasis on their professional relations, the resulting network gives more weight to their careers in government and the army. Their respective networks, however, diverge significantly. Yuan Shikai is connected to all the military leaders of the early republican period, except Wu Peifu. Many were his protégés. His professional relations include only a few political or intellectual figures such as Sun Zhongshan, Hu Hanmin, Wang Jingwei, or Cai Yuanpei. Yuan shares these figures with Jiang Jieshi's professional network. They are the connecting points, though indirectly, between Jiang and Yuan. Jiang's network includes two major military figures on the nationalist side — He Yingqin and Zhang Fakui — and a host of less central individuals. What distinguishes the network based on professional relations is the greater diversity one can see in the multiple sub-networks built around individuals (Hu Shi, Wu Peifu, Zhang Xueliang) who are only remotely connected to Jiang and Yuan. This observation also holds true for CCP leaders. Mao Zedong's professional relations place him as a secondary figure in the network, and he connects mostly to the triad formed by Zhou Enlai, Liu Shaoqi, and Zhu De. Yet again, this reflects the nature of post-1949 relations.

Military relations provide a good case to compare the networks of cooccurrences and annotations. We compare the network based on annotated military relations with the corresponding network of cooccurrences (based on military attributes) that we built in section 2 and filtered down to the 36 annotated biographies. In the network based on annotations, the number of nodes and edges decreases greatly, from 227 nodes and 811 edges to 101 nodes and 225 edges (Table 7. This is mostly due to the fact that we have focused only on the relations involving the bionodes. More significantly, the lower number of edges and the lower clustering coefficient suggest that military operations played but a limited part in the biographies of military elites. The BDRC emphasizes their involvement in a wider range of relations instead. The power and prestige resulted less from their military deeds — victories or defeats on the battlefield — than from other sources of influence (political negotiations, professional contacts, recommendations).

Table 7. Comparative analysis, military networks (global metrics).

Who are the most important players in these two networks? In the network of cooccurrences based on military attributes, betweenness centrality places Jiang Jieshi as the domineering broker, connecting a host of second-rank military and non-military actors (including Sun Zhongshan). In the network of annotations, Jiang remains central, but at the same level with other military leaders. Moreover, non-military actors are relegated to a secondary position. Quite interestingly, Yuan Shikai becomes a minor and more marginal broker with a very limited range of military relations.

Figure 5. Military networks: A. Based on attributes; B. Based on annotated relations (sample of 36 biographies). Note: Size of nodes is proportionate to betweenness centrality.

To conclude this section, the analysis based on annotated relations in biographies reveals that individual mentions in the BDRC are not just cooccurrences – i.e., names connected through textual links – but also refer to historical actors who actually came into contact in the course of their life. There is, however, a substantial number of cases of indirect relations (25% of all annotations) and even of dropped names that argue for the need to exercise greater caution about considering any cooccurrence as the expression of an actual historical relation. On the other hand, the BDRC features a great variety of actual links within and between the biographies. The relational patterns we delineated through PCA and SNA also demonstrated the intermingling of multiple relationships among individuals, which complicates the previous typologies based on sentiment analysis. Clearly, annotations provide a necessary and efficient way to add historical substance to the analysis of networks simply based on cooccurrences.

5 Concluding Remarks

Reading the BDRC through the lens of social network analysis may seem like putting old wine in a new bottle. The major biases of this work have been established in previous academic reviews.[32] Our contribution is not to reassert these biases in terms of content (problems of sampling and representativeness) but rather to uncover the underlying structure of the book, namely the hidden relations between individual biographies and between the individuals therein. This allows us to (1) extract a "collective portrait" of the entire population based on their individual characteristics (attributes); (2) assess whether and how far cooccurrences were constitutive of networks; and (3) propose an approach that defines and qualifies relationships much more accurately.

We demonstrated that in the BDRC, political and professional relations far outweigh the "three sames" (native place, education, trade/business) commonly accepted in the historical literature. This goes against the grain, but we argue that this is not due solely to the nature of the elites selected in the

[32] J. K. Fairbank, "Biographical Dictionary of Republican China. Volume I, Howard L. Boorman, Editor. Richard C. Howard, Associate Editor. (New York: Columbia University Press. 1967)," *The American Historical Review* 73, no. 2 (December 1, 1967): 565–66 ; Lucien Bianco, "Howard L. Boorman, editor, Richard C. Howard, associate editor, Biographical Dictionary of Republican China, vol. 1.," *Annales* 23, no. 5 (1968): 1133–35; D. C. Twitchett, review of *Review of Biographical Dictionary of Republican China*, by Howard L. Boorman and Richard C. Howard, *Political Science Quarterly* 84, no. 4 (1969): 650–52.

BDRC.[33] The persistence of these types of relations, especially native place ties, was undeniable in Republican China. Chinese society, however, departed increasingly from the patterns studied for late imperial China. There was a flurry of new types of social organization that offered the possibility for individuals to get involved in multiple groups and networks. Political parties are a prime example of a completely novel type of organization, but professional or cultural associations also provided numerous arenas based on non-partisan grounds. While it is difficult to escape the conventional categorization (politician, merchant, military, etc.) that historians use to define elites — some individuals do fit in such categories — the relational profiles we have revealed challenge the relevance of such narrow categorizations for the Republican elites. The complex web of relations in which the individuals in the BDRC were enmeshed cut across such categories and their careers often followed more than one path, sometimes in parallel.

From a methodological perspective, we have also demonstrated that much original knowledge can be gained from biographies through this approach. On the one hand, network analysis reconnects the individualized and self-contained biographies and open pathways through the BDRC mosaic. It falls short of creating a global narrative, but it revisits the nature and function of a biographical dictionary. The relations we have unveiled can open a new way of navigating through the BDRC in a digital version, such as we are planning to release. On the other hand, network analysis allowed us to put the BDRC to a truth test. It reveals the tensions in this work between a largely densely connected “main component” — featuring a group of leading elites in the political and military fields — and different subsets of individuals, and even various disconnected individuals.

The BDRC contains a large but finite volume of biographical data. We processed only what was in the text, with no addition of external information. Yet the implementation of data mining and annotation methods based on NLP, followed by exploration with network analysis and PCA, allowed us to identify and trace patterns of relationships, to question some assumptions about the types of relations among this composite elite population, and to breathe life into the stock of knowledge contained in the BDRC. The set of manual annotations of just a small sample of the biographies proved highly instructive, as a learning experience for historians to “define” the nature of relations in a text. People are multifaceted and it proved very challenging to reduce the nature of relations to a single word. This also demonstrates the need for human intervention at every

[33] Boorman and Howard, *Biographical Dictionary of Republican China*, viii.

step, from close reading to defining terms and expressions, to annotating the text.

This is an on-going experiment, but we believe that the models that we trained on our set of manual annotations offer a enormous potential for moving toward automatic annotations of large biographical corpora. It paves the way for the exploration of similar corpora such as the main English language dictionaries and the numerous Chinese language works published both before and after 1949.

6 Appendix: List of 36 Biographies

Name	Chinese	Wade-Giles
Zhang Fakui	張發奎	Chang Fa-k'uei
Zhang Xueliang	張學良	Chang Hsueh-liang
Zhang Xun	張勳	Chang Hsün
Zhang Binglin	章炳麟	Chang Ping-lin
Zhang Zuolin	張作霖	Chang Tso-lin
Chen Duxiu	陳獨秀	Ch'en Tu-hsiu
Jiang Jieshi	蔣介石	Chiang Kai-shek
Zhou Enlai	周恩來	Chou En-lai
Zhou Shuren	周樹人	Chou Shu-jen
Zhu De	朱德	Chu Teh
Feng Yuxiang	馮玉祥	Feng Yü-hsiang
He Long	賀龍	Ho Lung
He Yingqin	何應欽	Ho Ying-ch'in
Xu Shichang	徐世昌	Hsü Shih-ch'ang
Hu Hanmin	胡漢民	Hu Han-min
Hu Shi	胡適	Hu Shih
Huang Xing	黃興	Huang Hsing
Kong Xiangxi	孔祥熙	H. H. K'ung
Guo Moruo	郭沫若	Kuo Mo-jo
Li Liejun	李烈鈞	Li Lieh-chün
Li Shizeng	李石曾	Li Shih-tseng
Li Dazhao	李大釗	Li Ta-chao
Li Zongren	李宗仁	Li Tsung-jen
Li Yuanhong	黎元洪	Li Yuan-hung
Liang Qichao	梁啓超	Liang Ch'i-ch'ao

Liu Shaoqi	劉少奇	Liu Shao-ch'i
Mao Zedong	毛澤東	Mao Tse-tung
Song Ziwen	宋子文	T. V. Soong
Sun Zhongshan	孫中山	Sun Yat-sen
Cai Yuanpei	蔡元培	Ts'ai Yuan-p'ei
Duan Qirui	段祺瑞	Tuan Ch'i-jui
Wang Jingwei	汪精衛	Wang Ching-wei
Wu Peifu	吳佩孚	Wu P'ei-fu
Ye Ting	葉挺	Yeh T'ing
Yan Xishan	閻錫山	Yen Hsi-shan
Yuan Shikai	袁世凱	Yuan Shih-k'ai

7 References

Aragon, Pablo, David Laniado, Andreas Kaltenbrunner, and Yana Volkovich. "Biographical Social Networks on Wikipedia: A Cross-Cultural Study of Links That Made History." In *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration - WikiSym '12*, 1. Linz, Austria: ACM Press, 2012.

Armand, Cécile. "Foreign Clubs with Chinese Flavor: The Rotary Club of Shanghai and the Politics of Language." In *Knowledge, Power, and Networks: Elites in Transition in Modern China*, edited by Cécile Armand, Christian Henriot, and Huei-min Sun. Leiden: Brill, 2021.

Belsky, Richard. *Localities at the Center: Native Place, Space, and Power in Late Imperial Beijing*. Cambridge, Mass.: Harvard University Asia Center, 2005.

Bianco, Lucien. "Howard L. Boorman, editor, Richard C. Howard, associate editor, Biographical Dictionary of Republican China, vol. 1." *Annales* 23, no. 5 (1968): 1133–35.

Bieler, Stacey. *"Patriots" or "Traitors"?: A History of American-Educated Chinese Students*. New York: Routledge, 2003.

Blouin, Baptiste, Pierre Magistry, and Nora Van den Bosch. "Creating Biographical Networks from Chinese and English Wikipedia." *JHNR*, n.d.

Boorman, Howard L, and Richard C Howard. *Biographical Dictionary of Republican China*. New York: Columbia University Press, 1967.

Camp, Matje van de, and Antal van den Bosch. “A Link to the Past: Constructing Historical Social Networks.” In *Proceedings of the 2Nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis*, 61–69. WASSA ’11. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011.

———. “The Socialist Network.” *Decision Support Systems* 53, no. 4 (November 2012): 761–69.

Ch’ên, Jerome. *The Military-Gentry Coalition: China under the Warlords*. Toronto: University of Toronto-York University Joint Centre on Modern East Asia, 1979.

Fairbank, J. K. “Biographical Dictionary of Republican China. Volume I, Ai-Ch’Ü. Howard L. Boorman, Editor. Richard C. Howard, Associate Editor. (New York: Columbia University Press. 1967. Pp. Xv, 483. $20.00).” *The American Historical Review* 73, no. 2 (December 1, 1967): 565–66.

Goodman, Bryna. *Native Place, City, and Nation : Regional Networks and Identities in Shanghai, 1853-1937*. Berkeley: University of California Press, 1995.

Hershatter, Gail, Berkeley University of California, and Area Global and International Archive. *Women in China’s Long Twentieth Century*. Berkeley: Global, Area, and International Archive : University of California Press, 2007.

Klie, Jan-Christoph, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. “The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation,” n.d., 5.

Lary, Diana. *Region and Nation: The Kwangsi Clique in Chinese Politics, 1925-1937*. London; New York: Cambridge University Press, 1974.

Levine, Marilyn. “Post WWI Chinese Revolutionary Leaders in Europe.” *Journal of Historical Network Research* xx, no. xx (n.d.): xx–xx.

Levine, Marilyn Avra. *The Found Generation: Chinese Communists in Europe during the Twenties.* Seattle, Wash.: University of Washington, 1993.

Liu Xiaoqin. “Minguo Liumei Shetuan Yu Liumei Sheng de Shehui Wangluo -- Yi Chengzhihui Zhang Boling Fenxi Wei Zhongxin (Les Réseaux Sociaux et Les Clus d’étudiants Chinois Aux Etats-Unis Sous La République : Une Étude Centrée Sur La Chengzhihui et Zhang Boling).” *Huaqiao Huaren Lishi Yanjiu*, no. 4 (2019): 88-95.

McCord, Edward Allen. *The Power of the Gun: The Emergence of Modern Chinese Warlordism*. Taipei: SMC Pub., 1997.

Mittler, Barbara, Michael Hockx, and Joan Judge, eds. *A Space of Their Own: Women and the Periodical Press in China's Long Twentieth Century*. Cambridge: Cambridge University Press, forthcoming.

Rowe, William T. *Hankow: Commerce and Society in a Chinese City, 1796-1889*. Stanford, Calif.: Stanford University Press, 1984.

Rudolf, Henrike. "Structures of Empowerment: A Network Exploration of the Collective Biographies of Women Activists in Twentieth- Century China," 25. Aix en Provence, 2019.

———. "Structures of Empowerment: A Network Exploration of Women Activists' Collective Biographies in Twentieth-Century China." In *Knowledge, Power, and Networks: Elites in Transition in Modern China*, edited by Cécile Armand, Christian Henriot, and Huei-min Sun. Leiden: Brill, 2021.

Shemo, Connie Anne. *The Chinese Medical Ministries of Kang Cheng and Shi Meiyu, 1872-1937: On a Cross-Cultural Frontier of Gender, Race, and Nation*. Bethlehem: Lehigh University Press, 2011.

Smith, Steve and Taylor & Francis. *A Road Is Made: Communism in Shanghai 1920-1927*, Honolulu, University of Hawaii Press, 2018

Su, Gang, Allan Kuchinsky, John H. Morris, David J. States, and Fan Meng. "GLay: Community Structure Analysis of Biological Networks." *Bioinformatics* 26, no. 24 (December 15, 2010): 3135–37.

Tamper, Minna, Eero Hyvönen, and Petri Leskinen. "Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research." EasyChair Preprints. EasyChair Preprints. EasyChair, April 8, 2019.

Twitchett, D. C. Review of *Review of Biographical Dictionary of Republican China*, by Howard L. Boorman and Richard C. Howard. *Political Science Quarterly* 84, no. 4 (1969): 650–52.

Wang, Y.C. *Chinese Intellectuals and the West, 1872-1949*. Chapel Hill, University of North Carolina Press, 1966.

Warren, Christopher N., Daniel Shore, Jessica Otis, Lawrence Wang, Mike Finegold, and Cosma Shalizi. "Six Degrees of Francis Bacon: A Statistical Method for Reconstructing Large Historical Social Networks." *Digital Humanities Quarterly* 010, no. 3 (July 12, 2016).