Eminent Chinese of the Shenbao (1872-1891). A digital investigation of news reporting and newspaper-making in late imperial China

Christian Henriot

Eminent Chinese of the *Shenbao* (1872-1891). A digital investigation of news reporting and newspaper-making in late imperial China

Christian Henriot
Aix-Marseille Université

Newspapers have long played a crucial role in historical research, especially in the field of Chinese history. Access to archives in China has historically been challenging, making newspapers like the Shanghai-based *Shenbao* an essential alternative resource.

The conditions of accessing Chinese newspapers have undergone profound changes in the past four decades, significantly impacting research practices. Advancements in technology have led to increased accessibility to newspapers in various formats. Chronologically, the original print version of the *Shenbao* transitioned into microfilm (1969), reprints (1987), scanned images (2008), and finally, a full-text digital format (2010).[1] However, the release of the full-text digital version only marginally improved research due to substantial limitations imposed by the primary providers in mainland China. Online queries yield numerous results, but each must be examined individually. Furthermore, the ability to copy or download only a small portion of selected articles hinders research efforts. Additionally, attempts to employ web scraping or implement Natural Language Processing (NLP) tools are met with immediate blocking by robots. In essence, Chinese full-text newspapers offered by mainland companies do not fully exploit the potential of NLP methodologies, as successfully demonstrated in historical newspaper research in the United States and Europe.[2] In the ENP-China project, we identified a Taiwan-based company willing to provide us with the XML text files of the *Shenbao*. This collaboration allowed us to extensively implement NLP methodologies for historical research.

Exploring the extensive wealth of information contained within the *Shenbao* presents a formidable challenge. Nonetheless, we now possess the means to develop strategies for a deeper understanding of the data within the newspaper's pages. During its 78-year existence

[1] The first two full-text digital versions of the *Shenbao* were released at one year interval by the Airusheng Company (Erudition) in Beijing in 2010 [https://www.eastview.com/resources/newspapers/shen-bao/] and by the Qingpingguo Company (Green Apple) in Changsha in 2011 [http://www.egreenapple.com/english/contents/136/372.html].

[2] Thomas Lansdall-Welfare et al., "Content Analysis of 150 Years of British Periodicals," *Proceedings of the National Academy of Sciences* 114, no. 4 (January 24, 2017): E457–65; *Digitised Newspapers – A New Eldorado for Historians?: Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitization*, *Digitised Newspapers – A New Eldorado for Historians?* (Berlin: De Gruyter Oldenbourg, 2022); Thomas Lansdall-Welfare et al., "The Actors of History: Narrative Network Analysis Reveals the Institutions of Power in British Society Between 1800-1950," in *Advances in Intelligent Data Analysis XVI*, ed. Niall Adams, Allan Tucker, and David Weston, vol. 10584 (Cham: Springer International Publishing, 2017), 186–97; Maud Ehrmann et al., "Language Resources for Historical Newspapers: The Impresso Collection," Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020.

(1872-1949), the *Shenbao* published a staggering 256,000 issues, comprising over two million articles. These articles encompassed a wide array of content, including brief telegrams (both official and foreign), news reports, literary works, and various supplements covering topics such as automobiles, New Year celebrations, and cinema. The newspaper's structure, including its sections, evolved over time. Significant changes occurred, particularly after 1897, prior to which all texts were presented on the same level.[3] Section names were altered, and some sections even duplicated (e.g., 'National News 1' and 'National News 2'). However, in the digital version we received, these section distinctions were lost. Section names were simply recorded as 'text,' lacking the metadata elements attached to individual articles, as seen in collections like the Chinese historical newspaper collection by ProQuest or in the *Dongfang zazhi* (Eastern Miscellany).[4]

In this paper, I aim to adapt the 'Pinagot approach' to the extensive study of historical actors within the early *Shenbao* newspaper.[5] I consider the articles published in the *Shenbao* as individual files containing the life stories of previously unknown individuals. The pages of the newspaper, in turn, serve as the archive where these files are stored. However, unlike Alain Corbin, I do not begin with a randomly selected individual from the newspaper. My objective is to encompass them all. I am particularly interested in identifying the individuals who appeared in the *Shenbao* during its first twenty years of existence, especially those mentioned repeatedly. I seek to understand how these individuals were connected to each other, their affiliations with various institutions, and the nature of their involvements.

Why focus on the first twenty years? I have chosen this period deliberately to cast a wide net over a time without significant events such as the first Sino-Japanese War. My aim is to scrutinize news reporting in a newspaper still in the process of development.[6] While I employ the term 'eminent' to describe the individuals that I seek to examine, my actual interest encompasses the entire population that appeared in the *Shenbao*. I anticipate that the newspaper's pages will feature a substantial number of individuals who may not be considered 'eminent.'

For all individuals, 'eminent' or otherwise, I pose the same questions:

- Why were they deemed newsworthy? Among those who may not be 'eminent,' who were the 'non-eminent' individuals who found their way into the *Shenbao*?
- What events were they associated with? What is the dynamic between the event itself and the individual's role within it? Which aspect takes precedence?
- To which institutions were they linked or attached in the news reporting? If they held formal positions within these institutions, what were those positions, and what roles did they fulfil?

---

[3] Mittler, *A Newspaper for China ?,* 94-95

4 ProQuest Historical Newspapers™ - Chinese Newspapers Collection. We acquired the collection from ProQuest along with the data mining rights by contract. The *Dongfang zazhi* is available online but not for individual access. Institutions can subscribe, but there is no possibility of data mining.

[5] This refers to Alain Corbin's book, *The life of an unknow*, in which the author explains that he started from an archival file selected randomly by pulling a folder from the stacks. Alain Corbin, *The Life of an Unknown: The Rediscovered World of a Clog Maker in Nineteenth-Century France* (New York: Columbia University Press, 2001).

[6] In the intervening year that this study covers, China was not involved in any major international dispute, except the brief Sino-French War (1884-85). Neither were there major domestic events that could have brought the *Shenbao* to focus more specifically on such events.

My second objective in this paper is to outline a comprehensive workflow, spanning from formulating an initial query to constructing a reference corpus, then proceeding to filter and select the final dataset. Subsequently, I employ a combination of methodologies to process the raw text of the articles, facilitating the exploration and analysis of the extracted data. My ultimate goal is to pinpoint more specific research questions, identify relevant sub-datasets, and iteratively refine them until I produce data that can be subjected to various analytical lenses.

My approach aligns with the paradigm of datafication, a term describing a perspective that views elements in the world as sources of data to be systematically examined for correlations, yielding insights into human behaviour and societal matters.[7] Applied to the realm of history, datafication translates into the generation of data points extracted from historical newspapers, offering an unstructured mapping of individuals and their social connections.

In this paper, I navigate through the following steps, which are discussed more comprehensively in the hermeneutics sections:

- Constructing a dataset through automated document extraction.
- Extracting actors (individuals, institutions, etc.) and associated data.
- Identifying and disambiguating actors using linked data from four major databases (MCBD, IMH collection, Wikipedia, CGED-Q).
- Conducting statistical analysis and graph visualization of the data.
- Analysing collected newspaper articles through topic modelling.

The *Shenbao* in history

The *Shenbao* was established in 1872 by John Major (1841-1908), a British entrepreneur with extraordinary foresight and business acumen.[8] It was the first modern newspaper in Chinese language and the most important one at least up to 1905 when competitors started to emerge.[9] From the beginning, John Major targeted a Chinese readership, especially the large population of literati that concentrated in the Jiangnan region.[10] Although he was the owner and editor, Major left to Chinese collaborators the decisions to select news items and to write editorials and articles. Yet, he was also keenly involved, first because he wanted to make the *Shenbao* a profitable venture, with multiple side productions such as its illustrated companion, *Dianshizhai huabao* (點石齋畫報)[11] ; second because his objective was to create a public arena to involve the educated population in civic discussions, while at the same time introducing knowledge and information through news reporting[12].

---

[7] Marcus Burkhardt et al., eds., *Interrogating Datafication: Towards a Praxeology of Data* (transcript publishing, 2022); Slavko Splichal, *Datafication of Public Opinion and the Public Sphere: How Extraction Replaced Expression of Opinion* (London, UK: Anthem Press, 2022); Karin van Es and Mirko Tobias Schäfer, "Introduction: New Brave World," in *The Datafied Society*, ed. Karin van Es and Mirko Tobias Schäfer, Studying Culture through Data (Amsterdam University Press, 2017), 13–22.

[8] Rudolf Wagner, "The Early Chinese Newspapers and the Chinese Public Sphere.," *European Journal of East Asian Studies* 1, no. 1 (March 2001): 1.

[9] Rudolf G. Wagner, "The '*Shenbao*' in Crisis: The International Environment and the Conflict between Guo Songtao and the '*Shenbao*.,'" *Late Imperial China* 20, no. 1 (juin 1999): 107–38.

[10] Barbara Mittler, *A Newspaper for China? Power, Identity, and Change in Shanghai's News Media, 1872-1912*, Harvard East Asian Studies Monographs 226 (Cambridge (Mass.): Harvard University Asia Center, 2004), 2-4

[11] Xiaoqing Ye, *The Dianshizhai Pictorial: Shanghai Urban Life, 1884-1898* (Ann Arbor: Center for Chinese Studies the University of Michigan, 2003).

[12] Mittler, *A Newspaper for China?,* 13-14, 38

There was no previous history of journalism in China. In fact, the very profession of journalist or reporter simply did not exist. It developed at a later stage along with the emergence of modern newspapers and the establishment of departments of journalism (新聞學係) in a few Chinese universities in the 1930s.[13] Throughout the period under study, the *Shenbao* relied on literati who came to specialize in newspaper writing and informants based in the various localities in the Jiangnan region, official gazettes, and local institutions, as we shall see below. Thus, the *Shenbao* created the first matrix of professional journalism in China. In terms of content, during the first two decades, it published a mix of national, international, and local news reports, stories (fiction), translations, excerpts from official gazettes without a proper hierarchy. The *Shenbao* remained unstructured until the end of the 1890s when more discernible sections were introduced.

The history of the *Shenbao* and of its creator, John Major, is still very much in a limbo. Most studies have concentrated on its published content. In English language, a whole spate of works was produced under the impetus of and by the late Rudolf Wagner.[14] The two major studies of the *Shenbao* include Barbara Mittler's *A Newspaper for China?: Power, Identity, and Change in Shanghai's News Media, 1872-1912*[15] and Weipin Tsai's *Reading Shenbao . Nationalism, Consumerism, and Individuality in China 1919–37*[16], to which one can add two unpublished dissertations, Terry Narramore's "Making the news in Shanghai: Shen Bao and the politics of newspaper journalism, 1912-1937"[17] and Natascha Gentz's "Die Anfänge des Journalismus in China (1860 – 1911)"[18]. None of these works, however, touched on the history of the *Shenbao* as a publishing company, its operations, and its staff for lack of access to its archives.

The historiography in Chinese is not just richer, it is massive. It is not my purpose here to cover it all. I will focus solely on monographs and exclude academic papers. If we take the production of books, the current historiography amounts to 19 titles. Indeed, a significant portion of these works consists of compilations of authentic documents sourced from the *Shenbao,* mostly focused on a given region or place,[19] on an organization, e.g., the Ningbo Sojourners'

[13] Qiliang He, *Newspapers and the Journalistic Public in Republican China: 1917 as a Significant Year of Journalism*, 2018.;Stephen R. MacKinnon, "Toward a History of the Chinese Press in the Republican Period," *Modern China* 23, no. 1 (1997): 11–12. On the development of the modern Chinese press, see Hanqi Fang, ed., *A History of Journalism in China*, 10 vols. (Singapore: Silkroad Press, 2014).

[14] Rudolf Wagner, *Joining the Global Public : Word, Image, and City in Early Chinese Newspapers, 1870-1910* (Albany NY: State University of New York Press, 2007); Rudolf G. Wagner, "The Role of the Foreign Community in the Chinese Public Sphere.," *China Quarterly*, no. 142 (juin 1995): 423; Rudolf G. Wagner, "The '*Shenbao* ' in Crisis: The International Environment and the Conflict between Guo Songtao and the '*Shenbao* .,'" *Late Imperial China* 20, no. 1 (juin 1999): 107–38.

[15] Mittler, *A Newspaper for China?*.

[16] Weipin Tsai, *Reading Shenbao: Nationalism, Consumerism and Individuality in China, 1919-37* (Houndmills, Basingstoke: Palgrave Macmillan, 2010).

[17] Terry Narramore, "Making the News in Shanghai: Shen Bao and the Politics of Newspaper Journalism, 1912-1937" (Doctoral dissertation, Sydney, Australian National University, 1989).

[18] Natascha Gentz, "Die Anfänge des Journalismus in China (1860 - 1911)" (Doctoral dissertation, Heidelberg University, 1998).

[19] Guangxi Zhuangzu Zizhiqu tongzhiguan 广西壮族自治区通志馆 and Guangxi zhuangzu zizhiqu tushuguan 广西壮族自治区图书馆, *Shenbao Guangxi ziliao suoyin* 申报广西资料索引(An index of materials about Guangxi in the Shenbao)) (Nanning: Guangxi renmin chubanshe, 1992); Taiwan yinhang and Jingji yanjiushi, *Qingji Shenbao Taiwan jishi jilu (1872-1887)* 清季申報臺灣紀事輯錄 (A chronicle of events in Taiwan by the Shenbao in the Qing Dynasty) (Taizhong: Taiwan sheng wenxian weiyuan hui, 1994); Zhongjia Lin et al. 林忠佳, *Shenbao Guangdong ziliao xuanji* 申報廣東資料選集 (An anthology of Guangdong materials in the Shenbao) (Guangzhou: Guangdong sheng dang'anguan Shenbao Guangdong ziliao xuanji bianjizu, 1995); Pudong Xinqu dang'anju and Shanghai shi Pudong Xinqu wenshi xuehui, *Shenbao zhong de Pudong* 申報中的浦東 *(Pudong*

Association[20] and the 1911 Revolution in Guangxi[21]. This is clearly an enduring genre in Chinese historiography. The second largest group is made up of subject monographs that have appeared since 2000 on Sino-Korean relations[22], the advertising of Chinese classics[23], on the merchants from Cixi[24], on the relations between Qing officials and newspapers[25], on *Shenbao* reports and editorials[26]. There are also two studies that examine the literary content of the *Shenbao*.[27] Finally, the last genre consists of historical studies based on advertising materials from the 1920-1940s in the *Shenbao*[28]. Several doctoral dissertations were also written with the *Shenbao* as the main source and object of study. Like the published monographs, they usually examine a topic through the lens of newspaper reporting[29].

---

in the Shenbao), (Shanghai: Sanlian shudian, 2019; Xiangqun Li 李向群, *Jindai Xiamen lishi ziliao huikan: Shenbao jiwen* 近代厦门历史资料汇刊：申报纪闻 (Modern Xiamen Historical Data Collection: Shenbao Records), 2020.

[20] Ningbo shi dang'anguan 宁波市档案馆, *Shenbao Ningbo shiliao ji* 《申报》宁波史料集 (A collection of historical materials on Ninbo in the Shenbao) (Ningbo: Ningbo chubanshe, 2013); Hua Changhui, ed. 华长慧, *Shenbao Ningbo lühu tongxiang shetuan shiliao* 申报宁波旅沪同乡社团史料 (Historical Materials on Shanghai Ningbo native-place Associations in the Shenbao) (Ningbo: Ningbo chubanshe, 2009).

[21] Zhennan Huang 黄振南 and Qinhui Jiang 蒋钦挥, *Shenbao Guangxi xinhai geming ziliao xuanbian* 申報"广西辛亥革命资料选编 (Selected Materials from the Shenbao on the Revolution of 1911 in Guangxi) (Guilin: Guangxi shifan daxue chubanshe, 2012).

[22] Yuanhua Shi 石源华, *Shenbao youguan Hanguo duli yundong ji zhong-han guanxi ziliao xuanbian: 1910-1949* 申报有关韩国独立运动暨中韩关系史料选编 (1910-1949) (Selected Compilation of Historical Materials on the Korean Independence Movement and Sino-Korean Relations in the Shenbao) (Beijing: Renmin wenxue chubanshe, 2000).

[23] Shengdong Lin, *Zhongguo jinxiandai jingdian guanggao chuangyi pingxi: qishiqi nian* 中国近现代经典广告创意评析：<申报>七十七年 (Comments and Analysis on Modern and Contemporary Chinese Classics Advertisement Creativity: Seventy Seven years of the Shenbao) (Nanjing: Dongnan daxue chubanshe, 2005).

[24] Dizhen Xu, *Shanghai tan shiye xia de Cixi shangren: "Shenbao" sanbei shangbang shiliao jicheng* 上海滩视野下的慈溪商人：《申报》三北商帮史料集成 (Cixi Merchants viewed from the Shanghai Bund: A Collection of Historical Materials on the Sanbei Merchant Group in the Shenbao) (Beijing: Dangdai zhongguo chubanshe, 2012).

[25] Ning Lu 卢宁, *Zaoqi Shenbao yu wa qing zheng fu: Jindai zhuanxing shiye zhong baozhi yu guali guanxi de kaocha* 早期"申报"与晚清政府：近代转型视野中报纸与官吏关系的考察 (The early Shenbao and the Late Qing Government: An Investigation of the Relationship between Newspapers and Officials from the Perspective of Modern Transformation) (Shanghai: Shanghai kexue jishu wenxian chubanshe, 2012).

[26] Shuqiang Song 宋书强, Zhaolu Yin 殷昭鲁, and Feifei Zhao 赵飞飞, *Shenbao baodao yu pinglun* 申报报道与评论 (Shenbao reports and editorials) (Nanjing, Nanjing daxue chubanshe, 2019).

[27] Weiwei Pan 潘薇薇, *Cong Shenbao guanggao kan zhongguo jindai xiaoshuo yundong* 从申报广告看中国近代小说运动 (The Movement of Chinese Modern Novels seen from Shenbao Advertisements) (Shanghai: Dongfang chuban zhongxin, 2015); Hongyan Hua 花宏艳, *Shenbao kanzai jiutishi yanjiu (1872-1949)* 申报刊载旧体诗研究(1872-1949) (A study of Old Style Poetry in the Shenbao), (Nanjing: Fenghuang chubanshe 2018).

[28] Runian Wang 王儒年, *Yuwang de xiangxiang: 1920-1930 niandai guanggao de wenhuashi yanjiu* 欲望的想像：1920-1930 年代申报广告的文化史研究 (The Imagination of Desire: A Study in the Cultural History of Shenbao Advertisements in the 1920s and 1930s) (Shanghai: Shanghai renmin chubanshe, 2007); Ju'ai Pang 庞菊爱, *Kua wenhua guanggao yu shimin wenhua de bianqian: 1910-1930 nian shen bao kua wenhua guanggao yanjiu* 跨文化广告与市民文化的变迁 : 1910-1930 年申报跨文化广告研究 (Cross-cultural advertisements and the changes of citizen culture: a study of cross-cultural advertisements in the Shenbao from 1910 to 1930) (Shanghai: Shanghai jiaotong daxue chubanshe, 2011); Deming Chen 陈德明, *Yuanqu de huihuang: (1929-1949) haipai yangao mizhu zi tuji* 远去的辉煌 : 申报(1929-1949) 海派广告美术字图集 (Distant Glory: (1929-1949) A collection of Artistic Character Advertisements in the Shenbao) (Shanghai: Shanghai daxue chubanshe, 2019.

[29] Zhuo Feng 冯卓, "Qingmo minchu shenbao cihui yanjiu" 清末民初《申报》词汇研究 (A Study on the Vocabulary of the Shenbao in the Late Qing and the Early Republic)" (Doctoral dissertation, Jilin University, 2021); Yongsheng Liu 刘永生, "Shenbao de dui ri yulun yanjiu"《申报》的对日舆论研究（1931.9-1937.12）

All the work achieved until now has been based on searching for information through close reading. The longer the period or the more extensive the topic, the more likely articles will be missed, or sampling will become unavoidable. As a long-time user of the *Shenbao* myself, I can measure the difference between reading through a full decade of the newspaper to gather information on a given topic — Shanghai Municipal Government — as I have in 1982 on microfilms and doing the same today using digital methodologies.[30] It took me three months back then just focusing on the Local news section (本阜新聞). Even if I trust that I gathered a lot of relevant data to support my analysis, I must have missed elements that could have enriched my research. For my study of prostitution, I sampled the *Shenbao* to read a whole year every ten years.[31] It was only for the last stretch of my research on death in Shanghai that I was able to search the *Shenbao* digitally, albeit with the limitations discussed above.[32] Today, using a mix of computational methodologies, various keywords and word embeddings, it would not take more than a day to build a more complete corpus and to produce various datasets for analysis. This changes the conditions of historical research, and it challenges past approaches of historical newspapers. It offers the possibility to re-visit the dualism that Barbara Mittler once drew between the *Shenbao* as text and the *Shenbao* as source.[33]

Corpus building and NER extraction.

My quest for « eminent Chinese" started with a search based on two very common terms in the type of classical Chinese used in the *Shenbao*: 之 (zhi) and 也 (ye). Although I could have used a sample build randomly by an algorithmic method, there was little point to include all the possible types of texts (such as advertisements, company announcements, etc.). The main target was news articles, which the two terms were most likely to identify as they usually appeared in a "narrative". This method produced respectively 123,274 and 71,651 results (194,925) for

(A study of Shenbao editorials on Japan) (Doctoral dissertation, Hunan Normal University, 2008); Li Liu 刘莉, "Zhou Shoujuan zhubian shiqi shenbao·ziyoutan xiaoshuo yanjiu" 周瘦鹃主编时期申报·自由谈小说研究 (Research on the Novels in the "Free Talk" Section of the Shenbao during Zhou Shoujuan's Editorship) (Doctoral dissertation, Fudan University, 2010); Migming Dan 单明明, "Shenbao shiye zhong de makesi xueshuo"《申报》视野中的马克思学说 (Marxist Theory through the Shenbao) (Doctoral dissertation, Central Party School 2017); Liqin Zhang 张立勤, "1927-1937 nian minying baoye jingying yanjiu" 1927-1937 年民营报业经营研究 (Research on the Management of Private Newspapers in 1927-1937)" (Doctoral dissertation, Fudan University, 2012); Xiaokai Zhu 朱晓凯, "Shenbao yu Zhong-Fa zhanzheng yanjiu"申报与中法战争研究 (A Study of the Shenbao" and the Sino-French War) (Doctoral dissertation, Anhui University, 2015); Runian Wang 王儒年, "Shenbao guanggao yu shanghai shimin de xiaofei zhuyi yishi xingtai" 申报广告与上海市民的消费主义意识形态 (Advertisements of the Shenbao and the Consumerism Ideology of Shanghai Residents) (Doctoral dissertation, Shanghai Normal University, 2004); Bohong Xiao 肖鸿波, "Shenbao 77 nian tiyu baodao yanjiu" 申报 77 年体育报道研究（1872-1949）(A Study of 77 Years of Sports Reporting in the Shenbao) (Doctoral dissertation, Fudan University, 2011); 谢圣明, Shengming Xie, "Chuanboxue shiye xia shenbao yu zhongguo meishu xiandaihua jincheng" 传播学视野下申报与中国美术现代化进程（1872-1937）(Shenbao and the Modernization Process of Chinese Art from the Perspective of Communication Studies) (Doctoral dissertation, Zhejiang University, 2014); Xueqin Gao 高学琴, "Shenbao shehui guanggao yanjiu"申报社会广告研究 (Research on Social Advertising in the Shenbao) (Doctoral dissertation, Wuhan University, 2019).

[30] Christian Henriot, *Shanghai, 1927-1937: Municipal Power, Locality, and Modernization* (Berkeley: University of California Press, 1993).

[31] Christian Henriot, *Prostitution and sexuality in Shanghai: a social history 1849-1949* (Cambridge, UK; New York: Cambridge University Press, 2001).

[32] Christian Henriot, *Scythe and the City: A Social History of Death in Shanghai* (Stanford: Stanford University Press, 2016).

[33] Mittler, *A Newspaper for China?*, 4.

the 1872-1892 period, which I boiled down to 130,485 unique documents after removing the duplicates. It should be noted here that the work of separating articles as individual documents was not done properly by the company that transformed the print version of the *Shenbao* into a digital version. Therefore, it is not rare to find several articles in a document, which produces extra-long articles. I discuss below how I addressed this issue.

The next step was to retrieve the full text for all the unique documents and to apply NER to extract all the named entities. The results are necessarily biased by the fact that the segmentation of articles is far from ideal. The first way to obviate this issue was to exclude the extra-long articles that often corresponded to pages of advertisements, pages of telegrams, or excerpts from the official gazette (京报 Beijing Gazette)[34]. The reason for eliminating the official gazette was that the excerpts concerned mostly regulations, not news items *per se*. This may have eliminated a few relevant papers and some individuals, but given the wealth of available data, this cannot have not introduced any severe bias.

Second, because I wanted to have as much as possible individual articles, I filtered the original corpus based strictly on the length of the articles. This is based on previous work done on the length of articles in the *Shenbao* to identify and deal with the issue of article separation.[35] Although this is a gross measure, we found that generally articles with 500 or less characters are more likely to be "genuine" individual articles. There were still cases where two or more articles were compiled together, but close reading established that they provided the same type of information, as I discuss below, and their inclusion did not distort the results of the analysis of the extracted data. The filtering of all the articles with 500 characters or less produced a corpus of 87,997 articles.

The implementation of NER on the 500-character corpus produced 355,699 entities. I removed all the entities that contained only one character, which brought down the final tally to 286,134 entities. For the sake of analysis, I retained only four main entities: persons (117,028 PER), institutions (36,365 ORG), locations (29,551 LOC), and geopolitical entities (100,210 GPE). Of course, these are the raw results before initiating a verification of the content and the cleaning of unsuitable or erroneous named entities (wrong classification, false positives, etc.). The initial curation of the extracted data was done manually by reviewing the four sets of named entities. This task included the normalization of names (name variants [e.g., 上海法總, 古法領事署, 埠法領事署, 法界領事署] or character variants [e.g., 恒,恆], misspellings, etc.).

In the case of persons, I removed all the names with more than three characters. By doing so, I am aware that I removed individuals whose names were either genuine long names of Chinese (in fact names of Manchus and Mongols transliterated into Chinese characters), names of foreigners (for those who adopted a transliteration of their name rather than a genuine Chinese name), or compounded names (unseparated names, names unseparated from common words, erroneous classification). The long names represented a little more than 2,500 cases that I decided to leave aside for further processing at a later stage.

[34] Mittler, *A Newspaper for China*, 177-181

[35] Christian Henriot, "Forging Bonds and Building Factories: The Networked World of Shanghai's Industrial Elite," in *Modern China in Flux: Networks, Mobility, and Transformation* (Berlin: De Gruyter, 2024, forthcoming).

To ensure that the two-character and three-character names that I had extracted were the names of persons, which the simple metric could not guarantee, I extracted all the surnames in the sample (54,255 entries with 2,349 unique full names) and the list of 846 Chinese surnames obtained from the CGED-Q data on Qing officials in 1850-1864 and 1900-1912.[36] This removed 1,700 entries that were either text without a surname or text with a surname imbedded only in second position in the name character string. I filtered out all these entries, which left a reference file with 48,280 named individuals. In this sample, however, I had different types of names depending on the nature of the name. The dataset included in fact five populations that required specific processing:

- A. Persons with a complete name (3- and 2-character names).
- B. Persons with a complete or incomplete name that includes the 阿 (a) character.
- C. Persons with a complete or incomplete name that includes the 君 (jun) character.
- D. Persons with their surname followed by an official title.
- E. Persons named only through their title.

The 'A' population was the only one that did not pose any issues, even though there was always the possibility of two distinct individuals sharing the same full name. The number of surnames in China is very limited, especially for the most common surnames. In the CGED-Q population of Qing officials, for example, the top 10 surnames accounted for 38.3% of the records of officials with surnames. The top 20 surnames accounted for approximately half of the records, and the top 200 accounted for 95.1%."[37]

The population that came closest to A was the C population with names that included the 君 character, which can roughly be translated as "sir". It is a courtesy formula to designate a male person. Yet there were three naming configurations. Generally, the 君 character was inserted between the surname and the given name and all it took was to remove it to obtain a neat full name. Sometimes, however, the character was added at the end of the full name. There was no trouble removing it for three-character names, but when there was a string of three characters (Surname + character + 君), this was more troublesome because the 君 character could also be part of the given name. All such cases had to be processed by manual verification in the textual context. Finally, there were cases where the 君 character appeared with just a surname. In most cases, this was due to another mention following the initial mention of the full name. Yet, there were also cases when there was no mention of the full name. I kept these cases in the sample.

The D population raised a challenge that I was able to solve only partially. Since the information provided only a surname and a title (e.g., Governor Wang), the objective was to identify the person based on his title and the textual context, and to find the full name. The degree of difficulty was proportional to the rank: the lower the rank, the more difficult to identify the individual. For example, there was no issue in identifying all the governors general or viceroys (zongdu) who ruled over a province or a pair of provinces (liangguang, liangjiang, etc.). The Shanghai county magistrates or the Shanghai or Customs *daotai* (circuit intendant) were also easy to identify. Yet, many other similar officials were mentioned, who were administering other areas in the Jiangnan region or even further away in China. I extracted all such cases in their larger textual context to find a mention of the full name. Yet in most cases,

[36] I am especially grateful to the Lee-Campbell Research Group at Hong Kong University of Science and technology for sharing their dataset of Qing officials. Campbell, Cameron Dougall; Chen, Bijia; Ren, Yuxue; Lee, James, 2019, "China Government Employee Database-Qing (CGED-Q) Jinshenlu 1900-1912 Public Release", https://doi.org/10.14711/dataset/E9GKRS, DataSpace@HKUST, V1.

[37] Cameron Campbell and Bijia Chen, "Nominative Linkage of Records of Officials in the China Government Employee Dataset-Qing (CGED-Q)," *Historical Life Course Studies* 12 (September 8, 2022): 239.

the *Shenbao* referred only to the surname and the title. I kept these names in their original format, even if there is a limited risk of having more than one person under the same name. This practice was very common: Ge Fanfu, a magistrate of the Mixed Courts in the city was mentioned under his full name in 300 articles, but in 1,944 articles he appeared only under his surname and action (tongzhuan 同轉).

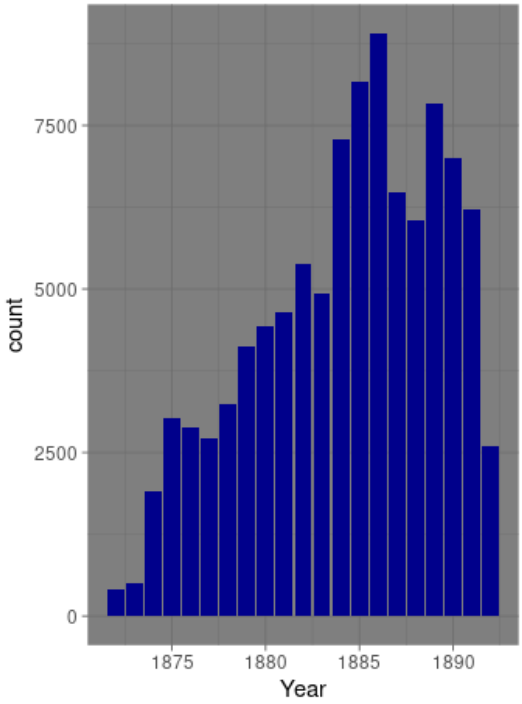
The B population was undoubtedly the most troublesome one for several reasons. The inclusion of the 阿 character tends to point to individuals, both men and women, that belonged to the common people. Elite males in China had distinctive characters for their personal name, which counterbalanced the narrow range of surnames. The very fact of having just one character as an individual marker increased the chance of different people sharing the same surname considerably. This is compounded by the fact that the second character was often a common character or simply a number. As we shall see later, due to the type of news in which these individuals appeared, some even appeared without a surname. Because NER extraction itself sometimes separated a given name from its related surname, I did a systematic search through concordance to identify such cases and to reconstruct the full name. Yet, even after this verification, I ended up with many names that corresponded to different individuals (like having a dozen John Smith in an English newspaper). The only possibility to sort out these individuals could be to use time (date of publication) and the related institutions, albeit with an imperfect result.

After removing the duplicates that resulted from joining the articles and the list of validated names from the five name files and further data cleaning, I was left with 50,525 unique documents that contained 57,950 unique names. From the discussion above, it should now be clear that what we have in the articles are the representations of individuals (physical persons), not concrete individuals. In a number of cases, though quantitatively and qualitatively limited, there will be several distinct physical persons under the same name. It will appear later that there are ways to disambiguate such cases, though only partially. This is not the purpose of this paper, but the methodology that we implemented in data analysis will provide clues about solving these issues. Let us just say here that the ambiguous cases do not alter or distort the analysis and the results.

A macroscopic analysis of the data

The distribution of the 50,525 unique documents in the corpus shows a pattern of regular increase through the twenty years under study. The first two years are clearly underrepresented, which may reflect the nature of articles during the first two years (more reliance on official gazettes that I have excluded). This increase may also be linked to the number of pages in the newspaper. The number of articles became more stable after 1884, despite irregularities. An analysis by type of names failed to reveal specific patterns. We can only note that the *Shenbao* reported on the same types of population in the first twenty years of its existence.

Figure 1. Distribution of the selected articles by year.



There was of course a huge inequality among the 57,950 names that appeared in the news reporting of the *Shenbao* in terms of mentions. The vast majority — 45,313 — appeared only once. In fact, only 3.1 percent of all names appeared more than five times in the *Shenbao*, and only 0.5 percent appeared twenty times or more. Since our study covers a period of twenty years, it means that even if the distribution of mentions was not even throughout the period and some mentions may have been concentrated in a short period, it is obvious that the historian is likely and unluckily facing a huge population of nobodies for the most part. It is also obvious that only those with more than mentions, if they were concentrated during a given period, referred to individuals with a certain degree of eminence. The truly eminent ones, however, were those in the list of names mentioned more than twenty times, even if a few are artificial statistical constructs as we shall see later.

Table 1. Distribution of names by frequency in the *Shenbao* (1872-1891).

Mention	Number	Percentage
1	45313	78,2%
2	6632	11,4%
3	2247	3,9%
4	1087	1,9%
5	587	1,0%
6-20	1791	3,1%
>20	293	0,5%
Total	57950	100,0%

The two most important figures were local officials, Mo Xiangzhi (莫祥芝) and Pei Dazhong (裴大中), both Shanghai county magistrates, respectively in 1876-1879 and 1887-1890. We

can add to this roster another Shanghai county magistrate, Li Guangdan (黎光旦) who served in 1882-1884. In-between, we find Li Fuxiang 李傅相[Li Hongzhang] who was by then at the peak of his career. In 1870, Li became Viceroy of Zhili and Beiyang Trade Minister until his death in 1901. He was a major national figure who wielded considerable power in foreign policy. Although he was no longer in the Jiangnan area, he kept an eye on the enterprises he had launched and, moreover, he continued to play a leading role at the national level as Viceroy of Zhili and Beiyang Trade Minister. As we go down the list of names, it becomes more difficult to identify individuals without much research. Li Boxiang (李伯相) was a high-ranking official who worked alongside Li Fuxiang in foreign affairs. Gong Yangqu (龔仰蘧) and Yuan Haiguan (袁海觀) were both Qing officials, the latter a Shanghai daotai. Cai Eryuan (蔡二源) was a would-be official who gave up officialdom and converted to business, while Liu Shengsan (劉省三) was a military official. Ge Fanfu (葛蕃甫) and Song Eryi 宋二尹 served as Chinese judge assessors at the French or British Mixed Court. They were mentioned exclusively in relation to judicial decisions. Wang Rongpei (王榮培) was a Chinese detective (baotan 包探) attached to the British Mixed Court. Kong Yin (孔殷) remains a mystery at this stage. The name *fangbo* (方伯) is not an actual person, but several persons. It was the title for the provincial treasurer (布政使) that was simply disconnected from the name of the individuals who held the position. The reader will have noted that I left aside a few names, such as Chen Ajiu (陳阿九), Wang A'er (王阿二) or Wang Asan (王阿三), even if they ranked high because there are typical examples of names that each represented several individuals by the same name.

Table 2. The most frequent names of individuals in the dataset (1872-1891).

Name	Mentions	Name	Mentions	Name	Mentions
莫祥芝	620	王榮培	102	阿昌	66
裴大中	605	王阿大	99	嚴佑之	63
李傅相	484	阿公	93	張阿二	63
陳阿九	341	張香濤	91	王阿金	63
王阿二	245	秦少卿	90	張阿福	62
方伯	218	龔仰蘧	87	楊石泉	62
李伯相	180	嚴頌眉	83	張阿三	61
黎光旦	172	阿金	82	祝融	61
宋二尹	147	嚴佑	80	阿四	60
葛蕃甫	136	朱森庭	80	曾沅圃	59
王阿三	133	王阿四	77	曾襲侯	56
陸元鼎	133	王洪緒	72	柴樛	55
袁海觀	128	陳竹坪	72	葛同轉	55
顧阿六	124	沈仲復	71	波臣	54
康阿順	121	孤拔	70	曾文正	53
劉省三	116	施少欽	70	李鴻章	53
阿三	115	張皇	69	阿富	53
蔡二源	114	宋莘樂	68	劉坤一	52
蘇垣	109	阿寶	68	羅少耕	52
阿二	109	王阿福	67	彭雪琴	51
孔殷	104	張香帥	66	何瑞福	50

The 57,657 names of individuals in the dataset were linked to 5,967 organizations. The initial number was much higher, but this was due to name variants. The main reason was the instability of the terms used to designate institutions, in particular foreign institutions like consulates, the mixed courts, companies, etc. This is a recurrent difficulty in all sources and throughout the life of the *Shenbao*. The standardization of names was done through a mix of automatic correction and manual curation. Some organizations were of course more prominent than others. I selected the organizations mentioned more than 100 times in twenty years (Table 3).

Table 3. The most frequent organizations in the dataset (1872-1891).

Institution	Mentions	Institution	Mentions	Institution	Mentions
英界會審公堂	940	英總領事署	114	法國公司	72
法界會審公堂	621	保甲總局	109	滬北棲流公所	72
招商輪船局	473	字林西字報	103	兩江總督	69
工部局	208	文報局	102	江海關	66
上海縣署	207	英租界巡捕房	98	吏部	64
仁濟醫院	160	洋藥釐捐局	94	虹口巡捕房	59
西字報	153	翰林院	81	上海北市絲業	58
巡防總局	135	香港西字報	80	法廷	57
法國租界巡捕	133	同仁醫院	78	都察院	54
太古洋行	132	江南製造局	78	中國朝廷	53
怡和洋行	124	公董局	73	三菱公司	52

The first obvious observation we can make is the overwhelming dominance of public institutions and, secondarily, of a few prominent trading and shipping companies. The two institutions that top the list, which we could merge, are the Mixed Court in the International Settlement (940) [英界會審公堂] and in the French Concession (621) [法界會審公堂]. Together, they represent ten percent of the total number of mentions in the *Shenbao*. The mixed courts basically addressed issues of social disorder in the foreign settlements that involved disputes among the Chinese and disputes between Chinese and foreigners.[38] The magnitude of this judicial presence already tells us that the *Shenbao* relied on these institutions as a source of information in the first twenty years of its existence. In the same order of idea, the office of the county magistrate (上海縣署) ranks next, although we may also assume that not every mention concerned a legal dispute. In fact, the same pattern can also be inferred from the high number of mentions of the police bureaus, police stations, *baojia* bureau, etc. that for the most part concerned issues of social order. This is a topic that I explore further below.

Apart from the courts and the Shanghai County (xian) office, the next most important institutions were the Chinese Police bureau (巡防總局) (135), the Police bureau of the French Concession (法國租界巡捕房) (133), the British consulate general (英總領事署) (114), the Baojia Bureau (保甲總局) (109), the Police bureau of the International Settlement (英租界巡捕房) (98), the Bureau for the Management of Opium Contributions (洋藥釐捐局) (94), and the Hanlin Academy (翰林院) (81). Next came mostly ministries (吏部,都察院), police stations, and other local or regional administrations (牙釐總局,兩江總督). The central place of these public institutions in the news reporting of the *Shenbao* tends to highlight an initial focus on the action initiated by official authorities, be they Chinese or foreign. It also reflects the structure of power in Shanghai where most of these institutions were located.

Outside the realm of public institutions, the most important organizations were a single shipping company, the China Merchants' Steamship Navigation Company (hereafter CMSNC) (招商輪船局) (680). Its high number of mentions, however, is most likely related to notifications about the departure or arrival of steamships. Further below in the same category, we find John Swire and Sons (太古洋行) (132), Jardine Matheson Company (怡和洋行) (124), the Jiangnan Arsenal (江南製造局) (which was a state-run facility) (78), and the French

[38] Thomas B. Stephens, *Order and Discipline in China: The Shanghai Mixed Court, 1911-27*, Asian Law Series (Seattle: University of Washington Press).

Messageries Maritimes (法國公司) (72). Perhaps more surprising was the presence of major charity organizations such as the Shantung Road Hospital [仁濟醫院] (160), the St. Luke's Hospital [同仁醫院] and the Sinza Refuge [滬北棲流公所] (171). The two hospitals operated independently as out-patient clinics, but the Sinza Refuge was linked to the county magistrate and increasingly became a place of confinement tied to the Mixed Court of the International settlement.[39] I did not include in the previous discussion three organizations, despite their very high score. They include two newspapers, respectively "Western newspaper" and North China Herald and a "Telegraph bureau". The *Shenbao* did not mention them as institutions *per se*, but as sources of information. In fact, as was the case for persons, most organizations appeared only once (4,413) or twice (786), 3-5 times (429) or 6-9 times (146). Altogether, the lower mentions represented 96.8 percent of the population of organizations.

To refine my analysis, I applied a two-level typology to organizations. Type 1 is based on the full name of the organization and on the one or two main defining characters at the end of the name. Certain terms create a confusion, such as "Bureau" (局) which at the time designated either a public institution, a sub-level in a public institution, a charity organization, or a civic organization (reconstruction bureau). The typology paints a diverse and detailed landscape. Indeed, bureaus and mixed courts top the list, but we can see that among business ventures, foreign good firms (from small outlets to import-export companies), shops, and banks were quite prominent, along with shipping companies and general companies (公司). There is a strong component of Chinese administrative bodies (prefecture, circuit intendent, departments, courts, etc.). Although some were mentions of industrial companies abroad, there was a good crop of factories and workshops in Shanghai and other locations in China. There was a more limited presence of establishments of leisure and entertainment. Finally, whereas we found a high number of temples, there was hardly any mention of churches.

Table 4. Typology of organizations mentioned in the *Shenbao* (1872-1892).

Type 1	Number	Type 1	Number	Type 1	Number
Bureau	2820	Army	462	Telegraph	176
Mixed court	1643	Hospital	336	Department	132
Foreign goods company	1004	Consulate	324	Court	121
Shop	926	Prefecture	319	Association	94
Newspaper	758	Government	302	Opium den	92
Police	686	Xian	302	Customs	76
Charity	610	Gongsuo	281	Entertainment	70
Office	584	Ministry	256	Baojia	61
Temple	572	Workshop	253	Bang	46
Company	545	Daotai	251	Publisher	41
Shipping	545	Bank	219	Post	39
Academy	510	Viceroy	217	French Municipal Council	27
Camp	502	Shanghai Municipal Council	209	Militia	21

[39] On the Sinza Refuge, see Qiuyun Lin 林秋云, "Bianzhi de Cishan: Wanqing Hubei Shuliu Gongsuo Chutan 變質" 的慈善: 晚清滬北棲流公所初探 (A 'Metaphor' of Charity: Preliminary Study of the Sinza Refuge in the Late Qing Dynasty)," *Qingshi Yanjiu (The Qing History Journal)*, no. 4 (2017): 84-98.; Kōsuke Takahashi 高橋孝助, "Kohoku seiryu guzo no seiritsu -- Shanhai sokai no zendō" 滬北棲流公所の成立--上海租界の善堂 (The Establishment of the Sinza Refuge in the Shanghai Concession), Bulletin of Miyagi University of Education 宮城教育大学紀要 第 1 分冊, 人文科学・社会科学, no. 19 (1984): 261-278-; E.S. Elliston, *Ninety-Five Years a Shanghai Hospital, 1844-1938 : Chinese Hospital, Shantung Road Hospital, the Lester Chinese Hospital*, n.d.

If we examine the higher-level typology (Type 2), we find of course a very high number of public institutions (6,369), mostly Chinese but also foreign organizations established in China, along with a very substantial number of foreign institutions (diplomatic representations, foreign governments, etc.) (2,150). Military organizations (troops, camps) were counted separately due to their nature (964), but they also belonged to the realm of public institutions. The other main types included business ventures (everything from shops to companies) (3,749), civic organizations (guilds, cultural associations, etc.) (1,969) and educational institutions (512). There were additionally 758 "sources" (all forms of Chinese and foreign newspapers) and 275 entities that I was unable to categorize from their sole name.

This preliminary study of the named entities in the *Shenbao* gave us some initial clues about who and what was present in the pages of the newspaper. It also raises questions about why certain individuals and organizations were so visible, why so many names of individuals appear with such a limited number of organizations, what the nature of the organizations tells us about the population, and how we can explore the nature of the relations between these individuals and these organizations. To this end — connecting individuals and institutions — I chose to apply network analysis to the dataset.

The web of actors in the *Shenbao*

Network analysis is a method that visualizes and measures the links that exist between nodes in a network. In the present case, I built an affiliation network (two-mode network) that comprised two types of nodes, persons, and organizations. The edge list represents the connections between the persons and organizations that appeared in the same article. My purpose is to situate the persons in the web of relations that they formed through their connection to the same institutions. The nature of the relation is not known *a priori*. Since I study the mentions of named entities in a two-mode network, I do not consider the links as relationships in the true sense of the word. These links only reflect the encounter of a person or persons with an institution within a news report. In some cases, this may be a genuine relationship, such as a position in the institution, but in most cases, as we shall see, the mentions reflect a form of involvement due to specific circumstances. In the analysis below, I rely only on the number of degrees as a measure of the relative importance of the nodes.

The network built from the dataset is very large, with 33,805 edges and 25,837 nodes. It is made up of one large main component and 1,568 small components. A cursory examination of the small components reveals four main patterns: small ego-networks around an institution (shop, charity, bureau, etc.), double ego-networks (nearly the same people connected to two institutions), contiguous networks (not star-like but street like networks), and dyads. In this paper, I chose to focus my analysis on the main component of the network with 20,524 nodes and 29,797 edges. It is still a very large network with a diameter of 21 and a very low density. Each node in the network is connected on average to three other nodes (2.904 neighbours). The distribution of nodes by degree is largely influenced by the extreme cases represented by the mixed courts. This pattern of degree distribution reveals a widespread distribution of news items that appeared only once and focused on a singular group of people.

To analyse the network, I applied the pruning method that consists in removing nodes based on the number of degrees to assess the level of coherence of the network and to highlight the

core elements of the network.[40] The range for degrees goes from 1 to 2,097 (British Mixed Court 英界會審公堂). I removed the nodes through successive decreasing threshold:

Table 5. Pruning table.

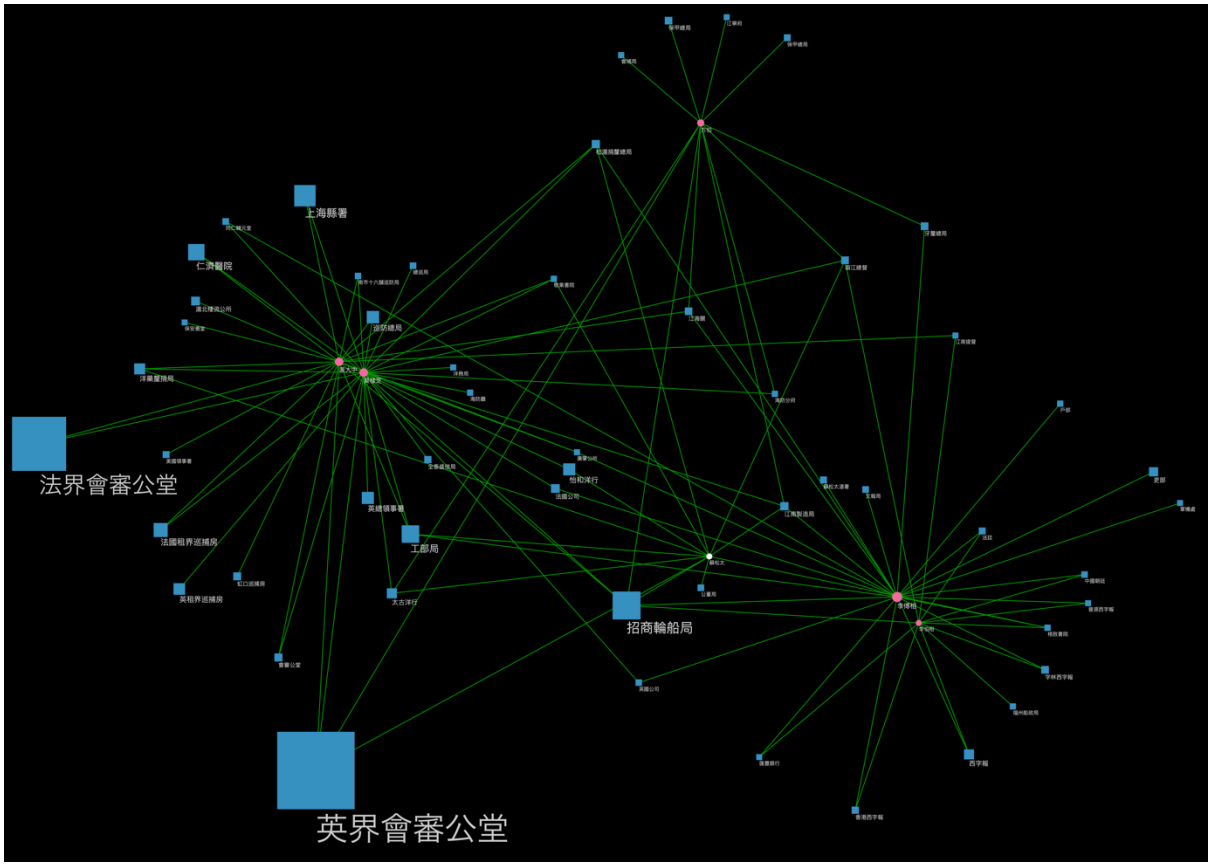
Degrees	Nodes	Edges
Main component	20,646	30,131
2 degrees and more	9,256	18,741
5 degrees and more	1,636	4,564
10 degrees and more	521	1,798
20 degrees and more	197	727
40 degrees and more	81	204
50 degrees and more	60	107
80 degrees and more	32	52
100 degrees and more	26	39

Source : EminNet5

The pruning of degrees alters the network substantially. When all nodes with only one degree are removed, the network shrinks almost by one half (-45 percent for degrees) and slightly less in terms of edges (-38 percent). At five degrees, the network shrinks further on all accounts. The number of nodes drops to 1,636 (-91.1 percent) and edges drop to 4,564 (-84.9 percent). At ten degrees, only 2.5 percent of degrees and 6 percent of edges remain in the main component. The same rate of decrease happens at each new step of pruning. At 40, we can see the core structure of the network with four main groups of individuals who are connected to a diverse range of institutions, although public institutions clearly dominate. This is basically the same after further pruning. Only the number of institutions decreases. The network retains a high level of consistency through the pruning process. At the highest number of degrees (100 or more), it is made of 26 nodes and 39 edges. At 100, only three individuals remain with 23 institutions that include mostly public institutions, except for five business ventures and two charities.

[40] Levine, Marilyn, "Revolutionary roads: An Integrative Analysis of Utilizing a Chinese Biographical Database", Cécile Armand, Christian Henriot, and Huei-min Sun, eds., *Knowledge, Power, and Networks. Elites in Transition in Modern China* (Leiden: Brill, 2022), 181-230.

Figure 2. Affiliation network of persons and institutions with 50 degrees or more.



Source: EminMCd50

What we can see through the two-mode network is not just the same statistical results that I discussed previously, we can see the connections between the individuals (or their representation) and the institutions. We can also see that only a very small number of individuals connected these institutions, mostly due to their positions and role in public affairs. The institutions involved are not the same. Li Fuxiang (李傅相) appears in exclusive relation with the Ministry of personnel and the Opium Bureau. The institutions that he shares with the Shanghai county magistrate were all related to shipping and shipyards, his long-time concern, and the Shanghai Municipal Council. The two county magistrate, Pei Dazhong (裴大中) and Mo Xiangzhi (莫祥芝) share eleven institutions and are mentioned exclusively in relation with four and three institutions respectively. Pei appears in relation to the Shanghai Municipal Police, Jardine Matheson, and the Sinza Refuge, whereas Mo seems to deal more with the British consulate, the Liangjiang viceroy (兩江總督), and the Swire & Sons Company. I excluded from my analysis all the newspapers that appeared in the articles as a source of information. They were significant, but they implied no connection to individuals as institutions.

The 50-degree network confirms Li Fuxiang's (李傅相) connections to the higher level of the Qing state and its bureaucracy, as well as with regional officials. Li appears in this network with connections that parallel those of Li Boxiang (李傅相), his long-time affiliate. The provincial treasurer (方伯) appears connected almost exclusively to Chinese public institutions

at the local level, except for links to the Swire & Sons Company and the Mixed Court of the International Settlement. The two county magistrates appear here in six fields of activity: police (both Chinese police, French police, and Shanghai Municipal Police), diplomacy (British and American consulates, Bureau of foreign affairs 洋務局), charity (four organisations), justice (mixed courts, xian court), local administration (Shanghai Municipal Council, Jiangsu governor), and business (Jardine Matheson, Swire & Sons, CMSNC). For the latter, however, it is still unclear at this stage why there was such a connection.

Network analysis provides insights in particular groups. These groups are not natural groups. They are groups defined by the typology that I used to categorize institutions. The hypothesis behind such an exploration is to examine whether certain groups had a higher degree of connection as documented by the *Shenbao*. In many cases, the network simply does not exist. This is the case, for example, for the guilds and *gongsuo*. Although one can extract a network with 396 nodes and 358 edges from the general network, there is no main component but a collection of 56 small components. The largest sub-networks are the ego-networks of the Ningbo Guild (四明公所), the Hunan Guild (湖南會館) and Hubei Guild (湖北會館) that share a few individuals, the Anhui Guild (安徽會館), and the Shanghai Silk Guild (上海北市絲業會館). In sum, in the news reporting there appears to be no connection between the various guilds.

The same is almost true for Charity organizations. They seem to form a large network with 1,516 nodes and 2,500 edges, but it is made up of 34 components. The main component only has 311 nodes and 358 edges with limited connections between the various ego-networks of charities. A few individuals play the role of brokers between these organizations, such as Yan Youzhi (嚴佑之) for the Renji shantang and Yangzhou chouzhen gongsuo (仁濟善堂/揚州籌賑公所); Li Guangdan (黎光旦) for the Bao'an shantang, Renji shantang, and Tongren fuyuantang (保安善堂/仁濟善堂/同仁輔元堂); Shi Shanchang 施善昌 for the Hubei Silk Guild Fundraising Office, Shanghai Fundraising Office, and Renji shantang (北市縣業會館籌賑公所/上海籌賑公所/仁濟善堂), etc. In several cases, the connecting person is an official (莫祥芝, 裴大中). In other cases, what constitutes a network at first sight quickly dissolves when criteria like the number of degrees are applied. This is the case for the nodes categorized as "Military institution". They form a network with 84 components. Its main component has 659 nodes and 959 edges. In this network, apart from the CMSNC that occupies a central place in terms of degree, the main nodes are the French army (法軍), the Chinese Army (華軍), the Shanghai Right Camp (上海提右營) (Chinese), the Liangjiang viceroy, and the Changjiang River Squadron (長江水師). Yet, with a filter set at 5 degrees, the network shrinks to 65 nodes and dissolves into 31 components.

The mixed courts network is one of the largest with 3,476 nodes and 3,711 edges, but they form a main component that consists in two gigantic ego-networks. Apart from the two main nodes, British Mixed Court (英界會審公堂) and French Mixed Court (法界會審公堂), we also find mentions of undefined Mixed Court (會審公堂) and the joint mention of the two mixed courts (英法會審公堂). The network includes almost no other institution directly connected to the mixed courts (e.g., German Mixed Court 德國領事公廨, two undefined French institutions 法公 (probably Mixed Court) and 法會 (unclear), a Chinese shop 椿如莊, a Chinese army 曾制軍, and a circuit 常鎮道. In other words, the news reporting about the mixed courts was all about legal affairs that involved only individuals who were linked to other institutions. Indeed, if we include the individuals and institutions at step 2 (two degrees from

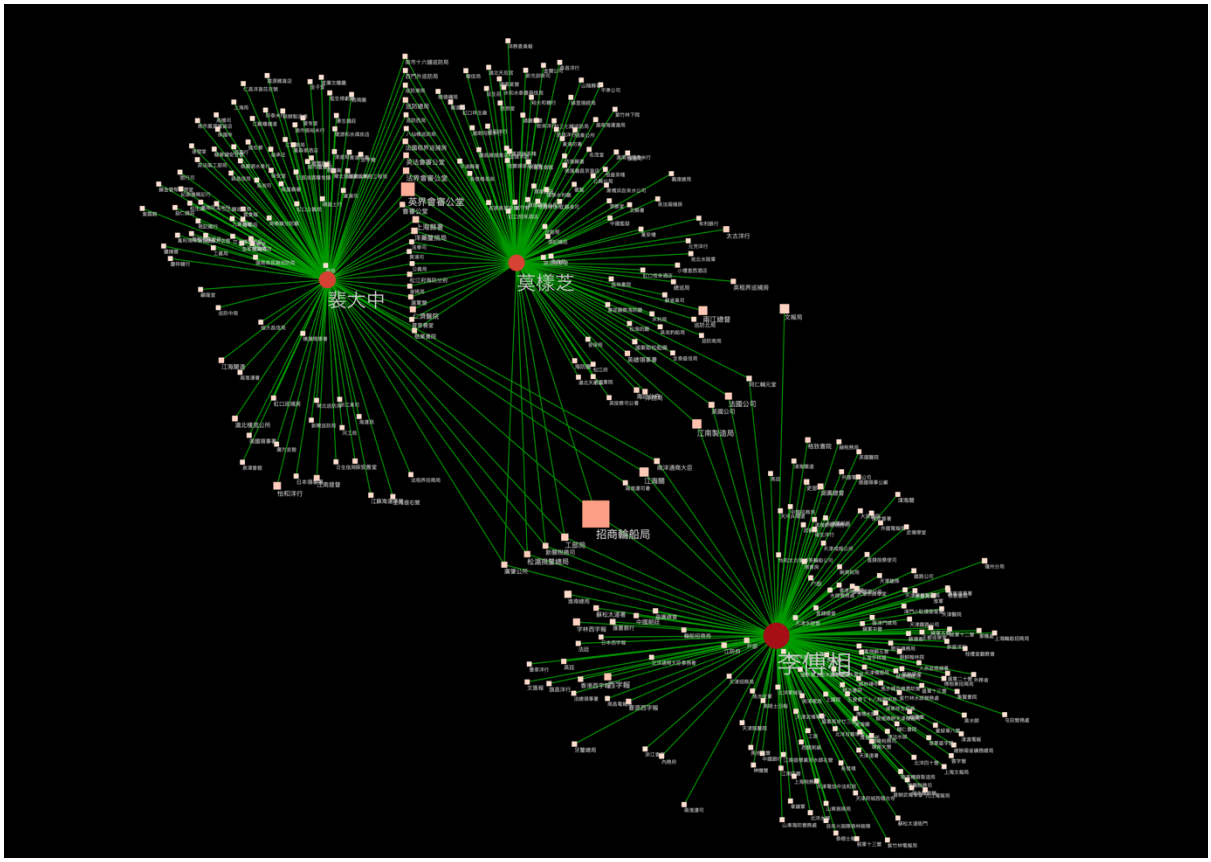
the mixed courts), the network expands to 4,925 nodes and 7,767 edges, in both cases 25 percent of the degrees and edges of the main component.

The Shanghai *xian* office network (763 nodes and 1,375 edges) presents a different profile. In this network, the most important nodes are institutions, especially the police bureaus and again the two mixed courts. The most prominent individuals are of course the county magistrates and other officials who served at the mixed courts. Business ventures appear in number (188), but they represent separate affairs with only a few that are connected through Chinese officials. If I take out the *xian* office node, the network splits into 16 components, and if I take out again Pei Dazhong and Mo Xiangzhi, the network splits into 127 components. Clearly, there was an obvious triangle between the *xian* office and the two main magistrates that connected most of the institutions and individuals in the network. This also confirms the nature of the news where people appeared when they encountered the local authorities. It is significant that Li Fuxiang disappears from this network once the *xian* office is removed.

The last example I want to address are companies (business ventures) and the network that they formed (2,993 nodes and 4,445 edges, with 177 components). Within this network, I examined more specifically the foreign goods companies, a subset of 1,914 nodes and 2,046 edges, with 76 components. The main component consists of 1,491 nodes and 1,690 edges. It constitutes a very coherent network in which the central nodes are the main trading companies, Jardine Matheson, Swire Company, and Sons, and the French Messageries Maritimes. Secondary nodes include the David Sassoon, Sons & Company (沙遜洋行), J. P. Bisset & Company (長利洋行), Shewan, Tomes & Company (旗昌洋行), David Sassoon & Company (老沙遜洋行), G. C. Schwabe & Company (公平洋行), Cowie & Company (高易洋行), and Siemssen & Company (禪臣洋行). In this network, individuals are much less prominent, even if they make the connections between the companies. The highest degree for an individual is 8 (馬爾沙, Foreigner, British, probably Marshall), but a total of 40 individuals have a degree of four or more. In other words, the foreign goods companies exhibit a fair level of connections in the news reporting of the *Shenbao* that goes beyond the realm of justice and commercial disputes.

I chose to concentrate my analysis on the individuals that were more important in the entire network. First, I selected the edge list of the individuals that received 20 or more mentions in the *Shenbao*. This approach contributed to removing the cases of institutions with a high degree, but weak links with most individuals in their network. The network initially had 1,406 nodes and 2,286 edges after removing all the A-names (阿) who are not real unique individuals. The three dominant figures are Mo Xiangzhi, Pei Dazhong, and Li Fuxiang. Their pattern of connection shows how much their positions and functions shaped their network. This is obvious for the two county magistrate, Mo Xiangzhi and Pei Dazhong, who share a great number of institutions linked to policing the city, adjudicating judicial matters (Mixed courts, *xian* office), and handling tax levies. Li Fuxiang is connected mostly to the county magistrate through a limited range of institutions, much of it shipping companies, especially the CMSNC, and the Jiangnan shipyard. Each individual, however, is connected to a much larger set of institutions that are unique to each. The network of these three individuals represents 373 nodes (28 percent of the total) and 409 edges. Beyond these dominant figures, we also find a few other officials (Cai Eryuan [蔡二源], Li Guangdan [黎光旦], Provincial Treasurer [方伯]) but also a police detective (Wang Rongpei 王榮培), a British Mixed Court official (Luo Shaogeng 羅少耕), and four merchant philanthropists (Shi Shaoqin 施少欽, Shi Shanchang 施善昌, Yan Youzhi 嚴佑之, Li Qiuping 李秋坪).

Figure 3. Affiliation network of persons and organizations with a degree above 20.



Source: Emin20MCnoa3topmen

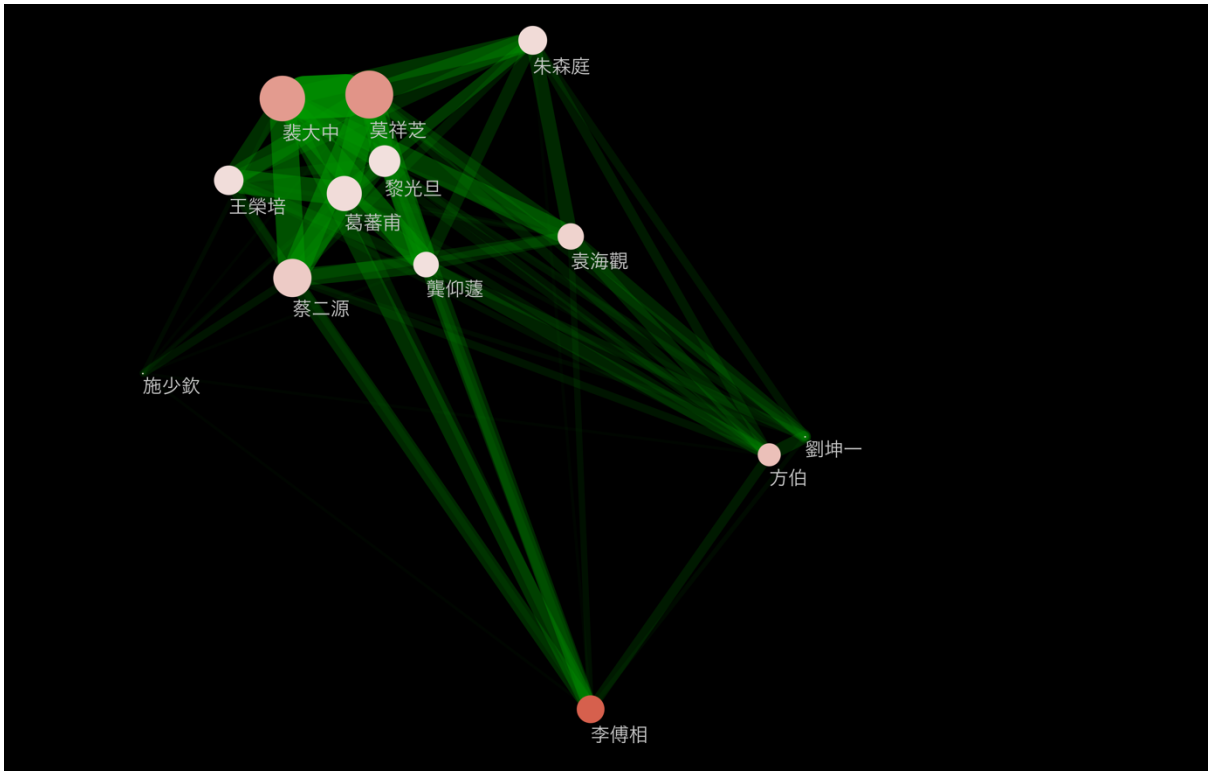
Public institutions are extremely present in this network with 491 nodes (37 percent) and their distribution is quite widespread among all individuals. Business ventures are also quite diffuse, though with a lesser concentration. Some individuals are not or not much connected to business ventures (Li Yiqing 李亦青, Wu Bozhuang 吳柏莊, Zhang Lianquan 張廉泉, Yang Shiquan 楊石泉 Chen Xuyuan, 陳煦元). We can assume that these individuals were officials as their other connections show. Military units (army, camp, etc.) are the next most significant type that exhibits a clear pattern of concentration around Li Fuxiang and the provincial treasurer (fangbo). On the opposite, police are mostly concentrated around the county magistrates and judicial officials. The other types of institutions (charities, civic associations, or workshops do not appear much in this network and do not present any particular pattern.

I selected the individuals with a degree equal or superior to 40. I assumed that the most highly connected individuals represented a nexus of power due not just to the number of ties, but to the nature and variety of institutions they were connected to. This network has 853 nodes and 1,247 edges. It includes 28 individuals who are connected to 825 institutional nodes. Public institutions are overwhelmingly represented in this network (46 percent if we include military institutions). The next two most important category of nodes are business ventures (26 percent) and civic institutions (13 percent). This defines the realm of action of the most central individuals who often shared the same qualities (high officials). The most central individuals are all connected to most other individuals, except Shi Shanchang (施善昌) or Yan Youzhi (嚴佑之) local philanthropists, and less with Wang Rongpei (王榮培), a police detective） and Ge Fanfu (葛蕃甫), a Mixed court magistrate. Civic institutions are quite dispersed, with

individuals that have no or few connections in this realm (She Zhongfu 沈仲復, Li Boxiang 李伯相, Qin Shaoqing 秦少卿, Gong Yangqu 龔仰蘧). Business is also quite diffuse, with a clear concentration around a small number of individuals. This highlights a pattern of connection to officials in their judicial capacity, which also means that the reason for their presence on the pages of the *Shenbao* was most likely a commercial dispute, not a common interest of any kind. We observe the highest concentration around Mo Xiangzhi (莫祥芝), Pei Dazhong (裴大中), Li Fuxiang (李傅相), Ge Fanfu (葛蕃甫), Qin Shaoqing (秦少卿), Song Eryi (宋二尹). It is interesting to note that Li Fuxiang appears mostly in connection with shipyards, shipping and railway companies, fields in which he was involved throughout his life. Shops, foreign goods companies, and guilds display no particular pattern.

In the next section, I transform the affiliation network for eminent Chinese (two-mode network) into two one-mode networks with weights. In the person-to-person network, I selected only the individuals with a weight of at least two and extracted the main component from the 256 connected components. The high number of components in the initial network points to the fact that some individuals appeared only in a single news item or in the same news type, limited to one or very few articles. This also confirms the absence of connection between many individuals who appeared in the *Shenbao*. The complete network, however, still included too many individuals to make it a relevant level of analysis. I further filtered based on two criteria, either the weight (above 15) or betweenness centrality (above 0.025). There was no significant difference as the same main figures — officials — appeared in both networks. The only difference was the presence of philanthropists in the network based on betweenness, which points to the specific role that these individuals played in connection with not just charities, but with official institutions that handled the poor.

Figure 4. One mode network of individuals with a weight above 15.



Note : On this graph, the width of edges is proportional to weigh, the size of nodes is proportional to degree centrality, the colour of nodes reflects betweenness centrality.

I implemented the same approach to organizations. I selected only the organizations with a degree equal or above 2. The initial network was made up of 313 components. After removing the newspapers (considered as a source), the main component had 1,591 nodes and 5,934 edges with most of the data from the initial network. I used pruning to delineate increasingly readable networks. I selected the nodes with a weight of 20, which produced a network with 35 nodes and 399 edges. At this level, the network displayed the core structure of institutions that the previous analysis of the affiliation network had established. Except for two charity organizations and two trading companies, the network included only public institutions or state-run enterprises. It also shows that the strongest connections existed between the two mixed courts and between them — especially the Mixed Court of the International Settlement— and the Shanghai Merchants' Steamship Navigation Company. At the next level (weight = 40) there remained only one charity organization, the Shantung Road Hospital [仁濟醫院] and one shipping company (CMSNC). All the other nodes were institutions of justice, police, or municipal administration. To make the story short, one-mode networks of this size are difficult to read unless one goes up the weight measure to trim down to the bare bones almost. But in this case, we do not learn much more than what we learned from the analysis of the two-mode network through pruning.

Finally, I applied PCA analysis on the one-mode network of individuals. The PCA analysis was distorted by the presence of names with very high scores that did not correspond to genuine distinct individuals. I chose to remove all the a-names that tended to bias the calculation of indices, even if they did not represent a unique individual. The most prominent individuals that we found in network analysis are situated in the same order and place in the PCA graph, which translates the fact that their centrality measures all combined separate them quite distinctively from the rest of the individuals. They received practically the same high measures in all three centralities in dimension 1 of the PCA. Li Fuxiang is the only one that presented a different profile due to his betweenness centrality, which points to his being mentioned much more than the others with links to a wider range of institutions. This contributed to make him a broker in journalistic terms. Without the main outliers, the following persons emerged: Wang Rongpei (王榮培), Jiangsu governor (蘇松太), Ge Fanfu (葛蕃甫), Yan Songmei (嚴頌眉), Li Chunzhai (黎蒓齋), Gong Yangqu (龔仰蘧), Zhu Senting (朱森庭), Shen Bingcheng (沈秉成), Yuan Haiguan (袁海觀), and the provincial treasurer. Some are not a surprise and two are actually titles (方伯, 蘇松太). These individuals present very different profiles, with regular police or judicial officials (Wang Rongpei, Yan Songmei, Ge Fanfu), local officials (Yuan Haiguan), and foreign affairs officials (Gong Yangqu, Li Chunzhai). What the PCA tends to point to is that individuals are much less important *per se* than the institutions in which they held positions and operated. Yet, it can also be said that in the same positions, we have very different profiles, with Mo Xiangzhi (he held the same job twice) and Pei Dazhong as very high profile county magistrates, whereas Li Guangdan, Lu Yuanding (陸元鼎), Ye Tingjuan (葉廷眷), and Song Ting (松亭), in the same position, were much less visible figures or even not visible at all .

News reporting focused very much on a small group of institutions linked by "news" — here we can say "affairs" — the institutions of justice (mixed courts) are most central. This should not be read as these institutions linking the other institutions as such — by which I mean there was no interaction between the institutions connected to the mixed courts, except in the case when these institutions were also involved in the management of public order. This concerns first the Shanghai *xian* office that participated in the operation of the mixed courts at various

levels (Chinese coadjutor, shared affairs, etc.). The county magistrate was the highest Chinese authority in the city when it came to social order. He had his own police force that appears prominently in the network. The other major institutions were the police bureaus of the two settlements, including the local police stations. Had they been grouped together under a single heading for each settlement, it would have placed them near the mixed courts. Yet, for the sake of preserving the diversity of mentions of the police in the *Shenbao*, I chose to keep them separate. Their sheer presence attests to the attention of news reporting to criminal affairs, delinquency, and social disputes. The second realm that the *Shenbao* reported on were the Chinese institutions that administered the Jiangnan area, either in terms of policing or in terms of taxation. Private firms are not significant, due to the extensive range of such entities and the rarity of their mentions, except for the shipping and trading companies (some carried out both activities). It remains to use seen whether this was linked to substantial news reporting or merely commercial announcements.

To conclude this section, we need to revisit what network analysis did in our case. By taking all the actors together within the same period of twenty years, what the affiliation network did was to establish their relative presence within the journalistic space-time of the *Shenbao*. Within this space, the successive county magistrate never met, even if they appear within the same network. Does this present a bias or even a form of a-historicity? I do not believe so. First, we observed the two types of actors together, both individuals and organizations, as the *Shenbao* reported on them during its first twenty years. This allowed me to show the patterns that shaped the relative presence of certain individuals and organizations. Second, network analysis outlined the matrix of power that dominated the city in which judicial and police institutions played a central role. Third, this matrix of power was clearly biased by the ways in which the *Shenbao* sourced itself with a focus on certain type of events to the expense of the uneventful day-to-day activities of public institutions, companies, associations, etc. We are left with a view of the city where imperial officials were unrelenting in taming a fractured and undisciplined local society.

From actors to action

In the following section, I use topic modelling with two main purposes: to situate the persons and their actions in their context and to differentiate between the types of articles and the nature of the events they were related to. The dataset that I selected for topic modelling includes a total of 41,723 articles with less than 501 characters ranging from 1872 to 1892. Topic modelling requires the text to be cooked down to tokens. Tokenisation constitutes a sensitive operation due to the nature of language in the *Shenbao* that presents several issues. The *Shenbao* had no punctuation marks in this period, making it impossible to simply segment the articles into sentences, which in turn makes it difficult to segment into tokens. I processed the articles through our tokenizer to produce the basic tokenization.

Tokenization for the dataset produced more than one million tokens. After removing the stop words — based on a list that I had to enrich manually — the final word tally amounted to 979,416 tokens. There is a wide range of algorithms to implement topic modelling. I initially explored two different workflows, one based on BERT Topics, and one based on the Latent Dirichlet Allocation (LDA) algorithm.[41] In the initial run, I built a 10-topic and a 15-topic model for the sake of comparison and selected a single approach to build successive models.

[41] David M. Blei, Andrew Y. Ng, and Micheal I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, no. 3 (2003): 993–1022.

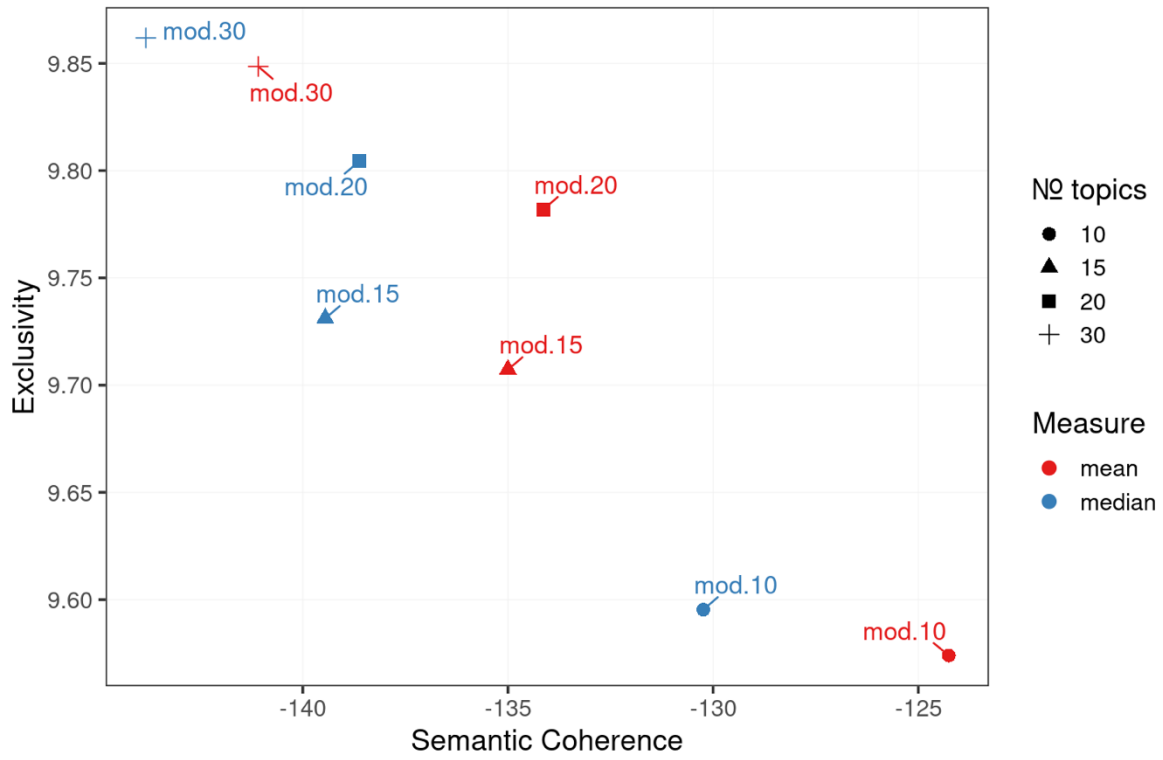
There was no substantial difference between BERT Topic and LDA. Both highlighted the prevalence of certain main topics and their evolution over time. Eventually, I settled for LDA mostly because I could control the whole workflow from NER to tokenization and topic modelling within the same environment. In concrete terms, I used the *stm* R package to enable topic modelling on my corpus.[42]

Although the *stm* package provides a set of metrics to help with defining the optimal number of topics, there is no objective hard benchmark that can tell which model to choose. In fact, this is not a problem because different models can provide increased granularity in shaping the topics from the documents. I built five models with 10, 15, 20, 30, and 50 topics. I proceeded step by step by adding new models and comparing their respective position in terms of exclusivity and semantic coherence. After comparing the results from the four models and their degree of connections/disconnection, as well as the range of topics each produced, I settled for a systematic comparison of the 15-topic and 20-topic models that seemed to provide more adequate and relevant topics. Depending on the actual content of the documents and their degree of heterogeneity/homogeneity, many topics may just produce subsets of prominent topics. This is what happened with the 50-topic model that I eventually discarded. The 10-topic model produced a relevant distribution, yet one that was dissonant with the other models in terms of exclusivity and semantic coherence. In the final stage, I decided to work with only three models with 15, 20, and 30 topics. There was a great deal of overlap between the 15- and 20-topic models. The latter refined some of the topics present in the 15-topic model. I found the same degree of consistency between the 20-topic and 30-topic models, with clear overlaps with the extra five topics in the 20-topic model. The 30-topic model, however, provided no more than a fine-grained distribution from the initial set of topics.

The 15 -topic model and 20-topic model shared 13 labels, which indicates that almost all topics in the former were included in the latter. The 20-topic model was able to differentiate better two specific topics: ‘Ministries and Shipping’. The other topics that the 20-topic model singled out were in fact topics that refined categories in the 15-topic model without adding much information. The final combined topics for the two models — after re-labelling — show a great deal of convergence. There are only three topics that are specific to the 20-topic model: ‘Local officials’, ‘Ministry officials’, ‘Miscellaneous’ (this catches mostly incidents such as fires).

[42] Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley, “Stm: An R Package for Structural Topic Models,” *Journal of Statistical Software* 91, no. 1 (October 31, 2019): 1–40.

Figure 5. Correlation graph of the topic models.



To examine the results from the different models, I followed the same script. First, I selected the ten top words in each topic to determine *a priori* the nature of the topic and to attribute a label to it. Second, I studied the graphs and statistics that the *stm* package provided to situate the relative importance of each topic in the corpus. Third, I selected ten representative documents from each topic for close reading. Reading the documents allowed me to refine the labels and to provide a succinct description of each.

Table 6. List of the 15 topics and the additional topics added by each model.

TPnb15	Topic15	Topic 20+	Topic 30 +
1	Miscellaneous news	Charity Announcement	Academies
2	Criminality	Inconsistent	Banditry
3	Relief operations	Local officials & Social order	Charity & medicine
4	Mixed Courts & Police	Local officials	Companies & Business
5	Examinations	Mixed Court French	Foreigners & consulates
6	Publishers Ads	Police	Inconsistent

6	Relief operations	Shop disputes	Li Hongzhi
7	Literary texts	Social incidents	Opium & Women
8	Foreign affairs	Women & Family	Opium trade & smuggling
9	Mixed Court Women		Police Settlements
10	Xian officials		Shanghai towns
11	Shipping		
12	Chinese army		
14	Local officials		
15	Mixed courts		

The topics generated by either model are quite on point. The close reading of the articles generally confirmed the labels, although the close reading allowed for merging some of the labels when the content seemed very close. I initially had three 'Inconsistent' labels (one in M15, two in M20). Close reading showed that this label included articles on miscellaneous small social incidents (mostly fires, capsized boats, coffins, but also list of names for examinations). In short, I have three topics that are quite close since they all address issues of petty criminality, money or debt disputes, mistreatment of women, etc. in Shanghai. The difference is the institutions involved (Mixed courts vs. county magistrate) or the focus. I have a fourth topic (criminality) on the same issues, but most likely in locations outside of Shanghai. There are two topics that are useless (Advertisements) or irrelevant (Literary texts) for my purpose. The topic on 'Foreign affairs' could be useful to weed out articles that do not cover Chinese topics in general, even if China as a country could be concerned. The lists of officials can be useful to locate people and their institutions and positions, although this would require retrieving systematically the same type of articles, based on the title. The same applies to the 'Examination' topics, although the usage may not be obvious since the information in the articles is brief and gives no background context.

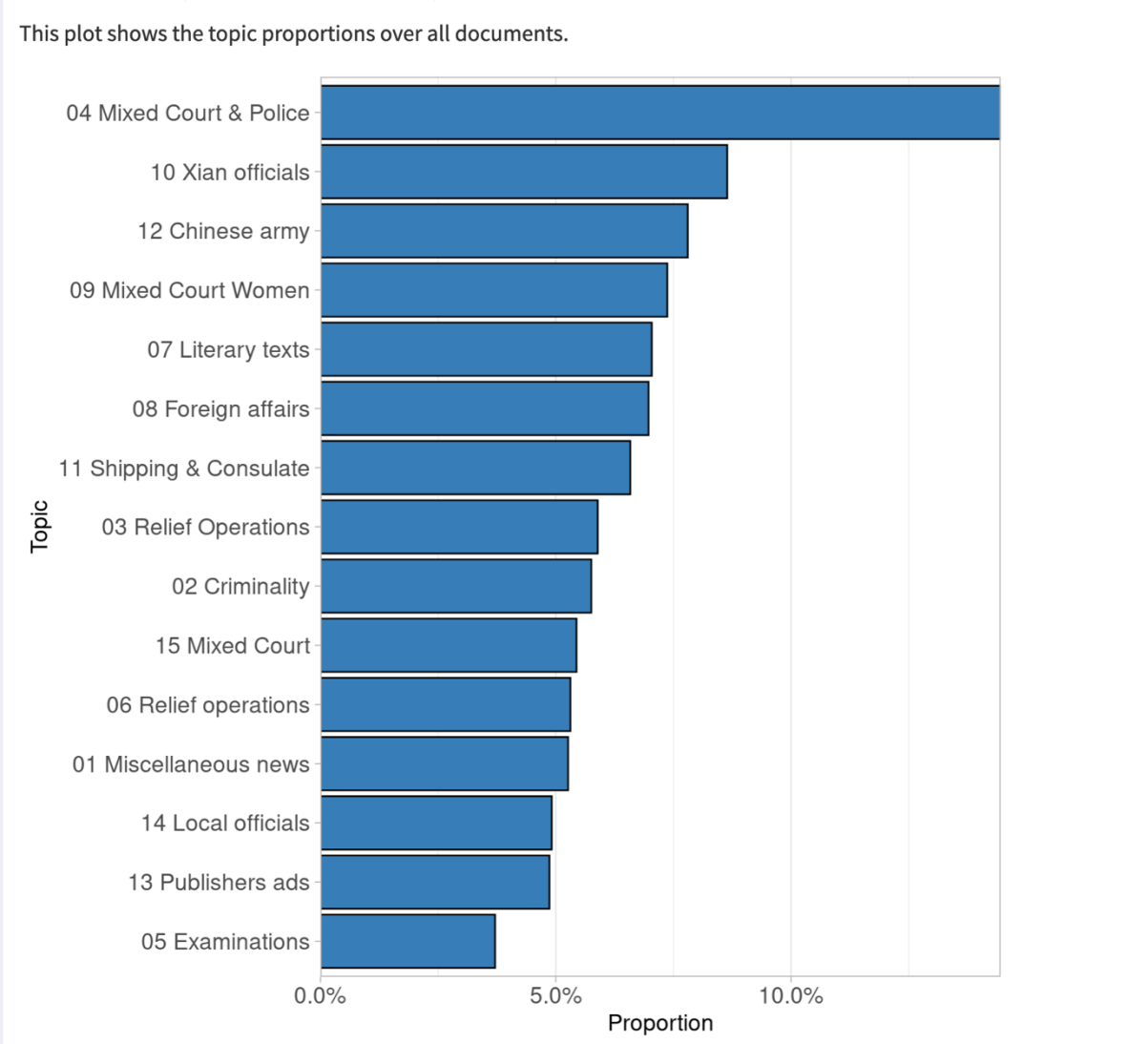
What insights can we gain from applying topic modelling to our corpus of 41,000 articles? How does this shed light on 'eminent Chinese' in the *Shenbao*, as well as the newspaper's construction and its utility as a historical source? Topic modelling serves as a tool to delve into texts before actual reading, allowing us to identify key patterns and trends within the selected corpus. At first glance, this approach may appear to challenge the fundamental principles of historical close reading practices. However, it should be viewed as complementary to established methodologies. By generating topics based on word frequency within the documents, topic modelling effectively unveils the content of the text and offers historians a unique means to interrogate the findings and uncover potential challenges posed by the nature of the texts. The emphasis on word frequency underscores structural aspects of how the documents were composed, even if certain terms may not hold intrinsic significance. This, in turn, provides valuable insights into the underlying dynamics of the texts.

One such example is the frequency of *zilin xizibao* 字林西字報 and its variants (Western language newspaper) in several topics. Somehow, as I did in network analysis, I considered this as a form of bias because these Western-language newspaper were the source of many news items. I could categorize it as a stop word and remove it. Yet by doing so, I would also eliminate a facet of the topics for which the Western language newspapers were significant. I faced the same issue for terms with little meaning at face value, like temporal marker (*ciri* (following day) 次日, *mingri* (tomorrow) 今日, *tongri* (same day) 同日, etc.) or numbers. Yet again, their sheer presence in certain topics, while not helping with labelling the topics, still proved useful when reading a sample of the documents to do the actual labelling.

Topic modelling seems to simplify brutally the content of the documents and to reduce the expression of the topics to their bare bones (even if one can of course change the number of terms that qualify a topic). In practice, and despite the nature of the articles in the *Shenbao,* I found that ten words were enough to get a preliminary sense of the topic and to proceed to preliminary labelling. The graph below shows the proportion of the top ten words in each topic, and it is the combination of frequency and distribution that gives sense to labelling the topic. Obviously, this is not enough, but it confirms the incredible potential of topic modelling to explore a large corpus.

Let me first state the obvious. The main topic that overrides all the other is the issue of social order and justice. Although one can find articles of a different nature — different stories — in each of the following topics, they all point to the interaction of individuals with the local judicial system. The 'Mixed Courts' topic is the most prevalent in the corpus, with its associate topics, 'Mixed Court & Police' and 'Mixed Court & Women'. By and large, this is about the same set of issues, but with a greater emphasis on the role of the police in one case and on the involvement of women in the other. The main 'Mixed Courts' topic covers a wide range of issues, including delinquency, petty crimes, but also commercial disputes. Two other topics fit in the same vein, 'Xian Officials' and 'Criminality'. In the first one, the only difference is that these are similar issues of social disorder and disputes, but these were brought before the county magistrate(s) instead of before the mixed courts. I should note right away that the county magistrates were often involved directly or indirectly in mixed court proceedings. Yet topic modelling detected the difference and delineated a specific topic for affairs that concerned only the county magistrate. The 'Criminality' topic refers to both criminal behaviour in the city and acts of violence and banditry in the larger area around Shanghai, involving other types of officials. The wealth of materials on issues of social disorder could be read as a reflection of the unstable nature of local society after the opening of the city to foreign trade and residency. It is more likely, however, that this reflected before all how the newspaper operated to collect newsworthy information.

Figure 6. Distribution and share of the topics over all documents in the 15-topic model.



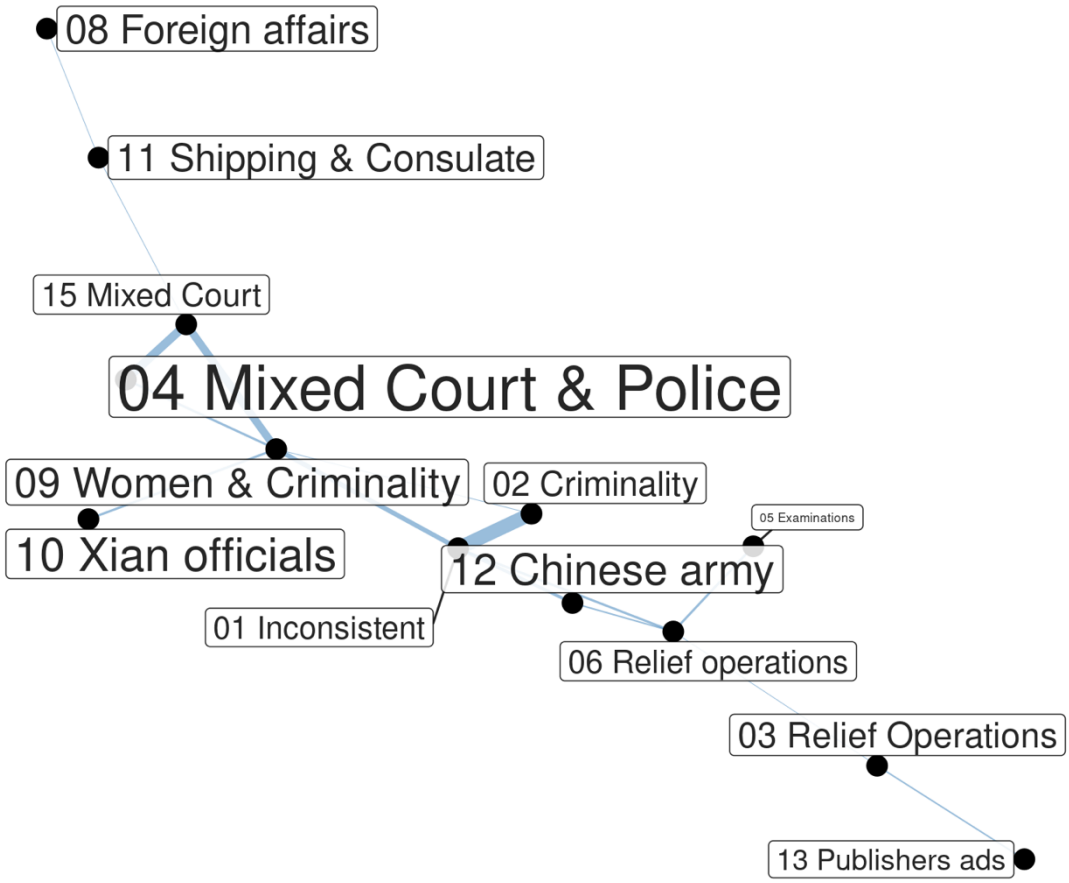
Note: The Mixed Court topic (04) in almost 20% of all documents, followed by Xian officials (8%), Chinese army, and again Mixed Court Women (7.5%). A good number of social order topics are present in at least or more than 5% (Criminality, MC). Relief operations are also equally prominent.

Except for the two 'Relief Operations' topics, the other topics cover issues that are not or only loosely connected. The 'Local Officials' topic concerns different aspects that involved local Chinese officials outside of Shanghai acting in various capacities (birthday, temple repair, etc.) or about their movements (transfer). In a parallel register, the Chinese Army topic also involves a good number of officials, both civil and military, often in relation to reviews and ceremonies in the Jiangnan area. The 'Shipping & Consulate topic' focuses not just on navigation companies (including regular press announcements), but also issues of customs and official control, hence the relative presence of consular officials. A whole set of documents appears to be reports on international affairs or news from foreign countries, as documented in the 'Foreign Affairs' topic. Two topics also crystallized around themes that spoke to the culture of literati, one on 'Examinations', with news on sessions or results, and one broadly defined as 'Literary Texts', mostly contributions by readers. The two 'Relief Operations' cover the same theme, but in different ways. One was focused much more on Northern China, notably the Tianjin area, while the other one concerned mostly Shanghai and its surrounding area or charity

actions initiated in the city. In fact, many such articles were calls for financial support in the form of press announcements. Finally, there remain two topics that provide little substance. One is made up of documents that had little in common (Miscellaneous News), while the other one drew entirely on advertisements by publishers.

In the 20-topic model, what I found was mostly a refinement of the topics discussed above. A good example is the 'Mixed Court French' topic that pulled out articles more directly linked to the Mixed Court in the French Concession. In the same way, the 'Police' topic assembles articles from the larger 'Mixed Courts' topic, though with a focus on interventions by the police. The model also created a second 'Local Officials' topic that traced further circles in the realm of social disorder with a topic on 'Shop Disputes', 'Women & Family' (mostly about women kidnapping, sale, and opium), 'Local Officials & Social Order', and 'Social Incidents' (this was more about house fires and similar issues). The 20-topic model did not alter substantially the distribution and nature of the main themes uncovered by the 15-topic model. There is a high degree of consistency, but it contributes to highlighting where one could look at for more specific inroads into the initial topics. The only new topic was a set of articles that concerned officials of imperial ministries. If we look at the additional topics in the 30-topic model, there is a mix of what I call refined topics and new topics. Under the former, 'Charity & medicine' is clearly an offshoot of 'Relief Operations', 'Banditry' is a subset of 'Criminality' with a focus on areas external to Shanghai, 'Foreigners & Consulates' emerges from the original 'Shipping & Consulates', with a focus on Westerners, companies and business (also present in the 20-topic model). The 'Police' topic branches into a more precise topic on the police of the foreign settlements. Four topics seem to be more original, even if they were somehow subsumed in the previous models. The 30-topic model introduces two topics where opium figures in association with issues of trade or women (opium dens). It also sets apart a topic that deals with the academies of classical learning and not unsurprisingly a topic whose centre is Li Fuxiang.

Figure 7. Correlation graph of the 15 topics in the 15-topic model.

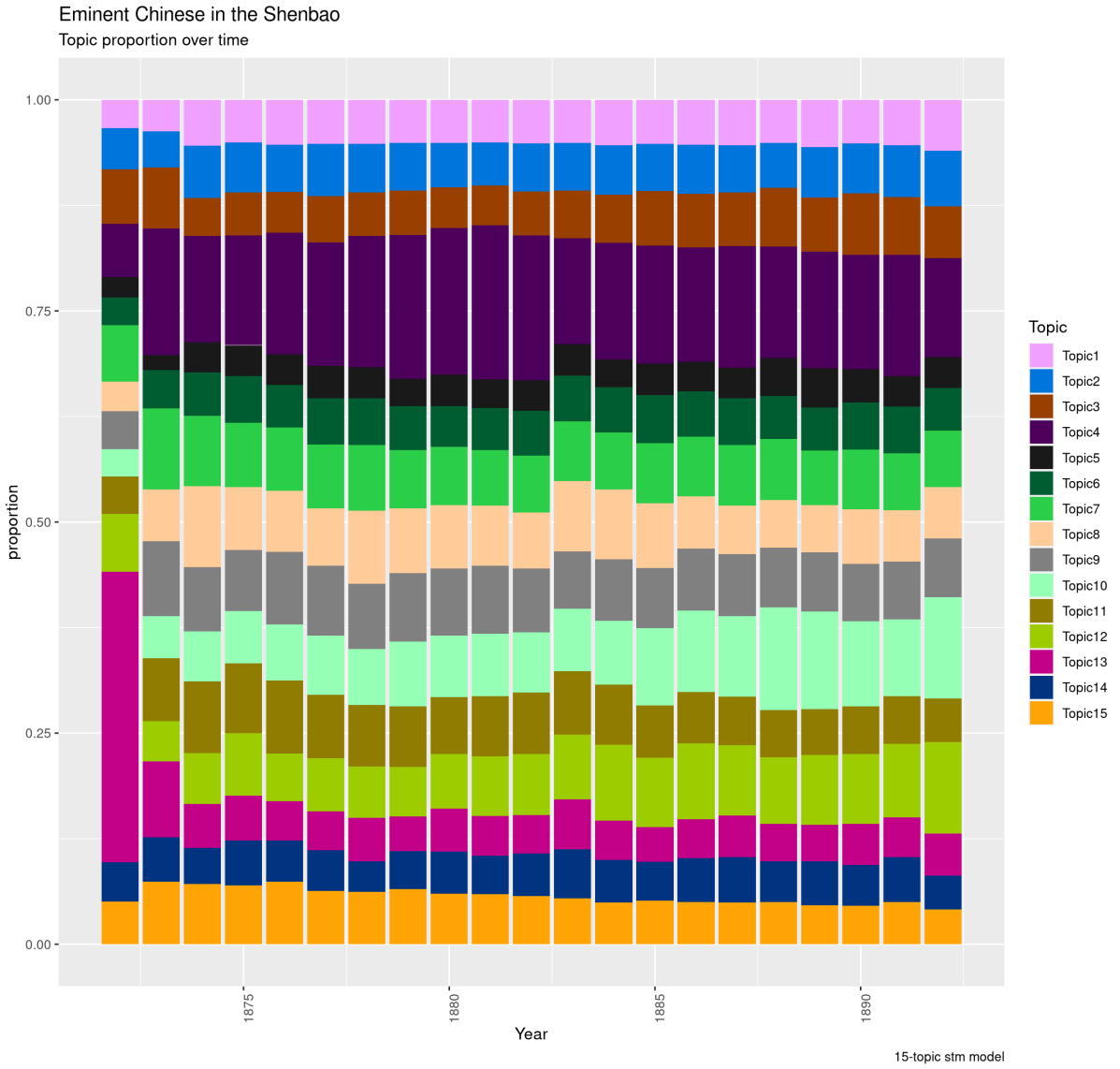


The three correlation graphs for the 15-, 20- and 30-topic models show a consistent pattern of semantic proximity between the same topics (see Figure 7). The topics that involve the Mixed Courts, the police, criminality, banditry, women, and by extension opium and foreigners share greater similarity than the grouping between relief operations or, in the 30-topic model, Li Fuxiang, foreign affairs, and foreigners. It also establishes that depending on the degree of granularity, one can see shifts in how the topic model traced dividing lines, such as with 'Shipping & Consulates' (15-topic), which translates into three distinct topics, with 'Companies & Business', 'Shipping', and 'Foreigners & Consulates'. Yet, it also appears that 'Companies & Business' was semantically linked to the Mixed Courts grouping. This points to the prevalence of disputes brought before the courts in the news about commercial companies.

There were a few noticeable changes over time in the relative importance of topics. One cannot point to a very drastic evolution, but rather to smooth trends. I shall take the 15-topic and the 20-topic models as examples, just because even beyond ten different colours, a graph becomes difficult to decipher. The two graphs show the proportion of each topic in a given year. Let us first point out what seems to be an anomaly. In 1872, the share of advertising by publishers seems very high, but it is not representative. This was the very start of the newspaper, with only a few months of publication at the start of the newspaper and, unsurprisingly, it was filled

with advertising that probably also included advertisements by the *Shenbao* itself. We can see that three topics experienced a certain growth, Topic 10 (Xian officials), especially at the end of the period, Topic 12 (Chinese Army) also increased by the end of the period and Topic 4, with an increase around 1880. The other topics hardly changed, except for a small decline of the general 'Mixed Courts' topic. The 20-topic graph did present a more diverse evolution. It singled out Topic 10 (Criminality) as a consistently increasing autonomous topic, while Topic 19 (Mixed Court French) also showed ups and down, but at a fairly high level throughout the period. In parallel, Topic 12 (Mixed Court & Police) experienced a noticeable increase, especially at the end of the period. On the opposite, Topic 8 (Local officials & Criminality) lost in importance substantially in the last years. We can make two preliminary observations about the use of different models: the shift from 15 topics to 20 topics highlights better the trends that shaped the news thanks to a more precise focus on sub-topics. On the other hand, there was by and large a certain stability in the nature of the types of news that the *Shenbao* chose to publish.

Figure 8. Topic proportion over time (1872-1892).



We can interpret these results from two different perspectives. Whatever the model used, the distribution of topics gives us some insight in how the editors of the *Shenbao* worked and on the process of news-making in the first twenty years of the newspaper. The *Shenbao* faced two constraints: on the one hand, it was a novel editorial enterprise that could only rely on existing cultural and social forces. In the absence of professional journalists or reporters, the editors relied on the highly educated literati who had not made it into the imperial bureaucracy. The general development of modern printing in Shanghai had already given birth to a host of new publishing ventures that offered new professional outlets for the educated elites.[43] The modern press was such an outlet. On the other hand, the *Shenbao* had to meet and build its audience, which it readily found in the large community of male literati in the Jiangnan area who were active both in official positions in the imperial bureaucracy and those — by far the largest number — involved in local communities at all levels of urban and rural settlements. These two factors can explain how the *Shenbao* in its early life organized and prioritized the collection of information. It should be obvious by now that official institutions were a major provider of newsworthy information, simply because they dealt with, regulated on, or passed judgement on everyday life events. This is especially true of the judicial institutions, mixed courts or Chinese courts, and the police, because they were very accessible public venues where the literati in search of "events" could obtain ready-made "facts" that could feed the newspaper. This would explain the overarching presence of these institutions in our topics which in turn tainted the profile of the ordinary people, especially women as Mittler's close reading previously spotted[44]. Yet, there was also another side to this preference for official sources, namely informing the literate public about decision-making, local decrees, official ceremonies, and all that mattered to the activities of the imperial elites in their respective communities. Whether recruited by and employed by the *Shenbao* as professional writers or informally incorporated in the news collection circuit, the individuals who contributed to the newspaper all belonged to the realm of a shared political culture. Officialdom and its prestige and authority remained paramount.

On the other hand, we can also take the topics as indicators about the individuals to be found in the *Shenbao* and figure out which eminent Chinese — and non-eminent Chinese — populated the pages of the newspaper, and why they were there. Whereas statistical analysis and network analysis told us something about "how many" and what connections might exist, topic modelling shed light on the reasons behind the presence of individuals in the news. Among the eminent Chinese, the most prominent individuals were those who officiated in an official capacity and who adjudicated almost daily on the life of other individuals, companies, and communities, etc. This includes the whole ladder of local officials such as county magistrates, prefects, *daotai* (circuit intendant), and viceroys (general governor), but also high military officers and foreign diplomats (consuls), though the latter often appeared only through a title, not a proper name. We can also assume that the cases of commercial disputes or business-related news that involved merchants and businessmen also brought to the fore the ebb and flow of economic activity and the tensions within this community. Yet network analysis has already established that these were very individual affairs. Conversely, there seems to be little mention of the leading business figures as members or leaders of the merchant organizations such as guilds and *gongsuo*. Our eminent Chinese, therefore, belonged primarily to a well-defined circle of high and middle-level civil and military officials, both in Shanghai and in the surrounding towns, who came into the news reporting due to their service in public

[43] Catherine Yeh, *Shanghai Love: Courtesans, Intellectuals, and Entertainment Culture, 1850-1910* (Seattle: University of Washington Press, 2006), 179–81.
[44] Mittler, *A Newspaper for China?,* 290

institutions. In other words, the social spectrum of elites as represented in the *Shenbao* in its early history was quite narrow, both in terms of categories of elites, but also in terms of concrete individuals. On the other hand, the topics that I have identified point to the presence of a vast army of common people who appeared in the pages of the *Shenbao* when their life intersected with an official institution, mostly the police and the courts. This may of course include people with varying social status, but as we already observed from the analysis of the names, the vast majority were just ordinary men and women.

The *Shenbao* is considered as a major source to explore the social history in Shanghai. For the period under scrutiny, it is almost the only source as it had very few competitors before 1892. The study above has shown that it had its intrinsic biases, not because of an explicit editorial choice, not for particular political reason, and not because it was run by a British merchant, but simply because in its early phase the newspaper was crafted by and for a particular segment of the elites. At this stage, even if a British-educated merchant initiated the creation of the *Shenbao*, its most direct actors — the literati — news reporting. There was yet no investigative journalism, no social surveys, no independent sources of information in China proper (except the telegraph and burgeoning foreign news agencies). The natural conduit to which the literati *cum* reporter naturally turned to was the small nexus of public institutions through which they could narrate the novelty as well as the uncertainties of urban life, and by extension everyday life in the Jiangnan area. The *Shenbao* provides a unique vantage point from which to step into the role of the elites in the 1870s-1890s, largely in Shanghai, secondarily in the cities and towns around, but it needs to be said that there was a definite framing of these elites.

The case of Xian Officials

The examination of the 'Xian Officials' topic revealed a focus on articles dedicated to the activities of county magistrates, emphasizing their role in maintaining social order. In this subsequent round of topic modeling, I chose to reprocess the articles within the 'Xian Officials' topic to gain deeper insights into their content. Instead of including the entire dataset, I set a threshold for topic proportion at 50 percent and above,[1] excluding articles falling below this threshold. The resulting dataset comprised 1,919 articles, and I conducted tests for topic modeling with four levels (10, 15, 20, 25), ultimately selecting the 15-topic model as the most suitable.

This sub-dataset processing further refined the original 'Xian Officials' topic into more specific themes. Themes related to social order remained predominant across most topics, with the 'Police Court Chinese' topic ranking highest in terms of proportions for individual documents (49 percent). In the table below, several topics from the complete dataset, such as 'Relief operations' and 'Charity,' still appear, likely corresponding to the actual involvement of county magistrates in these areas. Nevertheless, topic modeling introduced new topics that delineate specific domains of official engagement, offering a more precise understanding of the daily responsibilities of county officials. The table below labels the topics based on their primary associated terms (10 terms each) and lists the ten articles in each topic with the highest proportion score. I then systematically reviewed five to ten articles per topic, depending on the level of homogeneity and correspondence observed with the given topic.

Table 7. Topics in the 15-topic model for the ‘Xian Officials’ topic.

Topic	Label	Proportion
9	09 Army Ship Inspection	0.118
7	07 Police Court Chinese	0.116
6	06 County magistrates	0.101
1	01 Bandits Ships Sailors	0.084
12	12 Xian Official Court	0.064
14	14 Relief Operations	0.060
8	08 Liumang Violence	0.059
3	03 Police Rescue Wounded	0.056
2	02 Sima Court Shops	0.055
4	04 County magistrate Office	0.055
5	05 Police French Opium	0.055
10	10 Money Girl	0.050
11	11 Charity *Shenbao* Shandong	0.045
13	13 Police Thieves	0.043
15	15 Mixed Court IS	0.038

This close examination of articles facilitated a deeper understanding of their content and the reasons for their inclusion within the same topic. Importantly, this analysis did not alter the initial topic labels, as the sets of most common words within each topic generally sufficed to grasp their themes. However, it is essential to note that the articles within each topic were not mutually exclusive, and their boundaries sometimes overlapped. What distinguished articles' alignment with one topic over another often depended on the mix of terms found within the articles. For instance, while the 'County magistrates' topic in this subset included articles that could have fit into other topics, the lines between topics occasionally blurred.

To illustrate the composition of topics within specific documents, I selected the first 15 documents and assessed their primary topic associations. This approach provides a clearer depiction of the distribution of topics within individual documents.

Figure 9. Topic distribution among the first fifteen documents of the sample.

The visual representation of topic associations highlights that certain documents are strongly aligned with specific topics, while others exhibit a more mixed profile. Nonetheless, each document contains elements that elucidate its alignment with particular topics. For instance, the 'Police Court Chinese' topic primarily encompasses formal matters presented before the county magistrate, while the 'County magistrates' topic may include similar issues reported to the county magistrate or with mentions of officials before formal proceedings. In contrast, the 'County magistrate Office' topic primarily pertains to news regarding officials' movements, visits, or assuming duty. The 'Bandit Ship Sailors' and 'Liumang Violence' topics address misconduct or criminal actions, particularly involving specific groups such as sailors.

Figure 10 provides a word cloud visualization of the top 50 words in each topic, offering a more vivid representation of differences and the impact of word distribution on topic definitions. Word clouds serve as one method to visually depict what can also be presented in tabular form.

Figure 10. Word frequency for nine topics in the form of word clouds.

Notably, the boundaries between topics are not rigid, and some overlap is observed, particularly due to the stronger connections between certain topics. Additionally, a clear dichotomy emerges, with some topics primarily focused on city-based issues of crime and delinquency, while others centre on shipping-related matters. While historiography, especially in the West, has placed much emphasis on the 'Shanghai daotai,' the magistrate in direct contact with foreigners, the *Shenbao* shows the crucial role played by the county magistrate in everyday life in the city. The prevalence of topics related to criminal activities involving diverse groups implies a multifaceted dimension to his duties beyond routine administrative tasks. This shifts the historical perspective from the foreign settlements to the Chinese-administered city in the early history of treaty-port Shanghai.

Conclusion

This study of eminent Chinese in the *Shenbao* took several parallel pathways into the large dataset of articles extracted almost randomly from the first twenty years of publication of the newspaper. The close reading of the approximately 50,000 articles in our sample exceeds the capacity of human cognition. In the course of this exploration, therefore, I analysed the texts through different computational methods that ranged from named entity extraction to topic modelling. These methods transform the texts into data points, such as individuals and organizations, while omitting the narratives that constituted the core content of the original articles. This could be interpreted as a radical transformation of the stories that the *Shenbao* was expected to present to its readers, removing some of their original character. Yet, any close reading, even if done with time, method, and concentration, can simply never reach the level of completeness that computational methods can achieve when it comes to getting to the actors in the stories or even focusing on sub-datasets built around the topics that can be identified in the pages of the newspaper. Although the original stories are no longer there, the main trends — topics — provide an understanding of the news stuff that the *Shenbao* was made of.

The nature of elites — the eminent Chinese that I am after — in the *Shenbao* is quite well delineated. The main actors that we encountered were officials of the imperial state. No other category comes close to the same level of presence in the pages of the newspaper. Some merchants do appear at a relatively high level, but only in their role as philanthropists, not so much as representative of guilds or *gongsuo*. The overrepresentation of officials in the *Shenbao* was not really the expression of a pro-state bias as such. It was the result of the news-making process. The *Shenbao* depended on various sources to feed its pages: telegrams, local or foreign Western newspapers, Peking gazette. Yet these sources covered only official, national, and international news. To report on what happened on the ground and to publish news that made the newspaper relevant to its readership and to increase its readership. The main market of the *Shenbao* was the Shanghai area and the broader Jiangnan region. It had to meet the needs or expectations of this privileged segment of the population. Those who worked in the newspaper had to find and shape what was newsworthy, what could pick the interest of the readers. They had to invent the newspapers at the same time as they were inventing themselves as “newsmen”. The vision of the city that emerges is not the ‘fantastic Shanghai’ that several historians have chosen to play up, but a place filled with tensions that the representative were busy trying to contain.

The abundance of news concerning social conflicts and disorder directly resulted from the newspaper's main sources, which were the key justice and police institutions in the city. Whether in the foreign settlements with mixed courts, in the walled city or Zhabei with county magistrates, there was a continuous stream of minor incidents such as disputes, fights, robberies, kidnappings, and fraud. On a daily basis, the courts and the police were actively involved in responding to these events, either at the forefront of police actions, such as investigations and arrests, or in the final sentencing decisions by the courts. These institutions provided readily available stories that depicted everyday life, particularly the lives of common people and those on the fringes of local society. These were truly "news" in the most fundamental sense of the word—events that deviated from the norm. The courts handled a wide range of minor conflicts and offenses that were an inherent part of city life, far more numerous than what could be covered in the newspaper's pages. The magistrates and the police were constantly generating a continuous stream of stories, which the literati writing for the *Shenbao* could easily draw upon to supply the newspaper. Often, they would even directly borrow from court records.

The *Shenbao* newspaper covered a wide range of social actors, but many individuals, tens of thousands in fact, were mentioned purely by chance or because they encountered unfortunate situations. They were essentially just names in brief news items that summarized events in a few words. These individuals often appeared in the news due to problems they encountered, either as victims or perpetrators, which led to their involvement with the law and the courts. There were some people who appeared in the news for reasons other than legal issues, but we know very little about them. They were connected to announcements or meetings of various institutions like guilds and companies, but the newspaper provided limited information about their identities and roles. Despite Shanghai being a hub for trade and production, companies did not often make it to the headlines. Shipping companies were an exception, mainly because they frequently issued announcements and interacted with authorities. In the first twenty years of the *Shenbao*, the newspaper portrayed a society dominated by Qing officials who governed a dynamic but tension-filled community marked by constant incidents and disputes. Eminent Chinese figures were undoubtedly prominent and recognizable within this sea of non-elite and mostly anonymous individuals. Nonetheless, the news items themselves offer a rare glimpse into life in early treaty-port Shanghai.

Is the *Shenbao* primarily a text or a source? The newspaper was a commercial enterprise with a mission to produce and sell news. Its owner had goals beyond making money, but what matters most is how the newspaper actually operated. It was published to inform its readers about current events, not to look back in time or create a historical record for future generations. While the newspaper's owners and editors may have eventually recognized its historical significance, the day-to-day task of publishing a continuous stream of news likely took precedence. The computational methods I applied rely on the text of the *Shenbao*—essentially treating the newspaper as a vast collection of text—to examine what insights a historian can extract from a massive number of articles spanning twenty years, without reading them all in detail. These methods allowed me to go beyond the text and delve into the mechanics of news production, helping me understand what kind of source the *Shenbao* represents, what it can provide, and what it will never reveal in its pages. This also highlights the potential of computational methods to sift through countless documents and precisely define the corpus that will undergo systematic examination.

References

Armand, Cécile, Christian Henriot, and Huei-min Sun, eds. *Knowledge, Power, and Networks. Elites in Transition in Modern China*. Leiden: Brill, 2022.

Blei, David M., Andrew Y. Ng, and Micheal I. Jordan. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, no. 3 (2003): 993–1022.

Burkhardt, Marcus, Daniela van Geenen, Carolin Gerlitz, Sam Hind, Timo Kaerlein, Danny Lämmerhirt, and Axel Volmar, eds. *Interrogating Datafication: Towards a Praxeology of Data*. transcript publishing, 2022.

Campbell, Cameron, and Bijia Chen. "Nominative Linkage of Records of Officials in the China Government Employee Dataset-Qing (CGED-Q)." *Historical Life Course Studies* 12 (September 8, 2022): 233–59.

Chen, Deming 陈德明. *Yuanqu de huihuang: (1929-1949) haipai guanggao meishu zitu ji 远去的辉煌：申报(1929-1949) 海派广告美术字图集 (Distant Glory: (1929-1949) A*

collection of Artistic Character Advertisements in the Shenbao). Shanghai: Shanghai daxue chubanshe, 2019.
Corbin, Alain. *The Life of an Unknown: The Rediscovered World of a Clog Maker in Nineteenth-Century France*. New York: Columbia University Press, 2001.
Dan, Mingming 单明明. "Shenbao shiye zhong de makesi xueshuo 申报视野中的马克思学说 (Marxist Theory through the Shenbao)." Doctoral dissertation, Central Party School, 2017.
Digitised Newspapers – A New Eldorado for Historians?: Tools, Methodology, Epistemology, and the Changing Practices of Writing History in the Context of Historical Newspapers Mass Digitization. *Digitised Newspapers – A New Eldorado for Historians?* Berlin: De Gruyter Oldenbourg, 2022.
Ehrmann, Maud, Matteo Romanello, Simon Clematide, Phillip Benjamin Ströbel, and Raphaël Barman. "Language Resources for Historical Newspapers: The Impresso Collection." *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, 2020, 958–68.
Elliston, E.S. *Ninety-Five Years a Shanghai Hospital, 1844-1938 : Chinese Hospital, Shantung Road Hospital, the Lester Chinese Hospital*, n.d.
Es, Karin van, and Mirko Tobias Schäfer. "Introduction: New Brave World." In *The Datafied Society*, edited by Karin van Es and Mirko Tobias Schäfer, 13–22. Studying Culture through Data. Amsterdam University Press, 2017.
Fang, Hanqi, ed. *A History of Journalism in China*. 10 vols. Singapore: Silkroad Press, 2014.
Feng, Zhuo 冯卓. "Qingmo minchu shenbao cihui yanjiu 清末民初《申报》词汇研究 (A Study on the Vocabulary of the Shenbao in the Late Qing and the Early Republic)." Doctoral dissertation, Jilin University, 2021.
Gao, Xueqin 高学琴. "Shenbao shehui guanggao yanjiu《申报》社会广告研究 (Research on Social Advertising in the Shenbao)." Doctoral dissertation, Wuhan University, 2019.
Gentz, Natascha. "Die Anfänge des Journalismus in China (1860 - 1911)." Doctoral dissertation, Heidelberg University, 1998.
Guangxi Zhuangzu Zizhiqu tong zhi guan and Guangxi Zhuangzu Zizhiqu tu shu guan. *Shen bao Guangxi zi liao suo yin*. Nanning: Guangxi renmin chubanshe, 1992.
He, Qiliang. *Newspapers and the Journalistic Public in Republican China: 1917 as a Significant Year of Journalism*, 2018.
Henriot, Christian. "Forging Bonds and Building Factories: The Networked World of Shanghai's Industrial Elite." In *Modern China in Flux: Networks, Mobility, and Transformation*. Berlin: De Gruyter, 2024.
———. *Prostitution and sexuality in Shanghai: a social history 1849-1949*. Cambridge, UK; New York: Cambridge University Press, 2001.
———. *Scythe and the City: A Social History of Death in Shanghai*. Stanford: Stanford University Press, 2016.
———. *Shanghai, 1927-1937: Municipal Power, Locality, and Modernization*. Berkeley: University of California Press, 1993.
Hua, Changhui 华长慧, ed. *Shenbao Ningbo lühu tongxiang shetuan shiliao 《申报》宁波旅沪同乡社团史料 (Historical Materials on Shanghai Ningbo native-place Associations in the Shenbao)*. Ningbo: Ningbo chubanshe, 2009.
Hua, Hongyan 花宏艳. *Shenbao Kanzai Jiutishi Yanjiu (1872-1949) 申报刊载旧体诗研究 (1872-1949) (A Study of Old Style Poetry in the Shenbao)*. Nanjing: Fenghuang chubanshe, 2018.

Huang, Zhennan 黄振南, Qinhui 蒋钦挥 Jiang, and Chao Xu. *Shenbao Guangxi xinhai geming ziliao xuanbian 申報"广西辛亥革命资料选编 (Selected Materials from the Shenbao on the Revolution of 1911 in Guangxi)*. Guilin: Guangxi shifan daxue chubanshe, 2012.

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, and Nello Cristianini. "The Actors of History: Narrative Network Analysis Reveals the Institutions of Power in British Society Between 1800-1950." In *Advances in Intelligent Data Analysis XVI*, edited by Niall Adams, Allan Tucker, and David Weston, 10584:186–97. Cham: Springer International Publishing, 2017.

Lansdall-Welfare, Thomas, Saatviga Sudhahar, James Thompson, Justin Lewis, FindMyPast Newspaper Team, and Nello Cristianini. "Content Analysis of 150 Years of British Periodicals." *Proceedings of the National Academy of Sciences* 114, no. 4 (January 24, 2017): E457–65.

Li, Xiangqun 李向群. *Jindai Xiamen lishi ziliao huikan: Shenbao jiwen 近代厦门历史资料汇刊：申报纪闻 (Modern Xiamen Historical Data Collection: Shenbao Records)*. Xiamen: Xiamen daxue chubanshe, 2020.

Lin, Qiuyun 林秋云. "Bianzhi de Cishan: Wanqing Hubei Shuliu Gongsuo Chutan 變質" 的慈善: 晚清滬北棲流公所初探 (A 'Metaphor' of Charity: Preliminary Study of the Sinza Refuge in the Late Qing Dynasty)." *Qingshi Yanjiu (The Qing History Journal)*, no. 4 (2017): 84-98.

Lin, Shengdong 林升栋. *Zhongguo jinxiandai jingdian guanggao chuangyi pingxi: qishiqi nian 中国近现代经典广告创意评析：<申报>七十七年 (Comments and Analysis on Modern and Contemporary Chinese Classics Advertisement Creativity: Seventy years of the Shenbao)*. Nanjing: Dongnan daxue chubanshe, 2005.

Lin, Zhongjia 林忠佳, Tianxi Zhang, Guangdong sheng dang an guan, and "Shen bao" Guangdong zi liao xuan ji bian ji zu. *Shenbao Guangdong ziliao xuanji 申報廣東資料選集 (An anthology of Guangdong materials in the Shenbao)*. Guangzhou: Guangdong Sheng dang an guan Shen bao Guangdong zi liao xuan ji bian ji zu, 1995.

Liu, Li 刘莉. "Zhou Shoujuan zhubian shiqi shenbao·ziyoutan xiaoshuo yanjiu 周瘦鹃主编时期《申报·自由谈》小说研究 (Research on the Novels in the 'Free Talk' Section of the Shenbao during Zhou Shoujuan's Editorship)." Doctoral dissertation, Fudan University, 2010.

Liu, Yongsheng 刘永生. 2008. "Shenbao de Dui Ri Yulun Yanjiu 申报的对日舆论研究（1931.9-1937.12）(A Study of Shenbao Editorials on Japan)." Doctoral dissertation, Changsha: Hunan Normal University.

Lu, Ning 卢宁. *Zaoqi shenbao yu wanqing zhengfu: Jindai zhuanxing shiye zhong baozhi yu guanli guanxi de kaocha 早期"申报"与晚清政府：近代转型视野中报纸与官吏关系的考察 (Early The early Shenbao and the Late Qing Government: An Investigation of the Relationship between Newspapers and Officials from the Perspective of Modern Transformation)*. Shanghai: Shanghai kexue jishu wenxian chubanshe, 2012.

MacKinnon, Stephen R. "Toward a History of the Chinese Press in the Republican Period." *Modern China* 23, no. 1 (1997): 3–32.

Mittler, Barbara. *A Newspaper for China?: Power, Identity, and Change in Shanghai's News Media, 1872-1912*. Harvard East Asian Studies Monographs 226. Cambridge (Mass.): Harvard University Asia Center ; Distributed by Harvard University Press, 2004.

Narramore, Terry. "Making the News in Shanghai: Shen Bao and the Politics of Newspaper Journalism, 1912-1937." Doctoral dissertation, Australian National University, 1989.

Ningbo shi dang'anguan 宁波市档案馆. *Shenbao Ningbo shiliao ji 《申报》宁波史料集 (A collection of historical materials on Ninbo in the Shenbao)*. Ningbo: Ningbo chubanshe, 2013.

Pan, Weiwei 潘薇薇. *Cong shenbao guanggao kan zhongguo jindai xiaoshuo yundong 从申报广告看中国近代小说运动 (The Movement of Chinese Modern Novels seen from Shenbao Advertisements)*. Shanghai: Dongfang chuban zhongxin, 2015.

Pang, Ju'ai 庞菊爱. *Kua wenhua guanggao yu shimin wenhua de bianqian: 1910-1930 nian shen bao kua wenhua guanggao yanjiu 跨文化广告与市民文化的变迁：1910-1930 年申报跨文化广告研究 (Cross-cultural advertisements and the changes of citizen culture: a study of cross-cultural advertisements in the Shenbao from 1910 to 1930)*. Shanghai: Shanghai jiaotong daxue chubanshe, 2011.

Pudong xinqu dang'anju 上海市浦东新区档案局 and Pudong Xinqu wenshi xuehui 上海市浦东新区文史学会. *Shenbao zhong de Pudong 申報中的浦東 (Pudong in the Shenbao)*. Shanghai: Sanlian shudian, 2019.

Roberts, Margaret E., Brandon M. Stewart, and Dustin Tingley. "Stm: An R Package for Structural Topic Models." *Journal of Statistical Software* 91, no. 1 (October 31, 2019): 1–40.

Shi, Yuanhua and Da Han Minguo lin shi zheng fu jiu zhi guan li chu. *Shen bao you guan Han'guo du li yun dong ji Zhong han guan xi zi liao xuan bian: 1910-1949*. Beijing: Ren min wen xue chu ban she, 2000.

Song, Shuqiang 宋书强, Zhaolu 殷昭鲁 Yin, and Feifei 赵飞飞 Zhao. *Shenbao baodao yu pinglun 《申报》报道与评论 (Shenbao reports and editorials)*. Nanjing: Nanjing daxue chubanshe, 2019.

Splichal, Slavko. *Datafication of Public Opinion and the Public Sphere: How Extraction Replaced Expression of Opinion*. London, UK: Anthem Press, 2022.

Stephens, Thomas B. *Order and Discipline in China: The Shanghai Mixed Court, 1911-27*. Asian Law Series. Seattle: University of Washington Press. Accessed February 24, 2023.

Taiwan yinhang and Jingji yanjiushi. *Qingji Shenbao Taiwan jishi jilu (872-1887) 清季申報臺灣紀事輯錄 (A chronicle of events in Taiwan in the Qing Dynasty)*. Taizhong: Taiwan sheng wenxian weiyuanhui, 1994.

Takahashi, Kōsuke 高橋孝助. "Kohoku Seiryu Guzo No Seiritsu -- Shanhai Sokai No Zendō 滬北棲流公所の成立--上海租界の善堂 (The Establishment of the Sinza Refuge in the Shanghai Concession)." *Bulletin of Miyagi University of Education 宮城教育大学紀要* 第 1 分冊, 人文科学・社会科学, no. 19 (1984): 261–78.

Tsai, Weipin. *Reading Shenbao: Nationalism, Consumerism and Individuality in China, 1919-37*. Houndmills, Basingstoke: Palgrave Macmillan, 2010.

Wagner, Rudolf. *Joining the Global Public : Word, Image, and City in Early Chinese Newspapers, 1870-1910*. Albany NY: State University of New York Press, 2007.

———. "The Early Chinese Newspapers and the Chinese Public Sphere." *European Journal of East Asian Studies* 1, no. 1 (March 2001): 1.

Wagner, Rudolf G. "The Role of the Foreign Community in the Chinese Public Sphere." *China Quarterly*, no. 142 (juin 1995): 423.

———. "The 'Shenbao' in Crisis: The International Environment and the Conflict between Guo Songtao and the 'Shenbao.'" *Late Imperial China* 20, no. 1 (juin 1999): 107–38.

Wang, Runian 王儒年. "Shenbao guanggao yu shanghai shimin de xiaofei zhuyi yishi xingtai" 申报广告与上海市民的消费主义意识形态 (Advertisements of the Shenbao

and the Consumerism Ideology of Shanghai Residents).” Doctoral dissertation, Shanghai Normal University, 2004.

———. *Yuwang de xiangxiang: 1920-1930 niandai guang ao de wenhuashi yanjiu 欲望的想像 : 1920-1930 年代申报广告的文化史研究 (The Imagination of Desire: A Study in the Cultural History of Advertising Advertisements in the 1920s and 1930s)*. Shanghai: Shanghai renmin chubanshe, 2007.

Xiao, Bohong 肖鸿波. “Shenbao 77 nian tiyu baodao yanjiu”《申报》77 年体育报道研究（1872-1949）(A Study of 77 Years of Sports Reporting in the Shenbao).” Doctoral dissertation, Fudan University, 2011.

Xie, Shengming 谢圣明. 2014. “Chuanboxue Shiye Xia Shenbao Yu Zhongguo Meishu Xiandaihua Jincheng” 传播学视野下《申报》与中国美术现代化进程（1872-1937）(Shenbao and the Modernization Process of Chinese Art from the Perspective of Communication Studies).” Doctoral dissertation, Hangzhou: Zhejiang University.

Xu, Dizhen 徐娣珍. *Shanghaitan shiye xia de Cixi shangren: “Shen bao” sanbei shangbang shiliao jicheng 上海滩视野下的慈溪商人 : 《申报》三北商帮史料集成 (Cixi Merchants viewed from the Shanghai Bund: A Collection of Historical Materials on the Sanbei Merchant Group in the Shenbao)*. Beijing: Dangdai Zhongguo chubanshe, 2012.

Ye, Xiaoqing. *The Dianshizhai Pictorial : Shanghai Urban Life, 1884-1898*. Ann Arbor: Center for Chinese Studies the University of Michigan, 2003.

Yeh, Catherine. *Shanghai Love : Courtesans, Intellectuals, and Entertainment Culture, 1850-1910*. Seattle: University of Washington Press, 2006.

Zhang, Liqin 张立勤. 2012. “1927-1937 Nian Minying Baoye Jingying Yanjiu 1927-1937 年民营报业经营研究 (Research on the Management of Private Newspapers in 1927-1937).” Doctoral dissertation, Shanghai: Fudan University.

Zhu, Xiaokai 朱晓凯. “‘Shenbao yu zhong-fa zhanzheng yanjiu’《申报》与中法战争研究 (A Study of the Shenbao" and the Sino-French War.” Doctoral dissertation, Anhui University, 2015.

Hermeneutics

This document appears in the online version of the main paper as the "hermeneurics" layer, which we cannot reproduce on a conventional pdf document. We make it available here as a supplement disconnected from the relevant sections in the main text.

The distribution of nodes by degree is largely influenced by the extreme cases represented by the Mixed Courts. I obtain the same results for the whole network or for the main component of the network.

```
0% 10% 20% 30% 40% 50% 60% 70% 80% 90% 100%
1   1   1   1   1   1   2   2   3   4  2097
```

One-mode Network for Organizations

The basic network has 3,636 nodes and 67,285 edges, and it is made of 233 components. I selected only the nodes with at least two degrees. This reduced the network to 2,369 nodes and 6,853 edges. The main component retained has 1,642 nodes and 6,366 edges, with a diameter of 9 and a radius of 5. The characteristic path length is 3.322. The average number of neighbors is relatively high at 7.754. Network sensity is 0.005. I examined the network in by removing the nodes in successive steps at increasing levels of weight.

Nodes with weight = 1. The networks has 2,197 nodes and 41,656 edges. It has 12 components. The network is not readable as a hairball.
Nodes with weight = 3. The networks has 1,342 nodes and 23,530 edges. It has 12 components. The network is not readable as a hairball.
Nodes with weight = 4. The networks has 902 nodes and 15,585 edges. It has 10 components. The network is not readable as a hairball.
Nodes with weight = 5. The networks has 597 nodes and 10,670 edges. It has 10 components. The network is not readable as a hairball.
Nodes with weight = 10. The networks has 125 nodes and 2,883 edges. It has 2 components. The network is not readable as a hairball.
Nodes with weight = 15. The networks has 60 nodes and 1,116 edges. It has two components. It is starting to make sense, but the density in the inner core makes it difficult to read. The institutions with the highest betweenness degree are by decreasing order 招商輪船局 (0.03261478), 英界會審公堂, 工部局, 江海關, 法界會審公堂 (0.1970726), 松滬捐釐局, 法國租界巡捕房, 廣肇公所, 怡和洋行, 兩江總督, 法國公司, 江南製造局, 上海縣暑 (0.0123001). All nodes are interconnected at one step.
Nodes with weight = 20. The network has 39 nodes and 590 edges. It has one component. It is starting to make sense, but the density in the inner core makes it difficult to read. Diameter is 2. All nodes are interconnected at one step.
Nodes with weight = 30. The network has 21 nodes and 187 edges. It has one component. We can see that the main nodes are very much public institutions, except for the Shantung Road Hospital (仁濟醫院). Diameter is 2. All nodes are interconnected at one step.
Nodes with weight = 40. The networks has 14 nodes and 83 edges. It has 1 component. Diameter is 2. All nodes are interconnected at one step.

I also examined whether community detection would provide relevant clusters. I used the glay algorithm in Cytoscape. Community detection create4 94 clusters, which points to the heterogeneity of the network. The three largest clusters form hairball in themselves. I was able to qualify only a small number of clusters:

- Cluster 1 : it is a very dense network, with sub-clusters. The Shilin Buddhist Temple(師林禪寺) constitutes almost an ego-network in itself. We also find a lot of newspapers, which confirms that these nodes should be removed from the start. They are not actors *per se*, but "sources". It has an intrinsic value, but it somehow introduces a bias by creating connections that are even less real than the mentions of actors in articles.
- Cluster 3: it is made up mostly of Chinese local public offices, with a significant number of foreign goods companies. The 兩江營務處 and 發審局 are the central nodes that connect major parts of this cluster. Yet it retains the shape of a string network.
- Cluster 7: it is quite obviously the cluster of justice and police institutions
- Cluster 8: it is made up mostly of Chinese institutions, a lot at a high level (Liangjiang governor general 兩江總督), both locally and higher up (Ministry of Finance 戶部). Yet it seems to be a cluster centered on issues of maritime affairs, with the Customs, the Jiangnan Arsenal, Tax offices, maritime defense bureaus, but also some academies.

PCA analysis (Institutions)

Principal Component Analysis (PCA) was employed to scrutinise the data within the one-mode network of organisations. I constructed this network in R, utilising the igraph library. The process involved compiling indices for Degree, Eigenvector, Closeness, and Betweenness centralities, resulting in a file encompassing 2,408 nodes, each annotated with these four measures of centrality. These measures were then subjected to PCA.

The implementation of PCA was conducted using the R [FactoMineR](http://factominer.free.fr/) library, coupled with its Factoshiny interface (Lê 2008).

Notably, the data's variance is predominantly encapsulated within Dimension 1 (accounting for 68.52% of the variance) and Dimension 2 (comprising 24.57% of the variance).

Eigenvalues

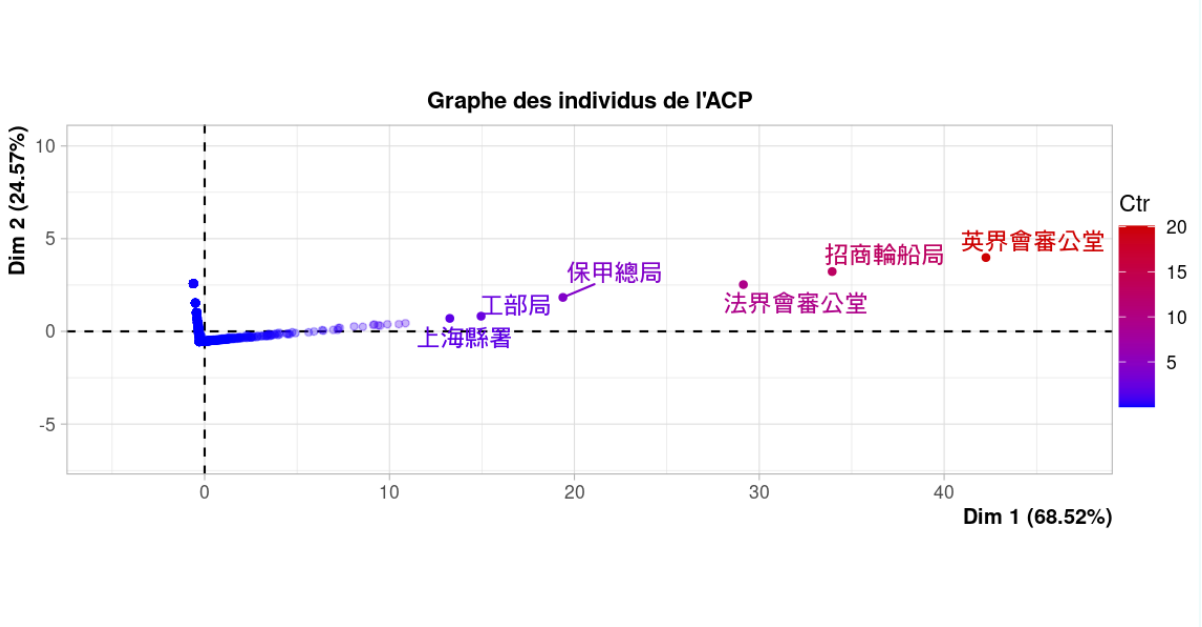
	Dim.1	Dim.2	Dim.3	Dim.4
Variance	2.741	0.983	0.226	0.050
% of var.	68.524	24.572	5.648	1.256
Cumulative % of var.	68.524	93.096	98.744	100.00

The graphical representation of variables indicates comprehensive projection for all variables, albeit with a marginally reduced projection for betweenness centrality. A robust correlation is evident between eigenvalue and betweenness, whilst degree centrality emerges as a pivotal factor in delineating Dimension 1. Conversely, closeness centrality asserts its relevance and significance predominantly in Dimension 2.

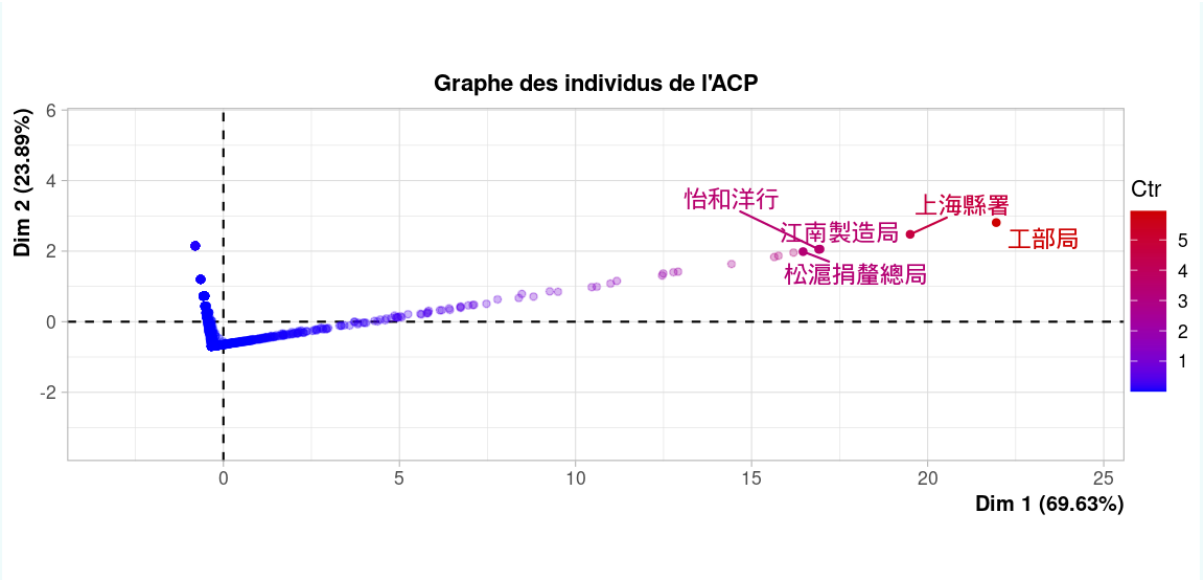
Variables								
	Dim.1	ctr	cos2	ctr	cos2	Dim.3	ctr	cos2
DegreeMC_n	0.982	35.149	0.963	0.148	0.001	0.058	1.491	0.003
EigMC	0.946	32.627	0.894	0.254	0.002	0.299	39.558	0.089
Betweenness	0.924	31.177	0.855	0.806	0.008	-0.365	58.844	0.133
ClosenessMC	0.169	1.047	0.029	98.792	0.971	0.016	0.107	0.000

Graph of individuals (Institutions)

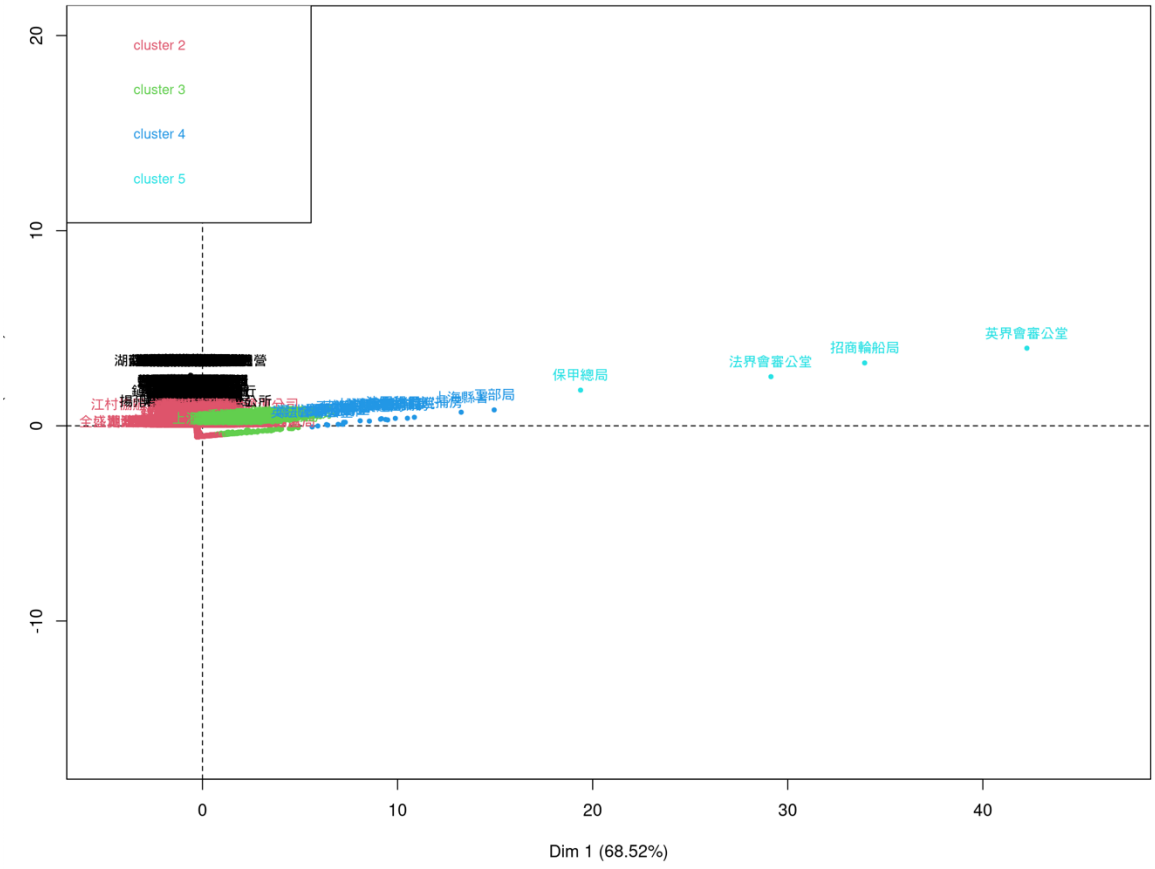
For the 'Contribution' parameter, I selected a value of 6. The distribution of institutions is markedly skewed due to the presence of outliers, resulting in the aggregation of most other institutions into a homogenous group. The PCA distinctly segregates local-level (Shanghai) institutions along Dimension 2. Within this local-level cohort, there is a discernible, more pronounced correlation among the outliers in this group (such as the mixed courts and CMSNC) and entities like the xian magistrate (上海縣暑), Shanghai Municipal Council (工部局), and the Baojia Bureau (保甲總局). Employing cos2 for analysis does not alter this observed configuration.



Upon the exclusion of the four principal outliers, the overall distribution remained largely unchanged. This is attributable to the fact that the majority of nodes contribute minimally and are collectively positioned in a negative correlation on both Dimension 1 and Dimension 2.



Upon examining various classification levels (ranging from 4 to 8), I noted a lack of substantive relevance in augmenting the number of classes. This increase did not significantly alter the nature of the highly structured clusters nor the less significant ones. Consequently, I opted for a five-class structure, which yielded the graph presented below.



The delineation of nodes into five classes distinctly segregates the four principal outliers within cluster 5, a pattern consistent across varying numbers of classes. Cluster 4

amalgamates Chinese institutions predominantly involved with customs, taxation, and the Chinese police.

The association between the cluster variable and the quantitative variables underscores the paramount role of betweenness centrality in defining cluster 5. Conversely, the eigenvalue in cluster 4 brings to prominence institutions interconnected with members of cluster 5. Degree centrality holds significance in cluster 3, although there is considerable overlap with cluster 2. Cluster 1, for the most part, remains indistinct and undifferentiated.

	DegreeMC_n	EigMC	BetweennessMC	ClosenessMC
Eta2	0.8945649	0.7761904	0.9258594	0.8349257
P-value	0	0	0	0

Description of each cluster by quantitative variables

Clust. 1		ClosenessMC	BetweennessMC	EigMC	DegreeMC_n
	v.test Mean	44.825651	-3.627987	-5.177998	-6.129161
	in category	3.624431e-01	1.851852e-02	1.706968e-21	4.770038e-04
	Overall	8.377262e-02	1.645067e+03	8.180086e-03	2.446491e-03
	mean sd in	1.289044e-01	1.469248e-01	2.728480e-20	1.498514e-04
	category	1.586926e-01	1.157458e+04	4.032627e-02	8.202464e-03
	Overall sd	0.000000e+00	2.856400e-04	2.242791e-07	8.834381e-10
	p.value				
Clust. 2		BetweennessMC	EigMC	DegreeMC_n	ClosenessMC
	v.test Mean	-7.264599	-8.633439	-10.155779	-39.032844
	in category	6.281618e+02	3.969566e-03	1.439046e-03	8.860775e-03
	Overall	1.645067e+03	8.180086e-03	2.446491e-03	8.377262e-02
	mean sd in	1.553849e+03	6.699250e-03	1.588681e-03	2.888689e-02
	category	1.157458e+04	4.032627e-02	8.202464e-03	1.586926e-01
	Overall sd	3.741473e-13	5.953428e-18	3.123065e-24	0.000000e+00
	p.value				
Clust. 3		DegreeMC_n	EigMC	BetweennessMC	ClosenessMC
	v.test Mean	17.199083	11.575891	9.442549	-5.127246
	in category	1.695604e-02	5.619175e-02	1.288591e+04	8.812975e-05
	Overall	2.446491e-03	8.180086e-03	1.645067e+03	8.377262e-02
	mean sd in	7.700841e-03	3.861118e-02	9.181137e+03	5.829761e-06
	category	8.202464e-03	4.032627e-02	1.157458e+04	1.586926e-01
	Overall sd	2.697652e-66	5.460103e-31	3.638180e-21	2.940110e-07
	p.value				
Clust. 4		EigMC	DegreeMC_n	BetweennessMC	ClosenessMC
	v.test Mean	27.751813	26.162801	11.994356	-2.307217
	in category	2.639637e-01	5.149454e-02	3.337546e+04	8.944714e-05
	Overall	8.180086e-03	2.446491e-03	1.645067e+03	8.377262e-02
	mean sd in	1.095409e-01	1.075334e-02	1.232360e+04	1.811750e-06
	category	4.032627e-02	8.202464e-03	1.157458e+04	1.586926e-01
	Overall sd	1.656855e-169	7.047639e-151	3.803648e-33	2.104274e-02
	p.value				
Clust. 5		BetweennessMC	DegreeMC_n	EigMC	
	v.test Mean	44.48831	33.68921	30.59926	
	in category	2.589513e+05	1.405276e-01	6.247724e-01	
	Overall	1.645067e+03	2.446491e-03	8.180086e-03	
	mean sd in	4.759604e+04	3.528968e-02	3.289517e-01	

	category Overall sd p.value	1.157458e+04 0.000000e+00	8.202464e-03 8.317351e-249	4.032627e-02 1.252121e-205	

Topic modeling

To construct the file for tokenisation, I commenced with the primary file containing validated names (emin_Names_MainAll), supplementing it with the full text of related articles. Upon eliminating duplicates, a file comprising 50,565 documents was compiled. I integrated this full-text file into the tokenizer developed in collaboration with Huang Hen-hsen at the Institute of Computing, Academia Sinica. The foundational pre-trained transformer model for the word segmenter is a Roberta model for Chinese, employing whole word masking (https://huggingface.co/hfl/chinese-roberta-wwm-ext). This model was subsequently fine-tuned to perform word segmentation (token classification with BIS tags) on a meticulously selected subset of the Shenbao, which had been manually tokenised.

The training set encompasses 663 articles, totalling 445,825 sinographs. A development set, comprising 24 articles (6,184 sinographs), facilitated the selection of the optimal training epoch and random seed. The model's efficacy was assessed on a test set of 54 articles (12,876 sinographs) drawn from 10 distinct time periods, ensuring a balanced representation of sinographs from each period. Our word segmenter achieved an accuracy score of 84% and an F1-score of 82% on this test set.

The resultant file contains 41,667 rows, each encapsulating tokens from all documents. However, 8,832 rows lacked full text. I isolated these rows and matched them against the full-text file. While not all IDs corresponded, this process yielded 7,181 rows with complete text. Subsequent to the removal of entries related to the Peking Gazette and advertisements, the file was reduced to 6,054 rows. This data was then tokenised and incorporated into the principal token files.

In the final stages, I refined the file by excising stopwords not included in standard lists (based on a distant reading of the results) and eliminating the remaining few rows of advertisements. Each iteration of stopword removal resulted in the creation of a new column for tokens.

Miscellaneous

The harmonisation of names was achieved through a combination of automated correction and meticulous manual curation. This process entailed aligning the names listed under organisations with those in an evolving index, cultivated through successive rounds of data processing. This index encompasses all identified variants of names alongside their standardised counterparts. Names failing to correspond with any in the index necessitate manual verification and normalisation. Each newly identified pair of terms is methodically incorporated into the reference index.

The Shenbao, during this period, lacked punctuation marks, rendering the segmentation of articles into sentences unfeasible. This, in turn, complicated the segmentation into tokens. Initially, the model processed inputs of 256 characters directly. To mitigate the issue of tokens potentially spanning two blocks of 256 characters, a stride of 10 characters was introduced at the inference stage.

Of the 20 topics identified by the model, only three are distinct to it: Local Officials, Ministry Officials, and Miscellaneous (the latter predominantly encompasses incidents such as fires). In response to these findings, I undertook two actions:

1. I re-categorised the topics to render them more generic, based on my analysis of a representative sample comprising 5 articles per topic per model (140 unique articles in total, with 33 duplicates and 1 triplicate, as the two models occasionally identified the same articles in their 5-article samples).
2. I determined that Model 15 (M15) is presently the most suitable option.