Essays in Political Economy

Inauguraldissertation zur Erlangung des akademischen Grades
eines Doktors der Wirtschaftswissenschaften der
Universitaet Mannheim

Chia-Yu Tsai

2018 Spring

Dean of the Economics department:
Prof. Dr. Jochen Streb

Thesis advisors:
Prof. Dr. Eckhard Janeba
Prof. Dr. Antonio Ciccone

External committee member:
Prof. Dr. Ulrich J. Wagner

Oral defense date: 23.05.2018

Contents

List of Figures

List of Tables

For my mother and my homeland Taiwan.

Acknowledgments

I would like to thank all the people who contributed in some way to the thesis. First and foremost, I would like to express my sincere gratitude to my supervisors Prof. Dr. Eckhard Janeba and Prof. Dr. Antonio Ciccone for the continuous support of my Ph.D study and for allowing me to grow as an economist. Their advice on both research as well as on my career have been priceless. Besides my supervisors, I would like to thank Prof. Dr. Galina Zudenkova and Prof. Dr. Jin-Tan Liu for their insightful comments on my research and encouragement. My sincere thanks also goes to Prof. Dr. Ming-Jen Lin and Prof. Dr. Chun-Fang Chiang, who offered me an opportunity for a research stay at the National Taiwan University and provided constructive comments on my research. I would especially like to thank Prof. Dr. Ruben Durante for inspiring me to conduct research in Media Economics during my study at Yale University.

I would like to acknowledge the University of Mannheim's Graduate School of Economic and Social Sciences funded by the German Research Foundation (DFG). My graduate experience benefited greatly from the course work, from the opportunity to serve as a teaching assistant under Prof. Dr. Henrik Orzen, and from the financial support for attending conferences and international research stays. Especially, I would like to thank Marion Lehnert, Dr. Dagmar Röttsches, Dr. Sandro Holzheimer, and Claudius Werry for always being supportive and kind to me.

Last but not least, I would like to thank my family for supporting me along the journey of my research work. In particular, I would like to thank my mom for her everyday company on Skype. I would also like to express my gratitude to my grandfather who supported me when I decided to study abroad. Finally, I would like to thank my fellow classmates and friends for the numerous thought-provoking discussions and conversations, for motivating me to strive towards my goal, and for all the fun we have had in the past six years.

Chapter 0

Introduction

The thesis contains three chapters in Political Economy with a focus on text mining in Chinese. In Chapter 1, I show that mass media can affect citizens' subjective feeling of national identity based on a case study in Taiwan. To do so, I first construct a measure of media ideology to assess the attitude towards China of 12 Taiwanese newspapers from 2000 to 2014. Then, I study how a change in ownership influenced the attitude of two mainstream newspapers. In particular, I employ a difference-in-difference method to show that after two Taiwanese newspapers were sold to a businessman who has great business interest in China and advocates unification with China, their ideologies became more pro-China. Finally, I explore the relationship between media ideology and citizens' subjective feeling of national identity. Adopting an instrumental variable approach with individual and time fixed effects, I find evidence that media ideology has significant effects on national identity. Coupled with the ownership effects, this implies that individual national identity can be susceptible to the influence of directly or indirectly captured media.

In Chapter 2, I examine protest in China and focus on the power structure of county leaders–whose promotion hinges on maintaining social stability in the county–and study how power and leaders' characteristics affect protest frequency at the county level. Using text mining to compile two novel data sets on protests and leadership, I demonstrate that power has negative effects on the frequency of protests. Next, I show that power influences protests by both reducing the incentive for collective action and raising grievances. I first concentrate on environmental protests and find that given the same level of air pollution, a proxy for the root of grievances, people are less likely to protest under the rule of a pow-

erful leader. Then, I show that power has opposite effects on protest frequency in terms of leaders' personal characteristics–under the rule of a "local" leader, whose hometown is located within the same prefecture as his/her ruling county, higher level of power contributes to more protests, whereas fewer protests are observed when an "outsider" gains more power. The results suggest that the discrepancy is due to reduced welfare under the rule of powerful "local" leaders, who enjoy better connections to superior authorities but are less competent than "outsiders".

In Chapter 3, I discuss the construction of my media ideology measure in Chapter 1, with details in word segmentation methods, keyword extraction algorithms, word association measures, and similarity metrics. The core concept of my measure of media ideology is to compare the use of language in Taiwanese newspapers with the one in press conference transcripts of the Taiwan Affairs Office (TAO), the official Chinese government institution in charge of policies related to Taiwan. Based on the comparison, my measure evaluates how closely each Taiwanese newspaper aligns with TAO. If a Taiwanese newspaper shares similar views with TAO, then it is assigned a high score of my measure and considered more pro-China. In line with expectations, the values of media ideology measure of the most pro-China Taiwanese newspapers, the *United Daily News* (聯合報) and the *China Times* (中國時報), are almost always above the average of all Taiwanese newspapers in the database and much higher than the ones of the least pro-China Taiwanese newspapers, the *Apple Daily* (蘋果日報) and the *Liberty Times* (自由時報).

The thesis is structured as follows: The main findings are presented in Chapters 1 to 3. Additional results are included in the Appendix A, and all the references are in References.

Chapter 1

Mass Media and National Identity

1.1 Introduction

Mass media is often considered an essential part in nation-building policies (Karthigesu, 1988; Postill, 2008). However, empirical evidence is scarce. Can mass media really affect citizens' subjective feeling of national identity? How does mass media influence citizens' subjective feeling of national identity and how large is the effect? To explore these questions, I conduct a case study in Taiwan.

Taiwan provides an excellent context for investigating how mass media affect citizens' subjective feeling of national identity. On the one hand, feelings of national identity have shifted substantially over the past two decades. Due to the special cross-Strait relations between Taiwan and China, the citizens of Taiwan may identify themselves as Chinese, Taiwanese, or both. According to surveys from the Election Study Center administered by National Chengchi University (Figure 1.1), the share of people identifying themselves as Taiwanese has grown from less than 20% to ca. 60%, while the share of people identifying themselves as Chinese has declined from more than 25% to ca. 4% over the past 20 years. It is also worth noting that ca. 30-40% of people in Taiwan think of themselves as both Taiwanese and Chinese, suggesting the national identity of many is still in a transition stage. On the other hand, the media industry in Taiwan also changed dramatically. Since the lift of martial law in 1987 and subsequent media liberalization, Taiwanese media have

transformed from state-controlled companies in favor of Chinese nationalism[1] to market-oriented ones with diverse views on the cross-Strait relations. As a result, Taiwan became one of only two Asian countries that enjoy full press freedom (Freedom House, 2015). The concurrent two phenomena naturally lead to the question of whether the diversity of media ideologies affected the changes in Taiwanese citizens' subjective feeling of national identity.

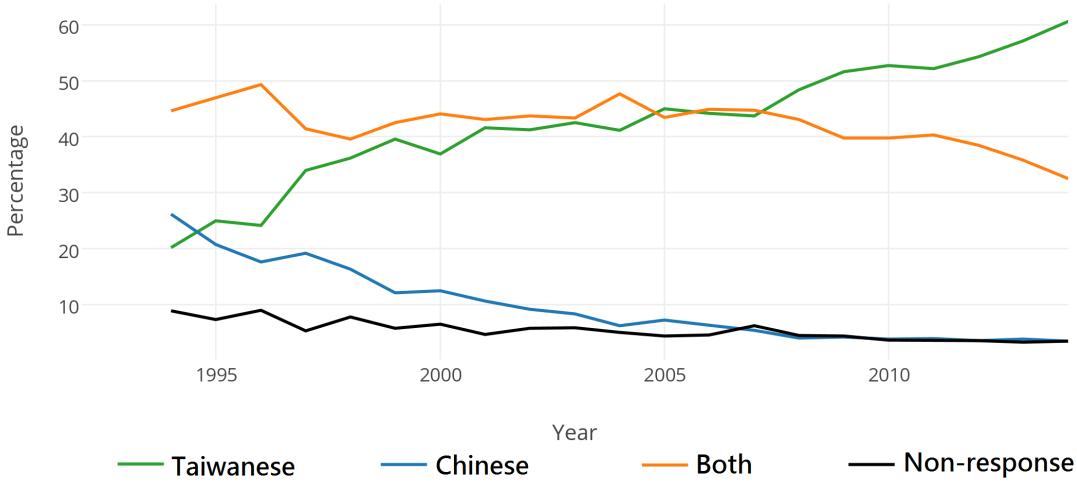


Figure 1.1: Changes in the Taiwanese/Chinese identity of Taiwanese from 1992 to 2014).

Source: Election Study Center, National Chengchi University (NCCU)

To study the relationship between the stance of Taiwanese newspapers towards China and Taiwanese citizens' subjective feeling of national identity, I first measure how pro-China each Taiwanese newspaper is and how its stance varies over time. Built on Gentzkow and Shapiro (2010) and Murphy and Westbury (2013), my approach basically compares the language used in the headlines of Taiwanese newspapers with the language used in press re-

[1]Generally speaking, Chinese nationalism refers to the idea that all people within the territories of China should build a single nation, perceive themselves as Chinese, and unite to fight against foreigners, such as Japanese and Westerners. Nevertheless, the content of Chinese nationalism differs between Taiwan and China and varies over time. Before 1987 in Taiwan, the "China" in Chinese nationalism refers to the Republic of China (ROC) represented by the Kuomintang (KMT) government, in contrast to the People's Republic of China (PRC) represented by the Chinese Communist Party (CCP). The KMT government has deliberately constructed and strengthened Chinese nationalism to "legitimate its dictatorship" after it lost a civil war to CCP in 1949 (Hughes & Stone, 1999).

leases of the Chinese government institution in charge of the policies towards Taiwan, the Taiwan Affairs Office (TAO). Based on this comparison, I place each Taiwanese newspaper on a one-dimensional ideological line that ranges from less to more pro China. A more pro-China newspaper is more likely to focus on positive facts regarding the cooperation or even unification of Taiwan with China (e.g. potential economic gains), provide greater coverage of positive news about China (e.g. economic development in China), frame news stories in ways that favor China and the Chinese government, and attack politicians, scholars, and protesters in Taiwan and elsewhere with positions that are less pro-China. To evaluate the pro-China stance of each Taiwanese newspaper in all these dimensions, the measure of media ideology considers diction, issue coverage, and context. In particular, I first examine press conference transcripts from TAO and compile monthly lexicons to represent the Chinese government's stance on topics about China-Taiwan relations. Then, I build a second order co-occurrence model (Murphy & Westbury, 2013) to compare TAO's use of the words in this lexicon with Taiwanese newspapers. If a Taiwanese newspaper systematically uses the same words in a similar fashion as TAO does, and diction, issue coverage and context also resemble TAO press conference transcripts, and the newspaper is classified as ideologically close to the Chinese government.

My measure of the pro-China stance of Taiwanese newspapers is novel in two aspects. First, different from measures of media slant in the existing literature, it is based on international politics, not on domestic political competition. That is, the measure does not examine bias towards a certain domestic political party, for example, whether U.S. newspapers are more pro-Democratic or pro-Republican. Instead, it analyzes the extent to which Taiwanese newspapers agree with the position of the Chinese government on issues related to China and Taiwan. Such agreement with the Chinese government is more relevant for the study of subjective feeling of national identity in Taiwan as issues on cross-Strait relations are more likely to affect Taiwanese readers' sense of national identity than other issues. Also, it is easier to measure the degree of agreement with the official Chinese positions directly than to examine how Taiwanese newspapers defend domestic political parties that take more pro-China positions. The relationship with China is arguably the most important issue in Taiwan and has a huge potential impact on its future. Nevertheless, due to its subtlety, domestic political parties often choose an ambiguous stance on the pro-China to

anti-China spectrum. Both the ruling party and the opposition party intentionally avoid political disputes on the relationship with China and tend to support the status quo. Besides, domestic politics encompasses various issues and creates numerous dimensions on media ideology. Thus, instead of measuring media ideology based on domestic politics, it is more direct and salient to examine media ideology in the way I do. The second novel aspect of my measure of media slant is that it is one of the first objective measures available for the Chinese language. The existing literature of mass media is generally focused on the English language and text analysis techniques are therefore mostly developed for English. I borrow techniques of Chinese text analysis from natural language processing to develop one of the first measures of media ideology in Chinese.

I use my new measure of media slant to investigate ownership effects on media ideology. According to the supply-side story in the literature, media ideology can be influenced by owners, and a change of ownership usually comes with a modification of media ideology. In November 2008, two mainstream newspapers, the *China Times* and the *Commercial Times*, were sold to a businessman who has great business interests in China and openly supports unification with China. Anecdotes suggest that the two newspapers became more pro-China after the acquisition. To verify the anecdotes, I adopt a difference-in-difference approach, where the treatment group is the *China Times* and the *Commercial Times* and the control group consists of all other Taiwanese newspapers. The results are consistent with the anecdote. Compared to the control group, the two newspapers in the treatment group have become 14.7-26.7% more slanted towards China after the change of ownership, which is both statistically and economically significant.

Finally, I also examine the effects of media ideology on the subjective feeling of national identity of Taiwanese newspaper readers. Nation is "an imagined community" (Anderson, 1991) and socially constructed based on history, traditions, language, culture, etc. National identity is the subjective feeling of a sense of belonging to a nation, regardless of one's citizenship (Guibernau, 2004). National identification involves the process of identifying with one's nation and differentiating with other nations (Smith, 1991). Mass media has a part in shaping the "imagined community" as individuals understand and interpret their nation (and other nations) partly based on the information provided by the media (Schiller, 1985).

In the context of Taiwan, media ideology determines the degree to which news stories are refracted through the lens of Chinese/Taiwanese identity. Media may choose to underscore the similarities or differences between Taiwan and China. If people in Taiwan recognize a large gap between themselves and people in China, they tend to self-classify as a group different from Chinese and choose to be Taiwanese, instead of Chinese. For example, pro-China media tend to emphasize the concept of Great China and cover stories about traditional Chinese culture from which some Taiwanese culture originated. On the contrary, anti-China media are inclined to contrast China's authoritarian regime with Taiwan's democracy. In addition, media may portray Taiwan/China in a good or bad light and create an image of a country that one would (not) like to identify with. For example, pro-China media are more likely to highlight the economic development in China, use commendatory terms to describe China, and ignore the human rights situations in China. Moreover, media may play down or up areas where Taiwan outperforms China to minimize or maximize the difference between Taiwan and China. For instance, pro-China media may cover lots of news stories about legislative fights in Taiwan and praise China's administrative efficiency and economic growth, suggesting that Taiwan's democracy is not an advantage compared to China's autocracy.

To examine the effects of media ideology on the subjective feeling of national identity in Taiwan, I combine my measure of media ideology with rotating panel surveys of Taiwanese newspaper readers to see whether changes in media ideology affect their subjective feeling of national identity. There are several econometric challenges in doing so. First, the omitted variable problem is unavoidable. For example, it is difficult to measure cross-Strait relations between China and Taiwan, which is very likely to affect both media ideology and national identity and introduce bias to the results. Second, reverse causality cannot be ignored. The growing trend of Taiwanese identity can encourage profit-seeking newspapers to modify their stance towards China to meet the demand of their readers.

To address some of the econometric challenges, I employ fixed effects models with two instrumental variables. The fixed effects include both individual fixed effects and time fixed effects. Individual fixed effects capture any time-invariant individual characteristics, such as parents' subjective feeling of national identity, while time fixed effects explain general factors that influence all individuals over time, such as national elections in 2004, 2008,

and 2012. Since fixed effects cannot account for omitted time-variant individual factors and reverse causality, I further utilize two instrumental variables. The first instrumental variable is the media ideology of the newspaper read in the last period. I argue that the ideology of the last media choice is correlated with the ideology of the current media choice, and that the ideology of the last media choice only influences individual national identity through the ideology of the current media choice. The second instrument variable is the corpus size of a newspaper. I argue that corpus size is correlated with media ideology (corr = 0.702), but does not directly relate to individual national identity. The correlation is based on the fact that most words have low frequency in a corpus (Piantadosi, 2014) and that the sparsity of words is more evident in small corpora. Since the measure hinges on the comparison of word frequency, newspapers with small corpus size are more likely to be assigned low values in the measure of media ideology, regardless of their actual stance on the cross-Strait relations. The results from my fixed effects models with instrumental variables indicate that if a newspaper takes a more pro-China ideological position, the probability that readers self-identify as Taiwanese drops significantly.

When combined with the ownership effects on the *China Times* and the *Commercial Times*, my results suggest that it is possible for the Chinese government to alter individual national identity in Taiwan by directly buying or indirectly capturing Taiwanese media. As former Chinese party chief Jiang Zemin said in 1998, "Controlling Taiwanese media is equivalent to controlling 50% of public opinion in Taiwan. This control is more useful than sending army there." The chapter provides some suggestive evidence that this claim may not be exaggerated.

The remainder of this chapter is organized as follows. Section 1.2 contains a short literature review on media bias and national identity. Section 1.3 provides background on cross-Strait relations between China and Taiwan, contemporary national identification, domestic politics, and the newspaper industry in Taiwan. Section 1.4 describes the data on press conference transcripts from the Taiwan Affairs Office, the collection of newspaper headlines, and the surveys on national identity and newspaper reading. Section 1.5 explains my measure of the pro-China stance of Taiwanese newspapers. Section 1.6 covers empirical applications, including ownership effects and the influence of media ideology on individual national identity. Section 1.7 concludes.

1.2 Literature Review

This chapter is linked to the literature on media bias. Media can choose what facts or supportive evidence to be included or excluded in a piece of news (facts bias). Groseclose and Milyo (2005) provide one of the first objective measures of media slant based on facts bias. By comparing think tank citation patterns of a news outlet with the ones of Congressmen, they attach an adjusted Americans for Democratic Action (ADA) score to each news outlet. Nevertheless, this measure is heavily criticized because think tanks are cited for various reasons. For instance, it is likely that a Congressman cites an article from a think tank to strengthen his point, while journalists use the same citation for sarcasm. Media can also selectively cover certain issues and ignore others (issue bias). Larcinese, Puglisi, and Snyder Jr (2011) study the news coverage of economic issues and find that, compared to newspapers with Republican endorsement, newspapers with Democratic endorsement systematically cover more economic bad news during Republican presidency than during Democratic presidency.

The measure in this chapter is based on the ideas from Gentzkow and Shapiro (2010) and Murphy and Westbury (2013). Their methods consider various types of media bias, such as facts bias, issue bias, and framing bias, and provide a more comprehensive proxy for media ideology (Prat & Strömberg, 2013). Gentzkow and Shapiro (2010) measure the similarity of language between newspapers and Congressmen (Prat & Strömberg, 2013). They first examine the phrases used by Congressmen and make separate lists of Republican and Democratic vocabulary. Then, they study the resemblance of phrases between each newspaper and Democratic/Republican Congressmen, and construct a measure of relative biasedness towards Democrats or Republicans. Murphy and Westbury (2013) apply co-occurrence models to analyze contexts and construct objective measures of word meaning and associations.

There are two alternative explanations for media bias: demand-side and supply-side stories. A demand-side story states that media bias originates from media revenue (or profit) maximization and reflects reader preferences, while a supply-side story claims that media bias comes from preferences of journalists, owners, or governments.

This chapter is connected to the literature on the supply-side story. Djankov, McLiesh, Nenova, and Shleifer (2003) compile a cross-sectional dataset of 97 countries on media ownership and examine two alternative theories on state-owned media: public interest theory (a benevolent government cures market failure) and public choice theory (state-owned media manipulate information and undermine economic and political freedoms). They correlate media ownership with policy outcomes and find that the data support the latter theory. Durante and Knight (2012) examine the effects of partisan control on media ideology and study how viewers respond to the ideological change. Using data from Italy, they first demonstrate that state-owned televisions shift to the right when the government moves from the center-left to the center-right party. Then, they use individual survey data to show that viewers respond by shifting to channels closer to their own ideology and conclude that in a competitive media environment, the change of media control has limited influence on viewers' ideological exposure. For a comprehensive survey on media literature, refer to Prat and Strömberg (2013) and Strömberg (2015).

This chapter is also related to literature on nationalism. In social science and political science, there are two main perspectives on the origins of nationalism–primordialism and constructivism. On the one hand, primordialists argue that national identity is formed by predetermined characteristics and cannot be produced, modified, or transformed through manipulation. On the other hand, constructivists think national identity is not fixed. Nation is a socially constructed community, imagined by the people who perceive themselves as part of that group (Anderson, 1991), and national identity is constructed and transformed by history, language, and culture. In this chapter, I follow the idea of constructivism and assume national identity is changeable.

In contrast to the rich literature on national identity in social science and political science, there are only few papers in economics about national identity. The existing empirical literature centers on nation-building policies and emphasizes the potential benefits of a common identity. For example, Miguel (2004) studies nation-building reforms in two East-African countries by examining the effects on inter-ethnic cooperation in two neighboring districts, one in Tanzania and the other in Kenya. Using "colonial-era national boundary placement as a natural experiment," he shows that given similar colonization histories, ethnic composition, and geographical conditions between the two countries, the adoption

of nation-building policies in Tanzania leads to better economic outcomes, compared to Kenya, where no such policies were implemented.

In addition to studies on nation-building policies, there is a growing literature focused on the formation of national identity. Georgiadis and Manning (2013) use data from England and Wales' 2007 Citizenship Survey and investigate the determinants of national identity. Based on the seminal work of Akerlof and Kranton (2000), the first paper incorporating identity into utility functions, they construct a conceptual framework to predict correlations between national identity and its potential determinants. Their main finding is that people who feel well-treated and tolerated are more likely to identify with the society. Masella (2013) studies 21 countries and uses data from the World Values Survey (WVS) to measure national identity at the individual level. Examining the relationship between national identity and ethnicity diversity, he finds that in countries with high ethnic diversity, nationalist feelings are weaker in minority groups than in the majorities, while in countries with low ethnic diversity, the reverse is true. Aspachs-Bracons, Clots-Figueras, Costa-Font, and Masella (2008) investigate the 1983 educational reform in Spain that turned the Catalan education system into a bilingual system. They use survey data to show that compulsory language policies have a stronger impact on identity than non-compulsory language policies. Analyzing the effect of the same education reform on the process of national identity formation, Clots-Figueras and Masella (2013) exploit variation in the number of years of compulsory education under the new system and find that individuals with longer exposure to bilingual education are more likely to identify themselves as Catalan.

The goal of the chapter is to explore the effect of media ideology on the construction of national identity. The existing literature discussing the relationship between mass media and national identity generally focuses on descriptive analysis, not quantitative research. Karthigesu (1988) examines the role of Television Malaysia, a national mass media organization, in nation building during the post-colonial period, and studies the adverse effects of restrictive and manipulative programming on the democratic transition. Simeunovic (2009) analyzes secondary data from different studies and investigates the effect of mass media on European identity formation in the contemporary European media industry.

1.3 Background

1.3.1 Cross-Strait Relations and Contemporary National Identification in Taiwan

The term cross-Strait relations refers to the relations between the two political entities, the People's Republic of China (PRC, or China) and the Republic of China (ROC, or Taiwan). Since 1949, Taiwan and China have been separately governed, with governments on both sides claiming to be the only legitimate political power of China. After over 60 years of separation, cross-Strait relations are now focused on the issue of unification with China versus Taiwan independence. On the one hand, the PRC (or the Chinese government) insists on a "One China Policy" and claims that Taiwan is a province of China. The PRC threatens to use military force against Taiwan if Taiwan declares its own independence. On the other hand, the majority of Taiwanese people prefer to maintain the status quo. According to surveys from Election Study Center administered by National Chengchi University (Figure 1.2), over 50% of people favored the status quo in 2001; the proportion of status quo supporters has increased over time and reached almost 60% in 2013. It is also worth noting that since 2002, the proportion of people favoring Taiwan independence has exceeded the proportion of people favoring unification with China. The share of Taiwan independence supporters has maintained around 20% since 2003, but the majority still favors the status quo.

Since 2008, both governments have set aside political dispute and focused on economic prospects, bringing about more peaceful cross-Strait relations and greater economic integration. Nevertheless, Figure 1.1 shows that around the same time, the proportion of people who identify themselves exclusively as Taiwanese stood at 48.4% in 2008 and exceeded the proportion of people who identify themselves as both Taiwanese and Chinese for the first time (Election Study Center, NCCU). The proportion of Taiwanese identity continued to increase over time and reached 57.1% in 2013, suggesting a surge of Taiwanese nationalism. The relationship between economic integration and political disintegration has been studied by Chiang, Liu, and Wen (2014). They find that rising investment in China strength-

ens Taiwanese identification among unskilled workers but has had no effect among skilled workers, leading to stronger Taiwanese nationalism overall.

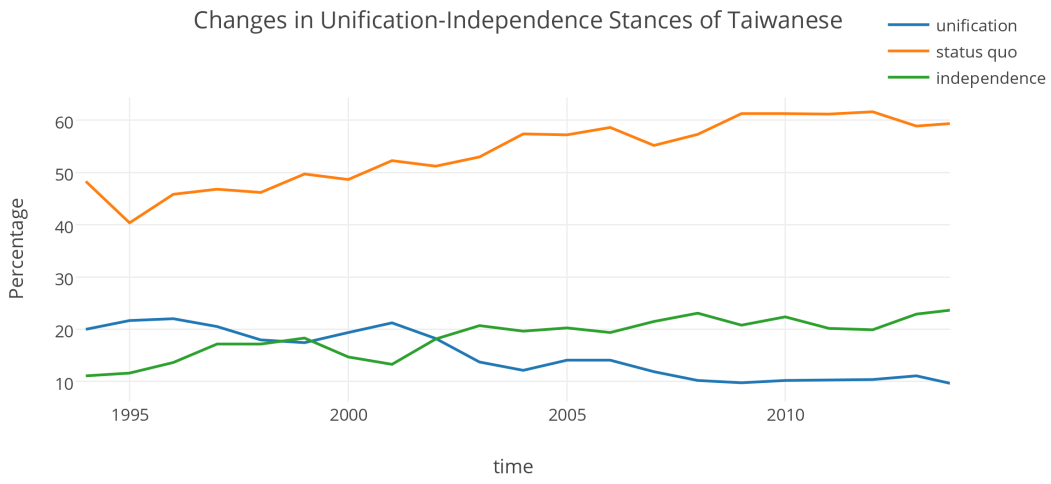


Figure 1.2: Changes in the unification-independence stances of Taiwanese from 1994 to 2014)

Source: Election Study Center, National Chengchi University (NCCU).
Originally, there are 7 options in the survey: (1) unification as soon as possible; (2) maintaining status quo and moving toward unification; (3) maintaining status quo and deciding at a later date; (4) maintaining status quo indefinitely; (5) maintaining status quo and moving toward independence; (6) independence as soon as possible; (7) no response. In the graph, I combine (1) and (2) as unification, (3) and (4) as status quo, (5) and (6) as independence, and ignore (7).

1.3.2 Contemporary Domestic Politics in Taiwan

During the martial law era (1949-1987), the formation of new political parties was prohibited. Taiwan was under a de facto single-party state system, where Kuomintang (KMT, or the Chinese Nationalist Party) had stayed in power for 38 years. However, after the lift of martial law, the freedom of assembly was no longer restricted. Currently, there are 337 political parties in Taiwan (Ministry of the Interior, Department of Civil Affairs, 2018), but only 5 parties hold places in the Legislative Yuan. Political parties are generally divided into two camps–the Pan-Blue Coalition and the Pan-Green Coalition. Headed by the KMT, the Pan-Blue Coalition adopts a softer stance towards China and promotes economic inte-

gration with China. The Pan-Blue Coalition used to support unification of Taiwan with China and claimed that Republic of China (ROC, the government in Taiwan), not People's Republic of China (PRC, the government in mainland China), was the only legitimate government of China. Nevertheless, given the rise of Taiwanese nationalism, the Pan-Blue Coalition now adopts a more conservative position in favor of the present status quo and opposes both immediate unification and immediate independence, as reflected by the KMT's "three noes" policy–no unification, no independence, and no use of force. On the other hand, the Pan-Green Coalition advocates Taiwanese nationalism and favors Taiwan independence over unification with China. However, since the majority of constituency support the present status quo, the leading party of the Pan-Green Coalition, the Democratic Progressive Party (DPP), does not adhere to Taiwan independence and holds a more moderate stance on cross-Strait relations to gain more votes in elections.

1.3.3 Newspaper Industry in Taiwan

Before the lift of martial law in 1987, press freedom in Taiwan was restricted and the mainstream media served as ideological apparatus of the state. The newspaper industry was highly regulated and controlled either directly or indirectly by the KMT regime. In particular, three restrictions, known as press bans, were imposed on the newspaper industry: (1) media operation control (entry permission); (2) pricing control; (3) media space control. In addition to strict press censorship, social and political dissidents were excluded from the newspaper industry (Hung, 2013).

The largest two newspapers, the *United Daily News* (UDN) and the *China Times*, were owned by core members of the KMT's Central Standing Committee, and UDN was recognized as party newspaper of the KMT. Until now, the two newspapers are still regarded as pan-Blue mass media. UDN retained its pro-KMT stance, supports President Ma's cross-strait political reconciliation policy and is optimistic about the growing economic cooperation between Taiwan and China, while the *China Times* changed from a moderate pro-KMT political stance to a radical one and became a strong unification advocate after the change in its ownership. In 2009, Yen-ming Tsai, a China-based food manufacturer, acquired the China Times Group, merging the *China Times*, a KMT-owned terrestrial

TV station, *CTV*, and another private TV station, *CTI*, into a media conglomerate. Tsai takes a clear-cut stand in favor of China. The media group often promotes Chinese officials and makes profits from Chinese product placement, which violates the current regulation (Hung, 2013).

Following the end of martial law in 1987, declining media control and press censorship led to a highly competitive print media industry (Fell, 2005). One of the most important newly founded newspapers is the *Liberty Times*. Established as a local newspaper by a Taiwanese business group, it was turned into a national newspaper in 1987 and gained substantial market share in the 1990s by using aggressive marketing strategies (Chyi & Huang, 2011). Nowadays, the *Liberty Times* has become a major pan-Green newspaper and takes a pro-independence political stance. Another important newspaper during post-martial law period is the *Apple Daily*. Launched by a Hong Kong-based media group in 2003, it actively adopts a reader-first policy and selects its news content to meet readers' demand (Ho & Ping Sun, 2008). Featuring tabloid journalism and sensational content, the *Apple Daily* is currently the most circulated newspaper in Taiwan (Chyi & Huang, 2011). Unlike the other three large newspapers, the political affiliation of the *Apple Daily* is ambiguous, neither pan-Blue nor pan-Green. Regarding the cross-Strait relations between Taiwan and China, the *Apple Daily* does not oppose unification of Taiwan with China (Apple Daily, 2012), nor does it protest against Taiwan independence. Nevertheless, it advocates the democratization of China and thus holds a political stance opposed to the Chinese government (Chinese Communist Party).

The political stance of a newspaper can also be observed by how the medium is treated by the Chinese government. For example, the Chinese government has refused to issue visas to journalists from the *Apple Daily* several times (International Federation of Journalists, 2014). Similarly, in "2013 Cross-Strait Media Prospects Forum" held in Beijing, the two largest print media in Taiwan, the *Liberty Times* and *Apple Daily*, were not invited (Apple Daily, 2013). These discrimination policies reveal their ideology perceived by the Chinese government.

1.4 Data

1.4.1 Press Conference Transcripts

My measure of media ideology requires a lexicon of pro-China ideology in China-Taiwan relations as a reference. I collect press conference transcripts of a Chinese government institution, the Taiwan Affairs Office of the State Council (TAO in short) from its official website[1]. The TAO is an administrative institution in charge of Taiwan affairs under the State Council of China. According to its official website, its main functions include "implementing and carrying out guidelines and policies related to Taiwan stipulated by the CPC Central Committee and the State Council" and "taking charge of the media and publicity work related to Taiwan and releasing news and information concerning Taiwan affairs." Since the TAO is responsible for conducting and publicizing China's policies towards Taiwan, I consider it a major official outlet of the Chinese government and use transcripts of its press conferences to build monthly lexicons of China's stance towards Taiwan. The press conferences take place irregularly. From 2000 to 2014, there are 211 TAO press conferences in total, 133 months with press conferences, and often one or two in each month. On average, I have 4,152 words each month to build a lexicon.

1.4.2 Newspaper Headlines

To measure media ideology, I collect newspaper headlines using the News Knowledge Management System (NKMS) administered by the Library of Legislative Yuan in Taiwan. The system consists of daily headlines of 34 newspapers since 1980; I study 12 of them that are also included in the TEDS surveys. I choose to analyze headlines to measure media ideology, not full news text, for three reasons. First, headlines are usually set by editors, not reporters, and more likely to reveal a newspaper's ideology. Second, readers tend to read headlines more often than to read full texts of news, so headlines are more influential than news content. Third, data availability and compilation difficulty impede full text analysis.

[1] http://www.gwytb.gov.cn/xwfbh/index.htm

The enormity of the data I use renders it impossible to collect it manually. Instead, I write Python programs to collect information on 18,383,058 headlines from 2000 to 2014, including newspaper names, headlines, dates, pages, sections, and sometimes reporters and regions. Following the literature (Gentzkow & Shapiro, 2010; Groseclose & Milyo, 2005), I remove editorials and reviews for two reasons. First, the language used by editors and opinion writers is usually sarcastic and difficult to tell its connotations with automatic text analysis. It can be that an editor uses many positive verbs to criticize "One China Policy". Second, news, editorials and reviews are utterly different from readers' perspective. News is supposed to be objective and truthful, while editorials and reviews are full of opinions and judgments. While editorials and reviews may reveal the ideology of a newspaper more directly and efficiently, they are less likely to have an impact on readers' thoughts and, in turn, influence their national identities. Therefore, I remove editorials and reviews in the data set. Then, I write additional code to detect repetitions in the data set. As I request a large amount of news, the NKMS system often responds with duplicates, which can seriously distort the result–it turns out that nearly 10% of the data I collect are repetitions. In the end, I collect 10,953,168 headlines. It is worth mentioning that the number of news stories from a newspaper in the NKMS system surges from 7.6 news stories per day in 2000 to 199.8 per day in 2014, suggesting that I need to be cautious about intertemporal analysis: measures in earlier years are built on relatively small text corpora and may be less stable compared to measures in later years. On average, there are 48,645 words per month per newspaper.

1.4.3 Surveys

I compile rotating panel surveys from Taiwan's Election and Democratization Studies (TEDS), each containing two periods, (2001, 2004P), (2004L, 2008L), (2008L, 2008P) and (2008P, 2012), where L stands for legislative elections and P stands for presidential elections. In the end, there are 7,820 observations in the sample, and the core questions are the following:

- Which newspaper do you read most often?

- In Taiwan, some people think they are Taiwanese. There are also some people who think that they are Chinese. Do you consider yourself to be a Taiwanese, a Chinese or both?

The dependent variable is a dummy variable equal to 1 if a person identifies him/herself exclusively as a Taiwanese, not a Chinese. Figure 1.3 displays the proportion of people who identify themselves as Taiwanese in the data set. The number increases from 41.9% in 2001 to 56.1% in 2012; the growing trend is comparable to the pattern of NCCU surveys mentioned in Section 1.3, where the proportion of Taiwanese identity increases from 41.6% in 2001 to 54.3% in 2012.

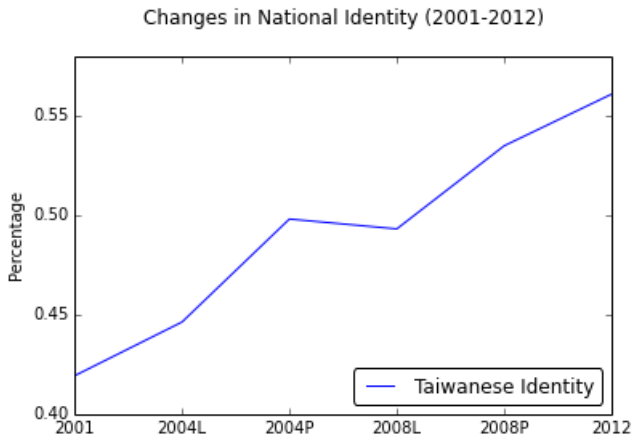


Figure 1.3: The share of Taiwanese identity from 2001 to 2012 in 4 rotating TEDS panel surveys, (2001, 2004P), (2004L, 2008L), (2008L, 2008P) and (2008P, 2012), where L stands for legislative elections and P stands for presidential elections.

The 12 newspapers distributed in Taiwan that I study comprise 98% of the market. The largest four newspapers are the *Apple Daily*, *Liberty Times*, *United Daily News*, and *China Times*. In terms of political affiliations, the *Liberty Times* favors the Pan-Green Coalition, the *United Daily News* and *China Times* are associated with the Pan-Blue Coalition, and the *Apple Daily* is not related to either camp. Regarding the current stance on cross-Strait relations between Taiwan and China, the *Liberty Times* clearly supports Taiwan independence, while the *China Times* openly advocates unification of Taiwan with China. The *United Daily News* holds an equivocal stance on the issue; nevertheless, given its long-lasting pro-KMT stance and Greater China ideological position, the *United Daily News*

is often classified as a pro-unification newspaper (Kuo, 2007; Kuo & Nakamura, 2005). The *Apple Daily* takes an ambiguous stance as well, resorting to the decision of Taiwanese people.

Figure 1.4 shows the changes in newspaper readership of the top four newspapers in the data set. The Pan-Blue affiliated newspapers, the *China Times* and the *United Daily News*, suffer from declining readership, especially the *China Times*, whose readership drops from 28.5% in 2001 to 14.8% in 2012. On the other hand, the Pan-Green affiliated newspaper, the *Liberty Times*, enjoys ca. 30% of readership over the period; the *Apple Daily* joins the market in 2003 and expands its readership from 12.0% in 2004 to 30.6% in 2012, and is currently the largest newspaper in Taiwan.

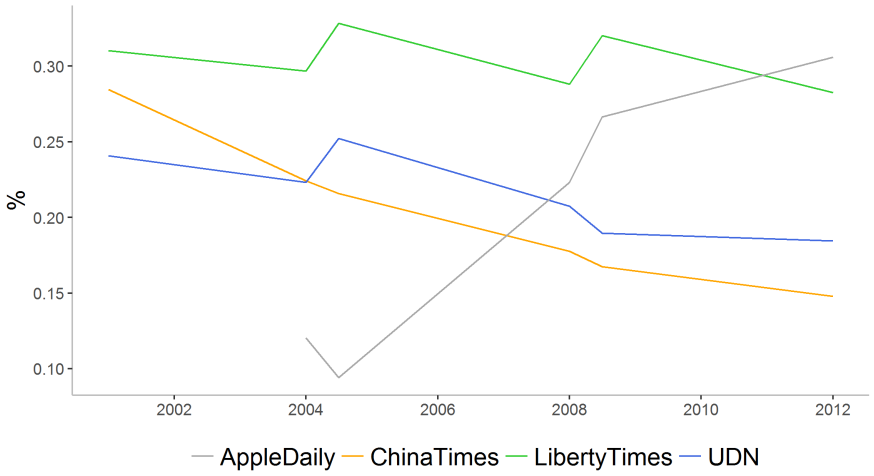


Figure 1.4: The change in the readership of the top four newspapers in 4 rotating TEDS panel surveys, (2001, 2004P), (2004L, 2008L), (2008L, 2008P) and (2008P, 2012), where L stands for legislative elections and P stands for presidential elections.

Note: The *Apple Daily* was introduced in 2003, so its readership in 2001 is 0.

The TEDS survey also contains questions about domestic politics and individual connections with China, such as partisanship, trips to China in the past 5 years, etc. Regarding partisanship, since the Pan-Green Coalition promotes Taiwanese nationalism, supposedly its supporters are more likely to favor Taiwanese identity, compared to supporters of the Pan-Blue Coalition. Thus, Pan-Green partisanship and Taiwanese identity should be positively correlated. In the data set, among 5,282 observations that recognize themselves as partisans,

50.1% of people support the KMT, the largest party in Pan-Blue Coalition, while 41.8% support the DPP, the largest party in the Pan-Green Coalition. Meanwhile, among the partisans, 49.0% identify themselves as Taiwanese, not Chinese, suggesting a discrepancy between national identification and partisanship. Firstly, this may reflect the election strategies of the leading parties in both camps as they tend to weaken their stance on political issues about cross-Strait relations. Secondly, this could also relate to multiple political issues in Taiwan–issues about China are important, but economic development, environmental protection, social security, etc. also play a role in individual political leanings. Lastly, and most importantly, this may indicate that national identification is a broader concept relative to partisanship. Many supporters from both camps think they are Taiwanese, not Chinese. In this chapter, I include partisanship as a control variable. Additionally, the TEDS survey includes the usual demographics like age, gender, education, ethnicity, religions, education, language, etc. For more discussions on demographics, refer to the appendix.

Table 1.1: Summary statistics of the combined TEDS panel surveys

Variable	N	Mean	SD
National identity:			
Taiwanese identity	7597	0.5	0.5
National envisioning:			
Taiwan independence	7223	0.23	0.42
unification with China	7223	0.17	0.37
Choice of newspapers:			
China Times	4914	0.19	0.39
United Daily News	4914	0.21	0.4
Liberty Times	4914	0.3	0.46
Apple Daily	4640	0.22	0.42
Commercial Times	4914	0.01	0.09
Commons Daily	4914	0	0.06
Economic Daily News	4914	0.01	0.1
Youth Daily News	4914	0	0
Taiwan Times	4914	0.01	0.09
United Evening News	4914	0.02	0.12
China Daily	4914	0.02	0.12

Keng Sheng Daily News	4914	0	0.05
Partisanship:			
KMT (Pan-Blue)	5282	0.5	0.5
DPP (Pan-Green)	5282	0.42	0.49
NP (Pan-Blue)	5282	0.02	0.13
PFP (Pan-Blue)	5282	0.05	0.21
TSU (Pan-Green)	5282	0.02	0.14
Age:			
Age	7820	47.89	15.79
Gender:			
Female	7820	0.49	0.5
Income:			
Monthly income	5651	15504.07	12607.88
Ethnicity of father:			
Min Nan	7744	0.75	0.43
Hakka	7744	0.13	0.34
Aboriginal	7744	0.01	0.09
Mainlander	7744	0.1	0.31
Foreigner	7744	0	0.05
Chinese	5020	0	0.02
Language spoken at home:			
Mandarin	7806	0.49	0.5
Min Nan	7806	0.7	0.46
Hakka	7806	0.06	0.24
Aboriginal	7806	0	0.06
Mainlander	7806	0	0.06
Foreigner	7806	0	0.04
Education:			
Elementary	7790	0.25	0.43
Junior high school	7790	0.13	0.34
High school	7790	0.27	0.45
Vocational school	7790	0.12	0.33
Above university	7790	0.22	0.41
Contact with China:			

Contact with Chinese people	6576	0.11	0.32
Travel in China (total times)	6569	0.7	2.37
Religion:			
None	7212	0.22	0.41
Buddhism	7212	0.42	0.49
Taoism	7212	0.3	0.46
Catholicism	7212	0.01	0.1
Christianity	7212	0.04	0.19
Islam	7212	0	0.03
Yiguandao	7212	0.02	0.14
Other	7212	0.01	0.1
Marital status:			
Married	7796	0.72	0.45
Divorced	7796	0.02	0.15
Single	7796	0.2	0.4

Note: (1) It happens that the data set do not include readers of *Youth Daily News*; I keep it just for consistency. (2) I compute monthly income from family income and number of family, i.e. monthly income = (family income)/(family N). (3) People often speak more than one language at home. As a result, the sum of mean for all the language variables is larger than one.

1.5 Measurement

1.5.1 Lexicon of Chinese Ideology

To construct the monthly lexicon, I first need to process the press conference transcripts and implement word segmentation. This task is non-trivial because, different from English, there are no explicit word boundary markers in Chinese. For example, "台中國小" can be segmented as "台中/國小" (Taichung Elementary School) or "台/中國/小" (Taiwan China small) and, thus, can be interpreted quite distinctively. While this task is easy for Mandarin speakers, it is challenging for computers. To implement word segmentation, I use an online service called Chinese Knowledge and Information Processing (CKIP, http://ckipsvr.iis.sinica.edu.tw/) provided by the Institute of Information Science and the Insti-

tute of Linguistics of Academia Sinica.[1] CKIP is built upon linguistic analysis and large Chinese text corpora, and is capable of out-of-vocabulary word identification and part-of-speech tagging; its recall, precision, and F score are ca. 93% - 97% (Ma & Chen, 2003). After word segmentation, I remove words appearing only once, which is a common procedure in text analysis. I aim to choose words that represent Chinese government's ideology towards Taiwan, and it is less likely that they appear only once. Then, I delete words with length smaller than two characters because it is often difficult to tell the meaning of one-character Chinese words. Next, I keep only content words (e.g. proper nouns). Presumably, non-content words such as prepositions, pronouns, and conjunctions are less meaningful in terms of building a lexicon. After that, I manually remove words that are common in general press conferences but do not represent Chinese government's ideology (e.g. reporters). Finally, I keep the 30 most frequent words in the lexicon.

Table 1.2 shows sample lists of words from November 2000 and December 2014 in the lexicons, respectively. The discrepancies between the lists suggest that the topics in 2000 are quite different from the ones in 2014, implying that the lexicons may capture the change of China's policies towards Taiwan. Figure 1.5 serves as an example. Figure 1.5a shows a declining frequency trend of "一個中國" (One China Policy) and "台獨" (Taiwan Independence) from 2000 to 2014, while Figure 1.5b displays a growing frequency trend of "經濟" (economy) and "發展" (development). This pattern suggests a shift from political issues to economic issues in China's policies towards Taiwan. I further apply an unsupervised LDA model[2] to examine the topics of TAO over time. Figure 1.6 displays the topics in TAO press conference transcripts from December 2000 to December 2014 and Table 1.3 summarizes the transition. Again, a similar trend is shown: the topics switch from politics to more diverse issues such as agriculture, tourism, and trade.

I also implement a popular keyword extraction method, TF-IDF (term frequency–inverse document frequency). In addition to the term frequency of a word, TF-IDF also considers whether the word appears in other press conferences. If the word is used in all press

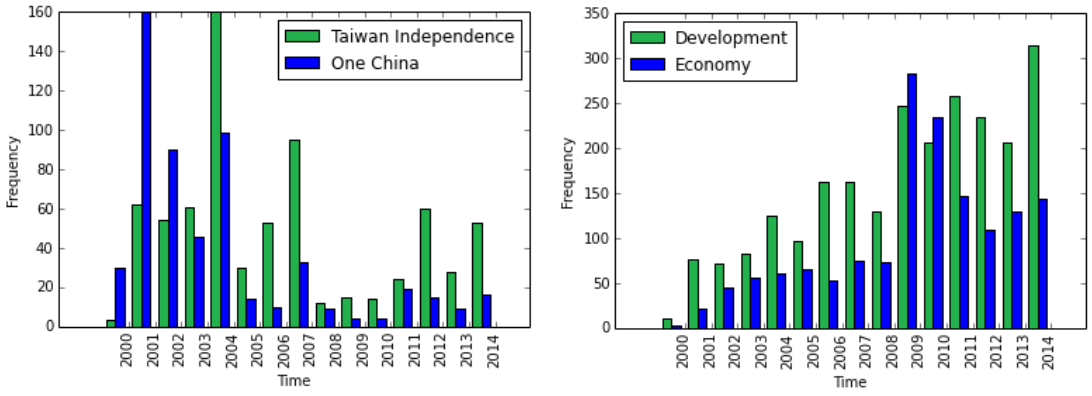
[1] I also use a modified N-gram algorithm and the results of word segmentation are very similar.

[2] Dirichlet parameters of the unsupervised model: 20 topics, topic-per-document distribution $\alpha = 0.1$, word-per-topic distribution $\beta = 0.01$, 2,000 iterations.

[1] Three Direct Links: shorthand for direct links of trade, mail, and air and shipping services across the Taiwan Strait.

Table 1.2: Lexicons constructed from Taiwan Affairs Office transcripts in November, 2000 and December, 2014, respectively.

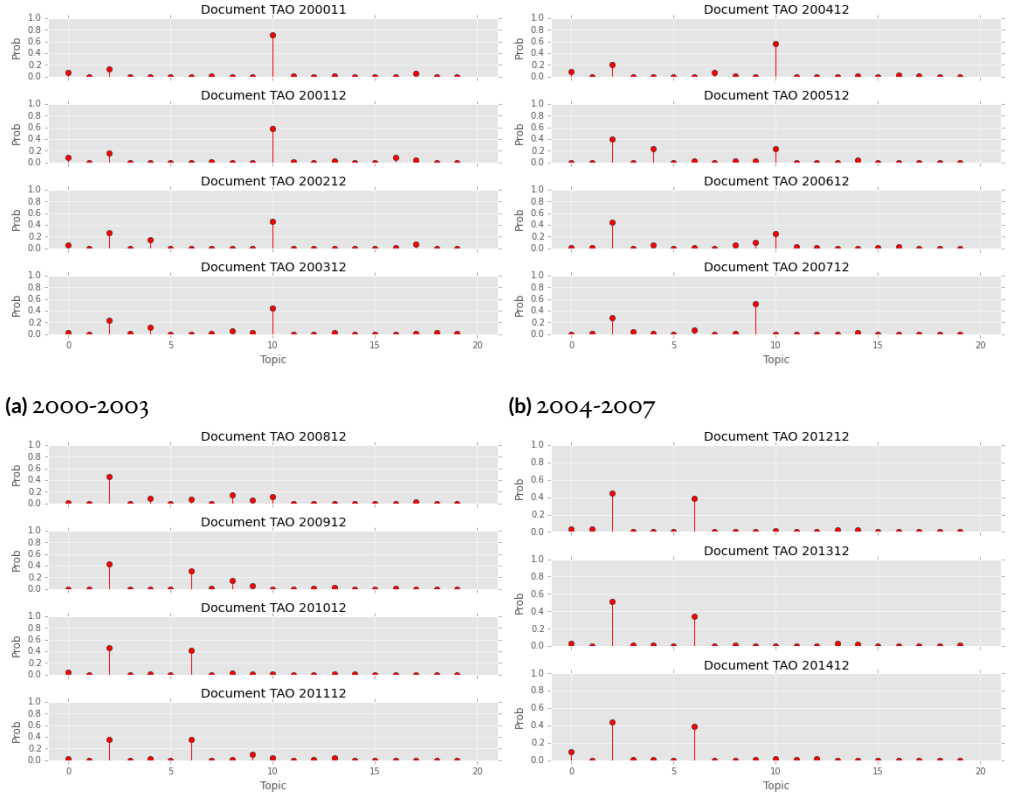
2000 Nov		*2014 Dec*	
word	translation	word	translation
台灣	Taiwan	台灣	Taiwan
兩岸	Cross-Strait	兩岸	Cross-Strait
大陸	Mainland	大陸	Mainland
關係	relations	關係	relations
發展	development	發展	development
共識	consensus	共識	consensus
張銘清	Zhang Mingqing	范麗青	Fan Liqing
三通[1]	Three Direct Links	合作	cooperation
原則	principle	和平	peace
當局	authorities	交流	exchange
黨派	partisan	經濟	economy
實現	achieve	同胞	compatriots
採訪	interview	共同	together
海峽	strait	創業	venture
開放	open	成果	achievement



(a) Political issues **(b)** Economic issues

Figure 1.5: The word frequency of selected words in Taiwan Affairs Office press conference transcripts from 2000 to 2014.

Note: Figure 1.5a displays the word frequency of "台獨" (Taiwan Independence) and "一個中國" (One China Policy), while Figure1.5b shows the frequency of "發展" (development) and "經濟" (economy).



(a) 2000-2003 **(b)** 2004-2007

(c) 2008-2011 **(d)** 2012-2014

Figure 1.6: Topics in the TAO press conference transcripts from 2000 to 2014.

Table 1.3: Summary of topics in TAO press conference transcripts from 2000 to 2014

2000-2003

2004-2007

2008-2014

Topic 10		Topic 9		Topic 2		Topic 6	
台灣	Taiwan	台灣	Taiwan	台灣	Taiwan		
兩岸	Cross-Strait			兩岸	Cross-Strait	兩岸	Cross-Strait
大陸	Mainland	大陸	Mainland	大陸	Mainland		
		合作	cooperation	合作	cooperation	合作	cooperation
		企業	enterprise			企業	enterprise
同胞	compatriot					同胞	compatriot
中國	China	農業	agriculture	交流	exchange	范麗青	Fan Li-qing
三通	Three Links	農產品	agricultural product	發展	development	協議	agreement
當局	authority	農民	farmer	共同	common	楊毅	Yang Yi
台獨	Taiwan independence	水果	fruit	旅遊	travel	關係	relations
原則	principle	李維一	Li Wei-yi	居民	resident	台資	Taiwan-funded

conferences, it may just be a common word and not representative of the current Chinese government's ideology towards Taiwan. Table 1.4 shows a list of words from November 2000 and December 2014 in the TF-IDF lexicon. As expected, there are fewer common words between the two months as TF-IDF underweights words that appear in multiple documents. Besides, the topics are more specific than the previous lexicon. For example, the TF-IDF method tends to pick up more proper nouns, such as "WTO", "奧運會" (Olympic Games), and "九合一" (nine-in-one elections). Despite the discrepancy between the two lexicons, the final measures of media ideology are highly correlated (corr = 0.7879). In the following sections, I will present the results from both lexicons.

Table 1.4: TF-IDF Lexicons constructed from Taiwan Affairs Office transcripts in November, 2000 and December, 2014, respectively.

Nov 2000		Dec 2014	
張銘清	Zhang Mingqing	范麗青	Fan Liqing
倒退	fall back	峰會	summit
李遠哲	Yuan T. Lee	紅利	dividend
三通	Three Links	創業	venture
愚蠢	silly	台灣青年	Taiwanese youth
秦慧珠	Qin Huizhu	法庭	court
偽造	forge	貿區	Free Trade Area
地步	extent	九合一	nine-in-one elections
黨派	partisan	兩黨	the two parties (CCP and KMT)
倒行逆施	perversely	抗日	resistance against Japan
拙劣	poorly	糾紛	dispute
信件	letter	企業家	entrepreneur
投降	surrender	勝利	victory
危險	danger	國共	KMT and CCP
駐點	on-site	光復	recovery
當局	authorities	前行	forward
奧運會	Olympic Games	中小企業	SMEs
早報	morning newspaper	柯文哲	Ko Wen-je
比賽	contest	青年	youth
衛星	satellite	利民	benefit the people

1.5.2 Measurement of Media Ideology

I modify the method from Gentzkow and Shapiro (2010) and Larcinese et al. (2011) and extend it based on Murphy and Westbury (2013). I refer to words in the lexicon as "keywords," and all the content words within the same sentences (or the same headlines) as "neighboring words". I start by extracting the keywords for my lexicon from TAO's press releases. Then, I compare TAO's use of each keyword and its neighboring words with that of Taiwanese newspapers. If a Taiwanese newspaper systematically uses the same set of words in a similar way as the Chinese government institution does, my measure of the pro-China slant of the newspaper will be high.

By taking neighboring words into consideration in addition to comparing keywords, my approach addresses at least four issues. Firstly, it eliminates ambiguities caused by homographs, i.e. words with the same form but different meanings. Take English words, "bank", for an instance. "Bank" can refer to a financial institution or a sloping raised land along a lake or river. Since the neighboring words of a financial institution differ from the ones of a sloping raised land, the actual meaning of "bank" can be determined and thus the noise from homograph ambiguity is lessened. Secondly, it helps avoid confusion caused by lexical differences between Taiwan and China. Although Taiwan and China both speak Mandarin, the use of language in two regions differs. There are many "false friends," i.e. words that look or sound similar, but have different meanings. For example, "小姐" means Miss in Taiwan, but refers to a prostitute in China. Nevertheless, through neighboring words, false friends from one another can be distinguished and disturbance in the measure is reduced. Thirdly, it alleviates the negative effects of word segmentation mistakes on the measurement of media ideology. As previously said, "台中國小" can be segmented as "台中/國小" (Taichung Elementary School) or "台/中國/小" (Taiwan China small). Suppose computers wrongly interpret "台中國小" as "台/中國/小" (Taiwan China small) and "中國" (China) is a keyword in the lexicon. The measure will not be affected as long as the neighboring words are about an elementary school and different from the neighboring words of "中國" (China) in TAO press conferences. Lastly, it detects different views on a topic. For example, TAO often mentions "台獨" (Taiwan Independence). If I study only keywords, then high frequency of "台獨" (Taiwan Independence) may imply that the Chinese government supports Taiwan independence, which is not true. The actual posi-

tion of Chinese government on Taiwan independence can be inferred from the neighboring words: the neighboring words of "台獨" (Taiwan Independence) are "抵制"(boycott), "危害"(endanger), "遏制"(containment), "阻礙"(obstruction), "破壞"(damage), etc. , showing a negative perspective. On average, there are 54.9 neighboring words per keyword.

An important feature of my approach is that it compares keywords and neighboring words in TAO's press releases and Taiwanese newspapers using a measure called positive pointwise mutual information (PPMI) instead of raw frequencies. PPMI compares the actual joint distribution of a keyword and its neighboring word to the joint distribution assuming independence of the keyword and the neighboring word. If the neighboring word often appears with the keyword, PPMI will be positive and it is called word co-occurrence. If the use of the neighboring word is independent of the keyword, PPMI will be zero. Using PPMI mitigates three potential problems. Firstly, the number of news stories is unbalanced between small newspapers and large newspapers. For example, 27.2 news stories from the *United Daily News* are included in the data base per day, while only 3.0 news stories from *Keng Sheng Daily News* are included in the data base per day. If I use word frequency to measure media ideology, then small newspapers are deemed less pro-China simply because of fewer news stories. Using PPMI lessens the influence since it uses probabilities instead of raw frequency. Secondly, the number of news stories grows from 90.7 news stories per day in 2000 to 2,398.0 news stories per day in 2014. Using PPMI reduces the concerns for intertemporal comparison of media ideology.

Figure 1.7 displays the average raw frequency and PPMI per keyword for each newspaper over time, where the legend is based on the total number of words ranging from 377 to 214,949 words for a newspaper in a month. Figure 1.7a demonstrates that newspapers with larger total number of words tend to have higher raw frequencies, and the total number of words is positively correlated to raw frequencies across newspapers over time (corr = 0.43). However, this pattern does not exist in Figure 1.7b. Newspapers with small corpus size can have high values of PPMI, and the total number of words is not correlated to PPMI across newspapers over time (cor = 0.09).

Lastly, press conference transcripts and newspapers headlines are utterly different types of text. Press conferences transcripts document complete conversation between a spokesperson and reporters, while newspaper headlines summarize news stories and convey informa-

tion with much shorter texts. It is hard to justify the comparison between press conference transcripts and newspaper headlines, but PPMI is more comparable across different types of texts than word frequency.

The third attribute of the measure is that I use cosine similarity to assess the distance between two vectors. Cosine similarity is commonly used in natural language processing and more appropriate than Euclidean distance in text classification. Another feature of cosine similarity is that its values always range between 0 and 1, facilitating the interpretation of the measure.

Formally, the measure can be written as follows:

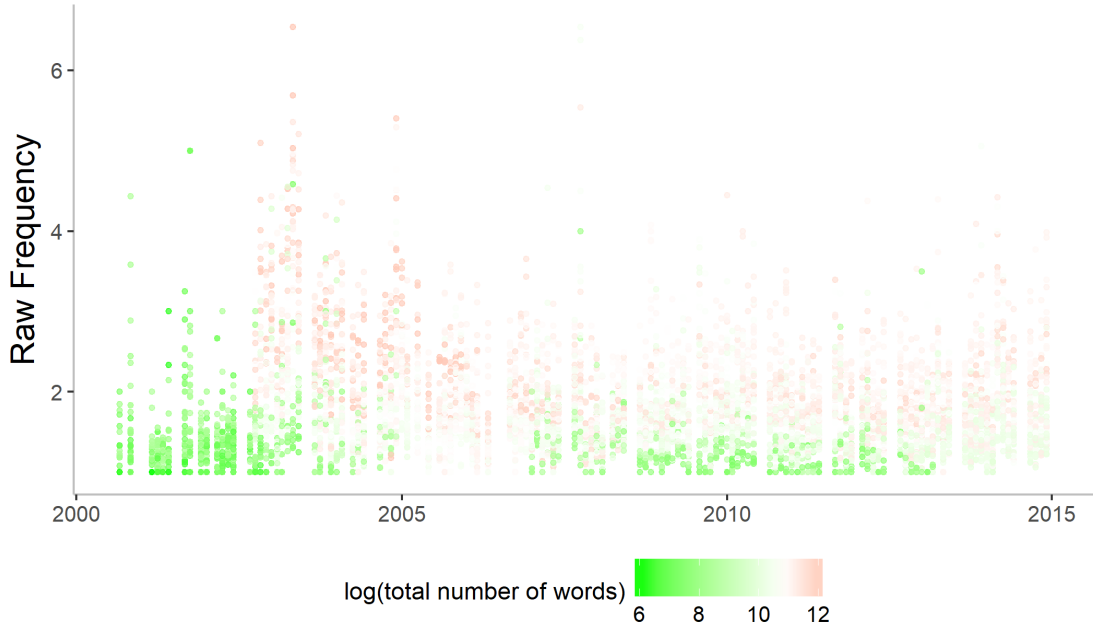
$$PPMI_{i,j,k} = \max\left\{log_2 \frac{p(i,j)}{p(i)p(j)}, 0\right\}$$

$$\text{PPMI}_{i,k} = (PPMI_{i,1,k}, PPMI_{i,2,k}, \ldots, PPMI_{i,J,k})$$

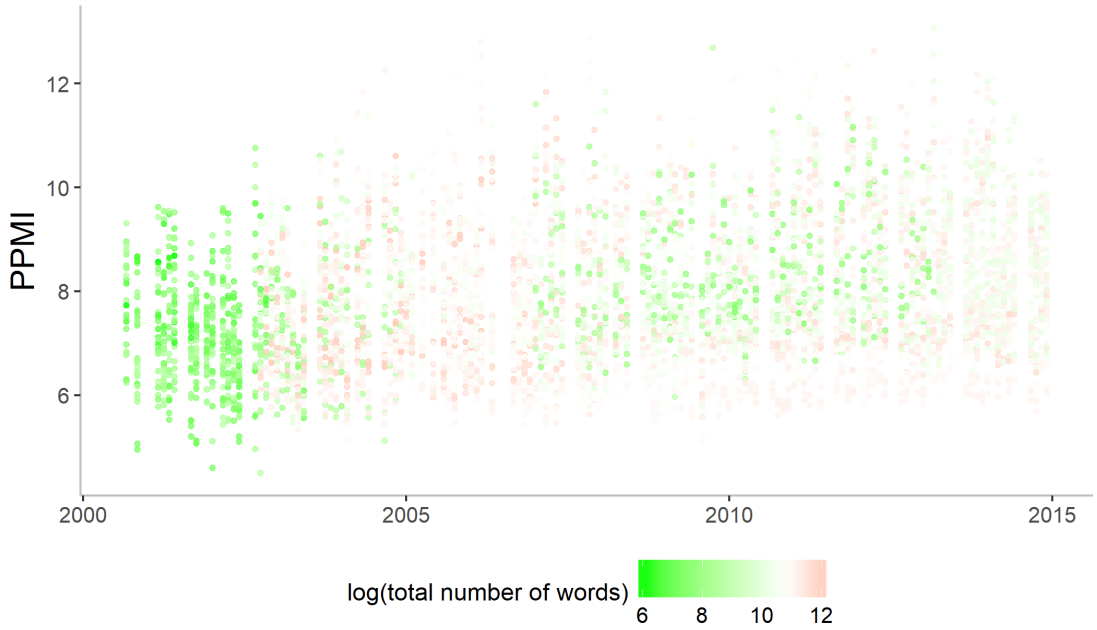
$$MI_k = \frac{1}{I}\sum_{i \in I} \cos(\vartheta_{i,k}) = \frac{1}{I}\sum_{i \in I} \frac{\text{PPMI}_{i,k}\text{PPMI}_{i,TAO}}{\|\text{PPMI}_{i,k}\|\|\text{PPMI}_{i,TAO}\|}$$

- i: keyword
- j: neighboring word
- k: newspaper
- $MI_k \in [0, 1]$

For each keyword i in the monthly lexicon, I construct a vector of its neighboring word's PPMI, building a 30 by J matrix for both TAO and each Taiwanese newspaper. Then, I calculate the cosine similarity between the matrices and average over all keywords in the monthly lexicon. If a Taiwanese newspaper systematically uses the same set of keywords with a similar context as TAO does, its measure of media ideology is high and the newspaper is ideologically close to the Chinese government.



(a) Raw Frequency



(b) PPMI

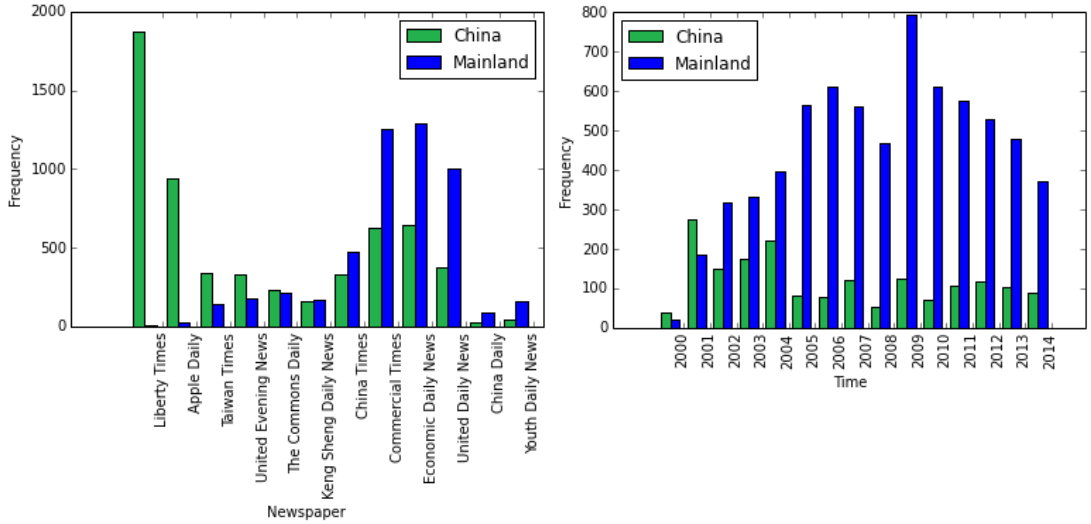
Figure 1.7: The average values of raw frequency and PPMI per keyword for each newspaper over time, where the legend is based on the corpus size.

My measure of media ideology aims to capture three aspects of texts: diction, issue coverage, and context. To illustrate the nuance of diction, consider the use of phrases "中國" (China) and "大陸" (mainland, or continent) when a newspaper refers to People's Republic of China (PRC). If one uses "中國" (China) to describe a piece of news about PRC but not about Taiwan, this implies that the newspaper does not consider Taiwan as part of China. Instead, if a newspaper uses "大陸" (mainland, or continent), then it does not exclude Taiwan from China because mainland and Taiwan can both be part of China. Take a piece of news from the *Liberty Times*, presumably the most anti-China newspaper, for example:

> 中國勞力成本大增模組廠紛回台生產
> Labor costs in China surge; many module manufacturing plants return to Taiwan.
> (Hong, 2012)

The *Liberty Times* assumes that Taiwan is not part of China and implicitly compares the labor costs between China and Taiwan. If the *Liberty Times* uses Mainland instead of China to refer to PRC, then it is inconclusive whether Taiwan is considered a part of China or not. This piece of news also shows that non-political news stories may reflect the political ideology of a newspaper. Figure 1.8a displays the word frequency of "中國" (China) and "大陸" (mainland, or continent) in newspaper headlines in 2014, excluding the combination of the two words, "中國大陸" (mainland China). A large discrepancy is observed between Pan-Blue affiliated newspapers and other newspapers. For example, the *United Daily News* uses "大陸" (mainland) for 1,006 times, compared to 378 times of its usage of "中國" (China). On the other hand, the *Liberty Times* uses "中國" (China) for 1,874 times, but only uses "大陸" (mainland) for 13 times. This difference is also shown in TAO press conference transcripts. Figure 1.8b shows that the Chinese government consistently calls itself "大陸" (mainland) instead of "中國" (China) when speaking to Taiwanese authorities and Taiwanese people.

The measure also reveals different patterns of issue coverage. Take the keyword, "平潭" (Pingtan), for instance. Pingtan stands for Pingtan Experimental Area in China and was planned to be constructed in the Haixi Special Economic Zone in 2009. A major goal of the zone is to strengthen cross-strait exchange and cooperation, and promote the strategic plan of peaceful reunification with Taiwan (The Construction Outline of West Coast Economic



(a) Newspapers **(b)** TAO

Figure 1.8: The word frequency of "中國" (China) and "大陸" (Mainland, or continent), excluding "中國大陸" (Mainland China).

Figure 1.8a shows 12 Taiwanese newspapers in 2014, while Figure 1.8b displays Taiwan Affairs Office press conference transcripts between 2000 and 2014.

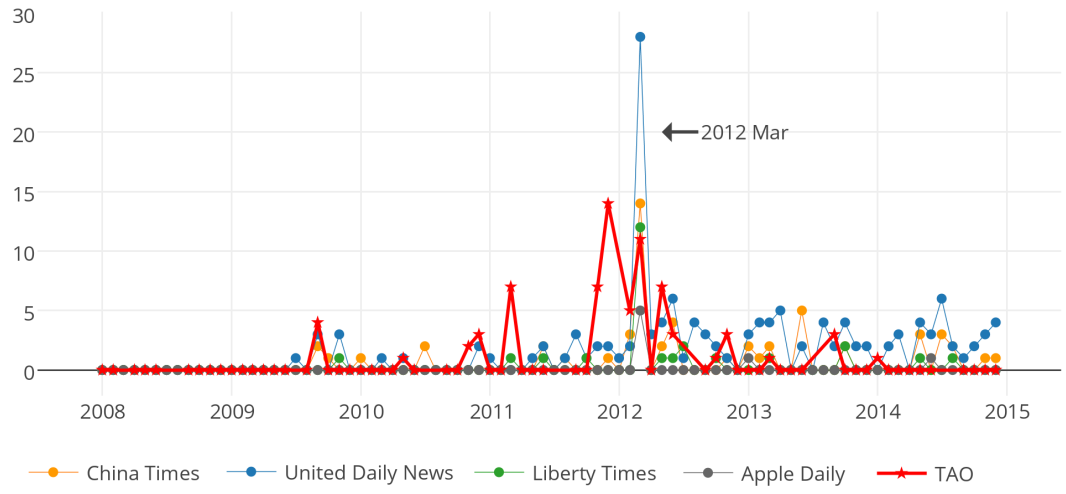


Figure 1.9: The word frequency of "平潭" (Pingtan) in TAO and the largest four Taiwanese newspapers from 2008 to 2014.

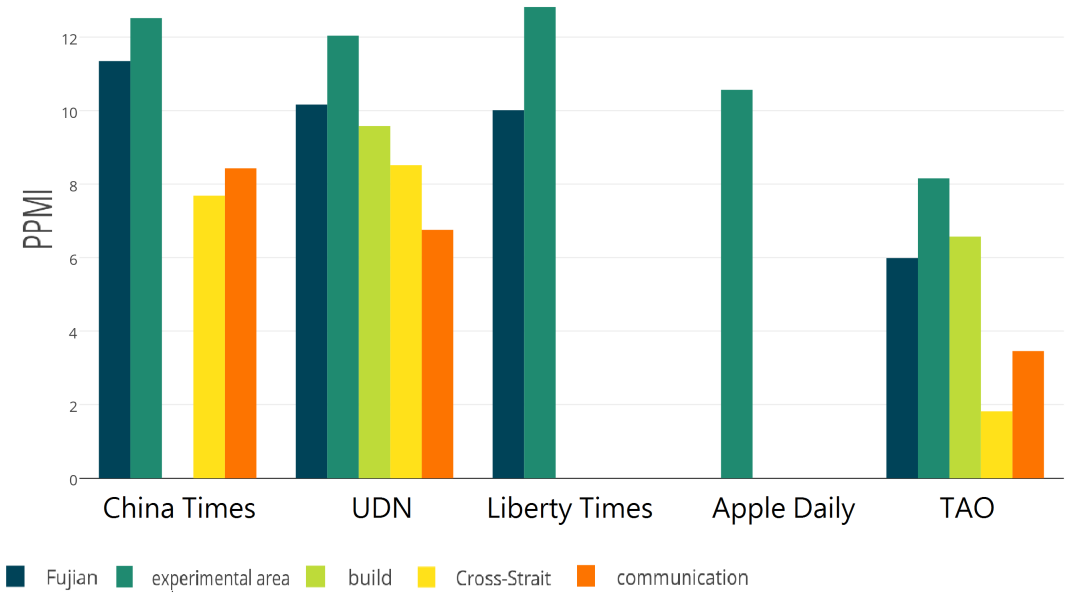


Figure 1.10: The values of PPMI of 5 neighboring words of "平潭" (Pingtan) in TAO and the largest four Taiwanese newspapers in March, 2012.

Zone in Fujian Province, 2010). Figure 1.9 shows the monthly frequency of "平潭" (Pingtan) during the period of 2008 to 2014. The thick line is the frequency in TAO transcripts, while the other four lines are the frequency of the largest four newspapers in Taiwan, respectively. Generally, the *China Times* and the *United Daily News*, the two pro-China newspapers, report on "平潭" (Pingtan) more often than the *Liberty Times* and the *Apple Daily*, the two anti-China newspapers, suggesting that newspapers with distinct preferences of China choose topics differently and that this behavior is captured by the measure.

However, this pattern does not always hold. For example, in March 2012, the frequency of "平潭" (Pingtan) is almost the same between the *China Times* (14 times) and the *Liberty Times* (12 times). In some other months even more "平潭" (Pingtan) is observed in the *Liberty Times* than in pro-China newspapers. Nevertheless, despite similar frequency of "平潭" (Pingtan) in March 2012, the views of the two newspapers on Pingtan are completely different. The *China Times* reports the comments of the Fujian Governor and focuses on the opportunities in Pingtan, while the *Liberty Times* warns violations of the law if Taiwanese citizens are recruited as Chinese public officials in Pingtan. Since my measure of me-

dia ideology considers the context of "平潭" (Pingtan), it is able to capture different views on the topic. Figure 1.10 shows PPMI of five neighboring words of "平潭" (Pingtan) for TAO and the largest four Taiwanese newspapers in March, 2012. Some neighboring words are neutral (Fujian, experimental area, and cross-strait), while others are positive (build and communication). Although anti-China newspapers mention "平潭" (Pingtan), they tend to use fewer positive neighboring words than TAO and pro-China newspapers do. As a result, in spite of similar keyword frequency, pro-China newspapers and anti-China newspapers are labeled with different media ideologies in terms of my measure. Only when a newspaper uses the same neighboring words as TAO does will its measure of pro-China stance be high.

It is worth mentioning that the measure can only capture the relative ideological position of a newspaper, not the absolute resemblance between each Taiwanese newspaper and the Chinese government. That is, the measure can tell if a newspaper shares more similar ideas with the Chinese government compared to other Taiwanese newspapers, but it does not show that the ideology of a newspaper is the same as the one of Chinese government. This is due to fundamental differences in the nature between the two types of text. It is possible that the measures of media ideology are low for all newspapers over time, suggesting that newspaper headlines may not be perfectly comparable to press conference transcripts.

Table 1.5 displays the descriptive statistics of the measured media ideology using both basic and TF-IDF lexicons, respectively. The measure is between 0 and 1 by construction; a higher value indicates more similar use of language between TAO and a Taiwanese newspaper, suggesting that the Taiwanese newspaper adopts a more pro-China ideology. In line with expectations, the *Apple Daily* and the *Liberty Times* are less pro-China than the *China Times* and the *United Daily News* are: the *Apple Daily* and the *Liberty Times* have an average of 0.089 and 0.143, respectively, while the *China Times* and the *United Daily News* have an average of 0.236 and 0.3, respectively. In particular, the media ideology of the *China Times* grows from 0.167 (0.054) in 2008 to 0.241 (0.061) in 2010, which underpins the argument that the *China Times* leans towards China after it was sold to a pro-unification businessman and included in the *Want Want China Times Group* in November 2008. For an extensive and formal discussion, see the next section.

Table 1.5: Descriptive statistics of media ideology

Newspaper	Mean	Std. Dev.	Min.	Max.	N
BASIC					
Apple Daily	0.089	0.037	0.008	0.191	111
Keng Sheng Daily News	0.108	0.088	0	0.416	119
Liberty Times	0.143	0.05	0.033	0.35	133
Taiwan Times	0.157	0.064	0.015	0.282	133
United Evening News	0.159	0.072	0	0.281	133
Commons Daily	0.165	0.061	0.019	0.358	133
China Daily	0.173	0.068	0.014	0.366	119
Youth Daily	0.227	0.088	0.02	0.376	133
China Times	0.236	0.086	0.063	0.45	133
Commercial Times	0.24	0.08	0.014	0.4	133
Economic Daily	0.249	0.088	0.017	0.409	133
United Daily News	0.3	0.082	0.058	0.458	133
all	0.189	0.095	0	0.458	1546
TFIDF					
Apple Daily	0.061	0.037	0	0.167	111
Keng Sheng Daily News	0.064	0.064	0	0.302	119
Liberty Times	0.091	0.052	0	0.315	133
Taiwan Times	0.097	0.058	0	0.273	133
United Evening News	0.099	0.06	0	0.305	133
Commons Daily	0.1	0.056	0	0.263	133
China Daily	0.105	0.055	0	0.292	119
Youth Daily	0.11	0.062	0	0.349	133
Commercial Times	0.12	0.061	0	0.289	133
Economic Daily	0.126	0.07	0	0.348	133
China Times	0.153	0.067	0.018	0.343	133
United Daily News	0.199	0.073	0.025	0.399	133
all	0.111	0.070	0	0.399	1546

1.6 Applications

1.6.1 Ownership Effects

1.6.1.1 Methodology

According to anecdotal evidence, the newspapers in the *Want Want China Times Group* have become more slanted in favor of China after they were sold to a pro-unification businessman in November 2008. To check on the anecdotal evidence, I use my measure of media ideology and adopt difference-in-difference approach, where the treatment group contains the newspapers in the *Want Want China Times Group* and the control group contains all other newspapers:

$$MI_{k,t} = \alpha_0 + \alpha_1 \mathrm{I}_{t \geq 2008\mathrm{II}} + \alpha_2 WW_k + \alpha_3 \mathrm{I}_{t \geq 2008\mathrm{II}} \cdot WW_k + \mathrm{X}_t + \eta_{k,t}$$

$MI_{k,t}$ is the media ideology of newspaper k at time t. WW_k is an indicator of newspapers in the *Want Want China Times Group*; among the 12 selected newspapers, the *China Times* and *Commercial Times* belong to the *Want Want China Times Group*. $\mathrm{I}_{t \geq 2008\mathrm{II}}$ is an indicator of all periods after November 2008, when the two newspapers were sold to the unification advocate. X_t is a vector of control variables, which includes monthly trade share with China, the total number of words in the headlines of a newspaper in a month, month, year, and month $\times$ year fixed effects.

I expect $\alpha_2 > 0$, i.e. on average, the ideology of newspapers in the *Want Want China Times Group* is closer to the Chinese government when compared to all other newspapers. Moreover, I expect a positive α_3, i.e. newspapers in the *Want Want China Times Group* have become more slanted towards China again after the change of ownership when compared to all other newspapers.

1.6.1.2 Results

Table 1.6 displays the baseline results. Column 1 considers only the main explanatory variables in an OLS model. As expected, the main coefficient of interest, $\mathrm{I}_{t \geq 2008\mathrm{II}} \cdot WW_k$, is

positive and statistically significant. Compared to newspapers in the control group, the media ideology of newspapers in the *Want Want China Times Group* increases by 0.032 after the ownership change. Column 2 also includes variables about (potential) structural change. In particular, structural change refers to the large discrepancy of media ideology between the periods before and after April 2005. One possible reason for the structural change is the improvement of cross-Strait relations. In April 2005, the KMT–the opposition party then–launched a groundbreaking trip to mainland China, which was the highest level of exchange between Chinese Communist Party and the KMT since 1945. The chairmen of the KMT expressed interest in improving relations with mainland China and reaffirmed the "One China principle." This change of relations between Taiwan and China could have led to a structural change of media ideology. After taking potential structural change into consideration, the coefficient of the main explanatory variable, $1_{t\geq 2008\text{II}} \cdot WW_k$ still remains highly significant, and the magnitude of the coefficient becomes even larger. Compared to newspapers in the control group, the media ideology of newspapers in the *Want Want China Times Group* increases by 0.056 after the ownership change, which is ca. 25.8% of the average media ideology of the treatment group before the ownership change. Column 3 adds newspaper fixed effects; the size of the coefficient remains the same and is still statistically significant. Column 4 further includes month and year fixed effects, and the coefficient is weakly significant and slightly increase to 0.058. Column 5 incorporates all time fixed effects variables, including month, year, and period fixed effects; the size of the coefficient does not change but the coefficient becomes (just) statistically insignificant with p-value at 10.8%.

In Table 1.7, I reproduce the baseline results with my measure of media ideology using TF-IDF algorithm for keyword extraction, where TF-IDF stands for Term Frequency and Inverse Document Frequency. The TF-IDF algorithm is designed to select keywords related to trendy topics in each period, but not keywords linked to popular topics across different periods. Despite the difference in keyword extraction, the coefficients are of the same magnitude as the ones in Table 1.6 and all statistically significant. In particular, this remains the case in column 5, which employs a fixed effects model with all time fixed effects. The baseline results from Table 1.6 and Table 1.7 show that the effect of ownership change is both economically and statistically significant. For example, the results in Table 1.6 indi-

cate that, compared to all other newspapers, the newspapers in the *Want Want China Times Group* have become ideologically closer to the Chinese government by ca. 14.7-26.7% after they were sold to a pro-unification businessman. In Table 1.7, the range for the same effects is between 31.8% and 52.7%.

Table 1.6: The baseline regressions for ownership effects on media ideology.

	(1) OLS	(2) OLS	(3) FE	(4) FE	(5) FE
$1(t \geq 200811)$	0.023^{***}	0.016^{***}	0.016	-0.016	0.079^{***}
	(0.005)	(0.004)	(0.014)	(0.011)	(0.014)
WW	0.010	-0.045^{***}	0.000	0.000	0.000
	(0.007)	(0.011)	(.)	(.)	(.)
$1(t \geq 200811) \cdot$ WW	0.032^{***}	0.056^{***}	0.056^{**}	0.058^{*}	0.056
	(0.009)	(0.009)	(0.024)	(0.032)	(0.032)
Total number of words	1.596^{***}	1.367^{***}	1.367^{***}	1.377^{***}	1.328^{***}
	(0.048)	(0.078)	(0.274)	(0.328)	(0.371)
Trade share: China	0.064	0.577^{***}	0.577^{***}	-0.520	0.000
	(0.045)	(0.076)	(0.126)	(0.299)	(.)
Structural change	No	Yes	Yes	Yes	Yes
Year/month fixed effects	No	No	No	Yes	Yes
Period fixed effects	No	No	No	No	Yes
Observations	1546	1546	1546	1546	1546
Adjusted R^2	0.51	0.72	0.53	0.59	0.71

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the baseline results of ownership effects on media ideology. The dependent variable is the media ideology measure constructed on basic keywords. Column 1 only includes the main explanatory variables in an OLS model. Column 2 adds variables about (potential) structural change. Column 3 adopts fixed effects model (newspaper fixed effects). Column 4 adds month and year fixed effects. Column 5 incorporates all time fixed effects variables, including month, year, and month $\times$ year fixed effects. Standard errors are clustered at the newspaper level.

Table 1.7: The baseline regressions for ownership effects on TF-IDF media ideology.

	(1) OLS	(2) OLS	(3) FE	(4) FE	(5) FE
$1(t \geq 200811)$	0.030***	0.024***	0.024**	-0.016	0.009
	(0.004)	(0.004)	(0.009)	(0.011)	(0.020)
WW	-0.009*	-0.030***	0.000	0.000	0.000
	(0.005)	(0.009)	(.)	(.)	(.)
$1(t \geq 200811) \cdot$ WW	0.035***	0.048***	0.048***	0.058*	0.048**
	(0.007)	(0.008)	(0.013)	(0.032)	(0.016)
Total number of words	1.079***	0.816***	0.816***	1.377***	0.861***
	(0.042)	(0.065)	(0.162)	(0.328)	(0.233)
Trade share: China	0.007	0.409***	0.409***	-0.520	0.000
	(0.033)	(0.059)	(0.083)	(0.299)	(.)
Structural change	No	Yes	Yes	Yes	Yes
Year/month fixed effects	No	No	No	Yes	Yes
Period fixed effects	No	No	No	No	Yes
Observations	1546	1546	1546	1546	1546
Adjusted R^2	0.44	0.52	0.35	0.59	0.63

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the baseline results of ownership effects on media ideology. The dependent variable is the media ideology measure constructed on TF-IDF keywords. Column 1 only includes the main explanatory variables in an OLS model. Column 2 adds variables about (potential) structural change. Column 3 adopts fixed effects model (newspaper fixed effects). Column 4 adds month and year fixed effects. Column 5 incorporates all time fixed effects variables, including month, year, and month $\times$ year fixed effects. Standard errors are clustered at the newspaper level.

1.6.1.3 Robustness Checks

One identifying assumption of my difference-in-difference approach is the common trend assumption, i.e. the average outcome of the treatment group and the one of the control group would follow the same trend in the absence of the treatment. In other words, the media ideology of newspapers in the control group must have the same trend as the one in the treatment group before newspapers in the treatment group were sold to the pro-unification businessman. Figure 1.11 displays the media ideology of the 12 selected newspapers from 2000 to 2014. While media ideology before November 2008 appears quite volatile, it does not seem to follow any particular trend.[1] I also apply synthetic control methods (Abadie & Gardeazabal, 2003) and compare the evolution of media ideology between each newspaper and its synthetic control counterpart.[2] Figure 1.12 displays the difference of predicted media ideology between a newspaper and its synthetic control counterpart, where the latter is an estimate of the counterfactual predicted media ideology in the absence of ownership change.[3] The results of the two newspapers in the *Want Want China Times Group* are quite different: the *China Times* seems to be affected by the ownership change with some time lags and a potential dip before the change, while the *Commercial Times* does not differ from its synthetic control counterpart. This may reflect the fact that the *Commercial Times* held a strong pro-China ideological position even before the ownership change.

To formally assess the existence of any pre-treatment trend that undermines the common trend assumption, I incorporate lead and lag variables into the model (Autor, 2003) as shown in Table 1.8. Column 1-4 use the media ideology measure constructed on basic keywords, while Column 5-8 use the one based on TF-IDF keywords. In Column 1-4, the results show no pre-treatment and post-treatment trends, and the coefficient on the main explanatory variable decreases to 0.044 and becomes more significant with leads and lags included into regressions. In Column 5-8, there are almost no pre-treatment trends, with

[1] It is worth noting that media ideology before April 2005 trends distinctively, compared to the ones after April 2005, confirming the structural change hypothesis.

[2] I minimize the mean squared prediction error (MSPE) over the periods after April 2005 with fully nested optimization.

[3] Due to limitations in the Stata program, only one newspaper is allowed in the treatment group, while the difference-in-difference method in the chapter assumes two newspapers the *China Times* and *Commercial Times* in the treatment group. This leads to some bias, but synthetic control method is applied here for informal visual inspection, rather than for formal assessment of common trends assumption.

Table 1.8: The regressions for ownership effects with leads and lags.

	basic				TFIDF			
	(1) FE	(2) FE	(3) FE	(4) FE	(5) FE	(6) FE	(7) FE	(8) FE
$1(t \geq 200811)$	0.079***	0.039**	0.155***	0.045	0.009	-0.028	0.070**	0.009
	(0.014)	(0.013)	(0.016)	(0.032)	(0.020)	(0.019)	(0.024)	(0.023)
F6: $1(t \geq 200811) \cdot WW$				-0.009				-0.011
				(0.012)				(0.017)
F5: $1(t \geq 200811) \cdot WW$				0.002				0.029*
				(0.027)				(0.014)
F4: $1(t \geq 200811) \cdot WW$				0.001				-0.040
				(0.039)				(0.054)
F3: $1(t \geq 200811) \cdot WW$				0.000				0.000
				(.)				(.)
F2: $1(t \geq 200811) \cdot WW$			-0.006	0.000			-0.022	0.000
			(0.032)	(.)			(0.041)	(.)
F1: $1(t \geq 200811) \cdot WW$		-0.006	-0.001	-0.001		-0.012	0.010	0.010
		(0.018)	(0.033)	(0.033)		(0.015)	(0.032)	(0.032)
$1(t \geq 200811) \cdot WW$	0.056	0.044***	0.044***	0.044***	0.048**	0.036**	0.036**	0.035**
	(0.032)	(0.012)	(0.012)	(0.012)	(0.016)	(0.016)	(0.016)	(0.016)
L1: $1(t \geq 200811) \cdot WW$		0.019	-0.007	-0.007		0.024	-0.013	-0.013
		(0.040)	(0.028)	(0.028)		(0.017)	(0.013)	(0.013)
L2: $1(t \geq 200811) \cdot WW$			0.027	0.005			0.038**	0.030
			(0.017)	(0.034)			(0.013)	(0.037)
L3: $1(t \geq 200811) \cdot WW$				0.006				-0.003
				(0.015)				(0.009)
L4: $1(t \geq 200811) \cdot WW$				-0.008				-0.034
				(0.039)				(0.046)
L5: $1(t \geq 200811) \cdot WW$				-0.000				0.069***
				(0.023)				(0.010)
L6: $1(t \geq 200811) \cdot WW$				0.026				-0.023**
				(0.025)				(0.011)
Structural change	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year/month fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Period fixed effects	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1546	1546	1546	1546	1546	1546	1546	1546
Adjusted R^2	0.71	0.71	0.71	0.71	0.63	0.63	0.63	0.63

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the regression results of ownership effects with lead and lag variables. The dependent variable in Column 1-4 is the media ideology measure constructed on basic keywords, while Column 5-8 show the one constructed on TF-IDF keywords. Standard errors are clustered at the newspaper level.

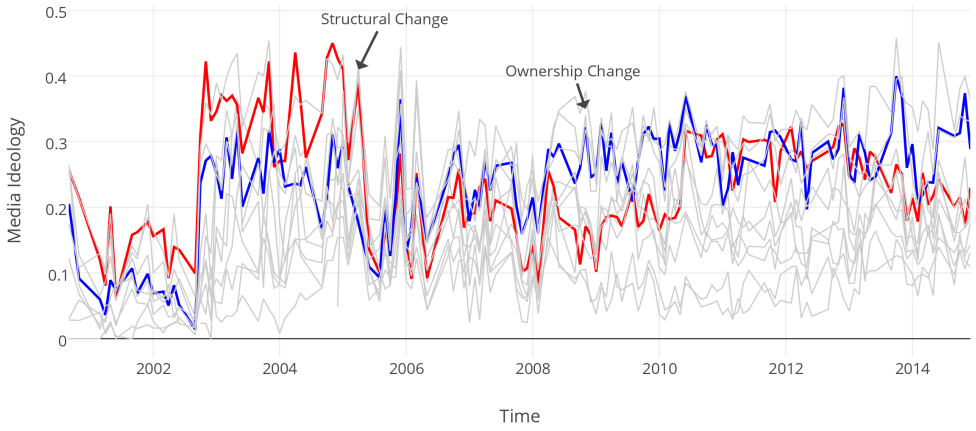


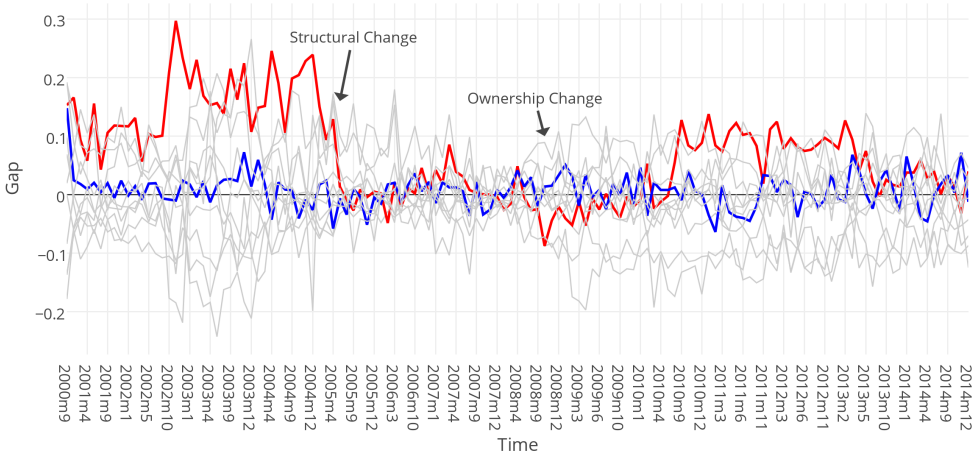
Figure 1.11: Media ideology of the selected newspapers from September 2000 to December 2014.

Note: The highlighted lines are newspapers from the *Want Want China Times Group*: the red line is the *China Times*, while the blue line is the *Commercial Times*.

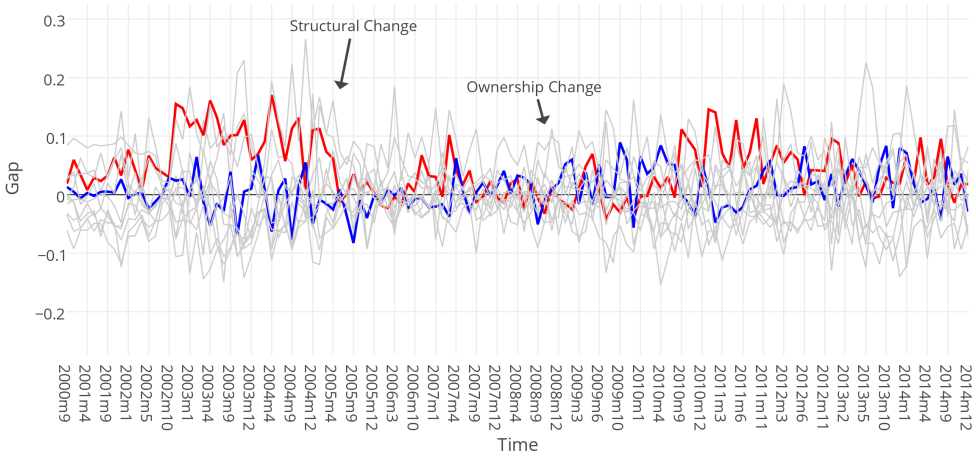
only one weakly significant lead variable. There may be some post-treatment time varying effects, but this does not weaken the identifying assumption of the difference-in-difference model and can be well explained: the ownership change increases the media ideology of newspapers in the *Want Want China Times Group* and its influence lasts longer than one period.

Table 1.9 and Table 1.10 include group-specific time trends to mitigate potential problems from violating the common trend assumption.[1] Column 1 uses the full sample, while Column 2-5 use only observations 6, 12, 18, or 24 months before and after the ownership change. Table 1.9 shows that the media ideology of newspapers in the treatment group follows a time trend that is indistinguishable from the one of newspapers in the control group and that the coefficient on the main explanatory variable $1(t \geq 200811) \cdot WW$ is generally significant. On the other hand, Table 1.10 shows some evidence of group-specific trends: the coefficients on $WW \cdot t$ in Column 1 and 3 are significant at 10% and 5% level, respectively.

[1] I also include newspaper-specific time trends instead of group-specific ones, and the results are quite similar.



(a) Basic



(b) TFIDF

Figure 1.12: Synthetic control method

Note: The y axis shows the gap of predicted media ideology between a newspaper and its synthetic counterpart, where the latter assumes no ownership change. The red line refers to the *China Times*, while the blue line denotes the *Commercial Times*.

Nevertheless, the magnitude of the coefficient is small and goes in different directions. Additionally, the main coefficient of interest is statistically significant in every column. By and large, the treatment group trend is either indistinguishable from the control group trend, or the difference between the trends is economically insignificant.

Table 1.9: The regressions for ownership effects on media ideology with group-specific time trends.

	(1) all	(2) +/-6m	(3) +/-12m	(4) +/-18m	(5) +/-24m
$1(t \geq 200811)$	-0.012	0.018^{*}	0.109^{***}	0.044^{**}	-0.005
	(0.016)	(0.009)	(0.015)	(0.015)	(0.013)
$1(t \geq 200811) \cdot WW$	0.044	0.044^{*}	0.034^{**}	0.023^{*}	0.019^{*}
	(0.028)	(0.021)	(0.011)	(0.011)	(0.010)
$WW \cdot t$	0.000	-0.001	-0.001	0.000	0.001
	(0.000)	(0.004)	(0.001)	(0.001)	(0.001)
Structural change	Yes	Yes	Yes	Yes	Yes
Year/month fixed effects	Yes	Yes	Yes	Yes	Yes
Period fixed effects	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes
Observations	1546	132	252	372	492
Adjusted R^2	0.71	0.44	0.63	0.57	0.59

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays regression results of ownership effects with group-specific time trends. The dependent variable is the media ideology measure constructed on basic keywords. Column 1 uses the full sample, while Column 2-5 use only observations 6, 12, 18, or 24 months before and after the ownership change. Standard errors are clustered at the newspaper level.

Finally, since I only have 12 newspapers (12 clusters), I apply a wild bootstrap approach (Cameron, Gelbach, & Miller, 2008) to adjust the standard errors. Column 1-4 in Table 1.11 use basic media ideology, while Column 5-8 employ TFIDF media ideology. Column 1-3, 5-7 reproduce the baseline regressions, while Column 4 and 8 include lead and lag variables.

Table 1.10: The regressions for ownership effects on TF-IDF media ideology with group-specific time trends.

	(1) all	(2) +/-6m	(3) +/-12m	(4) +/-18m	(5) +/-24m
$1(t \geq 200811)$	-0.006	-0.023*	0.079***	0.025**	-0.017*
	(0.015)	(0.011)	(0.010)	(0.010)	(0.009)
$1(t \geq 200811) \cdot WW$	0.038**	0.055**	0.046***	0.025*	0.030***
	(0.016)	(0.021)	(0.009)	(0.012)	(0.008)
$WW \cdot t$	0.000*	-0.003	-0.002**	-0.000	-0.000
	(0.000)	(0.003)	(0.001)	(0.001)	(0.001)
Structural change	Yes	Yes	Yes	Yes	Yes
Year/month fixed effects	Yes	Yes	Yes	Yes	Yes
Period fixed effects	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes
Observations	1546	132	252	372	492
Adjusted R^2	0.63	0.42	0.65	0.62	0.58

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays regression results of ownership effects with group-specific time trends. The dependent variable is the media ideology measure constructed on TF-IDF keywords. Column 1 uses the full sample, while Column 2-5 use only observations 6, 12, 18, or 24 months before and after the ownership change. Standard errors are clustered at the newspaper level.

The main coefficient of interest is of the same size as previous regressions, but only half of them remain significant. Table 1.12 incorporates treatment group specific time trends into the models. The results are consistent with Table 1.9 and Table 1.10: the coefficients on the main explanatory variable $\mathbb{1}(t \geq 200811) \cdot WW$ are mostly significant, and the coefficients on the treatment group specific trend are either statistically or economically insignificant.

1.6.2 Media Ideology and National Identity

1.6.2.1 Methodology

To examine the effects of newspaper ideology on readers' subjective feeling of national identity, I combine my measure of the media ideology of Taiwanese newspapers with rotating panel surveys from Taiwan's Election and Democratization Studies (TEDS). The baseline specification is a linear probability model as follows:

$$\Pr(Taiwan_{i,t} = 1|\Omega) = \beta_0 + \beta_1 \text{newspaper}_{i,t} + \beta_2 \text{MI}'_t \cdot \text{newspaper}_{i,t} + \beta_3 \text{X}_{i,t} + \mu_i + \delta_t + \varepsilon_{i,t}$$

$Taiwan_{i,t}$ is an indicator equal to 1 if individual i self-identifies as Taiwanese, not Chinese at time t. $\text{newspaper}_{i,t}$ is a vector of dummy variables $\{newspaper_{i,k,t}\}_{k=1}^{12}$ that newspaper k is the most commonly read newspaper by individual i at time t. MI_t is a vector of media ideology $\{MI_{k,t}\}_{k=1}^{12}$ constructed in previous sections. $\text{X}_{i,t}$ contains control variables, e.g. national envisioning preference[1], partisanship, parents' ethnicity, language used at home, district of residence, occupation, income, education, age, gender, and personal experience related to China (e.g. travel in China). μ_i represents individual fixed effects, and δ_t is time fixed effects. I expect $\beta_2 < 0$, i.e. if a person often reads a newspaper with an ideological position closer to the Chinese government, s/he is less likely to identify herself/himself as Taiwanese, not Chinese.

However, my linear probability model can be biased for several reasons. First, it is probable that I do not include all relevant variables, mostly because they are not available. For

[1] National envisioning refers to "統獨"(tondu), i.e. unification with China or Taiwan independence.

Table 1.11: The regressions of ownership effects using wild bootstrap.

	basic				TFIDF			
	(1) OLS	(2) FE	(3) FE	(4) FE	(5) OLS	(6) FE	(7) FE	(8) FE
$1(t \geq 200811)$	0.018	0.018	0.035*	0.030	0.025**	0.024**	-0.042*	-0.013
	(0.018)	(0.015)	(0.018)	(0.035)	(0.011)	(0.010)	(0.022)	(0.030)
WW	0.071	-0.051	0.088***	0.091**	0.030	-0.042	0.033**	0.037**
	(0.097)	(0.059)	(0.000)	(0.040)	(0.042)	(0.033)	(0.014)	(0.016)
F6: $1(t \geq 200811) \cdot$ WW				-0.009				-0.011
				(0.013)				(0.021)
F5: $1(t \geq 200811) \cdot$ WW				0.002				0.029*
				(0.227)				(0.017)
F4: $1(t \geq 200811) \cdot$ WW				0.001				-0.040
				(0.011)				(0.052)
F3: $1(t \geq 200811) \cdot$ WW				0.000				0.000
				(.)				(.)
F2: $1(t \geq 200811) \cdot$ WW				0.000				0.000
				(.)				(.)
F1: $1(t \geq 200811) \cdot$ WW				-0.001				0.010
				(0.005)				(0.026)
$1(t \geq 200811) \cdot$ WW	0.062	0.058	0.056	0.044**	0.051**	0.048**	0.048**	0.035
	(0.053)	(0.043)	(0.043)	(0.017)	(0.025)	(0.020)	(0.021)	(0.022)
L1: $1(t \geq 200811) \cdot$ WW				-0.007				-0.013
				(0.030)				(0.015)
L2: $1(t \geq 200811) \cdot$ WW				0.005				0.030
				(0.030)				(0.040)
L3: $1(t \geq 200811) \cdot$ WW				0.006				-0.003
				(0.017)				(0.008)
L4: $1(t \geq 200811) \cdot$ WW				-0.008				-0.034
				(0.019)				(0.044)
L5: $1(t \geq 200811) \cdot$ WW				-0.000				0.069***
				(0.003)				(0.000)
L6: $1(t \geq 200811) \cdot$ WW				0.026				-0.023
				(0.034)				(0.014)
Structural change	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Year/month fixed effects	No	No	Yes	Yes	No	No	Yes	Yes
Period fixed effects	No	No	Yes	Yes	No	No	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1546	1546	1546	1546	1546	1546	1546	1546
Adjusted R^2	0.68	0.72	0.82	0.82	0.49	0.52	0.72	0.72

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the regression results of ownership effects using wild bootstrap. The dependent variable in Column 1-4 is the media ideology measure constructed on basic keywords, while Column 5-8 use the one constructed on TF-IDF keywords. Column 1-3, 5-7 reproduce the baseline regressions, while Column 4 and 8 include lead and lag variables. Standard errors are clustered at the newspaper level.

Table 1.12: The regressions of ownership effects with group-specific time trends using wild bootstrap.

	basic			TFIDF		
	(1) all	(2) +/-12m	(3) +/-24m	(4) all	(5) +/-12m	(6) +/-24m
$1(t \geq 200811)$	0.036*	-0.060*	0.060**	-0.041**	-0.072***	0.034**
	(0.019)	(0.031)	(0.024)	(0.020)	(0.025)	(0.017)
WW	0.000	0.420	0.000	0.000	1.268	0.000
	(.)	(0.617)	(.)	(.)	(0.786)	(.)
$1(t \geq 200811) \cdot WW$	0.044	0.034**	0.019*	0.038*	0.046***	0.030***
	(0.037)	(0.015)	(0.010)	(0.021)	(0.000)	(0.000)
$WW \cdot t$	0.000	-0.001	0.001	0.000*	-0.002*	-0.000
	(0.000)	(0.001)	(0.001)	(0.000)	(0.001)	(0.001)
Structural change	Yes	Yes	Yes	Yes	Yes	Yes
Year/month fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Period fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Other control variables	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1546	252	492	1546	252	492
Adjusted R^2	0.82	0.87	0.86	0.72	0.77	0.72

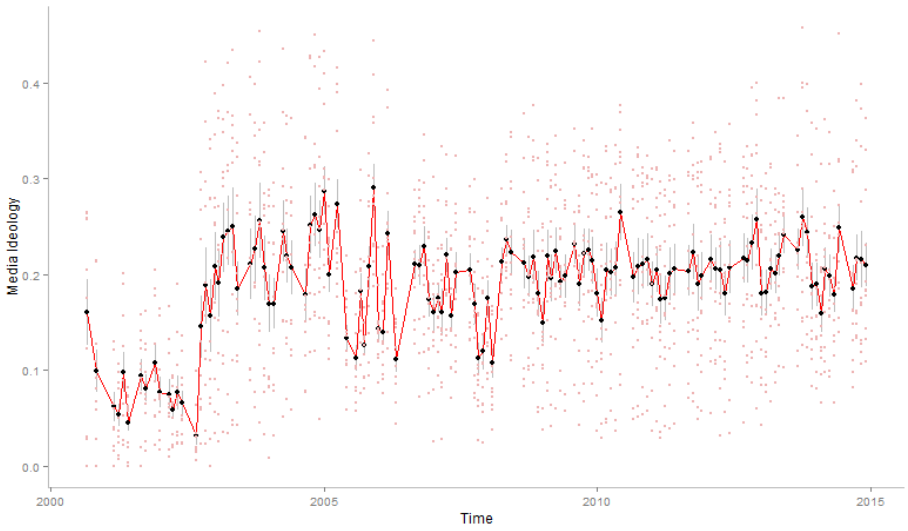
Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

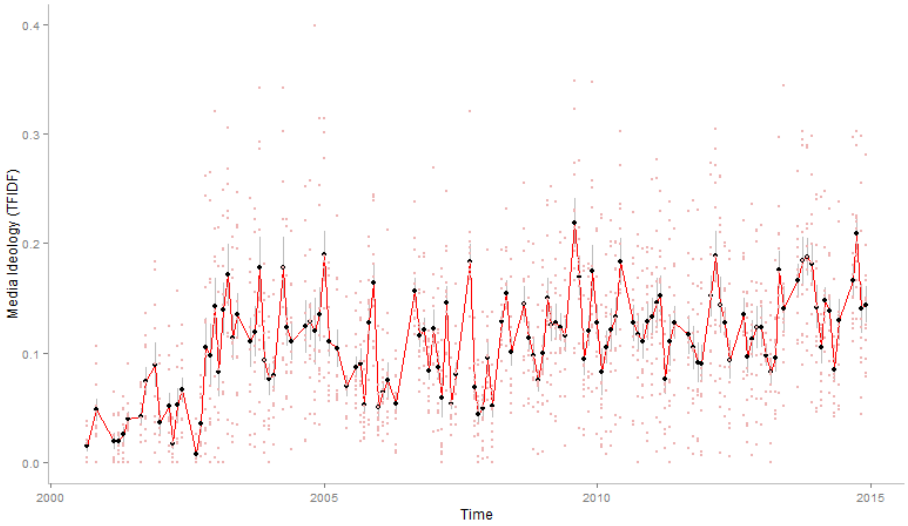
Note: The table displays the regression results of ownership effects using wild bootstrap. The dependent variable in Column 1-3 is the media ideology measure constructed on basic keywords, while Column 4-6 use the one constructed on TF-IDF keywords. Column 1 and 4 use the full sample; Column 2, 5 and 3, 6 use only observations 12, 24 months before and after the ownership change, respectively. Standard errors are clustered at the newspaper level.

example, I lack data on cross-Strait relations between China and Taiwan, which are difficult to measure and very likely to influence both media ideology and national identity. Cross-Strait relations have greatly improved since KMT, the largest Pan-Blue party, returned to power in 2008. It is possible that this improvement leads to both growing pro-China media ideology and higher Chinese identity. Figure 1.13 displays the average media ideology from 2000 to 2014, which shows a slightly upward trend of pro-China media ideology over time. The reverse can also be true–Taiwanese people may recognize differences between China and Taiwan through frequent exchange with Chinese people, motivating a stronger Taiwanese identity. Consequently, without taking these factors into account, the effects of media ideology on national identity may be spurious or overestimated. Second, instead of the hypothesis that media ideology influences national identity, it could be the reverse: national identity may affect media ideology. People tend to expose themselves to similar thoughts and ignore opposite opinions. Those who identify themselves as Taiwanese may choose to read newspapers criticizing China. As a result, growing Taiwanese identity can encourage profit-seeking newspapers to modify their stance against China to meet the demand of their readers.

To account for the above issues, I adopt a fixed effects model with two instrumental variables. The fixed effects include individual and time fixed effects. Individual fixed effects can capture any time-invariant individual characteristics, such as parents' subjective feeling of national identity, while time fixed effects explain general factors that influence all individuals over time, such as national elections in 2004, 2008, and 2012. Since fixed effects cannot explain omitted time-variant individual factors and reverse causality, I further utilize two instrumental variables. The first instrumental variable is the media ideology of the last media choice, $\text{MI}'_{t-1} \cdot \text{newspaper}_{i,t-1}$, i.e., the ideology of the newspaper individual i used to read at $t-1$. I argue that the ideology of the last media choice is correlated with the ideology of the current media choice, but does not have a direct impact on individual national identity, assuming that the ideology of the last media choice can only influence individual national identity through the ideology of the current media choice. The ideology of the last media choice can be positively or negatively correlated to the ideology of the current media choice. On the one hand, a reader tends to choose newspapers with ideology similar to oneself. If one's ideology is stable over time, so is the ideology of one's media choice. As a result, if a



(a) Basic



(b) TFIDF

Figure 1.13: Media ideology from 2000 to 2014

Note: The red line links the mean of media ideology of all newspapers in a month; the grey error bar depicts standard errors of media ideology in the relevant month; each dot is a value of media ideology of a newspaper in a month.

reader chose a pro-China newspaper in the last period, s/he is more likely to choose a pro-China newspaper in the current period. On the other hand, a reader may recognize the bias of the chosen newspaper and modify his/her media choice in the current period. Suppose a reader chose a pro-China newspaper in the last period. S/he may still choose a pro-China newspaper in the current period, but a less biased one compared to the last media choice. Therefore, although the ideology of the current media choice is presumably correlated to the one of the last media choice, the direction of the correlation is less obvious.

The second instrument variable is the total number of words in the headlines of newspaper k at time t, $n_{k,t}$. I argue that $n_{k,t}$ is correlated with media ideology, but does not directly relate to individual national identity. The first half of the argument can be supported by the high correlation between media ideology and the total number of words in headlines (corr = 0.702). According to Zipf's law, the r_{th} most frequent word has a frequency that is inversely proportional to r^{α} for $\alpha \approx 1$ (Piantadosi, 2014). This implies that the frequency decreases with rank very quickly and most words have low frequency in a corpus. Moreover, the sparsity of words may be more evident in small corpora and lead to lower values in the measure of media ideology. As a result, corpus size, i.e. the total number of words in headlines, affects the measure of media ideology, and a smaller corpus is related to a lower value of media ideology.

The second half of the argument is reasonable as well. A reader's national identity may be influenced by the newspaper s/he reads every day, but it is less likely to be affected by the total number of words in the headlines. The total number of words in the headlines of a newspaper is determined by two factors: the average length of a headline in a newspaper, and the total number of headlines included in the NKMS system. The former is influenced by the editors of the newspaper, while the latter may be affected by the actual total number of news stories produced by the newspaper and the administrator of the Legislative Yuan Library. One potential problem is that the administrator of the library may select more news stories from the newspapers that are politically affiliated with the ruling party. Yet, s/he can hardly (and probably will not) manipulate the total number of words in the headlines since it is also affected by the editors. Another potential problem is that a newspaper may set headlines of certain length to cater to readers with Taiwanese identity and increase its readership. However, it is hard to justify that readers with Taiwanese identity

prefer shorter or longer headlines. Besides, headline length depends on the information in the related news and contains some randomness, and thus is hard to control perfectly, not to mention the total number of words in all headlines. Therefore, both factors are unlikely to have an impact on individual national identity.

The modified model is as follows:

$$\Pr(Taiwan_{i,t} = 1|\Omega) = \beta_0 + \beta_1 \text{newspaper}_{i,t} + \beta_2 \widehat{\text{MI}'_t \cdot \text{newspaper}_{i,t}} + \beta_3 X_{i,t} + \mu_i + \delta_t + \varepsilon_{i,t}$$

where $\widehat{\text{MI}'_t \cdot \text{newspaper}_{i,t}}$ is the fitted value from the first stage regression.

1.6.2.2 Results

Table 1.13 displays the results from first stage regressions. Column 1 to 3 use media ideology from basic lexicons, while Column 4 to 6 use media ideology from TF-IDF lexicons. Column 1 and 4 apply a simple OLS model; Column 2 and 5 use fixed effects model and include current media choice and time fixed effects, while Column 3 and Column 6 further add last media choice into the model. Despite different size of coefficients on the focused explanatory variable, all models confirm that the ideology of the current media choice is positively correlated to the second instrumental variable, total number of words in the headlines of a newspaper. Meanwhile, all fixed effects models show that the ideology of the current media choice is negatively correlated to first instrumental variable, the ideology of the last media choice, implying that a reader tends to modify her/his media choice by choosing a less biased newspaper over time.

Column 1 to 3 in Table 1.14 and Table 1.15 display the baseline results from linear probability model using basic lexicons and TFIDF lexicons, respectively. The control variables include gender, ethnicity of parents, place of residence, age, occupation, education, language(s) spoken at home, marital status, religion, partisanship, China experience, and national envisioning; China experience refers to personal contact with China, such as studying, doing business, or traveling in China, while national envisioning indicates individual

Table 1.13: First stage regressions.

	basic			TFIDF		
	(1) OLS	(2) FE	(3) FE	(4) OLS	(5) FE	(6) FE
$L1$: MI · newspaper	-0.0130	-0.563***	-2.381***			
	(0.0222)	(0.107)	(0.266)			
$L1$: MI_{TFIDF} · newspaper				0.369***	-0.379***	-3.498***
				(0.0241)	(0.0915)	(0.599)
total number of words	0.774***	0.619***	1.624***	0.466***	0.372***	0.390***
	(0.0194)	(0.103)	(0.169)	(0.0181)	(0.0665)	(0.0395)
Media choice	No	Yes	Yes	No	Yes	Yes
L1: media choice	No	No	Yes	No	No	Yes
Time fixed effects	No	Yes	Yes	No	Yes	Yes
Observations	1968	1963	1760	1968	1963	1760

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the first stage regressions of media ideology effects on national identity. The dependent variable in Column 1-3 is the media ideology measure constructed on basic keywords, while Column 4-6 use the one constructed on TF-IDF keywords. Column 1 and 4 apply a simple OLS model; Column 2 and 5 use fixed effects model and include current media choice and time fixed effects, while Column 3 and Column 6 further add last media choice into the model. Standard errors are clustered at the individual level.

preference of "統獨" (tondu), i.e. unification with China or Taiwan independence. Current and last media choice and time fixed effects are also included in all columns.

The coefficient on the key independent variable, the ideology of the current media choice, is positive and mostly insignificant in OLS and random effects model, but becomes negative and significant once individual fixed effects are included in the model, implying the existence of reverse causality or omitted time-invariant individual characteristics. In Column 4 and 5, I employ an instrumental variable approach. The coefficient is still positive and insignificant in random effects models, but becomes much larger and significant in fixed effects models. This may reflect additional endogeneity problem, such as omitted time-variant variables. Overall, the results suggest that given an individual reads some newspaper, if the newspaper becomes more pro-China, the individual will be less likely to self-identify as a Taiwanese.[1]

1.6.2.3 Robustness Checks

Considering that readers may choose a newspaper based on its ideology from the past, not on its current ideology, I employ various versions of media ideology, such as media ideology from last year, MI^{-12}, moving average of media ideology over the past 6 months, $MI^{MA(6)}$, and moving average of media ideology over the past 12 months, $MI^{MA(12)}$. Table 1.16 and Table 1.17 display the results using basic lexicons and TFIDF lexicons, respectively. The results are consistent with Table 1.14 and Table 1.15, but the coefficient magnitude is smaller.

1.7 Conclusion and Discussion

In this chapter, I explore the relationship between media ideology and national identity in a case study of Taiwan. First, I construct an objective measure of media ideology by comparing the use of language between Taiwanese newspapers and a Chinese government institution, the Taiwan Affairs Office (TAO). If a Taiwanese newspaper systematically uses the same set of words in a similar fashion as TAO does, the measure will be higher, suggesting

[1] Logit models are not applied here because individual national identity seldom changes and logit models cannot converge.

Table 1.14: The linear probability models of media ideology on Taiwanese identity.

	(1) OLS	(2) RE	(3) FE	(4) RE + IV	(5) FE + IV
MI · newspaper	0.812*	0.629	-2.958*		
	(0.482)	(0.448)	(1.668)		
$\widehat{\text{MI} \cdot \text{newspaper}}$				1.144	-13.06**
				(0.725)	(5.917)
Media choice	Yes	Yes	Yes	Yes	Yes
L1: media choice	Yes	Yes	Yes	Yes	Yes
Partisanship	Yes	Yes	Yes	Yes	Yes
National envisioning	Yes	Yes	Yes	Yes	Yes
China experience	Yes	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes
Other individual characteristics	Yes	Yes	Yes	Yes	Yes
Observations	1181	1181	1184	1150	1153

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays linear probability models of media ideology on national identity. Column 1 uses an OLS model, Colunm 2 employs a random effects model, and Column 3 adopts a fixed effects model; Column 4 and 5 apply random and fixed effects models with instrumental variables, respectively. The dependent variable is Taiwanese identity, and the key explanatory variable is the media ideology measure constructed on basic keywords. The control variables include gender, ethnicity of parents, place of residence, age, occupation, education, language(s) spoken at home, marital status, religion, partisanship, China experience, and national envisioning. Current and last media choice and time fixed effects are also included in all columns. Standard errors are clustered at the individual level.

Table 1.15: The linear probability models of TF-IDF media ideology on Taiwanese identity.

	(1) OLS	(2) RE	(3) FE	(4) RE + IV	(5) FE + IV
MI_{TFIDF} · newspaper	1.106	0.665	-4.398**		
	(0.681)	(0.610)	(1.732)		
$\widehat{MI_{TFIDF} \cdot \text{newspaper}}$				1.891	-10.21**
				(1.266)	(4.908)
Media choice	Yes	Yes	Yes	Yes	Yes
L1: media choice	Yes	Yes	Yes	Yes	Yes
Partisanship	Yes	Yes	Yes	Yes	Yes
National envisioning	Yes	Yes	Yes	Yes	Yes
China experience	Yes	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes
Other individual characteristics	Yes	Yes	Yes	Yes	Yes
Observations	1181	1181	1184	1150	1153

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays linear probability models of media ideology on national identity. Column 1 uses an OLS model, Colunm 2 employs a random effects model, and Column 3 adopts a fixed effects model; Column 4 and 5 apply random and fixed effects models with instrumental variables, respectively. The dependent variable is Taiwanese identity, and the key explanatory variable is the media ideology measure constructed on TF-IDF keywords. The control variables include gender, ethnicity of parents, place of residence, age, occupation, education, language(s) spoken at home, marital status, religion, partisanship, China experience, and national envisioning. Current and last media choice and time fixed effects are also included in all columns. Standard errors are clustered at the individual level.

Table 1.16: The linear probability models using various versions of basic media ideology.

	(1) t-12	(2) t-12	(3) MA(6)	(4) MA(6)	(5) MA(12)	(6) MA(12)
MI^{-12} · newspaper	-5.310***					
	(1.783)					
$\widehat{MI^{-12}}$ · newspaper		-6.409**				
		(2.811)				
$MI^{MA(6)}$ · newspaper			-5.001**			
			(2.160)			
$\widehat{MI^{MA(6)}}$ · newspaper				-8.553*		
				(4.432)		
$MI^{MA(12)}$ · newspaper					-4.244**	
					(1.730)	
$\widehat{MI^{MA(12)}}$ · newspaper						-5.176**
						(1.976)
Media choice	Yes	Yes	Yes	Yes	Yes	Yes
L1: media choice	Yes	Yes	Yes	Yes	Yes	Yes
Partisanship	Yes	Yes	Yes	Yes	Yes	Yes
National envisioning	Yes	Yes	Yes	Yes	Yes	Yes
China experience	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Other individual characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1184	1086	1184	1151	1184	1152

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays linear probability models of basic media ideology on national identity. The dependent variable is Taiwanese identity, and the key explanatory variable is the media ideology measure constructed on basic keywords. Column 1 and 2 show the results using media ideology from last year, Column 3 and 4 adopt the moving average of media ideology over past 6 months, and Column 5 and 6 employ the moving average of media ideology over the past 12 months. The control variables include gender, ethnicity of parents, place of residence, age, occupation, education, language(s) spoken at home, marital status, religion, partisanship, China experience, and national envisioning. Current and last media choice and time fixed effects are also included in all columns. Standard errors are clustered at the individual level.

Table 1.17: The linear probability models using various versions of TFIDF media ideology.

	(1) t-12	(2) t-12	(3) MA(6)	(4) MA(6)	(5) MA(12)	(6) MA(12)
MI^{-12}_{TFIDF} · newspaper	-2.648**					
	(1.094)					
$\widehat{MI^{-12}_{TFIDF}}$ · newspaper		-2.988**				
		(1.248)				
$MI^{\mathcal{MA}(6)}_{TFIDF}$ · newspaper			-5.858***			
			(2.086)			
$\widehat{MI^{\mathcal{MA}(6)}_{TFIDF}}$ · newspaper				-8.244*		
				(4.285)		
$MI^{\mathcal{MA}(12)}_{TFIDF}$ · newspaper					-6.397**	
					(2.467)	
$\widehat{MI^{\mathcal{MA}(12)}_{TFIDF}}$ · newspaper						-5.548*
						(3.130)
Media choice	Yes	Yes	Yes	Yes	Yes	Yes
L1: media choice	Yes	Yes	Yes	Yes	Yes	Yes
Partisanship	Yes	Yes	Yes	Yes	Yes	Yes
National envisioning	Yes	Yes	Yes	Yes	Yes	Yes
China experience	Yes	Yes	Yes	Yes	Yes	Yes
Time fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Other individual characteristics	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1184	1086	1184	1151	1184	1152

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays linear probability models of TFIDF media ideology on national identity. The dependent variable is Taiwanese identity, and the key explanatory variable is the media ideology measure constructed on TFIDF keywords. Column 1 and 2 show the results using media ideology from last year, Column 3 and 4 adopt the moving average of media ideology over past 6 months, and Column 5 and 6 employ the moving average of media ideology over the past 12 months. The control variables include gender, ethnicity of parents, place of residence, age, occupation, education, language(s) spoken at home, marital status, religion, partisanship, China experience, and national envisioning. Current and last media choice and time fixed effects are also included in all columns. Standard errors are clustered at the individual level.

that the newspaper is ideologically close to the Chinese government. The average media ideologies of the largest four newspapers in Taiwan are in line with expectations: the measures of the *China Times* and the *United Daily News* are higher than the ones of the *Apple Daily* and the *Liberty Times*, indicating that the former two newspapers are ideologically closer to China. Next, I verify the anecdote that *China Times* and the *Commercial Times* became more pro-China after they were sold to a pro-unification businessman in November 2008. I adopt a difference-in-difference approach, where the treatment group includes newspapers in the *Want Want China Times Group* and the control group comprises all other Taiwanese newspapers. The *Want Want China Times Group* was purchased by a pro-China businessman in November 2008 and the anecdotal evidence suggests that newspapers in the media group have become more pro-China after the acquisition. The results are consistent with the anecdote: compared to the control group, newspapers in the *Want Want China Times Group* have become 14.7-26.7% more slanted towards China after the change of ownership. Finally, I combine the constructed measure of media ideology with rotating panel surveys from Taiwan's Election and Democratization Studies to examine the effects of media ideology on national identity. To account for potential endogeneity, I adopt an instrumental variable approach with individual and time fixed effects. The results show that given an individual reading a certain newspaper, if the newspaper adopts a more pro-China ideological position, then the probability that the individual self-identifies as Taiwanese becomes significantly lower.

There are two caveats worth noting. First, one implicit assumption of my measure of media ideology is that Taiwanese newspapers and the Chinese government use the same language, Chinese, so that their dictions are comparable. This ignores the cross-strait lexical difference between China and Taiwan. The measure may end up capturing the lexical similarity between the Chinese press conference transcripts and each Taiwanese newspaper, which is not related to political ideology. However, since the cross-strait lexical difference is relatively moderate, I assume that the difference is not as salient as issue coverage, diction, and context which are evaluated in the measure. Second, one assumption of the difference-in-difference method is stable unit treatment value assumption: no interference between units and no variation in the treatment. That is, I assume that the ownership change of newspapers in the *Want Want China Times Group* does not affect the media ideology of

newspapers in the control group and that the effects of ownership change are the same on the newspapers in the *Want Want China Times Group*. Both parts of the assumption are unlikely to be strictly true. On the one hand, the ownership change of a newspaper with a sizable market share may motivate other newspapers to adjust their ideology on the pro-China to anti-China spectrum. Still, I argue that the ownership change should not immediately influence newspapers in the control group. As shown in Table 1.9 and Table 1.10, the coefficient on the main explanatory variable $1_{t \geq 2008\text{II}} \cdot WW_k$ is significant even if the sample is limited to 6 months before and after the ownership change.[1] On the other hand, according to Figure 1.12, the ownership change may not uniformly affect the two newspapers in the *Want Want China Times Group*. To avoid the problem, I assume the *China Times* to be the only newspaper in the treatment group and reproduce the baseline regressions. The results are consistent with Table 1.6 and Table 1.7, and the coefficient on the main explanatory variable $1_{t \geq 2008\text{II}} \cdot WW_k$ is highly significant.

[1] I further reduce the sample to 1 month before and after the ownership change. With only 36 observations, the coefficient falls just short of statistical significance (p-value = 0.13).

Chapter 2

Protest and Power Structure in China

2.1 Introduction

Protests are related to social instability and perceived as a threat to authoritarian regimes. Previous studies have shown that protests can undermine autocratic stability and increase the likelihood of democratization (Celestino & Gleditsch, 2013; Hollyer, Rosendorff, & Vreeland, 2015; Ulfelder, 2005). In single-party regimes, such as China, the ruling party often claims to represent the interests of the people (Ulfelder, 2005). Manifesting public discontent with the government, mass protests challenge the claim and erode regime legitimacy. To justify its political monopoly and sustain the regime, maintaining social stability becomes one of the most important goals of the Chinese government. Protests in China have risen from 58,000 in 2003 to 180,000 in 2010 [1], which amounts to almost 500 protests a day in 2010 (Human Rights Watch, 2012; Sun, 2011). In the meanwhile, the cost of maintaining social stability has been increasing over time. In 2016, it was estimated that the Chinese government's budget for social stability maintenance (the official category is named police and public security) has reached over 900 billion RMB (ca. 135.8 billion US dollars), almost equivalent to the budget for national defense (Lui, 2016).[2] However, due to sensitivity of the issue, protest data can hardly be found (Cai, 2008), and the existing literature either

[1] Protests are classified as "mass incidents" by the Chinese government, and each mass incident is defined as "an illegal assembly" of more than 10 people (Gong & Scott, 2016).

[2] The official budget for police and public security has exceeded the budget for national security since 2011; the Chinese government has ceased to disclose detailed information on the budget for police and public security since 2015.

focuses exclusively on labor protests or only consists of a handful of observations for qualitative analysis (Chan, Backstrom, & Mason, 2014; X. Chen, 2017; Chou, 2011; Yanwei Li, Koppenjan, & Verweij, 2016; Yuan Li, 2014; O'Brien & Li, 2005). The chapter contributes to the existing literature on protests in China by assembling a comprehensive data set. More specifically, the chapter focuses on the role of political leaders and provides evidence on the causal relationship between power structure and protests. In addition, the chapter provides novel insights into the role of the past experience of leaders by showing that local leaders perform differently from outsiders.

Since the Tiananmen Square Massacre in 1989, protests in China have been characterized as transient, regional, and issue-specific events emerging from grievances against regional governments or firms. Protesters rarely coordinate activities across regions and do not appeal to central communist leaders for national reforms or democratization (Chan et al., 2014; Gong & Scott, 2016; O'Brien & Li, 2005; Tong & Lei, 2010). These protests often reveal people's discontent with regional governments and further affect regional leaders' political career. Maintaining social stability is one of the so-called "priority targets with veto power" (一票否决) of the Chinese government, which means that failing to meet the targets leads to a veto on a politician's career advancement (Y. Wang, 2014).[1] As a result, regional leaders have strong incentives to reduce protests. To analyze the influence of regional leaders on protests, I use text mining to compile a detailed data set on over 50,000 protests in 2014 and 2015 from a citizen media outlet, *Not the News* (非新聞), with information on dates, locations, pictures, videos, online discussion archives, the cause of a protest, the form of a protest, the identity of protesters, and whether a protest is repressed.

I propose that power structure among regional leaders plays a crucial role for protests. The number of protests depends on two factors: the degree of dissatisfaction (grievances) and the ease to engage in protests due to collective action problems, and county leaders can influence protests through both factors. On the one hand, protests stem from perceptions of injustice, and the root of grievances may relate to a certain policy and reveal public

[1]Performance targets are classified into three categories ordered by their significance: priority targets with veto power, hard targets, and soft targets. Priority targets are often political and necessary for career advancement. Hard targets are quantifiable economic goals, such as regional economic growth rate and fiscal revenue collection, whereas soft targets are non-quantifiable objective, for example, propaganda work (Edin, 2003; Göbel & Ong, 2012; Shih, Adolph, & Liu, 2012; Xu, 2011; Zuo, 2015).

discontent with the government. Under a regionally decentralized system (Xu, 2011), a regional leader can establish regional policies and directly impact the well-being of people in his/her jurisdiction, and the degree of influence depends on power structure. On the other hand, people are more likely to participate in a protest when they perceive a higher chance of success, or a political opportunity, which hinges on state strength, repression level, and elite division (Cragun Cragun, D., 2006; Klandermans & van Stekelenburg, 2013). In particular, when power is divided among elites, aggrieved citizens are more likely to find powerful allies to support their claims (Klandermans & van Stekelenburg, 2013; O'Brien, 2013; O'Brien & Li, 2006). If a county leader holds many public positions and dominates the region, the probability of finding a strong advocate is slim and citizens are less prone to challenge the leader. Because power structure influences protests through both grievances and collective action, whether power has positive or negative effects on protests remains an empirical question.

In this chapter, power structure is specified in the context of county leadership–how much power is held by a county party secretary, the most powerful leader in a county. I collect the biographical data of 2,714 county leaders from an online database, Local Leadership Database, provided by People's Daily Online (人民網地方領導資料庫) and further complement the data by Baidu Encyclopedia (百度百科). To proxy for power structure, I create a measure, *Position*, for the number of public positions held by a county leader at the same time. *Position* indicates the level of power held by a county leader: the more positions s/he has, the more powerful s/he becomes. Given the panel structure of the data set[1], *Position* can be interpreted as power concentration as well. If a county party secretary occupies many public positions, it not only implies a powerful leader, but also suggests a concentrated power structure in a county.

The empirical results show that power structure has negative effects on protests. On average, if a county leader gains one more position, the frequency of protests decreases by around one-fifth of the average number of protests in a county[2]. The results are both statistically and economically significant and imply that political opportunities play a prominent role in protests–concentrated power structure negatively influences the chance of success

[1]In addition, I assume fixed total number of public positions between 2014 and 2015 in a county.
[2]On average, there are 0.34 protests in a county in a month.

for a protest and deters aggrieved citizens from participating in a protest. I address the issue of endogeneity by adopting an instrumental variable approach with fixed effects. Reverse causality can be a potential threat since protests may influence promotion probabilities and alter power structure. I first use one-period lag of power structure to alleviate reverse causality. Then, I employ fixed effects models with two instrumental variables. The first instrumental variable is the average number of positions held by other county leaders in the same prefecture, *Average Position*. Since county leaders in the same prefecture are evaluated based on the comparison of one another by the same prefecture authority, *Average Position* can be interpreted as the level of power held by the competitors of a county leader and should be negatively correlated to *Position*. Furthermore, *Average Position* should not influence *Protest* through protests in other counties in the same prefecture because most protests are transient and arise from regional grievances. Besides, vast and comprehensive censorship prohibits communication across counties. Without the help of media and the Internet, a short-term, regional protest can hardly have an impact on protests in other counties. The second instrumental variable is the power structure twice lagged. I argue that the power structure twice lagged is correlated with the lagged power structure, but does not have a direct impact on current protest. Put differently, I argue that power structure two months ago can only influence current protests through power structure in the previous month.

I identify the underlying mechanisms by examining the two components of protests, grievances and collective action, respectively. First, I show that under the rule of a power leader, the difficulty of forming collective action is elevated and thus fewer protests take place. I focus on environmental protests and use pollution as a proxy for the root of grievances. The results show that given the same level of pollution, more concentrated power structure leads to fewer environmental protests. In other words, holding grievances constant, people are less likely to protest if county leaders are more powerful.

Second, I find that a "local" party secretary behaves differently from an "outsider" in terms of the level of grievances when they become more powerful. A party secretary is defined as "local" if his/her hometown and ruling county are located in the same prefec-

ture[1]. The empirical evidence shows that in counties ruled by outsiders, power structure negatively affects protests, but this negative relationship is not found in counties ruled by locals. One plausible explanation for the distinct empirical pattern is that there are more grievances when a local party secretary obtains more power. To verify the explanation, I analyze the level of air pollution for counties ruled by locals and outsiders respectively. The results show that when a local party secretary wields a lot of power in the government, the level of air pollution is significantly higher, whereas the effects of power on air pollution are negative and insignificant for outsiders. I also examine the causes of protests and find that, given the occurrence of protests, people tend to protest against the government, not firms or other entities, if a local party secretary becomes more powerful. Moreover, protests are more likely to arise from governmental policies, such as environmental protests and land protests, when power is more concentrated at the hands of a local leader.

To answer why there are more grievances when a local leader becomes more powerful, I analyze who gains more power. Examining the link between power and leaders' characteristics, I find that economic performance and workplace connection are positively correlated to power[2], consistent with the literature on political selection in China (T. Chen & Kung, 2016; Persson & Zhuravskaya, 2016; Shih et al., 2012). Despite the fact that economic performance and workplace connections are both crucial for promotion, the two factors exhibit different patterns for locals and outsiders: compared to outsiders, local party secretaries tend to fail in economic performance and have worked under the direct superior authorities before their terms in the office. If economic performance is a proxy for competence and workplace connections represent ties with superior authorities (T. Chen & Kung, 2016; Persson & Zhuravskaya, 2016), the empirical evidence implies that local leaders are less competent but have better connections with superior authorities.

[1]In most cases, a county leader comes from the province where his/ her ruling county belongs, so there are not many outsiders for analysis. Similarly, there are not enough observations if hometown connection is defined at the county level.

[2]Economic performance is proxied by nighttime light, which is frequently used as an indicator of GDP growth in the literature. Workplace connection is a dummy of previous work experience in a prefectural government. Since county leaders are promoted based on the evaluations of prefectural authorities, working in a prefectural government suggests a closer tie to prefectural authorities and may increase the chance of promotion.

In sum, the results demonstrate that power has negative effects on the frequency of protests. I also show that power plays a crucial role in the two major components of protests, grievances and the formation of collective action. Moreover, the effects differ in terms of leaders' hometown. Local leaders, who are less competent and rely on better connections with superior authorities, generate more grievances when they wield a lot of power in the government.

The remainder of this article is organized as follows. Section 2.2 covers related literature on protests and power structure. Section 2.3 describes the data, including two novel data sets on protests and county leaders in China. Section 2.4 explains the empirical strategy and analyzes the relationship between power and protest. Section 2.5 discusses the underlying mechanism. Section 2.6 provides some evidence on the correlation between power and county leaders' personal characteristics. Section 2.7 concludes the chapter.

2.2 Related Literature

The chapter contributes to several strands of literature. First, the chapter speaks to the rich literature on local capture in developing countries. Bardhan (2002) points out that without political participation of the disadvantaged, decentralization can hardly achieve local accountability under weak institutions, and regional governments are often captured by local elites, which may harm local business development. Given highly restricted opportunities for political participation in China, the results in this chapter are aligned with the theory of local capture and demonstrate that higher levels of air pollution are observed in counties ruled by local leaders, and the situation deteriorates if local leaders gain more power in the government. On the other hand, positive outcomes of local capture have been documented. Persson and Zhuravskaya (2016) classify provincial leaders based on their career background and find that "local" governors, who build their careers in their ruling provinces, provide more public goods in health care and education, while "outsiders" spend more on construction infrastructure. The authors provide suggestive evidence that the distinct behavior comes from local elite capture. Local capture has been discussed in other developing countries as well. Slinko, Yakovlev, and Zhuravskaya (2005) study regional regulatory capture in Russia and find that regulatory capture benefits politically connected

firms but is detrimental to firms without political ties. In addition, regulatory capture has negative effects on small-business growth and fiscal collection. Alatas et al. (2013) document that local elites with formal leadership positions tend to benefit more from government welfare programs than non-elites in Indonesia. Panda (2015) finds evidence in India that politically connected households are more likely to receive poverty-alleviating entitlements. Nath (2014) employs difference-in-difference models and finds that without re-elections, politicians in India tend to allocate resources on projects catering to the wealthy.

Second, the chapter contributes to the literature on protest in China by introducing a large comprehensive protest data set. Protest data in China are rarely available and the existing literature often focuses on labor protests, which are less sensitive to the Chinese government. Yuan Li (2014) uses labor disputes as a proxy for collective action and studies the downward accountability of provincial leaders to collective actions in China. The author examines policies with opposing preferences between the central authority and the people, say, One-Child Policy, and finds that more labor disputes lead to decreased fines on excessive fertility. One weakness of the paper is the ambiguous link between labor disputes and excessive fertility fines since, essentially, workers do not protest against One-Child Policy. In my data set, protests of different causes display distinct patterns, and One-Child Policy is rarely a reason for protest. Chan et al. (2014) and Cai (2008) discuss all kinds of protests at the provincial level, but the sample size is under 100 observations and not enough for robust empirical analysis. Using the data set on protests of different protest causes and protester identities at the county level, the chapter expands the literature by exploring previous unanswerable questions and documenting different patterns of protests in China.

Third, the chapter also relates to the literature on political selection in China. On the one hand, promotion is based on competency—the cadre evaluation system motivates regional leaders to pursue economic growth and regional leaders with outstanding economic performance are more likely to get promoted. Hongbin Li and Zhou (2005) and Y. Chen, Li, and Zhou (2005) provide empirical evidence that economic performance is positively correlated to the promotion probability of provincial leaders but negatively correlated to the likelihood of termination. On the other hand, political connections are crucial to career advancement. Shih et al. (2012) introduce a quantitative measure of factional ties with top leaders and find that connections to top leaders have a positive impact on elite rank-

ing. Keller (2015) uses social network analysis to investigate the informal network among Chinese elites and confirms that links to patrons increase the probability of becoming top leaders in the Communist Party. The key independent variable in the chapter, *Position*, can be perceived as a measure of "concurrent job promotions"–a change in *Position* indicates a minor promotion or demotion and captures the fluctuations of power structure within the term of office of a politician, while the traditional definition of promotion evaluates the career of a politician only at the end of his/her term. Despite different definitions, the two types of promotion share similar patterns. Consistent with the literature, empirical evidence shows that "concurrent job promotions" are also based on meritocracy and patronage: county leaders with better economic performance and workplace connections are more likely to obtain positions in the government.

Finally, the relationship between power structure and collective action has been discussed in the literature of sociology and political science. O'Brien and Li introduce the concept of rightful resistance and point out that protesters exploit elite divisions to locate suitable benefactors and use official rhetoric and slogans to advance their claims (O'Brien & Li, 2005, 2006). Rightful resistance relies on the fragmentation of state power and distinct interests of officials in different departments or at different levels of the governments. Accordingly, if a leader is dominant in a region and the power is concentrated in his/her hands, people can rarely find strong advocates to help address their grievances, leading to fewer protests. The chapter is built on the theory of rightful resistance and shows empirically that given the same level of grievances proxied by air pollution, there are fewer environmental protests if leaders wield a lot of power in the government. The chapter is also motivated by Somers (1993), which documents divergent regional patterns in the emergence of citizenship identities among working class in the 18th century of England. Under the same legal system in a country and within the same social class, the author finds that in arable regions, the working population perceived the law as "a form of social control", whereas working communities in pastoral regions expressed identities with various dimensions of citizenship rights and used the law to protect their rights. The author claims that the disparity stemmed from the difference in regional power structure–arable regions were dominated by powerful elites and the public sphere were controlled by the wealthy, while pastoral regions were featured by small farms and people were encouraged to participate in

the public sphere. Since rights-consciousness is essential to rightful resistance, the finding lends support to the theory of rightful resistance and indicates that regional power structure has an impact on protest by shaping the public sphere of a region.

2.3 Data

2.3.1 Protests

One major contribution of the chapter is to introduce two novel data sets. The first data set includes all kinds of protests in China from a citizen media outlet, *Not the News* (非新聞). Since 2012, *Not the News* had documented daily protests systematically throughout China from online sources and reported the data on Youtube, Twitter and other social media which are blocked in China till 13th June 2016. The founders of *Not the News*, Lu Yuyu and Li Tingyu, were awarded the Reporters Without Borders (RSF) Prize in 2016 for their commitment to the promotion of press freedom in China (Reporters Without Borders, 2016).[1]

The original data set consists of 50,129 protests from 2014 to 2015[2] with information on dates, locations, pictures, videos, online discussion archives, the cause of a protest, the form of a protest, the identity of protesters, and whether a protest is repressed. Based on the labels in the data set, I develop Python programs to compile a list of keywords for each type of protests and mark the causes, forms, and identities of participants for each unlabeled protest. Then, I classify protests into 11 types of protest causes, 7 types of protest forms, and 18 types of protesters. To match the protest data with the county leadership data, I keep only protests with place names that can be identified at the county level. That is, protests with no place information or with only information on provinces or prefectures are removed from the data set. Figure 2.1 displays the number of protests at different adminis-

[1] Lu Yuyu and Li Tingyu were found missing on 15th June 2016, and confirmed arrested by the police for "picking quarrels and provoking trouble" in late July 2016 (Qiao, Wong, & Mudie, 2016; Wong, Yang, & Mudie, 2016). Li Tingyu was given a suspended sentence, while Lu Yuyu was sentenced to four years in prison in 2017 (Ramzy, 2017).

[2] Although *Not the News* started collecting data from 2012, it has only released text data from the second half of 2013, and labeled the protest cause, protest method, or protester identities since November 21st, 2013. Therefore, I focus on protests from 2014 instead of 2012.

trative levels over time. 8.44% of the protests take place at the provincial level, 40.35% of the protests arise in prefectures, and 1.79% of the protests are not labeled with any place. In the end, the data set is trimmed down 50.58% of the observations.

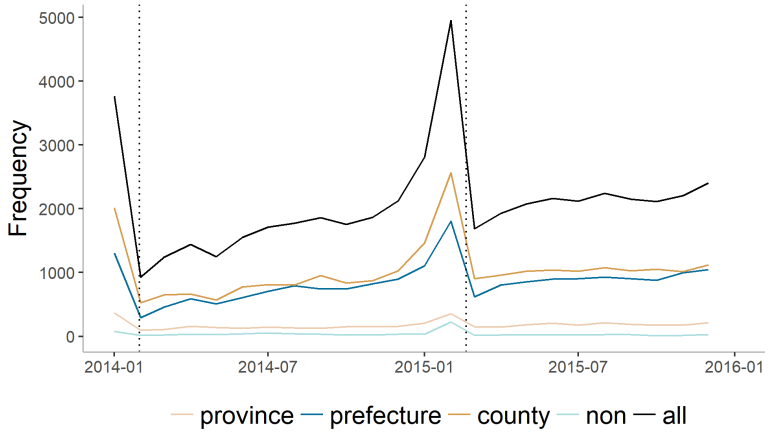
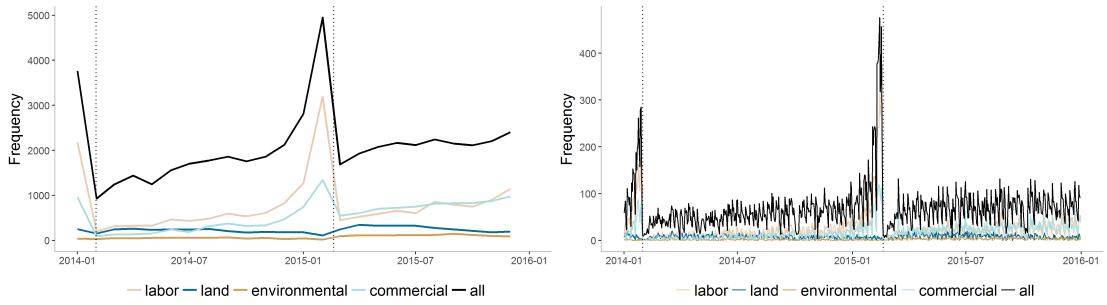


Figure 2.1: The number of protests at different administrative levels.

Overall, there are 24,774 protests at the county level from 2014 to 2015. On average, 0.44 protests occur in a month in a county[1], which translates into 33.94 protests per day all over China. Protests usually arise from labor (37%), land (28%), or commerce disputes (21%), protests are often initiated by workers (40%), consumers (16%), or peasants (11%), and the most popular protest forms are demonstrations (84%), strikes (10%), and conflicts (6%). Figure 2.2 displays the number of protests for the major protest causes in China, including labor disputes, land acquisition and resettlement, commercial disputes, and environmental damage. There are two spikes before the Chinese New Year on January 31st 2014 and February 19th 2015, which mainly come from protests about labor and commercial disputes. Traditionally, people would like to pay all debts by New Year's Eve. Besides, migrant workers and merchants prefer to go home with their earnings for the Chinese New Year. As a result, the number of protests related to labor and commercial disputes usually skyrocket before the Chinese New Year, while environmental and land protests remain the same level around the Chinese New Year.

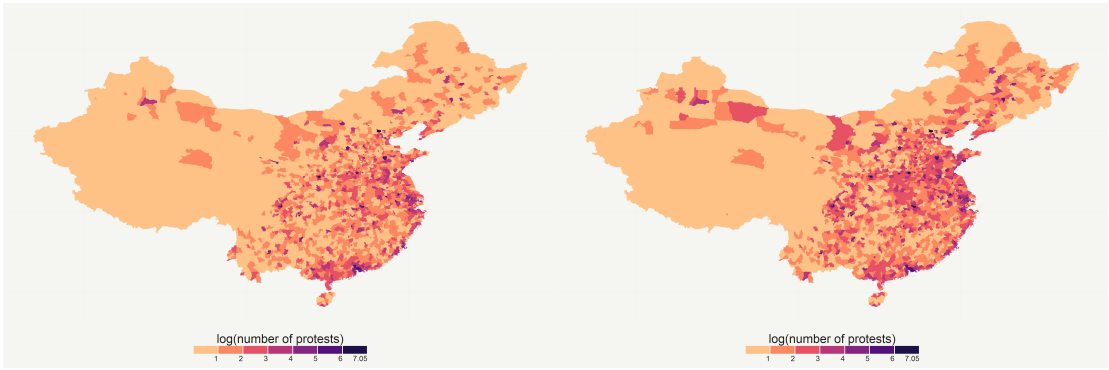
[1]There are 2,368 counties covered in the protest data set.



(a) Monthly **(b)** Daily

Figure 2.2: The number of protests for the major protest causes in China.

Note: The dashed lines represent the Chinese New Year on January 31st 2014 and February 19th 2015.

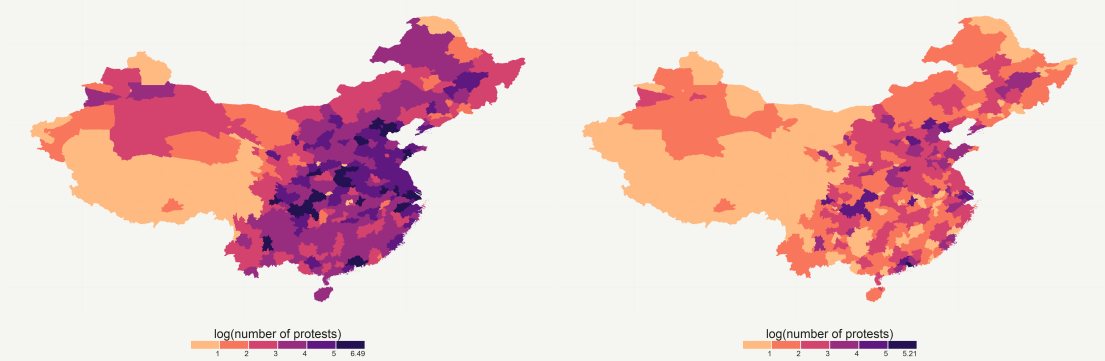


(a) 2014 **(b)** 2015

Figure 2.3: The log number of protests in 2014 and 2015, respectively.

Figure 2.3 illustrates the maps of log number of protests in 2014 and 2015 on the same scale. There are relatively more protests in the east (22%) and central south (22%), and relatively fewer protests in the northeast (10%) and northwest (12%), and considerable variations within each region. Figure 2.3a displays 21,266 protests in 2014, while Figure 2.3b shows 28,875 protests in 2015, which is 35.6% higher. The increase may reflect an accelerating trend in the frequency of protests, but also the improvement in data collection. To account for this issue, time fixed effects are included in all regression models.

To corroborate the credibility of the data from *Not the News* (NN), I compare labor protests from the dataset with the data from China Labour Bulletin (CLB), a Hong-Kong based non-governmental organization devoted to the promotion of workers' rights in China.[1] Figure 2.4 shows the maps of log number of protests from NN dataset and CLB dataset on the same scale. NN documents much more labor protests than CLB does: NN reports 19,056 labor protests in 2014 and 2015, while CLB archives 4,107 protests during the same period. Despite the discrepancy in terms of the quantity, the two datasets are highly correlated at 93.5% across regions over time, and the two maps suggest similar distributions of protests all over China.

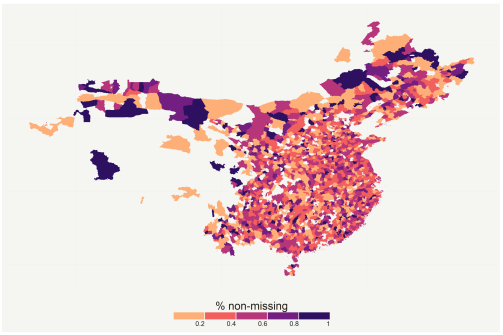


(a) *Not the News* **(b)** *China Labour Bulletin*

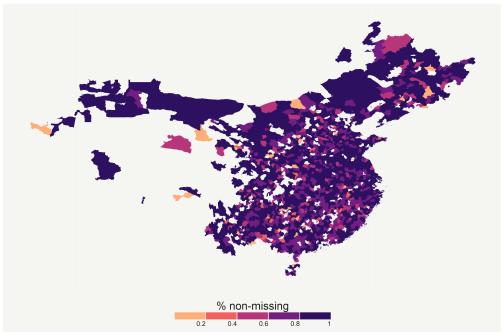
Figure 2.4: The total number of labor protests from *Not the News* and from *China Labour Bulletin*, respectively.

[1]Since China Labour Bulletin only records labor protests at the prefecture (city) level, I select only labor protests from the dataset of *Not the News* and re-compile the data at prefecture level to make the two datasets more comparable.

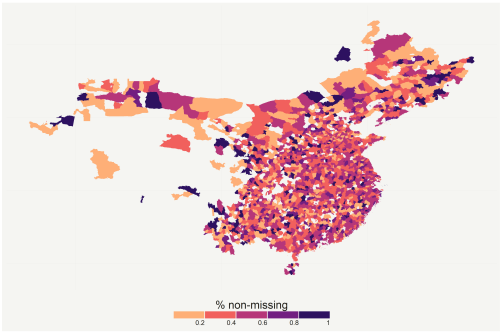
Figure 2.5 displays the share of protests with labels about protest causes, protester identities, or protest methods. Given the original labeling by *Not the News*, 36.94% protests are categorized into one or more protest causes, 84.48% protests are labeled with at least one type of protester identities, and 41.63% protests are tagged with some protest methods. In total, 87.81% protests are marked with at least one of the protest causes, protester identities, or protester methods, and Figure 2.5 does not seem to indicate any particular pattern regarding the labeling. After extending the labeling based on keyword matching, 99.42% protests are marked with at least one type of protest causes, protester identities, or protest methods. In the following sections, I always use the original manual labeling whenever possible, and only use the extended labeling if too few observations are available.



(a) Protest Causes



(b) Protester Identities



(c) Protest Methods

Figure 2.5: The percentage of protests with labels on protest causes, protester identities and protest methods, respectively.

2.3.2 County Leadership

The second novel data set comprises the biographical data of county leaders scraped from an online database, Local Leadership Database (人民網地方領導資料庫), provided by People's Daily Online and further complemented by the information from Baidu Encyclopedia (百度百科). The data set includes 2,714 county leaders with each leader's name, gender, ethnicity, birth year, ancestral home place, birthplace, education, and work history. The average age of county leaders is 48.9, and most of them are Han Chinese (89%) and male (94%); almost all county leaders hold a bachelor's degree (98%) and almost two thirds possess a master or doctoral degree (65%); more than half of the county leaders are from the central south or the east of China.

To proxy for power structure, I create a measure, *Position*, or the number of public positions held by a county leader at the same time. The measure is built on the work history in the biography and processed by self-written Python programs. I first identify the service time of each county leader. Since the data set only includes the information on current leaders, anyone who comes in office after 2015 or leaves office before 2014 is defined as missing. Then, given a politician takes up the post of a county party secretary (the position with the largest power), I count the number of other positions taken by the politician. There are multiple public positions in regional party organizations and the regional government, such as county party secretaries, county magistrates (the position with the second largest power in a county), bureau chiefs, governmental department heads, party school principals, etc. It is not uncommon for a county leader to be assigned more than one role and hold many public positions at the same time. In the data set, a county leader holds 1.45 positions on average, 11.80% of county leaders hold more than two positions, and the maximum number of positions held at the same time is 8.

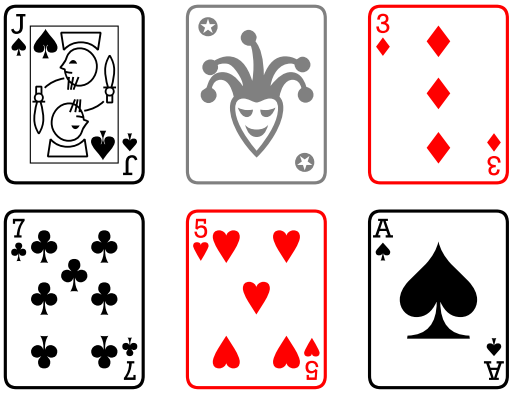
The reasons for a county leader to hold multiple positions vary. In some cases, a public position suddenly becomes vacant because the previous position holder resigns, passes away, or gets arrested, and a county leader temporarily takes the position until a suitable candidate for the position is determined. In some other cases, a county leader is appointed as the leader of a new administrative area established nearby or within the jurisdiction. In either case, the Chinese government fails to assign a non-incumbent politician to a vacant

public position within a short period. It can be that the incumbent leader is competent and well-connected to serve several posts at the same time; it can also be the lack of powerful candidates or competitors in the region. Nevertheless, both situations indicate a low level of political competition and results in a more concentrated power structure.

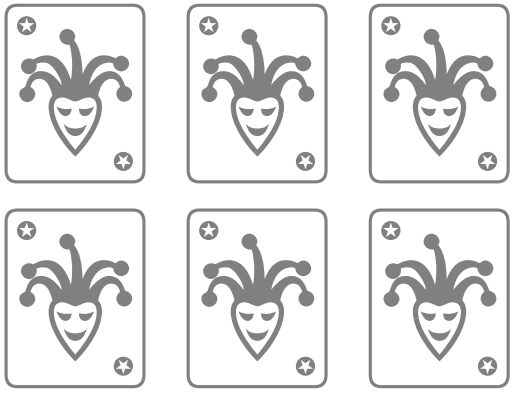
Table 2.1 illustrates the concept behind the measure. Suppose there are 6 official positions in a county, including the most powerful position, county party secretary, which is represented by a joker card. Under the most decentralized power structure (Table 2.1a), each position is taken by a different official and each official belongs to a different faction, denoted by the suit of a card. In this case, the county party secretary only occupies one position and possesses limited power. On the other hand, in the case of unitary power structure (Table 2.1b), all 6 positions are held by the county party secretary–the power is concentrated in the hands of the county party secretary with few checks and balances from other officials. As a result, given the total number of positions in a county[1], if the county party secretary is appointed more positions, the power structure becomes more concentrated in the region.

Table 2.1: Power structure

(a) Divided power structure



(b) Unitary power structure



[1]Although information on the total number of public positions in a county is not included in the data set, it is controlled by the county fixed effects in the model if assuming the number is constant within each county in 2014 and 2015.

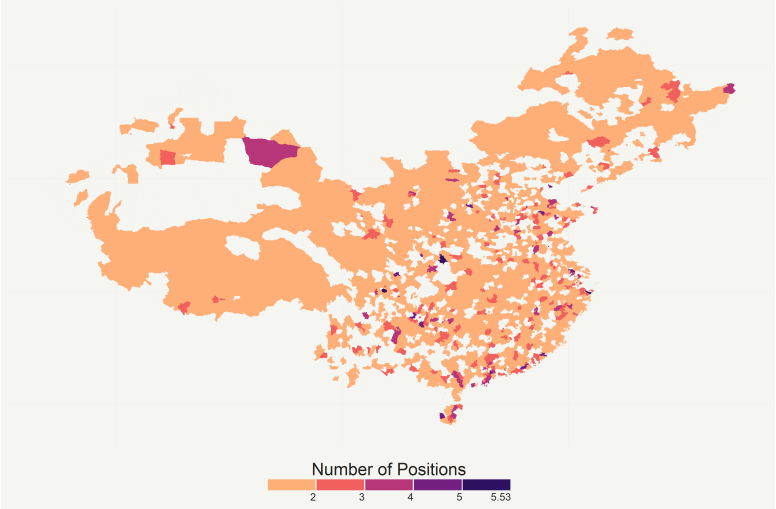


Figure 2.6: Number of positions

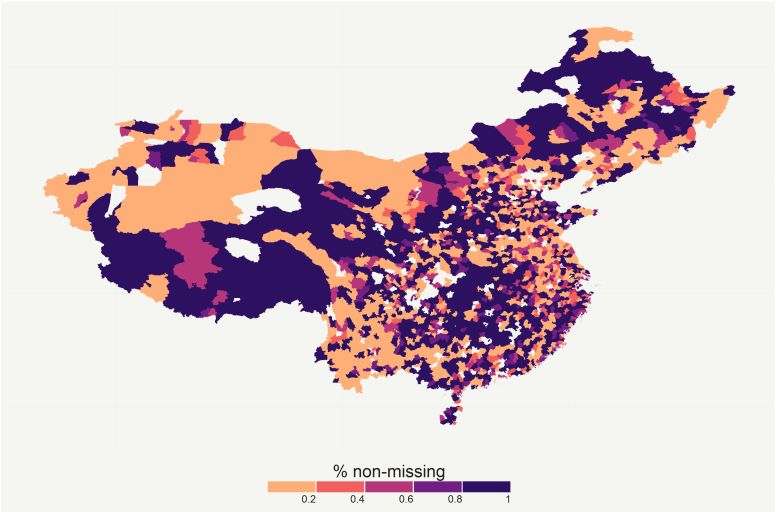


Figure 2.7: The percentage of positions with non-missing values.

Figure 2.6 maps the average number of positions held by each county party secretary from 2014 to 2015. The white area depicts missing values in the data set, which is 33.55% of all county party secretaries. The information on the number of positions held by a county leader is missing for two reasons. First, the leader's biography is incomplete in both Local Leadership Database and Baidu Encyclopedia, and the information on position number is unavailable. Second, Local Leadership Database only keeps track of the information on current regional leaders. If a regional leader is appointed as a county party secretary after 2016, then s/he is not included in the data set and labeled as missing to match the protest data set in 2014 and 2015. Figure 2.7 displays the percentage of positions with non-missing values all over China. In spite of the nonnegligible number of missing values, the counties with non-missing values are arguably random and located all over China. More importantly, as shown in the next section, the baseline regression model includes county fixed effects–the empirical results are based on the analysis within each county and what really matters is the relative number of positions, not the absolute number of positions.

2.3.3 Others

2.3.3.1 Precipitation

Precipitation is often used as a source of exogenous variations in protest participation (Madestam, Shoag, Veuger, & Yanagizawa-Drott, 2013; Wasow, 2017) because people are less likely to take to the streets to protest on rainy days. In this chapter, I include precipitation as a control variable and collect monthly global data on Global Precipitation Measurement (GPM) between April 2014 and December 2015 from the National Aeronautics and Space Administration (NASA). Combing the data on administrative areas from the Database of Global Administrative Areas (GADM), I use QGIS to calculate rainfall for each county in China.

It is worth mentioning that 36.8% of the observations are missing, resulting in a sizable reduction in the sample size. The considerable missing data come from two sources. First, NASA used to collect rainfall data under the Tropical Rainfall Measuring Mission (TRMM) from December 1st 1997 to June 1st 2015 and only started the GPM project from

March 12th 2014, so monthly GPM data before April 2014 are not available. Second, the administrative areas from the GADM database are not up to date and cannot be completely matched with the more current administrative structure in the protest and leadership data sets. Nevertheless, the two sources are arguably random and should not bias the regression results.

2.3.3.2 Nighttime Light

Economic performance can be interpreted as the opportunity cost of participating a protest (Dorsch, Dunz, & Maarek, 2015). That is, people are less willing to take to the streets given high opportunity costs, such as protesters' monetary income if they have to take a day off to participate in a protest. I include nighttime light as a proxy of regional GDP. In particular, I use monthly NPP-VIIRS nighttime light data by the Earth Observation Group of National Oceanic and Atmospheric Administration/National Geophysical Data Center (NOAA/NGDC), an updated version of the DMSP-OLS nighttime light data that provide images with higher resolution and serve as a better proxy of GDP and other socioeconomic indicators (Ou, Liu, Li, Li, & Li, 2015; Shi, Huang, et al., 2014; Shi, Yu, et al., 2014).

I use nighttime light instead of official GDP released by the Chinese government for two reasons. First, monthly county-level GDP data are unavailable.[1] Second, GDP data in China are notorious for measurement errors(Lü & Landry, 2014). Similar to precipitation data, nighttime light data contain quite substantial missing values: 27.8% observations are missing solely because of mismatch in the administrative areas.

2.3.3.3 Air Pollution

I collect data on air quality from Berman (2017), which consists of self-reported data on particulate matter 2.5 micron diameter $\mu g/m_3$ (PM 2.5) and other indicators of air pollution three times per day[2] from nearly 1,700 ground monitoring stations in China. I choose

[1] Yearly county-level GDP data can be found in China Statistical Yearbook for Regional Economy till 2013.

[2] The timing corresponds to morning rush hour, the end of a work day, and midnight.

Berman (2017) over remotely sensed data provided by NASA because the latter is susceptible to temporary weather conditions and can barely measure the severity of local pollution on the ground. Nevertheless, the more accurate ground data come with a cost–the ground monitoring stations are far from abundant and 76.6% counties are not covered. That being said, Figure 2.8 shows that the stations are widely distributed throughout China. More importantly, as mentioned previously, the empirical strategy of the chapter takes advantage of the panel data structure and the identification comes from within-county variations. Therefore, any time-constant county-specific characteristics should be explained by county fixed effects. The average value of PM 2.5 is 118.21 $\mu g/m_3$ in the data set, which is far beyond the limits in the European Union (25 $\mu g/m_3$), the United States (12 $\mu g/m_3$), and even China (35 $\mu g/m_3$) (Environment of European Commission, 2016; Ministry of Environmental Protection (China), 2012; US EPA, 2015).

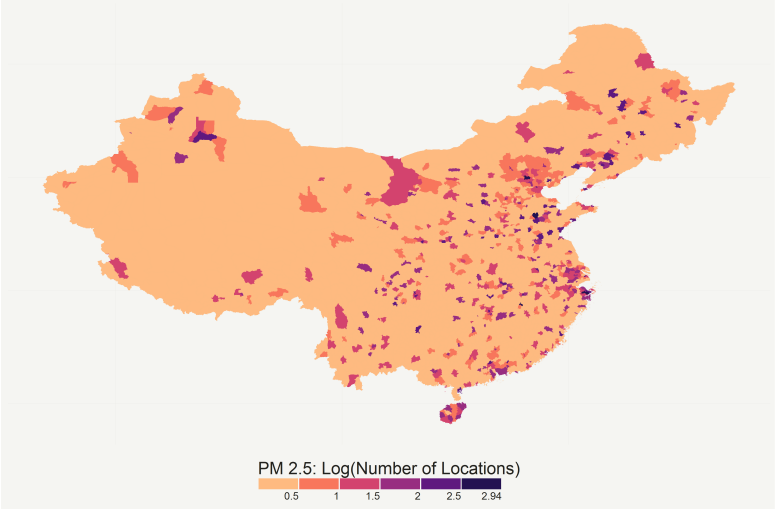


Figure 2.8: Ground monitoring stations

Table 2.2: Summary Statistics: County-level

	count	mean	sd	min	max
PROTEST					
# Protests	72360	0.34	0.90	0.00	23.00
Area: Northeast	72360	0.10	0.30	0.00	1.00
Area: Central South	72360	0.22	0.42	0.00	1.00

Area: East	72360	0.22	0.42	0.00	1.00
Area: North	72360	0.15	0.36	0.00	1.00
Area: Northwest	72360	0.12	0.33	0.00	1.00
Area: Southwest	72360	0.18	0.38	0.00	1.00
LEADERSHIP					
# Positions	39174	1.45	0.83	1.00	8.00
Duration	39980	29.62	20.25	0.00	170.00
Birth Year	56811	1965.47	3.76	1951.00	1982.00
Age	56811	48.88	3.81	31.08	64.00
Female	69234	0.06	0.23	0.00	1.00
Ancestral Place					
Ancestry: Northeast	38940	0.04	0.19	0.00	1.00
Ancestry: Central South	38940	0.27	0.45	0.00	1.00
Ancestry: East	38940	0.26	0.44	0.00	1.00
Ancestry: North	38940	0.16	0.37	0.00	1.00
Ancestry: Northwest	38940	0.11	0.31	0.00	1.00
Ancestry: Southwest	38940	0.15	0.36	0.00	1.00
Birthplace					
Birthplace: Northeast	37398	0.05	0.21	0.00	1.00
Birthplace: Central South	37398	0.27	0.45	0.00	1.00
Birthplace: East	37398	0.32	0.46	0.00	1.00
Birthplace: North	37398	0.13	0.33	0.00	1.00
Birthplace: Northwest	37398	0.10	0.29	0.00	1.00
Birthplace: Southwest	37398	0.14	0.35	0.00	1.00
Home					
Home County: Ancestry	38940	0.01	0.08	0.00	1.00
Home County: Birthplace	37398	0.01	0.09	0.00	1.00
Home Prefecture: Ancestry	35622	0.56	0.50	0.00	1.00
Home Prefecture: Birthplace	34107	0.57	0.49	0.00	1.00
Home Province: Ancestry	38940	0.82	0.38	0.00	1.00
Home Province: Birthplace	37398	0.84	0.37	0.00	1.00
Education					
EDU: Vocational	32010	0.02	0.14	0.00	1.00
EDU: Undergraduate	32010	0.33	0.47	0.00	1.00
EDU: Graduate	32010	0.65	0.48	0.00	1.00
Ethnicity					
Ethnicity: HA	49542	0.89	0.31	0.00	1.00

Ethnicity: ZH	49542	0.01	0.11	0.00	1.00
Ethnicity: HU	49542	0.01	0.08	0.00	1.00
Ethnicity: MA	49542	0.00	0.00	0.00	0.00
Ethnicity: UG	49542	0.00	0.00	0.00	0.00
Ethnicity: ZA	49542	0.01	0.12	0.00	1.00
Ethnicity: MG	49542	0.01	0.12	0.00	1.00
Ethnicity: Other	49542	0.06	0.24	0.00	1.00
OTHERS					
Precipitation	40689	93.77	93.00	0.00	881.90
Nighttime Light	46512	0.80	2.04	-0.19	38.79
Flood	72360	0.02	0.18	0.00	2.00
PM 2.5	13228	118.21	41.36	0.00	342.31
Observations	72360				

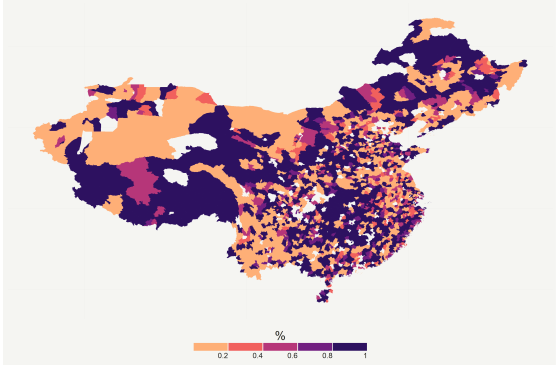
Table 2.3: Summary Statistics: Protest-level

	count	mean	sd	min	max
# Netizens	24774	3.41	5.97	0.00	198.00
# Protest Places	24774	1.00	0.02	1.00	2.20
# Protest Days	24774	1.01	0.08	1.00	2.00
State Power					
Repression*	24774	0.21	0.41	0.00	1.00
Law Enforcer*	23304	0.27	0.45	0.00	1.00
Casualties	24030	0.05	0.22	0.00	1.00
Gangland*	23304	0.03	0.18	0.00	1.00
Protest Cause					
Labor	9152	0.37	0.48	0.00	1.00
Land	9152	0.28	0.45	0.00	1.00
Commerce	9152	0.21	0.41	0.00	1.00
Environment	9152	0.06	0.24	0.00	1.00
Medic	9152	0.02	0.14	0.00	1.00
Government	9152	0.02	0.15	0.00	1.00
Chengguan	9152	0.02	0.12	0.00	1.00
Transport	9152	0.01	0.08	0.00	1.00
Property*	18864	0.24	0.43	0.00	1.00
Disaster*	18864	0.02	0.15	0.00	1.00

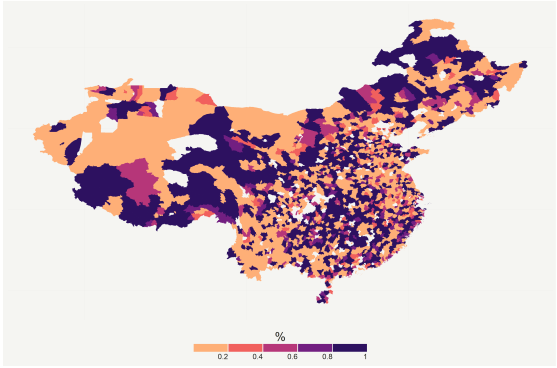
Other	9152	0.01	0.09	0.00	1.00
Protest Cause: Gov					
Gov	9152	0.38	0.48	0.00	1.00
nGov	9152	0.61	0.49	0.00	1.00
Gov: Policy	9152	0.34	0.47	0.00	1.00
Gov: nPolicy	9152	0.04	0.19	0.00	1.00
Protester Identity: Social Class					
Middle Class	20930	0.21	0.40	0.00	1.00
Lower Class	20930	0.66	0.47	0.00	1.00
Army	20930	0.00	0.06	0.00	1.00
Public Servant	20930	0.00	0.06	0.00	1.00
Protester Identity: Other					
School	20930	0.03	0.16	0.00	1.00
Retired, Laid-off	20930	0.00	0.06	0.00	1.00
Ethnic Minority*	23304	0.01	0.08	0.00	1.00
Human Rights Activist*	23304	0.01	0.09	0.00	1.00
Protest Method					
Demonstration	10314	0.84	0.37	0.00	1.00
Strike	10314	0.10	0.30	0.00	1.00
Conflict	10314	0.06	0.24	0.00	1.00
Blocking	10314	0.03	0.16	0.00	1.00
Petition	10314	0.01	0.09	0.00	1.00
Violence*	18274	0.03	0.18	0.00	1.00
Other*	18274	0.03	0.18	0.00	1.00
Observations	24774				

* In case of frequency < 50, I extended the protest classification based on discussion archives.

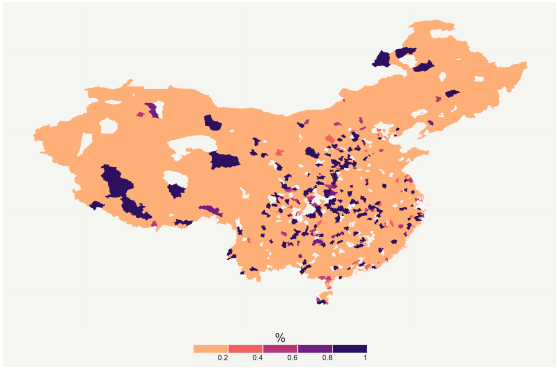
Figure 2.9 and Figure 2.10 illustrate the percentage of non-missing values in the data set. Sample 1 contains two major variables of interest, protests and positions, Sample 2 adds two control variables, precipitation and nighttime light, and Sample 3 further includes personal characteristics of county leaders, such as age. Overall, there are 35,631 non-missing observations in Sample 1 (55.3%), 23,583 non-missing observations in Sample 2 (36.6%), and 5,736 non-missing observations in Sample 3 (8.9%).



(a) Sample 1



(b) Sample 2



(c) Sample 3

Figure 2.9: The percentage of non-missing values in each sample.

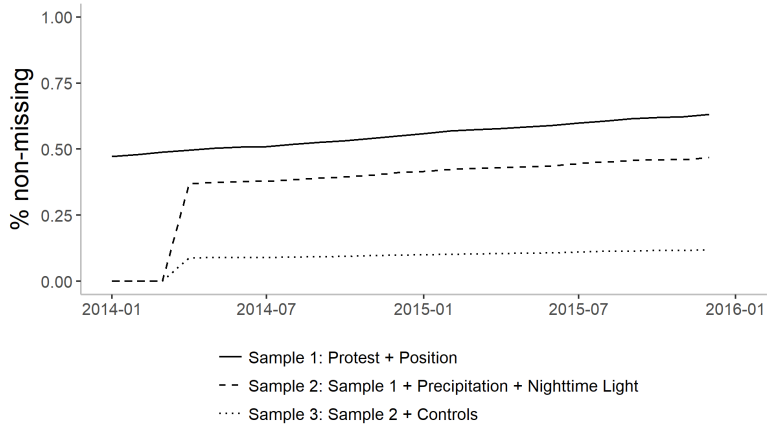


Figure 2.10: The percentage of non-missing values in each sample over time.

2.4 Does Power Affect Protest?

In this section, I investigate the relationship between power and protests. On the one hand, protests stem from grievances, and the root of grievances may relate to a certain policy and reveal public discontent with the government. Power structure can not only have an impact on policy choices, but also determine the degree of a county leader's influence on citizens. On the other hand, people are more likely to participate in a protest when they perceive a higher chance of success, or a political opportunity. If a county leader holds many public positions and dominates the region, the probability of finding a strong advocate is slim and citizens are less prone to challenge the leader. Because power structure influences protests through both grievances and collective action, whether power has positive or negative effects on protests remains an empirical question.

2.4.1 Empirical Strategy

Combining the data set on protests with the one on county leadership, I examine the relationship between power structure and protests in China using the following baseline regression model:

$$Protest_{i,t} = \beta_1 Position_{i,t-1} + \beta_2 \mathrm{X}_{i,t} + \mathrm{County}_i + \mathrm{Period}_t + \mathrm{Prefecture}_p \times \mathrm{Period}_t + \varepsilon_{i,t}$$

The dependent variable, $Protest_{i,t}$, is the number of protests in county i at time t, and the key independent variable, $Position_{i,t-1}$, is the number of positions a leader of county i holds at time t-1, representing the level of power or power concentration in a county. $\mathrm{X}_{i,t}$ contains precipitation, nighttime light, and personal characteristics of a county leader, such as age, ethnicity, gender, etc. Precipitation lowers people's willingness to take to the streets, while nighttime light, a proxy of GDP growth, correlates to the opportunity costs of joining a protest. Both variables should be negatively related to protests. The model also includes county fixed effects and period fixed effects. Moreover, I add a period fixed effect for each prefecture to capture the general time trend in each prefecture. Hence, the research question should be interpreted as follows: given a county leader holds some public positions in a county, if s/he is appointed one more public position in the next period and the power structure becomes more concentrated, will it reflect on the frequency of protests in the county?

Despite the inclusion of county and period fixed effects, endogeneity can still undermine the regression results. First, reverse causality can be problematic. Since maintaining social order is one of the "priority targets with veto power" (一票否决) for a county leader, protests are likely to influence the promotion opportunity of a county leader and indirectly alter the power structure in a county. Second, the issue of measurement errors is inevitable. The total number of protests in the data set is much smaller than the estimated number in 2010 (Sun, 2011) and cannot be verified as a representative sample because official information on protests is confidential to the public. In addition, the number of positions held by a county leader contains a non-negligible amount of missing values due to data limitations. Third, omitted variables can be a serious problem. For example, the popularity, or prestige, of a county leader is likely to have an impact on both power structure and protests, and because it is time-variant, the omitted variable problem cannot be solved by the inclusion of county or leader fixed effects.

To conquer the endogeneity, I first use one-period lag of power structure instead of the current period to alleviate reverse causality. Then, I employ fixed effects models with two instrumental variables. The first instrumental variable is the average number of positions held by other county leaders in the same prefecture, $\textit{Avg. Position}_{-i,t-1}$. Since county leaders in the same prefecture are evaluated based on the comparison of one another by the same prefecture authority, $\textit{Avg. Position}_{-i,t-1}$ can be interpreted as the level of power held by the competitors of a county leader and should be negatively correlated to $\textit{Position}_{i,t-1}$ (Huiping Li, Wang, & Zheng, 2016; Lü & Landry, 2014; Yu, Zhou, & Zhu, 2016). Figure 2.11 displays the distribution of the number of counties in a prefecture. On average, there are 9.68 counties in a prefecture and 5.45 counties with observed $\textit{Position}_{i,t-1}$. Furthermore, $\textit{Avg. Position}_{-i,t-1}$ should not influence $\textit{Protest}_{i,t}$ through $\textit{Protest}_{-i,t}$ for three reasons. First, information on protests is censored. Press freedom is very limited in China (Freedom House, 2017; Reporters Without Borders, 2017) and protests are rarely reported by domestic media. In addition, due to vast and comprehensive Internet censorship, online microblog posts and instant messages about sensitive topics are often quickly deleted and people can barely distribute information on protests across counties. Second, protests are usually driven by grievances from a specific region. Coordination across counties for a protest is extremely difficult under censorship and rarely happens. In the data set, only 11 protests occur in more than one county, which translates to 0.04% of 24,780 protests. Finally, most protests last only for a short period of time. In the data set, 98.98% of the protests take at most one day. The average area of a county is ca. 3,400 km^2, which is 1.3 times the area of Luxembourg. Without the help of media and the Internet, a short-term protest can hardly have an impact on protests in other counties. The second instrumental variable is the power structure twice lagged, $\textit{Position}_{i,t-2}$. I argue that the power structure twice lagged is correlated with the lagged power structure, but does not have a direct impact on current protest. In other words, I argue that power structure two months ago can only influence current protests through power structure in the previous month.

2.4.2 Results

Table 2.4 displays the baseline results. Column 1 runs an OLS regression, whereas Column 2-6 include county and period fixed effects. Column 1 and 2 assume the error terms to be

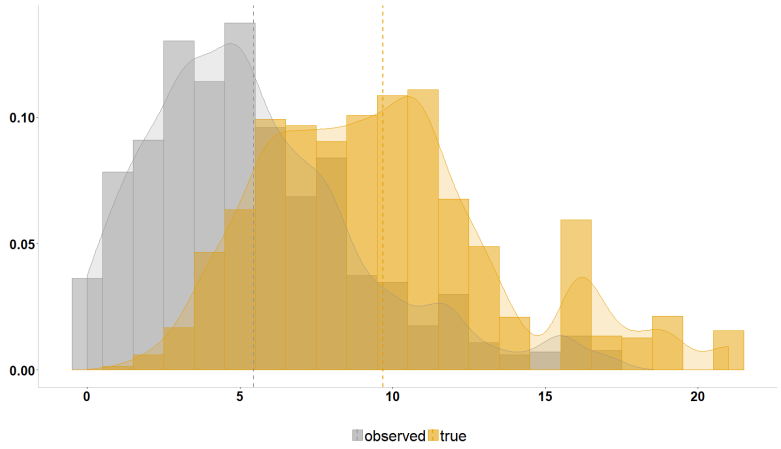


Figure 2.11: The number of counties in a prefecture.

independent and identically distributed. Column 3 relaxes the i.i.d. assumption and allows for arbitrary correlation between errors within each county, while Column 4-6 cluster the error terms at the prefecture level and incorporate a specific time trend for each prefecture into the model; Column 6 further adds a set of personal characteristics of county leaders into the model.

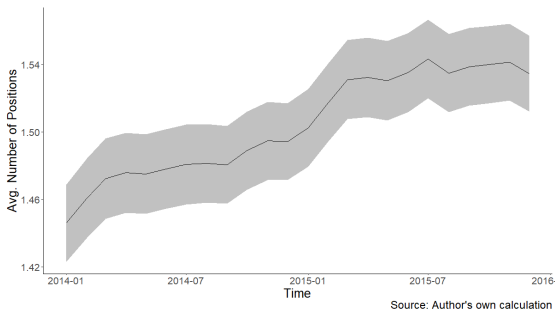


Figure 2.12: The average number of positions.

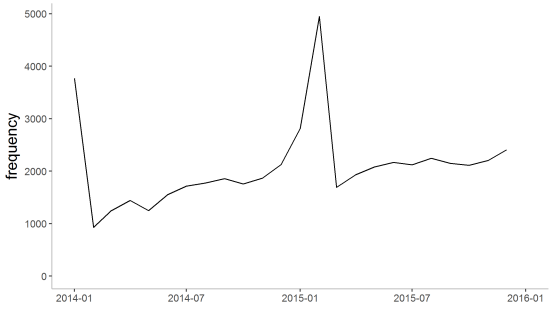


Figure 2.13: The total number of protests.

Figure 2.12 and Figure 2.13 display the average number of positions held by each county leader and the total number of protests, respectively. The two figures suggest that power is positively correlated with the frequency of protests, which is consistent with the positive coefficient of interest from the OLS model in Column 1. However, after including the county fixed effects and the month fixed effects, the coefficient of interest becomes nega-

tive. In Column 2-6, the effects of power structure on protest frequency are significant and negative, suggesting that people are less likely to protest under more concentrated power structure.

Table 2.4: The baseline regressions for power structure on protests.

	(1) iid	(2) iid	(3) County	(4) Prefecture	(5) Prefecture	(6) Prefecture
L.Position	0.106***	-0.0637**	-0.0637*	-0.0978**	-0.0906**	-0.108***
	(0.00593)	(0.0282)	(0.0368)	(0.0480)	(0.0361)	(0.0236)
Precipitation					-0.000569**	-0.000860**
					(0.000238)	(0.000416)
Nighttime Light					-0.0119	-0.0746***
					(0.0368)	(0.0199)
FE: County		V	V	V	V	V
FE: Time		V	V			
FE: Prefecture x Time				V	V	V
Controls						V
Observations	35134	35114	35114	34281	22033	4265
Adjusted R^2	0.009	0.316	0.316	0.325	0.337	0.274

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the baseline results of power structure on protests. The dependent variable is $Protest_{i,t}$, or the number of protests in a month in a county, and the key independent variable is $Position_{i,t-1}$, or the number of positions a county leader holds in a month. Column 1 includes only the main explanatory variable. Column 2 adds county and period fixed effects into the model. Column 3 relaxes the i.i.d. assumption and allows the error terms to be correlated within each county. Column 4 clusters the error terms at the prefecture level and incorporates a specific time trend for each prefecture. Column 5 adds two control variables, precipitation and nighttime light. Column 6 further includes other control variables, such as age, ethnicity, and gender of a county leader.

Next, I use two-stage least squares (2SLS) regression analysis to mitigate the endogeneity. Table 2.5 shows the results from first stage regressions. The dependent variable is a proxy of power, $Position_{i,t-1}$; the key independent variables are the instrumental variables, $Avg.\ Position_{-i,t-1}$ and $Position_{i,t-2}$. All models include county fixed effects and a specific time trend for each prefecture. Column 1 assumes the error terms to be independent and identically distributed. Column 2 relaxes the i.i.d. assumption and allows for arbitrary cor-

Table 2.5: First stage regressions.

	(1) iid	(2) County	(3) Prefecture	(4) Prefecture	(5) Prefecture
Avg. Position$_{-i}$	-0.343***	-0.343***	-0.343***	-0.307***	-0.202**
	(0.00695)	(0.0628)	(0.0926)	(0.0800)	(0.0821)
L.Position	0.785***	0.785***	0.785***	0.818***	0.872***
	(0.00343)	(0.0365)	(0.0460)	(0.0426)	(0.0381)
Precipitation				-0.00000706	-0.000000246
				(0.0000118)	(0.00000787)
Nighttime Light				-0.000474	-0.00147**
				(0.00102)	(0.000649)
FE: County	V	V	V	V	V
FE: Prefecture x Time	V	V	V	V	V
Controls					V
Observations	34149	34149	34149	21938	4243
Adjusted R^2	0.993	0.993	0.993	0.993	0.987

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the results from first stage regressions. The dependent variable is a proxy of power, $Position_{i,t}$, or the number of positions a county leader holds in a month; the two key independent variables are $Avg.\ Position_{-i,t}$, the average number of positions held by other county leaders in the same prefecture, and $Position_{i,t-1}$. Column 1-3 all include a specific time trend for each prefecture; Column 1 assumes i.i.d errors, while Column 2 and 3 cluster the error terms at county and prefecture level, respectively. Column 4 adds two control variables, precipitation and nighttime light. Column 5 further includes other control variables, such as age, ethnicity, and gender of a county leader.

relation between errors within each county, while Column 3-5 cluster the error terms at the prefecture level; Column 5 further adds a set of personal characteristics of county leaders into the model. In all models, the first instrumental variable $\mathit{Avg.\ Position}_{-i,t-1}$ is significantly and negatively correlated to $\mathit{Position}_{i,t-1}$, while the second instrumental variable, $\mathit{Position}_{i,t-2}$, is significantly and positively correlated to $\mathit{Position}_{i,t-1}$.

Table 2.6: Second stage regressions.

	(1) iid	(2) County	(3) Prefecture	(4) Prefecture	(5) Prefecture	(6) IV1	(7) IV2	(8) FE
L.Position	-0.0743**	-0.0743	-0.0768	-0.0699	-0.0780***	-0.461**	-0.0708***	-0.108***
	(0.0337)	(0.0470)	(0.0556)	(0.0471)	(0.0278)	(0.208)	(0.0249)	(0.0236)
Precipitation				-0.000447*	-0.000787*	-0.000977**	-0.000815*	-0.000860**
				(0.000251)	(0.000450)	(0.000439)	(0.000437)	(0.000416)
Nighttime Light				-0.0138	-0.0743***	-0.0740***	-0.0737***	-0.0746***
				(0.0333)	(0.0164)	(0.0205)	(0.0163)	(0.0199)
FE: County	V	V	V	V	V	V	V	V
FE: Time	V	V						
FE: Prefecture x Time			V	V	V	V	V	V
Controls					V	V	V	V
Observations	33664	33664	33619	21639	4169	4252	4181	4265
Adjusted R^2	0.305	0.305	0.324	0.336	0.273	0.269	0.273	0.274

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the results from second stage regressions. The dependent variable is $\mathit{Protest}_{i,t}$, or the number of protests in a month in a county; the key independent variable is the predicted value of $\mathit{Position}_{i,t-1}$, or the number of positions a county leader holds in a month from first stage regressions. Column 1-5 use both instrumental variables, $\mathit{Avg.\ Position}_{-i,t-1}$ and $\mathit{Position}_{i,t-2}$, while Column 6 and 7 employ each instrumental variable separately. Column 8 is the result from the baseline regression for reference. All models include county fixed effects and a specific time trend for each prefecture.

Table 2.6 presents the results from second stage regressions. The dependent variable is $\mathit{Protest}_{i,t}$, and the key independent variable is the predicted value of $\mathit{Position}_{i,t-1}$ from first stage regressions. Column 1 assumes the error terms to be independent and identically distributed. The coefficient of interest is statistically significant and its magnitude is slightly larger than the one in Column 2 in Table 2.4. Column 2 relaxes the i.i.d. assumption and allows for arbitrary correlation between errors within each county, while Column 3 and 4 cluster the error terms at the prefecture level and incorporate a specific time trend for each prefecture into the model. After relaxing the i.i.d. assumption, the standard errors becomes

larger and the coefficients of interest fall just short of statistical significance. Column 5 further adds a set of personal characteristics of county leaders into the model, and the coefficient of interest becomes highly significant at the p-value of 0.006. Column 6 and 7 use only the first and second instrumental variable, respectively, as a robustness check. The coefficient of interest in Column 6 becomes 5.9 times larger, but the coefficients of interest in Column 6 and 7 are statistically significant as well. Column 8 copies the result from Column 6 in Table 2.4 for comparison. Overall, the results are consistent with the results from baseline regressions–power structure has a negative effect on the frequency of protests.

In sum, the above results show that, on average, if the number of positions held by a county leader increases by one standard deviation, the number of protests decreases by 0.0039-0.0044, equivalent to 1.14%-1.28% of the average number of protests in a county. Given the discrete nature of positions, the results indicate that if a county leader obtains one more position, the frequency of protests in the county drops by 20.3% - 22.8%. In other words, more power or more concentrated power structure leads to fewer protests in a county, which is consistent with the prediction from political opportunity theory.

2.4.3 Robustness Checks

In this section, I first present some supportive evidence of the exclusion restriction assumption on the instrumental variable, $\textit{Avg. Position}_{-i,t-1}$. Then, I identify the sub-sample of regional leaders who also serve as county magistrates. I find that when a county leader holds the two most important positions in the county, county party secretary and county magistrate, the effects of power on protests are stronger.

2.4.3.1 Exclusion Restriction

To test the exclusion restriction assumption, I examine the direct effects of the instrumental variable, $\textit{Avg. Position}_{-i,t-1}$, on the frequency of protests. Table 2.7 shows that, except the OLS model in Column 1, the coefficients on $\textit{Avg. Position}_{-i,t-1}$ are insignificant in all other models, while the key independent variable, $\textit{Position}_{i,t-1}$, remains significant and the magnitude of coefficients is comparable to the baseline results. The results suggest that the in-

strumental variable, *Avg. Position*$_{-i,t-1}$, is not directly related to the frequency of protests, supporting the exclusion restriction assumption.

Table 2.7: Exclusion restriction.

	(1) iid	(2) iid	(3) County	(4) Prefecture	(5) Prefecture	(6) Prefecture
L.Position	0.0859***	-0.0720**	-0.0720*	-0.129***	-0.0884**	-0.0834***
	(0.00611)	(0.0281)	(0.0375)	(0.0446)	(0.0344)	(0.0202)
L.Avg. Position$_{-i}$	0.0946***	0.0136	0.0136	-0.119	0.00731	0.386
	(0.00955)	(0.0288)	(0.0315)	(0.0817)	(0.156)	(0.271)
Precipitation					-0.000534**	-0.000855**
					(0.000255)	(0.000427)
Nighttime Light					-0.0118	-0.0767***
					(0.0368)	(0.0191)
FE: County		V	V	V	V	V
FE: Time		V	V			
FE: Prefecture x Time				V	V	V
Controls						V
Observations	34256	34234	34234	34232	22007	4252
Adjusted R^2	0.012	0.307	0.307	0.325	0.337	0.274

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table examines the exclusion restriction assumption of the instrumental variable, *Avg. Position*$_{-i,t-1}$. The dependent variable is *Protest*$_{i,t}$, the number of protests in a month in a county, and the key independent variables are *Position*$_{i,t-1}$, the number of positions a county leader holds in a month, and *Avg. Position*$_{-i,t-1}$, the average number of positions held by other county leaders in the same prefecture. Column 1 includes only the main explanatory variables. Column 2 adds county and period fixed effects into the model. Column 3 relaxes the i.i.d. assumption and allows the error terms to be correlated within each county. Column 4 clusters the error terms at the prefecture level and incorporates a specific time trend for each prefecture. Column 5 adds two control variables, precipitation and nighttime light. Column 6 further includes other control variables, such as age, ethnicity, and gender of a county leader.

Furthermore, I examine whether *Avg. Protest*$_{-i}$, the average number of protests in other counties within the same prefecture, has a direct impact on the frequency of protests. As mentioned previously, due to strict censorship, small, transitory, regional protests can hardly be contagious and influence protests in other counties. Table 2.8 confirms the conjec-

ture: except Column 1, $Avg.\ Protest_{-i}$ does not have an impact on the frequency of protests, and the coefficients on the key independent variable, $Position_{i,t-1}$, are significant in all models.

Table 2.8: No spillover effefcts.

	(1) L0	(2) L1	(3) L3	(4) L6	(5) L0	(6) L1	(7) L3	(8) L6
L.Position	0.00888	-0.0900**	-0.0908**	-0.0907**	-0.0636*	-0.109***	-0.109***	-0.109***
	(0.0203)	(0.0364)	(0.0366)	(0.0362)	(0.0342)	(0.0237)	(0.0242)	(0.0236)
Avg.Protest$_{-i}$	-3.326**	0.0532	0.0969	0.0843	-1.080	0.0309	0.00475	-0.0116
	(1.593)	(0.0667)	(0.117)	(0.0745)	(1.007)	(0.0531)	(0.0655)	(0.0408)
Precipitation	-0.000176	-0.000545**	-0.000535**	-0.000552**	-0.000918***	-0.000885**	-0.000886**	-0.000916**
	(0.000124)	(0.000252)	(0.000253)	(0.000247)	(0.000337)	(0.000410)	(0.000398)	(0.000420)
Nighttime Light	-0.0659**	-0.0119	-0.0118	-0.0119	-0.0866***	-0.0754***	-0.0740***	-0.0757***
	(0.0271)	(0.0370)	(0.0370)	(0.0366)	(0.0117)	(0.0193)	(0.0205)	(0.0191)
FE: County	V	V	V	V	V	V	V	V
FE: Prefecture x Time	V	V	V	V	V	V	V	V
Controls					V	V	V	V
Observations	22016	22016	21990	21891	4261	4261	4259	4244
Adjusted R^2	0.577	0.337	0.338	0.337	0.325	0.274	0.274	0.276

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table examines the spillover effects of protests in the same prefecture. The dependent variable is $Protest_{i,t}$, the number of protests in a month in a county, and the key independent variables are $Position_{i,t-1}$, the number of positions a county leader holds in a month, and several lagged periods of $Avg.\ Protest_{-i}$, the average number of protests in other counties within in the same prefecture. Column 1-4 include the main independent variables, precipitation, nighttime light, county and period fixed effects, and a specific time trend for each prefecture, clustering the error terms at the prefecture level; Column 5-8 further add other control variables, such as age, ethnicity, and gender of a county leader.

2.4.3.2 County Magistrate

The key independent variable is the number of positions held by a county leader, and the positions include county party secretaries, county magistrates, heads of a governmental bureau, party school principals, etc. Nevertheless, each position represents a different level of power: a party school principal can hardly possess the same level of power as a county party secretary, and counting the number of positions without attaching any "power weight" to each position can be biased.

To address the issue, I identify a sub-sample of county leaders who also serve as county magistrates, the second powerful position, in the same county. If power indeed affects protests, then the effects should be stronger if a county leader holds the two most powerful positions at the same time. Of 35,639 county-period observations, county leaders hold both county party secretaries and county magistrates in only 184 observations, which translates to 0.5% of the observations and less than 8 county leaders each period on average. I then interact the indicator of the sub-sample with the key independent variable, $Position_{i,t-1}$. If the hypothesis is true and power does negatively influence protests, the coefficient on the interaction term should be negative and significant. Table 2.9 confirms the hypothesis.[1] Column 1-4 show that power has negative effects on protests and the effects are stronger if a county party secretary serves as a county magistrate at the same time. Column 5 reports the result from Column 5 in Table 2.4 for comparison.

2.5 Mechanism

In this section, I discuss the underlying mechanism of power influence on protests. I first disentangle the two essential components of protests–grievances and collective action–by focusing on environmental protests and using the level of pollution as a control for the root of grievances. Then, I document different patterns of power on protests between leaders from the locality and leaders from other regions by examining pollution, protest causes, and repression.

2.5.1 Grievances vs. Collective Action

Protests comprise two major elements–grievances and collective action, and county leaders can influence protests through both elements. Classical collective behavior theories claim that protests stem from indignation at injustice, and the root of grievances may relate to a certain policy and reveal public discontent with the government. For example, land protests arise from enforced land acquisition, while environmental protests reflect pub-

[1] I do not present the results from 2SLS regressions because the interaction term is dropped out due to stricter data requirement compared to fixed effects models.

Table 2.9: Dual jobs: county party secretary and county magistrate.

	(1) iid	(2) County	(3) Prefecture	(4) Prefecture	(5) Prefecture
L.Position	-0.0630**	-0.0630*	-0.0973**	-0.0893**	-0.0906**
	(0.0282)	(0.0368)	(0.0481)	(0.0361)	(0.0361)
L.Position × L.County Magistrate	-0.996	-0.996***	-0.703***	-0.918***	
	(1.091)	(0.0445)	(0.0481)	(0.0458)	
Precipitation				-0.000569**	-0.000569**
				(0.000238)	(0.000238)
Nighttime Light				-0.0120	-0.0119
				(0.0368)	(0.0368)
FE: County	V	V	V	V	V
FE: Time	V	V			
FE: Prefecture x Time			V	V	V
Observations	35114	35114	34281	22033	22033
Adjusted R^2	0.316	0.316	0.325	0.337	0.337

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the additional effects of power structure on protests if a county leader holds the two most important jobs in a county, county party secretary and county magistrate. The dependent variable is $Protest_{i,t}$, the number of protests in a month in a county, and the key independent variables are $Position_{i,t-1}$, the number of positions a county leader holds in a month, and the interaction of $Position_{i,t-1}$ and $County\ Magistrate_{i,t-1}$ if a county leader serves as a county magistrate as well. Column 1 includes the main explanatory variables, county and period fixed effects. Column 2 relaxes the i.i.d. assumption and allows the error terms to be correlated within each county. Column 3 clusters the error terms at the prefecture level and incorporates a specific time trend for each prefecture. Column 4 adds two control variables, precipitation and nighttime light. Column 5 reports the result from Column 5 in Table 2.4 for comparison.

lic dissatisfaction about weak regulation and governance on polluting industries. Under a regionally decentralized system (Xu, 2011), a county party secretary can establish regional policies and directly influence the well-being of people, and the degree of influence depends on power structure. On the one hand, the more power a county leader wields within the government, the more likely that s/he can use his/her clout to influence policies and the welfare in the county. On the other hand, power concentration reflects political competition in a county, which relates to the quality of a politician, social welfare, and economic performance (Ashworth, Geys, Heyndels, & Wille, 2014; Besley, Persson, & Sturm, 2010). In places with lower political competition, reduced welfare and more grievances are anticipated. Moreover, if power is highly concentrated, there is little separation of powers for checks and balances among different departments and bureaus in the government, considering the absence of judicial independence and opposition parties in China.

In contrast to classical collective behavior theories, resource mobilization theory argues that, to form a collective action, aggrieved people need resources and only take action when the expected payoff of a protest dominates the associated costs. People are more likely to participate in a protest when they perceive a higher chance of success, which hinges on state strength, repression level, and elite division (Cragun Cragun, D., 2006; Klandermans & van Stekelenburg, 2013). Particularly, a political opportunity is created when power is divided among elites. Aggrieved citizens seize the opportunity and find powerful allies to help promote their claims (Klandermans & van Stekelenburg, 2013; O'Brien, 2013; O'Brien & Li, 2006). If a county leader holds many public positions and dominates the region, the probability of finding a strong advocate is slim and citizens are less prone to challenge the leader. In the context of China, an authoritarian regime with poor human rights record (Amnesty International, 2017; Human Rights Watch, 2016), the risk of participating a protest is phenomenal. When aggrieved citizens disagree with current policies and strive to voice their claims, they tend to seek support from someone within the party organizations or the government. Without support from the powerful, citizens are much less likely to protest given the high risk of participating a protest in China.

To separate power's influence on the two components, I focus on environmental protests and use one of the most prominent air pollution indicators, particulate matter 2.5 micron diameter $\mu g/m_3$ (PM 2.5), as a proxy for the root of grievances. Figure 2.14 displays the level

of PM 2.5 over time. The average level of PM 2.5 is 118.21 $\mu g/m_3$ in the data set, which is far beyond the limits in the European Union (25 $\mu g/m_3$), the United States (12 $\mu g/m_3$), and even China (35 $\mu g/m_3$) (Environment of European Commission, 2016; Ministry of Environmental Protection (China), 2012; US EPA, 2015).

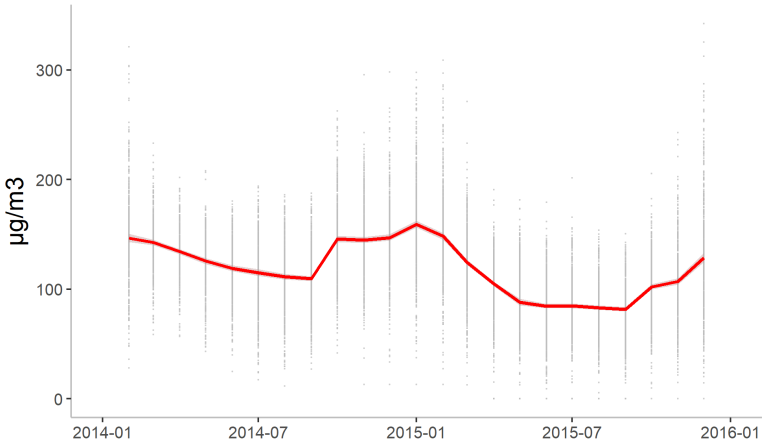


Figure 2.14: The level of PM 2.5 over time.

The regression model is shown as follows:

$$
\begin{aligned}
&\textit{Environmental Protest}_{i,t} \\
&= \beta_1 \textit{Position}_{i,t-1} + \beta_2 \textit{PM 2.5}_{i,t-6} + \beta_3 \textit{Position}_{i,t-1} \times \textit{PM 2.5}_{i,t-6} \\
&+ \beta_4 \mathrm{X}_{i,t} + \mathrm{County}_i + \mathrm{Period}_t + \mathrm{Prefecture}_p \times \mathrm{Period}_t + \varepsilon_{i,t}
\end{aligned}
$$

The dependent variable, $\textit{Environmental Protest}_{i,t}$, is an indicator of whether there is an environmental protest in county i at time t. The key independent variables are $\textit{Position}_{i,t-1}$, $\textit{PM 2.5}_{i,t-6}$, and the interaction term. I use 6-month lag of PM 2.5 because, presumably, air pollution is a long-term phenomenon that plagues residents for months or even years, and people rarely protest based on air pollution in the current month.[1] $\mathrm{X}_{i,t}$ includes precipitation and nighttime light in county i at time t[2]. In all specifications, county, period,

[1] I also use 3-month lag and 12-month lag of PM 2.5 as a robustness check, and the results are consistent with 6-month lag of PM 2.5.

[2] $\mathrm{X}_{i,t}$ does not include county leaders' characteristics because sample sizes are much smaller than the ones in the baseline models. The drop in sample size comes from the inclusion of PM 2.5.

and prefecture x period fixed effects are included, and the error terms are clustered at the prefecture level.

Table 2.10: The effects of power and air pollution on environmental protests.

	(1) FE	(2) 2SLS	(3) FE	(4) 2SLS
L.Position	-0.0789*	-0.131	-0.0561	-0.107
	(0.0437)	(0.0802)	(0.0445)	(0.0780)
L6.PM 2.5	0.000365**	0.000315**	0.000639***	0.000607**
	(0.000151)	(0.000150)	(0.000214)	(0.000215)
L.Position × L6.PM 2.5			-0.000177*	-0.000186*
			(0.0000973)	(0.000101)
Nighttime Light	-0.00272	-0.00259	-0.00243	-0.00229
	(0.00277)	(0.00328)	(0.00271)	(0.00319)
Precipitation	0.00000511	-0.0000414	0.00000859	-0.0000407
	(0.0000788)	(0.0000815)	(0.0000795)	(0.0000817)
FE: County	V	V	V	V
FE: Prefecture x Time	V	V	V	V
Observations	1030	995	1030	995
Adjusted R^2	-0.025	-0.018	-0.024	-0.017

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the effects of power and air pollution on the environmental protests. The dependent variable is $Environmental\ Protest_{i,t}$, the frequency of environmental protests in a month in a county, and the key independent variables are $Position_{i,t-1}$ and $PM\ 2.5_{i,t-6}$, the level of air pollution 6 months ago. Column 1 and 2 include the two key independent variables, while Column 3 and 4 add the interaction term.

Table 2.10 displays the effects of power and PM 2.5 on environmental protests. In Column 1 and 2, 6-month lag of PM 2.5 is positively correlated to the probability of environmental protests. In other words, more air pollution triggers more environmental protests. The coefficients of power are negative and the magnitude is comparable to the baseline models. Column 3 and 4 incorporate the interaction term of PM 2.5 and power. The coeffi-

cients of the interaction term are negative and marginally significant, suggesting that given the same level of grievances originated from air pollution, if a county leader wields a lot of power in the government, people are less likely to participate in environmental protests, which is consistent with the political opportunity theory. The coefficients of PM 2.5 stay positive and significant; the coefficients of power remain negative but become insignificant, implying that power affects protests mainly through the interaction term.

The result in Column 4 can be interpreted as follows: given the average level of PM 2.5, if the number of public positions held by a county leader increases by one standard deviation, the number of environmental protests decreases by 0.0012, translating to 20.28% of the average number of environmental protests in a county. The coefficients are both statistically and economically significant, and the results further confirm the hypothesis that power has substantial effects on protests by discouraging the formation of collective action.

2.5.2 Local vs. Outsider

In this section, I separate the sample into two subgroups–"local" and "outsider". "Local" is defined as a county leader whose ancestral home[1] is located in the same prefecture as his/her ruling county is, whereas "outsider" consists of all non-local leaders.[2] Figure 2.15 shows the geographical distribution of local leaders. Overall, there are 17,796 local observations, which accounts for 27.6% of the data set.

Table 2.11 shows the effects of power on protests for the two subgroups, "local" and "outsider". Column 1-4 present the results using the "local" sample, while Column 5-8 display the results using the "outsider" sample. In contrast to the baseline results using the full sample, power does not have negative effects on protests in counties ruled by local leaders. In other words, if a local leader becomes more powerful, people are not less likely to protest. Since power always negatively relates to protests by lowering the success chance of protests and reducing people's willingness to protest, the positive coefficients indicates that more power or more concentrated power structure generates a higher level of grievances and leads to more protests.

[1] An ancestral home refers to the place that a person's ancestors, usually parents, originally come from.
[2] "Outsider" also includes leaders with missing values on their ancestral home places.

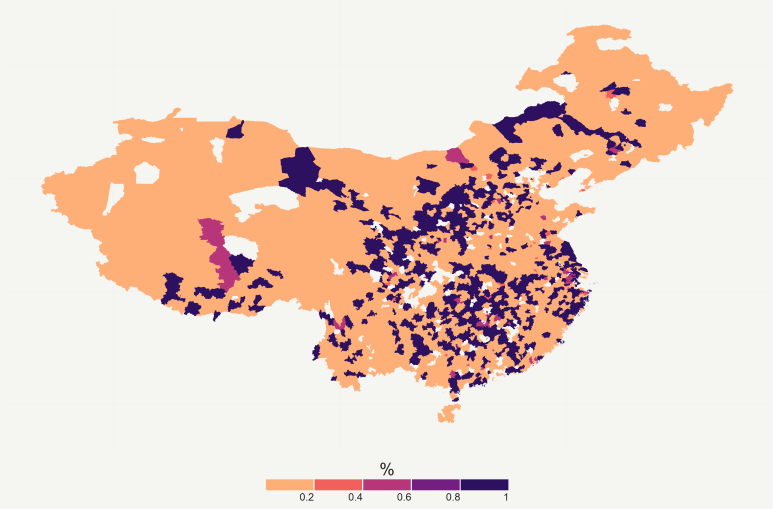


Figure 2.15: The geographical distribtion of local leaders.

Table 2.11: The effects of power structure on protest frequency by group.

	Local				Outsider			
	(1) FE	(2) IV1+2	(3) IV1	(4) IV2	(5) FE	(6) IV1+2	(7) IV1	(8) IV2
L.Position	0.0684	-0.0465	-0.595	0.166***	-0.110***	-0.0543***	-0.319**	-0.0517***
	(0.0619)	(0.186)	(0.670)	(0.0327)	(0.0235)	(0.00833)	(0.150)	(0.00928)
Precipitation	-0.00141	-0.00145	-0.00162	-0.00138	-0.000529*	-0.000378	-0.000604**	-0.000424
	(0.000987)	(0.000995)	(0.000985)	(0.001000)	(0.000260)	(0.000264)	(0.000289)	(0.000258)
Nighttime Light	0.181***	0.175***	0.184***	0.175***	-0.0924***	-0.0895***	-0.0923***	-0.0888***
	(0.0564)	(0.0530)	(0.0556)	(0.0533)	(0.00796)	(0.00915)	(0.00831)	(0.00912)
FE: County	V	V	V	V	V	V	V	V
FE: Prefecture x Time	V	V	V	V	V	V	V	V
Controls	V	V	V	V	V	V	V	V
Observations	1513	1476	1513	1476	1775	1721	1762	1733
Adjusted R^2	0.159	0.136	0.152	0.136	0.290	0.295	0.287	0.294

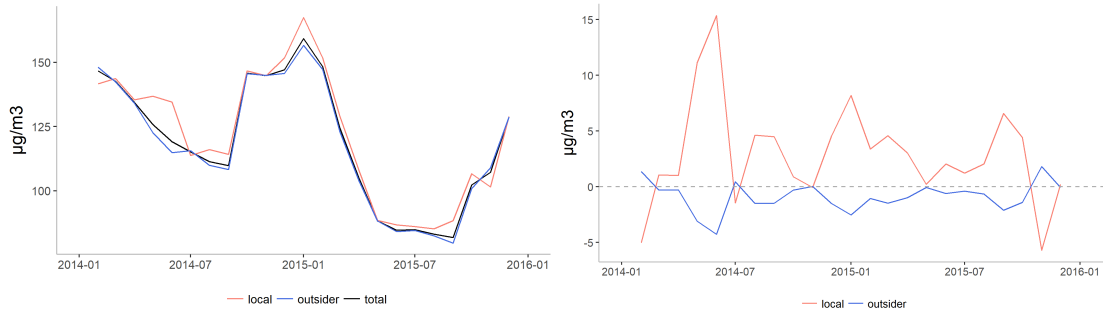
Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the effects of power on protests using "local" and "outsider" samples. A leader is classified as local if his/her hometown is located in the same prefecture as his/her ruling county. The dependent variable is $Protest_{i,t}$, the frequency of protests in a month in a county, and the key independent variable is $Position_{i,t-1}$ as a proxy of power. Column 1-4 use the local sample, while Column 5-8 present the results from the outsider sample. All models include county fixed effects, a specific time trend for each prefecture, and control variables.

2.5.2.1 Air Pollution

To corroborate the hypothesis that local leaders arouse more public indignation when they wield more power, I examine the effects of power on the root of grievances, which is proxied by the air pollution indicator, PM 2.5. Figure 2.16 shows the average level of PM 2.5 for the local/outsider sample and its deviation from the average level of the entire sample over time. Figure 2.16a shows that the average level of PM 2.5 for the local and outsider groups exhibit the same pattern, and the level of PM 2.5 is relatively high in winter with peaks around January. More importantly, Figure 2.16b illustrates that the level of PM 2.5 is above average in places under the rule of local leaders, while the level of PM 2.5 for the outsider sample is below average for most of the time.



(a) The average level.

(b) The deviation from the average.

Figure 2.16: PM 2.5 for local leaders and outsiders.

Table 2.12 displays the regression results. Column 1 and 2 feature the results for the local sample. The results show that the more power a local leader holds, the condition of air pollution in the county becomes more serious. In contrast to the local sample, the outsider sample depicts a different story–in counties rule by outsiders, the effects of power on air pollution is negative. In Table 2.12, I employ different indicators of air pollution as a robustness check. The results are consistent with Table 2.12 and confirm that under the rule of a local leader, the more power s/he gains, there are higher chances for the county people to suffer from air pollution.

Overall, the results from Table 2.12 suggest that in counties ruled by local leaders, an increase in public positions by one standard deviation leads to a surge of PM 2.5 by 1.23-1.75

Table 2.12: The effects of power structure on air pollution by group.

	Local		Outsider		All	
	(1) FE	(2) 2SLS	(3) FE	(4) 2SLS	(5) FE	(6) 2SLS
L.Position	21.83*	24.91***	-8.176	-33.17***	-5.070	-22.93
	(10.86)	(6.788)	(8.557)	(9.465)	(7.148)	(14.00)
FE: County	V	V	V	V	V	V
FE: Prefecture x Time	V	V	V	V	V	V
Observations	1105	1073	2867	2788	4596	4483
Adjusted R^2	0.880	0.881	0.830	0.828	0.845	0.845

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the effects of power on the level of air pollution using "local" and "outsider" samples. A leader is classified as local if his/her hometown is located in the same prefecture as his/her ruling county. The dependent variable is $PM\ 2.5_{i,t}$ as a proxy of air pollution in a month in a county, and the key independent variable is $Position_{i,t-1}$ as a proxy of power. Column 1 and 2 use the local sample, while Column 3 and 4 present the results from the outsider sample; Column 5 and 6 employ the whole sample for reference. All models include county fixed effects and a specific time trend for each prefecture.

Table 2.13: The effects of power structure on air pollution by group.

	Local				Outsider			
	(1) NO_2	(2) NO_2	(3) O_3	(4) O_3	(5) NO_2	(6) NO_2	(7) O_3	(8) O_3
L.Position	3.654***	5.430***	4.602	12.71***	-0.826	-5.954***	-4.294	-6.253
	(0.959)	(1.941)	(5.167)	(3.127)	(1.723)	(1.832)	(2.775)	(6.288)
FE: County	V	V	V	V	V	V	V	V
FE: Prefecture x Time	V	V	V	V	V	V	V	V
Observations	1103	1071	1105	1073	2878	2798	2785	2706
Adjusted R^2	0.807	0.804	0.758	0.764	0.790	0.788	0.702	0.701

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

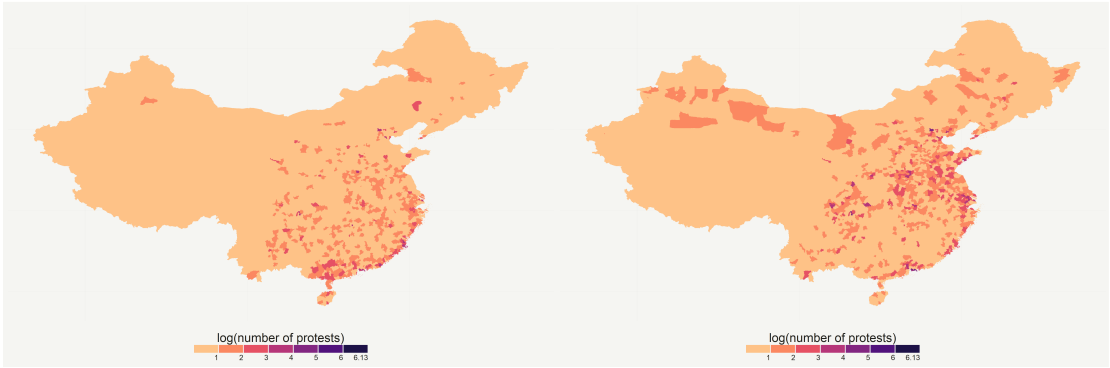
Note: The table displays the effects of power on the level of air pollution using "local" and "outsider" samples. A leader is classified as local if his/her hometown is located in the same prefecture as his/her ruling county. The dependent variables are $NO2_{i,t}$ and $O3_{i,t}$ as indicators of air pollution in a month in a county, and the key independent variable is $Position_{i,t-1}$ as a proxy of power. Column 1-4 use the local sample, while Column 5-8 present the results from the outsider sample. Column 1, 3, 5, and 7 employ fixed effects models, while Column 2, 4, 6, 8 adopt an instrumental variable approach. All models include county fixed effects and a specific time trend for each prefecture.

on average, translating to 1.04%-1.19% of the average level of air pollution in China. The empirical pattern indicates that the influence of power on grievances differ between locals and outsiders, which results in the discrepancy in the effects of power on protests.

2.5.2.2 Protest against the Government

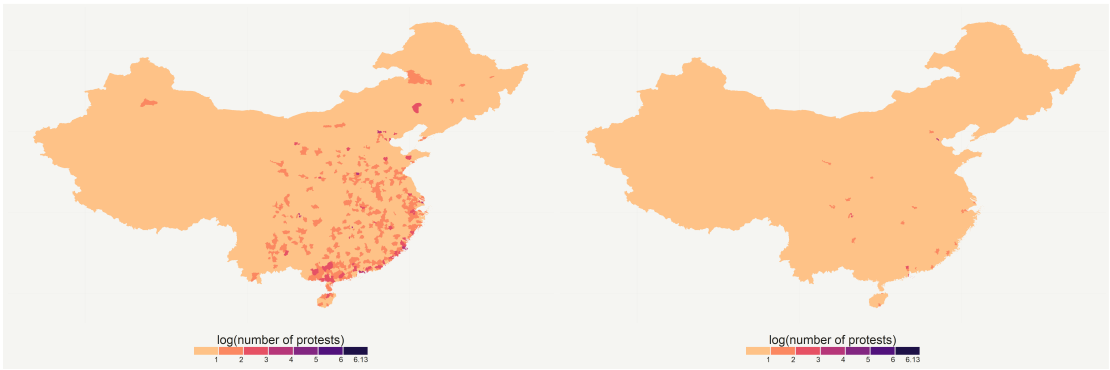
In this section, I explore the causes of protests to identify the root of grievances. More specifically, I classify protests in terms of protest targets. On the one hand, aggrieved citizens may protest against the government because of land acquisition, environmental pollution, dissatisfaction with the government or a state-owned enterprise, etc. On the other hand, people may protest against non-governmental entities. For example, workers assemble to express their anger at private firms and claim their wages in arrears; merchants go on strike to voice their discontent with suppliers about defaulted debts. Given that protesters express their resentment against the government, I further categorize the causes into two types: policy-related protests and policy-unrelated protests. Policy-related protests mostly

contain land protests and environmental protests, while policy-unrelated protests refer to bitter complaints against urban management officers or Chengguan (城管), regional governments, or even the communist party in some rare cases. Figure 2.17 maps the number of protests against the government and non-governmental entities, respectively. Overall, 38% of the protests are against the government, while 61% of the protests are against non-governmental entities. Figure 2.18 further maps the number of government-related protests. Among the protests against the government, 89% of the protests are related to a policy, while 11% of the protests are not policy-related.



(a) Government-related **(b)** Government-unrelated

Figure 2.17: The log number of protests against the government and non-governmental entities, respectively.



(a) Policy-related **(b)** Policy-unrelated

Figure 2.18: The log number of protests related to governmental policies and other governmental issues, respectively.

The model is defined as follows:

$$\Pr(\textit{Protest Type}_{i,t} = 1|\Omega) = \beta_1 \textit{Position}_{i,t-1} + \beta_2 \mathrm{X}_{i,t} + \mathrm{County}_i + \mathrm{Period}_t + \mathrm{Prefecture}_p \times \mathrm{Period}_t + \varepsilon_{i,t}$$

The dependent variable, $\Pr(\textit{Protest Type}_{i,t} = 1|\Omega)$, is an indicator of a protest type given Ω, i.e. one or more protests take place and at least one cause is identified at time t in county i. There are four protest types: whether a protest is against the government or against non-governmental entities, and for the former, whether a protest is related to a policy or not. The indicators are not mutually exclusive, i.e., a protest can be classified as both against the government[1] and against non-governmental entities.[2] The key independent variable is the proxy of power, $\textit{Position}_{i,t-1}$.

Table 2.14 shows the effects of power on protests by different protest targets.[3] Column 1-3 display the effects of power on protests against the government. Column 1 shows that an increase in a local leader's power brings about higher chances of a government-related protest, while Column 2 illustrates that an increase in an outsider's power reduces the likelihood of a government-related protest; Column 4-6 show similar effects of power on protests against non-governmental entities between locals and outsiders. The empirical evidence suggests that given one or more protests take place, people are more likely to protest against the government when local leaders wield a lot of power in the government, which is consistent with the hypothesis that grievances play a crucial role in the distinct pattern in Table 2.11 between locals and outsiders.

[1] State-own enterprises are classified as governmental entities.

[2] Precipitation and repression are not included in the controls for two reasons. First, since I focus on observations with at least one protest, the formation of collective action is not of major concern. The research question here is as follows: given that a protest takes place, how does power affect the probability that this protest contains a certain characteristic? Therefore, whether people protest is not a question anymore. Second, the sample size is often too small to generate plausible results, and the inclusion of precipitation and repression exacerbates the situation.

[3] All models in this section only use fixed effects because the data requirement on instrumental variable approach cannot be fulfilled for small samples. The samples shrinks considerably because I only consider observations with protests and at least one identified cause. In many cases, it is impossible to obtain reasonable results or even successfully run regressions with small samples.

Table 2.14: The effects of power on protests by protest targets.

	Gov			non-Gov		
	(1) Local	(2) Outsider	(3) All	(4) Local	(5) Outsider	(6) All
L.Position	0.501***	-0.116***	-0.115***	-0.499***	-0.0362**	-0.0262
	(0.00304)	(0.0159)	(0.0246)	(0.00209)	(0.0179)	(0.0212)
Nighttime Light	-0.0128	-0.000559	-0.00175	-0.0335	0.000932	-0.000137
	(0.0764)	(0.0125)	(0.0136)	(0.0527)	(0.00670)	(0.00674)
FE: County	V	V	V	V	V	V
FE: Prefecture x Time	V	V	V	V	V	V
Observations	235	656	1192	235	656	1192
Adjusted R^2	0.128	-0.001	0.034	0.043	0.119	0.127

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the effects of power on protests with different protest targets using "local" and "outsider" samples. A leader is classified as local if his/her hometown is located in the same prefecture as his/her ruling county. Each protest is categorized into a protest against the government and/or a protest against non-governmental entities. The dependent variables are $Gov_{i,t}$ and $non\text{-}Gov_{i,t}$, which are indicators of protests against the government or against non-governmental entities. The key independent variable is $Position_{i,t-1}$ as a proxy of power. Column 1 and 4 use the local sample, while Column 2 and 5 present the results from the outsider sample; Column 3 and 6 displays the full sample for reference. All models include county fixed effects and a specific time trend for each prefecture.

Table 2.15 displays the effects of power on protests against the government for policy-related and policy-unrelated issues. Column 1-3 focus on policy-related protests, whereas Column 4-6 show the results of policy-unrelated protests. Column 1 and 4 indicate that given that a county is ruled by a local leader and that at least one protest takes place and a protest cause is identified, people are more likely to protest for a policy-related issue, not policy-unrelated issue, while in counties governed by outsiders the opposite pattern is observed. The evidence suggests that when a local leader gains more power in the government and becomes more involved in regional policies, people are more likely to protest against the government for policy-related issues. This further supports the hypothesis that powerful local leaders generate more grievances, leading to more protest.

Table 2.15: The effects of power on policy-related and policy-unrelated protests.

	Gov: Policy			Gov: non-Policy		
	(1) Local	(2) Outsider	(3) All	(4) Local	(5) Outsider	(6) All
L.Position	1.001***	-0.138***	-0.132***	-0.500***	0.0197***	0.0155***
	(0.00275)	(0.0171)	(0.0286)	(0.000717)	(0.00506)	(0.00579)
Nighttime Light	-0.0289	0.00593	0.00406	0.00738	-0.00913*	-0.00892*
	(0.0692)	(0.00981)	(0.0107)	(0.0180)	(0.00497)	(0.00491)
FE: County	V	V	V	V	V	V
FE: Prefecture x Time	V	V	V	V	V	V
Observations	235	656	1192	235	656	1192
Adjusted R^2	-0.065	-0.024	-0.028	-0.411	-0.154	-0.234

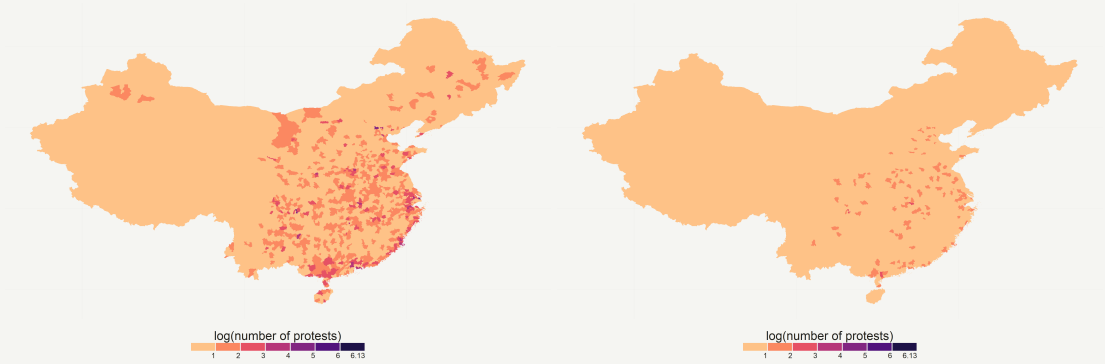
Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the effects of power on policy-related and policy-unrelated protests using "local" and "outsider" samples. A leader is classified as local if his/her hometown is located in the same prefecture as his/her ruling county. Each protest is categorized into a policy-related or/and policy-unrelated protest, given that it is a protest against the government. The dependent variables are $Gov:\ Policy_{i,t}$ and $Gov:\ non\text{-}Policy_{i,t}$, which are indicators of policy-related or policy-unrelated protests. The key independent variable is $Position_{i,t-1}$ as a proxy of power. Column 1 and 4 use the local sample, while Column 2 and 5 present the results from the outsider sample; Column 3 and 6 displays the full sample for reference. All models include county fixed effects and a specific time trend for each prefecture.

2.5.2.3 State Violence

In this section, I examine the effects of power on state violence to find out whether local leaders behave the same as outsiders in terms of repression. If their choices of repression are different, then it signals that local leaders may affect protests through the formation of collective action as well, given that repression deters aggrieved citizens from participating in protests. On the other hand, if the results show similar patterns between locals and outsiders, then it further supports the hypothesis that the discrepancy in the influence of power on protests is originated from grievances. Figure 2.19a maps the number of protests related to state violence. Since there are only 21 protests marked as relating to state violence in the original labeling by *Not the News*, I use the extended labeling instead. Overall, there are 5,246 protests related to state violence, which translate to 21% of the protests.



(a) State Violence **(b)** Casualties

Figure 2.19: The log number of protests related to state violence and casualties, respectively.

Table 2.16 presents the effects of power on the probability of repression, given that one or more protests take place. Column 1 and 2 display the results of local sample, Column 3 and 4 show the results of outsider sample, and Column 5 and 6 exhibit the results of all sample. Despite the effects of power on repression probability become insignificant in 2SLS models for the outsider sample, the coefficients are all negative and consistent with the "all" sample.

In addition, I use an indicator of casualties as the dependent variable, given that there is at least one protest in county i at time t. A protest is identified as involving casualties if key-

words related to injuries, deaths, or hospitals are mentioned, excluding any protest about medical malpractice or natural disasters. Presumably, casualties arise from repression and should exhibit similar patterns as repression. Figure 2.19b maps the number of protests involved casualties. Overall, there are 1,174 protests related to casualties, which translates to 5% of the protests. Table 2.17 presents the effects of power on casualties, given at least one protest happens. The results show negative effects of power on casualties, which is consistent with Table 2.16.

Table 2.16: The effects of power structure on repression by group.

	Local		Outsider		All	
	(1) FE	(2) 2SLS	(3) FE	(4) 2SLS	(5) FE	(6) 2SLS
L.Position	-0.180	-0.410*	-0.0380**	-0.00439	-0.0692***	-0.0524
	(0.176)	(0.214)	(0.0181)	(0.0181)	(0.0251)	(0.0330)
Nighttime Light	0.0410	0.0416	0.0162***	0.0128**	0.0173***	0.0135**
	(0.0377)	(0.0366)	(0.00614)	(0.00587)	(0.00657)	(0.00610)
FE: County	V	V	V	V	V	V
FE: Prefecture x Time	V	V	V	V	V	V
Observations	939	915	1993	1931	3969	3851
Adjusted R^2	-0.023	-0.023	-0.006	-0.010	0.006	0.003

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the effects of power on repression using "local" and "outsider" samples. A leader is classified as local if his/her hometown is located in the same prefecture as his/her ruling county. The dependent variable is $Repression_{i,t}$, an indicator of whether a protest is repressed, given at least one protest takes place. The key independent variable is $Position_{i,t-1}$ as a proxy of power. Column 1 and 2 use the local sample, while Column 3 and 4 present the results from the outsider sample; Column 5 and 6 employ the whole sample for reference. All models include county fixed effects and a specific time trend for each prefecture.

In general, Table 2.16 and Table 2.17 provide suggestive evidence that power lowers the probability of state violence. More importantly, local and outsider samples present similar empirical patterns of state violence, implying that the different effects of power on protests do not result from state violence. The evidence is consistent with the hypothesis that the discrepancy emerges from grievances, not the formation of collective action.

Table 2.17: The effects of power structure on casualties by group.

	Local		Outsider		All	
	(1) FE	(2) 2SLS	(3) FE	(4) 2SLS	(5) FE	(6) 2SLS
L.Position	-0.0341	-0.0133	-0.0550***	-0.0550**	-0.0622***	-0.0697***
	(0.182)	(0.121)	(0.0183)	(0.0259)	(0.0197)	(0.0261)
Nighttime Light	-0.00260	-0.00609	-0.00906*	-0.00917	-0.00657	-0.00686
	(0.0176)	(0.0161)	(0.00491)	(0.00555)	(0.00622)	(0.00665)
FE: County	V	V	V	V	V	V
FE: Prefecture x Time	V	V	V	V	V	V
Observations	851	831	1845	1788	3653	3544
Adjusted R^2	-0.016	-0.014	-0.082	-0.089	-0.011	-0.018

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the effects of power on casualties using "local" and "outsider" samples. A leader is classified as local if his/her hometown is located in the same prefecture as his/her ruling county. The dependent variable is $Casualties_{i,t}$, an indicator of whether there are casualties in a protest, given at least one protest takes place. The key independent variable is $Position_{i,t-1}$ as a proxy of power. Column 1 and 2 use the local sample, while Column 3 and 4 present the results from the outsider sample; Column 5 and 6 employ the whole sample for reference. All models include county fixed effects and a specific time trend for each prefecture.

2.6 Who Has More Power?

Section 2.5 demonstrates that local leaders arouse more social grievances and spark more protests when they wield a lot of power in the government. To investigate the phenomenon, I first need to answer the question: who has more power? In the following model, I examine the correlation between power and various characteristics of county leaders:

$$\begin{aligned} \textit{Position}_{i,t} = {} & \beta_1 \textit{Local}_i + \beta_2 \textit{Nighttime Light}_{i,t} + \beta_3 \text{Prefecture Experience}_i \\ & + \beta_4 \mathrm{X}_i + \beta_5 \mathrm{Z}_{i,t} \\ & + \text{Prefecture}_p + \text{Period}_t + \text{Prefecture}_p \times \text{Period}_t + \varepsilon_{i,t} \end{aligned}$$

The dependent variable is $\textit{Position}_{i,t}$, which can be understood as a measure of "concurrent job promotions". In contrast to the literature which studies promotion only after a politician leaves his/her current position, $\textit{Position}_{i,t}$ assesses the power granted to a politician when s/he serves as a county party secretary throughout the whole time, and a change in $\textit{Position}_{i,t}$ indicates a minor promotion or demotion within the term of office of a county leader. Two essential independent variables are $\textit{Nighttime Light}_{i,t}$, a proxy of regional GDP growth, and Prefecture Experience$_i$, a proxy of connections to superior authorities. In particular, Prefecture Experience$_i$ consists of three indicators–if a county leader has worked in a prefecture before his/her term of office ($\textit{Previous Prefecture Experience}_i$), is working in a prefecture during the term ($\textit{Current Prefecture Experience}_i$), and will work after the term ($\textit{Future Prefecture Experience}_i$). The most relevant one is $\textit{Previous Prefecture Experience}_i$, which is more likely to determine the promotion probability of a county leader and more comparable to the literature (T. Chen & Kung, 2016). Since each county leader is evaluated by the prefecture government, previous prefecture work experience helps build connections to superior authorities and elevates the likelihood of promotion. X_i includes time-invariant characteristics, such as gender, education, ethnicity, hometown, etc.; $\mathrm{Z}_{i,t}$ consists of time-varying characteristics: age and duration of the term.

Table 2.18 displays the empirical results. Column 1 uses a simple OLS model, while Column 2-4 add prefecture, period, and prefecture x period fixed effects, respectively. In all specified models, $\textit{Nighttime Light}_{i,t}$ and $\textit{Previous Prefecture Experience}_i$ are significantly

and positively correlated to $Position_{i,t}$, suggesting that higher economic growth and ties to superior authorities contribute to more positions within the term of office of a county leader.[1] Table 2.19 examines the local and outsider samples separately, and $Nighttime\ Light_{i,t}$ and $Previous\ Prefecture\ Experience_i$ still positively predict the likelihood of promotion. Despite using a different definition of promotion, the empirical results are consistent with the literature and show that "concurrent job promotions" are also based on meritocracy and patronage.

Although economic performance and workplace connections are both correlated to promotion, I find that the two factors exhibit different patterns for locals and outsiders. Table 2.20 and Table 2.21 show that being a local predicts lower levels of $Nighttime\ Light_{i,t}$ and higher chances of previous work experience in the prefecture government directly superior to his/her ruling county. In other words, compared to outsiders, local party secretaries are more likely to have work experience under the direct superior authorities before their terms in the office, but tend to fail in economic performance. If economic performance is a proxy for competence and workplace connections represent ties with superior authorities (T. Chen & Kung, 2016; Persson & Zhuravskaya, 2016), the empirical evidence implies that local leaders are less competent but have better connections with superior authorities.

2.7 Conclusion and Discussion

This chapter aims to provide empirical evidence on the relationship between protest and power structure among regional leaders in China. I first use text mining to compile a comprehensive data set on the information of over 50,000 protests, and collect biographical data of 2,714 county leaders and construct a measure to proxy for power, *Position*, or the number of public positions held by a county leader at the same time, from the curriculum vitae of county leaders. Next, I demonstrate the negative causal effects of regional power structure on protests. To overcome the endogeneity issue, I employ fixed effects models with two instruments. The first instrumental variable is the average number of positions held by other county leaders in the same prefecture, *Average Position*, which can be inter-

[1] Additionally, Column 2-4 present the positive effects of age, being a female, and higher education on $Position_{i,t}$; interestingly, the results indicate that being a local negatively correlates to $Position_{i,t}$.

Table 2.18: The determinants of power.

	(1) OLS	(2) FE	(3) FE	(4) FE
Nighttime Light	0.212***	0.215***	0.219***	0.238***
	(0.00636)	(0.00641)	(0.00629)	(0.00620)
Duration	-0.00256***	0.000305	0.000671	0.000733
	(0.000639)	(0.000637)	(0.000785)	(0.000949)
Age	0.00457*	0.00972***	0.00977***	0.0115***
	(0.00278)	(0.00242)	(0.00241)	(0.00291)
Female	0.372***	0.303***	0.304***	0.305***
	(0.0486)	(0.0497)	(0.0497)	(0.0585)
Higher Education	-0.104***	0.823***	0.835***	0.826***
	(0.0395)	(0.0912)	(0.0910)	(0.111)
Local	-0.164***	-0.171***	-0.172***	-0.163***
	(0.0197)	(0.0200)	(0.0199)	(0.0238)
Prefecture Work Experience				
Previous	0.390***	0.334***	0.333***	0.343***
	(0.0163)	(0.0273)	(0.0275)	(0.0334)
Current	0.561***	0.557***	0.553***	0.554***
	(0.0372)	(0.0342)	(0.0348)	(0.0413)
Future	0.0696***	-0.206***	-0.209***	-0.196***
	(0.0204)	(0.0257)	(0.0253)	(0.0296)
FE: Prefecture		V	V	
FE: Time			V	
FE: Prefecture x Time				V
Ethnicity	V	V	V	V
Ancestral Place	V	V	V	V
Birthplace	V	V	V	V
Observations	5592	5589	5589	3985
Adjusted R^2	0.308	0.761	0.761	0.464

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the determinants of power. The dependent variable is $Position_{i,t}$, a proxy of power. The key independent variables are $Nighttime\ Light_{i,t}$ and $Previous\ Prefecture\ Experience_i$, which represent economic performance and connections to superior authorities, respectively. Column 1 uses a simple OLS model, while Column 2-4 add prefecture, period, and prefecture x period fixed effects, respectively.

Table 2.19: The determinants of power by group.

	Local		Outsider		All	
	(1) FE	(2) FE	(3) FE	(4) FE	(5) FE	(6) FE
Nighttime Light	0.218***	0.236***	0.0318***	0.0331***	0.0426***	0.0441***
	(0.00575)	(0.00553)	(0.00416)	(0.00507)	(0.00467)	(0.00567)
Duration	-0.00393***	-0.00391***	0.00393***	0.00419***	0.00110	0.00136
	(0.000623)	(0.000750)	(0.00138)	(0.00161)	(0.00106)	(0.00128)
Age	0.0256***	0.0268***	-0.0144**	-0.0149**	0.00638**	0.00838**
	(0.00243)	(0.00284)	(0.00596)	(0.00681)	(0.00289)	(0.00337)
Female	0.794***	0.721***	0.388***	0.383***	0.165***	0.165***
	(0.0645)	(0.0699)	(0.0489)	(0.0566)	(0.0500)	(0.0584)
Higher Education	0	0	0	0	0	1.306***
	(.)	(.)	(.)	(.)	(.)	(0.111)
Prefecture Work Experience						
Previous	0.141***	0.134***	0.886***	0.846***	0.449***	0.452***
	(0.0254)	(0.0315)	(0.0631)	(0.0715)	(0.0329)	(0.0388)
Current	0.731***	0.725***	-0.224	-0.263	0.539***	0.544***
	(0.0349)	(0.0434)	(0.146)	(0.164)	(0.0385)	(0.0458)
Future	-0.177***	-0.177***	0.347***	0.452***	-0.0735**	-0.0595
	(0.0233)	(0.0304)	(0.0997)	(0.110)	(0.0330)	(0.0383)
FE: Prefecture	V		V		V	
FE: Time	V		V		V	
FE: Prefecture x Time		V		V		V
Ethnicity	V	V	V	V	V	V
Ancestral Place	V	V	V	V	V	V
Birthplace	V	V	V	V	V	V
Observations	2922	1713	3506	2011	6428	4848
Adjusted R^2	0.900	0.679	0.559	0.153	0.555	0.181

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the determinants of power using "local" and "outsider" samples. A leader is classified as local if his/her hometown is located in the same prefecture as his/her ruling county. The dependent variable is $Position_{i,t}$, a proxy of power. The key independent variables are $Nighttime\ Light_{i,t}$ and $Previous\ Prefecture\ Experience_i$, which represent economic performance and connections to superior authorities, respectively. Column 1 and 2 use the local sample, while Column 3 and 4 present the results from the outsider sample; Column 5 and 6 employ the whole sample for reference. All models include either prefecture and period fixed effects or a specific time trend for each prefecture.

Table 2.20: Previous prefecture experience.

	(1) OLS	(2) FE	(3) FE	(4) FE
Local	0.146***	0.151***	0.151***	0.151***
	(0.00352)	(0.00372)	(0.00372)	(0.00417)
FE: Prefecture		V	V	
FE: Time			V	
FE: Prefecture x Time				V
Observations	74556	74556	74556	70800
Adjusted R^2	0.023	0.390	0.390	0.192

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the effects of power on previous prefecture experience using "local" and "outsider" samples. A leader is classified as local if his/her hometown is located in the same prefecture as his/her ruling county. The dependent variable is an indicator of whether a county leader has worked in the prefecture government directly superior to his/her ruling county, which can be perceived as a proxy for workplace connections. The key independent variable is $Position_{i,t}$, a proxy of power. Column 1 uses a simple OLS model, while Column 2-4 add prefecture, period, and prefecture x period fixed effects, respectively.

Table 2.21: Nighttime light.

	(1) OLS	(2) FE	(3) FE	(4) FE
Local	-0.127***	-0.0826***	-0.0826***	-0.0820***
	(0.0248)	(0.0183)	(0.0182)	(0.0197)
FE: Prefecture		V	V	
FE: Time			V	
FE: Prefecture x Time				V
Observations	24072	24072	24072	22656
Adjusted R^2	0.001	0.578	0.581	0.429

Standard errors in parentheses

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Note: The table displays the effects of power on nighttime light using "local" and "outsider" samples. A leader is classified as local if his/her hometown is located in the same prefecture as his/her ruling county. The dependent variable is $Nighttime\ Light_{i,t}$, a proxy of economic performance. The key independent variable is $Position_{i,t}$, a proxy of power. Column 1 uses a simple OLS model, while Column 2-4 add prefecture, period, and prefecture x period fixed effects, respectively.

preted as the level of power held by the competitors of a county leader and is negatively correlated to *Position*. The second instrumental variable is two-period lag of power structure. The results from two-stage least squares (2SLS) regressions show that if a county leader obtains one more position and regional power structure becomes more concentrated, the number of protests in the region will drop by 20.3% - 22.8%. I also show that power affects protests through both increasing grievances and lowering the incentive for collective action. On the one hand, with regard to the formation of collective action, I examine environmental protests and use the level of air pollution as a proxy for the root of grievances. The results show that, given the same level of air pollution, people are less likely to protest under rule of a powerful leader, suggesting that power reduces the incentive for collective action. On the other hand, the distinct patterns between local leaders and outsiders indicate that power can induce more protests through evoking grievances of people. In particular, the empirical results show that under the rule of a local leader, whose hometown is located within the same prefecture as his/her ruling county, people are more likely to suffer from a higher level of air pollution and tend to express discontent against the government. Finally,

I provide some suggestive evidence that local leaders are less competent and depend more on connections to superior authorities than outsiders.

It is worth mentioning that the protest data set in this chapter may not be representative. Although it contains rich information on comprehensive protests all over China, the total number of recorded protests is much smaller than the prediction in 2010 (Sun, 2011). Moreover, the fact that the data were collected from the Internet suggests that extremely poor areas or places under strict control by the Communist Party, such as Xinjiang and Tibet, may suffer more from under-representation. The identification strategy relies on the panel structure and the results should not be affected, but external validity need to be taken into consideration for less populated, remote districts or places with highly strict censorship.

In addition, there is only one county leader in each county in the data set, which means that it is impossible to distinguish county leader fixed effects from county fixed effects. A related point to consider is that there is no information on the quality of county leaders. 98% of the county leaders hold a bachelor's degree or higher, so education is not informative in terms of quality comparison among county leaders. To partially address the issue, I use economic performance to evaluate county leaders' competence (Y. Chen et al., 2005; Hongbin Li & Zhou, 2005). The negative coefficients of nighttime light, which is a proxy of economic performance, in the baseline results can be interpreted as the opportunity cost of participating a protest. It can also be understood as a manifestation of county leaders' ability: the more capable a county leader is, the fewer protests there are in his ruling county. Furthermore, I examine the hometown relationship of each county leader and find that compared to outsiders, local leaders tend to rely more on connections to higher authorities but perform worse in economic performance. Nevertheless, the above pieces of evidence are quite indirect and merely suggestive. To pin down the relationship between politician quality and protest frequency, a more thorough evaluation on the quality of politicians is required and will be left for future studies.

Chapter 3

Constructing Media Ideology in Chinese: Methodology and an Application to Cross-Strait Relations

3.1 Introduction

The chapter explains a new measure of media ideology in Chinese with an application to cross-Strait relations. I modify the English-based methods from Gentzkow and Shapiro (2010) and Murphy and Westbury (2013) and construct one of the first measures of media ideology in Chinese. There are two major challenges regarding Chinese text analysis. First, most existing tools are developed in English and usually cannot be applied to Chinese text directly. For example, the most popular Python package in Natural Language Processing, NLTK, does not support all functions in Chinese, and in many cases, it is unavoidable to take extra steps to process Chinese text before making good use of NLTK. Second, it is much more complicated to implement word segmentation and identify words, the smallest meaningful units, in a piece of Chinese text than English text. There is no explicit word boundary marker in Chinese, such as a space between words in English, and this means that even a piece of short text can be interpreted as many different possible word compositions. This chapter discusses the difficulties in constructing media ideology in Chinese and experiments with different word segmentation methods, keyword extraction algorithms, word association measures, and similarity metrics.

Although my measure is developed in Chinese and devoted to cross-Strait relations, its methodology allows a wide range of possibilities for applications. In principal, my mea-

sure can be applied to any topic with a text corpus representing one extreme stance on the topic so that subjects of interest, such as newspapers, can be assigned a score evaluating how "extreme" the subjects are. The quality of the measure depend on two factors. First, the text corpus should correctly characterize the extreme position on the topic. Second, the chosen topic should be relevant to the subjects of interest so that the subjects of interest actually discuss the topic and are considered holding some positions regarding the topic. My measure of media ideology on the pro-China to anti-China position meets the two criteria. On the one hand, the most pro-China stance is characterized by the press conference transcripts of the Taiwan Affairs Office (TAO), which is the official Chinese government institution in charge of policies related to Taiwan and representing the position of the Chinese government on cross-Strait issues. On the other hand, cross-Strait relations are crucial to Taiwanese citizens and have been one of the most important topics in Taiwan since 1949. Therefore, Taiwanese newspapers cover lots of stories about the topic and usually have been seen as holding some specific stance on cross-Strait relations.

Figure 3.1 illustrates the intuition of my measure. The core concept of the measure is to compare the use of language in Taiwanese newspapers and in TAO press conference transcripts. Based on the comparison, each Taiwanese newspaper is assigned a score that evaluates how closely it aligns with TAO. If a Taiwanese newspaper receives a high score, then it is considered more pro-China than other newspapers. The construction of my measure requires various steps. I first extract keywords from TAO press conference transcripts to represent the Chinese government's policies towards Taiwan. Then, I compare the use of each keyword and its neighboring words by TAO with the use by each Taiwanese newspaper. If a Taiwanese newspaper systematically uses the same set of keywords in a similar way as TAO does, the measure of alignment with TAO will be high, implying that the Taiwanese newspaper is ideologically close to the Chinese government and considered more pro-China.

To measure the alignment of Taiwanese newspapers with the Chinese government, I first identify key issues in the relationship between China and Taiwan and the position of the Chinese government on these issues. To do so, I scrape all press conference transcripts from the official website of TAO and construct monthly lexicons that represent major issues regarding cross-Strait relations and the Chinese government's stance on these issues. To con-

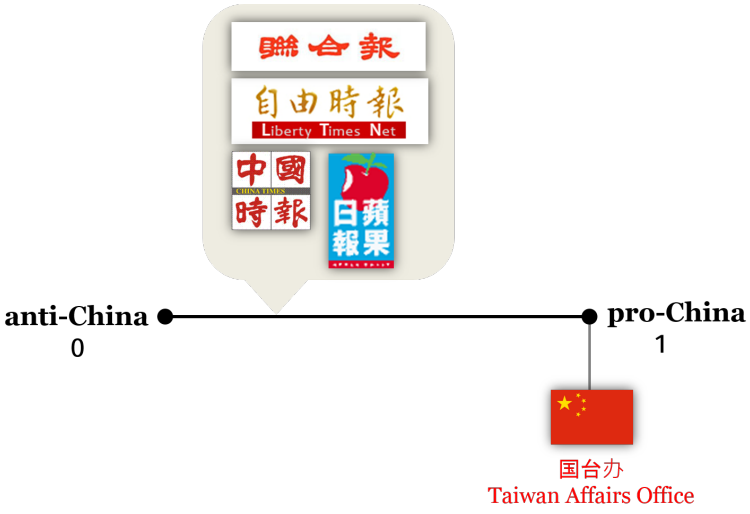


Figure 3.1: The intuition of the measurement

struct monthly lexicons, I first implement word segmentation on the TAO press conference transcripts. Word segmentation breaks written text into words, the smallest meaningful units, and is the first major task in text analysis. Unlike languages using the Latin alphabet, such as English, where words are usually separated by a blank space, Chinese has no explicit word delimiters and there are many possibilities to divide a chunk of Chinese text into words. For example, "台中國小" can be segmented into "台/中/國/小" (Taiwan, middle, country, small), "台/中國/小" (Taiwan, China, small), "台中/國小" (Taichung Elementary School), etc. To solve the problem, I adopt two methods for word segmentation in Chinese[1]. The first method is called Chinese Knowledge and Information Processing (CKIP)[2] and provided by the Institute of Information Science and the Institute of Linguistics of Academia Sinica. The second method is a popular Python package called jieba[3]. CKIP is better at capturing Taiwan-related topics, while jieba is much more efficient and can analyze enormous amount of text in a short period of time[4]. Word segmentation partially influences keyword selection–only ca. 60% of the keywords in the monthly lexicons

[1]The two packages for Chinese word segmentation also implement part-of-speech tagging at the same time.

[2]http://ckipsvr.iis.sinica.edu.tw/.

[3]https://github.com/fxsjy/jieba.

[4]Because CKIP is very slow and error-prone when analyzing large text, I choose jieba to perform word segmentation on millions of headlines from Taiwanese newspapers

are identical between CKIP and jieba. Nevertheless, the final measures of media ideology are highly correlated across Taiwanese newspapers over time, with correlation coefficient over 0.9[1].

Keyword extraction is implemented after word segmentation. I first pre-process the words from press conference transcripts and keep only content words (such as nouns, adjectives, verbs) that contain more than two characters, appear more than 5 times, and are not common phrases in general press conferences (eg. reporters)[2]. Then, I select keywords into the lexicons according to two algorithms. In the first algorithm, I simply pick the most frequent words each month. If a word is constantly used by the Chinese government and it is not a function word and not a commonly used word in general press releases, then the word is likely to relate to some fundamental issues concerning cross-Strait relations and, possibly, the Chinese government's stance on these issues. The second algorithm is TF-IDF, which stands for Term Frequency and Inverse Document Frequency. Term frequency is the number of occurrences of a word in a document, such as a piece of press conference transcript, while document frequency is the number of documents that contain the word. The TF-IDF score is proportional to word frequency and divided by document frequency as a weight to offset word frequency. Since the scarcity of words is considered, TF-IDF can select words that are specific to a month but rare in other months. If a word appears repeatedly in many months, then it is unlikely to be chosen by TF-IDF despite its high frequency. For example, the average word frequency of "原則" (principle) is comparable to "鮑威爾" (Powell, which refers to the 65th United States Secretary of State, Colin Luther Powell). However, because "原則" (principle) appears in most months (82.72%) and "鮑威爾" (Powell) only shows up in 2 months (1.23%), the average TF-IDF score of "原則" (principle) is merely 4% of the score of "鮑威爾" (Powell). The two algorithms are complements to each other–the first one selects keywords representing essential issues that are relatively stable over time, while the second one chooses keywords relating to issues that are very present in some months but not over longer time periods. Given the discrepancy

[1] All the correlation coefficients in the introduction are calculated by taking pairwise comparison between the measures from two different algorithms across all newspapers over the entire time period. For example, the correlation coefficient here comes from the comparison between the two word segmentation methods, CKIP and jieba.

[2] The list of words commonly used in general press conferences are manually selected from the text corpus.

between the two algorithms, it is not surprising that there are only ca. 22.3-25.9% common keywords on average. Despite these differences in the selection of keywords, the final measures of media ideology are highly correlated across Taiwanese newspapers over time (corr = 0.78).

Next, to understand the context of each keyword in the lexicon, I study the neighboring words around each keyword. Neighboring words are defined as all content words within the same sentence (or the same headline) as a keyword[1]. On average, I include ca. 40 neighboring words for each keyword. Taking context into consideration has several advantages. First of all, it can help detect different perspectives on a topic. For instance, one of the most commonly used words in the lexicon is "台獨" (Taiwan Independence), and the Chinese government's negative stance on Taiwan independence can be inferred from its neighboring words, such as "抵制"(boycott), "危害"(endanger), "遏制"(containment), "阻礙"(obstruction), "破壞"(damage), etc. Secondly, analyzing context helps avoid the confusion from lexicon differences between the official language in Taiwan, Taiwanese Mandarin (國語) and the official language in China, Putonhua (普通話). Although both languages are classified as Mandarin, there are many "false friends", i.e. words that sound or look similar but with different meanings, such as 小姐 (Miss in Taiwanese Mandarin vs. prostitute in Putonhua), and it is easier to distinguish one from another if neighboring words are considered. Thirdly, it helps alleviate the negative effects of word segmentation errors on the measure. Word segmentation is a complicated task and mistakes cannot be completely avoided. For example, "台中國小" should be segmented as "台中/國小" (Taichung Elementary School) but is wrongly interpreted as "台/中國/小" (Taiwan China small) by jieba. Since "中國" (China) is a keyword in the lexicon, the mistake may be perceived as a sign of similarity between a Taiwanese newspaper and TAO, and introduce a bias to the measure. Nevertheless, when context is taken into account, the measure should not be much affected because the neighboring words around an elementary school are very different from the ones around China. Lastly, it helps eliminate ambiguities from homographs, i.e. words with the same written form but different meanings, such as 易 (easy vs. change)

[1]The pre-processing for a neighbor word is similar to the one for a keyword except that it only requires its term frequency to be larger than 1

and 長 (long vs. elder). Since homographs tend to have different contexts, examining the neighboring words around a homograph can help identify the word and solve the issue.

To formally analyze the context, I calculate three different measures for word association. One of the most popular measures is Positive Pointwise Mutual Information (PPMI). PPMI compares the actual joint distribution of a keyword and a neighboring word to the joint distribution assuming independence of the keyword and the neighboring word. If the neighboring word often appears with the keyword, PPMI will be positive and there is so-called word co-occurrence. For example, suppose 台獨 (Taiwan Independence, denoted as K) has two neighboring words, 台灣 (Taiwan, denoted as N_a) and 決不允許 (never allow, denoted as N_b) with comparable average joint probabilities in the TAO press conference transcripts: $p(K, N_a) = 0.000755$ and $p(K, N_b) = 0.000753$.[1] However, because 台灣 (Taiwan) is much more frequent than 決不允許 (never allow) with average marginal probabilities $p(N_a) = 0.025817$ and $p(N_b) = 0.000753$, respectively, 台灣 (Taiwan) is expected to have a higher joint probability under the assumption of independence between 台獨 (Taiwan Independence) and 台灣 (Taiwan). As a result, the measure of PPMI for 台灣 (Taiwan) is much lower than the one for 決不允許 (never allow): $PPMI_{Na} = 2.630$ and $PPMI_{Nb} = 8.053$.

Using PPMI helps mitigate at least three potential problems. Firstly, the number of news stories included in the database is unbalanced between small newspapers and major newspapers. Since PPMI uses probabilities in its measure, small newspapers are not deemed less pro-China simply because of fewer news stories. Secondly, the number of news stories grows over time in the database, and adopting PPMI lessens the concerns for intertemporal comparison of media ideology. Lastly, PPMI performs better than raw frequency in terms of the comparison across different types of text, such as press conference transcripts and newspaper headlines.

In addition to PPMI, I also calculate two measures of word association as robustness checks. The first measure is the popular t-test statistics, which is conceptually very similar to PPMI. The null hypothesis of t-test is that a keyword and its neighboring word are independent. In this case, the t-test statistic should be zero. A higher t-test statistic indicates

[1] All the numbers in this paragraph are calculated by averaging the probabilities or PPMI in the monthly TAO press conference transcripts from 2000 to 2014.

greater likelihood of word association. Given the similarity between PPMI and t-test, the two measures of word association are highly correlated across Taiwanese newspapers (corr = 0.78), and the corresponding final measures of media ideology are highly correlated as well (corr = 0.85). The other measure is the raw frequency of a keyword and its neighboring word. Intuitively, when a neighboring word often appears with a keyword, the two words are more likely to be associated. The word association measure based on raw frequency is moderately correlated to the one based on PPMI across neighboring words for all keywords over time (corr = 0.52), and the corresponding measures of media ideology are very highly correlated (corr = 0.95).

Finally, I calculate several versions of similarity metrics to compare the patterns of keywords and neighboring words between the Chinese government and each Taiwanese newspaper. I first adopt the most common similarity metric, cosine similarity, which measures the cosine of the angle between two vectors and ranges between 0 and 1 for positive vectors. If a Taiwanese newspaper tends to use a keyword along with its neighboring words in a similar fashion as the Chinese government does, then the cosine similarity with regard to the keyword is high. For example, the Chinese government often uses the keyword 台獨 (Taiwan Independence) with 危險性 (risk), 鐵桿 (iron poles, a metaphor that describes someone who is hardcore and stubborn), and 老鞋 (old shoes, a metaphor for being old-fashioned), and the average PPMI for each neighboring word is 9.24, 8.05, 8.66, respectively[1]. The last two neighboring words are common phrases in China but rarely used in Taiwan, not to mention in the context of 台獨 (Taiwan Independence). Nevertheless, the most pro-China newspaper in Taiwan, the *United Daily News* (聯合報, UDN), also uses these neighboring words with 台獨 (Taiwan Independence) and the corresponding average PPMI is 9.41, 7.87, and 14.30. On the contrary, the least pro-China newspaper in Taiwan, the *Apple Daily* (蘋果日報), never uses these words with 台獨 (Taiwan Independence) and PPMI is zero for the three words. Overall, the average cosine similarity regarding the keyword 台獨 (Taiwan Independence) is 0.389 for the *United Daily News* and 0.173 for the

[1]All PPMIs in this paragraph are calculated by averaging the PPMI from 2000 to 2014, if 台　獨 (Taiwan Independence) is selected as a keyword in the month.

Apple Daily.[1] The final measure of media ideology is the average of cosine similarity metrics for all keywords in the lexicon.

In addition to cosine similarity, I also use Jaccard similarity and Dice similarity for comparison. The two measures are based on "the weighted number of overlapping features" with different normalization factors (Jurafsky & Martin, 2000), and the final measures of media ideology based on different similarity metrics are highly correlated for all newspapers over time (corr = 0.75-0.77).

My measure of media ideology is novel in two aspects. First, it is one of the first objective measures of media bias in Chinese. The existing literature primarily focuses on American politics and most analysis tools were developed in English, with only a few exceptions. Qin, Wu, and Strömberg (2016) collect the content of 117 urban newspapers in China from 1999 to 2010 and measure the degree of political control and/or commercial incentives based on the coverage of 9 content categories, such as mentions of political leaders, citations of official media outlets, reports on entertainment news, etc., and then apply principal component analysis to consolidate the categories into a one-dimensional measure. Yuan (2016) utilizes unsupervised learning to classify 21 newspapers in China into two groups based on hierarchical clustering. The measure I use is different from Qin et al. (2016) and Yuan (2016) in several aspects. Firstly, it captures the influence of the Chinese Communist Party in Taiwan, not China. Since Taiwan is a democracy with freedom of speech, the major concern is not direct media control by the Chinese government. In addition to the application context, the measurement itself is distinct from the two papers as well. Qin et al. (2016) consider only keywords to identify topics, while the measure in this paper takes both keywords and the corresponding context into account. Yuan (2016) is completely data-driven and relies on the quality of the input text, whereas the measure in this paper takes TAO as a reference for pro-China ideology and is robust to several different keyword extraction methods. The second novel aspect of the measure is that its ideological spectrum is based on international politics, not on domestic political competition. That is, the measure does not examine bias toward a certain domestic political party, such as a pro-Democratic or pro-

[1] The values of cosine similarity here are calculated by averaging the cosine similarity between a Taiwanese newspaper and the Chinese government from 2000 to 2014, if 台　獨 (Taiwan Independence) is selected as a keyword in the month.

Republican newspaper in the U.S.; instead, it analyzes the bias of newspapers in Taiwan towards official Chinese issues and positions.

The rest of paper proceeds as follows. Section 3.2 describes data sources. Section 3.3 explains the construction of the lexicons of issues regarding cross-Strait relations and the Chinese government's stance on these issues. Section 3.4 discusses the context of keywords in the lexicons. Sections 3.5 compares the language between TAO and each Taiwanese newspaper. Section 3.6 concludes.

3.2 Data

3.2.1 Press Conference Transcripts

The measure of media ideology requires a lexicon of Chinese ideology as a reference that I can compare with each newspaper. I collect press conference transcripts of a Chinese government institution, Taiwan Affairs Office from its official website (http://www.gwytb.gov.cn/xwfbh/index.htm). The Taiwan Affairs Office of the State Council, or Taiwan Affairs Office (TAO), is an administrative institution in charge of Taiwan affairs under the State Council of China. According to its official website, the main functions of Taiwan Affairs Office include "implementing and carrying out guidelines and policies related to Taiwan stipulated by the CPC Central Committee and the State Council" and "taking charge of the media and publicity work related to Taiwan and releasing news and information concerning Taiwan affairs." Since Taiwan Affairs Office is responsible for conducting and publicizing China's policies towards Taiwan, I consider TAO a major official outlet of the Chinese government and use transcripts of TAO press conferences to build monthly lexicons of China's ideology about Taiwan. The press conferences take place irregularly: from 2000 to 2017, there are 265 TAO press conferences in total, 162 months with press conferences, and often one or two in each month. On average, I have 4,211 words each month to build a lexicon.

3.2.2 Newspaper Headlines

To measure media ideology, I collect newspaper headlines from News Knowledge Management System (NKMS) administered by the Library of Legislative Yuan in Taiwan. The system consists of daily headlines of 34 newspapers since 1980; I study 12 of them that are also included in the TEDS surveys. I choose to analyze headlines to measure media ideology, not full news text, for three reasons. First, headlines are usually set by editors, not reporters, and more likely to reveal a newspaper's ideology. Second, readers tend to read headlines more often than to read full texts of news, so headlines are more influential than news content. Third, data availability and compilation difficulty impede full text analysis.

The enormity of data renders it impossible to collect them manually. Instead, I write Python programs to collect information on 18,383,058 headlines from 2000 to 2014, including newspaper names, headlines, dates, pages, sections, and sometimes reporters and regions. Following the literature (Gentzkow & Shapiro, 2010; Groseclose & Milyo, 2005), I remove editorials and reviews for two reasons. First, the language used by editors and opinion writers is usually sarcastic and difficult to tell its connotations with automatic text analysis. It can be that an editor uses many positive verbs to criticize "One China Policy". Second, news, editorials and reviews are utterly different from readers' perspective. News is supposed to be objective and truthful, while editorials and reviews are full of opinions and judgments. While editorials and reviews may reveal the ideology of a newspaper more directly and efficiently, they are less likely to have an impact on readers' thoughts and, in turn, influence their national identities. Therefore, I decide to remove editorials and review in the data set. Then, I write additional code to detect repetitions in the data set. If I request a large amount of news, the NKMS system often responds with many duplicates, which can seriously distort the result–it turns out that nearly 10% of the data are repetitions. In the end, I collect 10,953,168 headlines. It is worth mentioning that the number of news stories from a newspaper in the NKMS system surges from 7.6 news stories per day in 2000 to 199.8 per day in 2014, suggesting that we need to be cautious about intertemporal analysis: measures in earlier years are built on relatively small text corpora and may be less stable compared to measures in later years. On average, there are 48,645 words per month per newspaper.

3.3 Chinese Ideology Lexicons

To measure the degree of pro-China for each Taiwanese newspaper, I first construct a reference for Chinese preference. The reference contains a list of keywords extracted from the press conference transcripts of a Chinese government institution, the Taiwan Affairs Office (TAO), and represents China's policies regarding Taiwan. The following sections explain the word segmentation procedure and the keyword extraction process.

3.3.1 Word Segmentation

The first major task in text analysis is word segmentation, the process of breaking written text into words, the smallest meaningful units. Different from English, Chinese does not have explicit word boundary markers and word segmentation is far more challenging. For example, there are at least 6 possible word combinations for a short text of 4 characters, 台中國小, with quite different meanings:

台/中/國/小	(Taiwan / middle / country / small)
台/中/國小	(Taiwan / middle / elementary school)
台/中國/小	(Taiwan / China / small)
台中/國/小	(Taichung / country / small)
台中/國小	(Taichung / elementary school)
台中國/小	(Taichung Country / small)

Without proper word segmentation, text can be mistakenly interpreted and introduce errors to the measure of media ideology. To solve the problem, I employ two popular methods for word segmentation. The first method is called Chinese Knowledge and Information Processing (CKIP) and provided by the Institute of Information Science and the Institute

of Linguistics of Academia Sinica[1]. CKIP is built upon linguistic analysis and large Chinese text corpora, and capable of out-of-vocabulary word identification and part-of-speech tagging; its recall, precision, and F score are ca. 93% - 97% (Ma & Chen, 2003). The second method is a widely used Python package called jieba (Ren & Tian, 2017; Shao, Sennrich, Webber, & Fancellu, 2017; Xiu, Lan, Wu, & Lang, 2017; Zhang, Zeng, Jin, Yan, & Geng, 2017). Jieba adopts omni-segmentation, which lists all possible word combinations and "use dynamic programming to find the most probable combination based on word frequency" [2]. Nevertheless, since jieba is developed in simplified Chinese text corpora, it is relatively weak in capturing Taiwan-related topics.[3] For instance, jieba usually interprets 台中國小 as 台/中國/小 (Taiwan / China / small), while CKIP recognizes 台中 (Taichung) as a city in Taiwan and correctly breaks the term into 台中/國小 (Taichung / elementary school). Another example is 蔡英文 (Tsai Ing-wen), the president of Taiwan. While CKIP knows her and identifies the term as a name, jieba mistakenly perceives the term as 蔡/英文 (Tsai / English) because the last two characters happen to be the Chinese translation for "English".[4] Although CKIP is better at recognizing Taiwan-related topics, jieba is much more efficient and capable of processing a large chunk of text in a short period of time. Therefore, I adopt both methods and take jieba as a robustness check.

Figure 3.2 shows the percentage of common keywords between CKIP and jieba from 2000 to 2017. On average, 59.5% of the keywords in the monthly lexicons are identical between the two methods, suggesting that word segmentation indeed has an effect on keyword selection. Nevertheless, the final measures of media ideology under the two methods are very highly correlated across newspapers over time (corr = 0.95). Figure 3.3 displays the difference in media ideology between CKIP and jieba for all 12 newspapers from 2000 to 2014, and the values almost always lie within -0.1 and +0.1.

---

[1] I implement the CKIP word segmentation through an api provided by https:// github.com/ jason2506/ ckip.py

[2] https://github.com/fxsjy/jieba

[3] I always convert text into simplified Chinese when applying jieba and translate back to traditional Chinese to compare with Taiwanese newspapers. Jieba can execute word segmentation directly in traditional Chinese as well, but its performance is much worse than simplified Chinese, especially in terms of part-of-speech tagging. The conversion between simplified Chinese and traditional Chinese is implemented by a program called OpenCC (https://github.com/BYVoid/OpenCC).

[4] In some rare cases, jieba performs better in name recognition. For example, the accuracy of word segmentation for 李維一 (Li Weiyi) is 69.8% by jieba, but only 9.5% by CKIP.

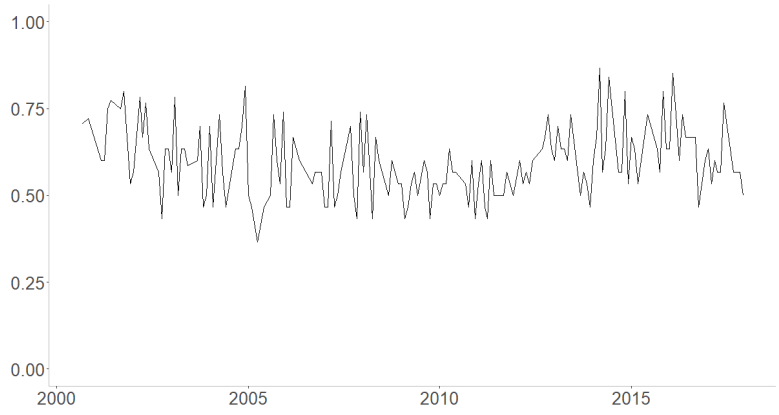


Figure 3.2: The percentage of common keywords between CKIP and jieba from 2000 to 2017.

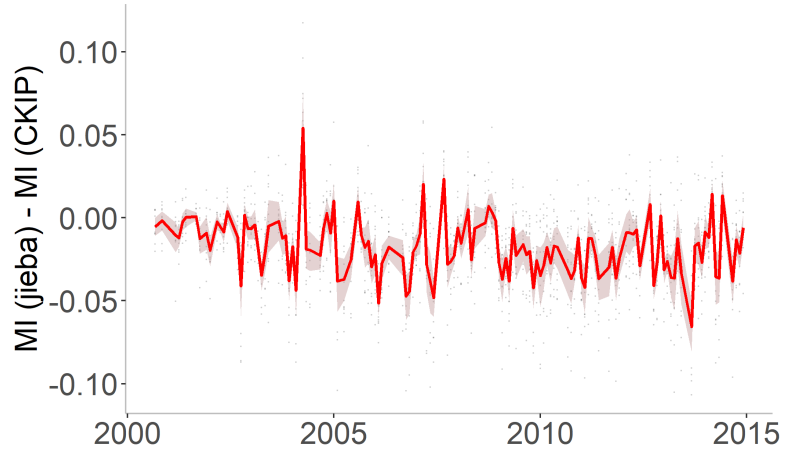


Figure 3.3: The difference in the final measures of media ideology under the two word segmentation methods, CKIP and jieba, from 2000 to 2014.

3.3.2 Keyword Extraction

Keyword extraction is implemented after word segmentation. I first remove words with less than 5 occurrences in a month. The purpose of the keywords is to describe China's major policies about Taiwan, and words with low frequency are unlikely to serve the purpose. Secondly, I discard words with only one character because there are many possible meanings for a single-character word and it is more informative to use words with two characters or more. For example, a single character "中" can be interpreted as "middle", "China", and "to hit", and it is much clearer when one more character is considered: "中間" (middle), "中國" (China), and "命中" (to make a hit). Then, I keep only content words, including nouns, verbs, and adjectives.[1] Content words contain information and are usually stressed in speech, while function words, such as determiners and prepositions, exist for grammatical reasons and are quite unlikely to characterize a policy. Finally, I manually remove words widely used in general press conferences, such as reporters, but do not relate to China's policies about Taiwan.

3.3.2.1 Term Frequency

Next, I select keywords from the processed text according two different algorithms. In the first algorithm, I sort the word list by term frequency and pick the 30 most frequent words each month[2]. If a content word is constantly used by the Chinese government and not commonly used in general press releases, then the word is likely to relate to some fundamental policies of the Chinese government concerning Taiwan.

[1] I use the part-of-speech taggers that come with the word segmentation packages. For CKIP, I include N (nouns), A (adjectives), Vi (intransitive verbs), Vt (transitive verbs), and FW (words in a foreign language), where foreign words usually stand for abbreviated proper nouns in Chinese text, such as WTO (World Trade Organization). On the other hand, for jieba, I include eng (words in a foreign language), j (abbreviated terms), i (idioms), l (slang terms), n (nouns), nr (a person's name), ns (place names), nt (institution name), nz (other proper nouns), nrt (foreign translated names), v (verbs), vn (words that can act as verbs and nouns), a (adjectives), an (words that can act as adjectives and nouns), t (adverbs about time), s (direction nouns), f (direction verbs), and b (distinguishing adjectives).

[2] In the earlier months, TAO did not hold press conferences as often and the size of text corpus in these months is too small to contain 30 words satisfying all the conditions mentioned in the last paragraph. As a result, in these months, the number of keywords is less than 30.

Table 3.1 lists the 20 most frequent keywords based on term frequency and CKIP word segmentation in each presidency from 2000 to 2017[1]. There were three presidencies and two rotations of parties during the period, from Democratic Progressive Party (DPP, May 2000 - April 2008), Kuomintang (KMT, May 2008 - April 2016), to DPP (May 2016 till now). KMT is considered more pro-China and has many contacts with the Chinese government, while DPP often relates to Taiwan Independence and emphasizes the subjectivity of Taiwan. Table 3.1 shows some consistency in China's attitude towards Taiwan. For example, "同胞" (compatriots), "發展" (development), and "合作" (cooperation) were constantly used throughout the three presidencies. More importantly, Table 3.1 reflects the change in China's policies about Taiwan. During the DPP presidencies, "台獨" (Taiwan Independence) was often brought up, while "協議" (agreement) and "經濟" (economy) were the main topics when KMT was in power. Table 3.2 displays the top 20 keywords based on jieba word segmentation and demonstrates the same pattern.

Figure 3.4 further exhibits this transition in topics across presidencies. On the one hand, Figure 3.4a shows a sharp drop in the use of "一個中國" (One China) and "台獨" (Taiwan Independence) after KMT won the presidency in 2008, and a rising trend after DPP regained power in 2016. On the other hand, Figure 3.4b displays a steep rise in the frequency of "經濟" (economy) and "協議" (agreement) in the KMT presidency, and a declining trend after DPP returned to power.

To study the keywords more systematically, I use a Python package called gensim (Řehůřek & Sojka, 2010) and apply an unsupervised topic model, Latent Dirichlet Allocation (LDA) (Blei, Ng, & Jordan, 2003), to examine the topics among keywords over time. LDA is a topic model that aims to discover the topics in a collection of unlabeled documents. It represents each document as a mixture of topics, and each topic is a probability distribu-

[1]I only have data on Taiwanese newspapers from 2000 to 2014, so the measures of media ideology range from 2000 to 2014 as well, despite that I also have data on TAO press conference transcripts up to date.

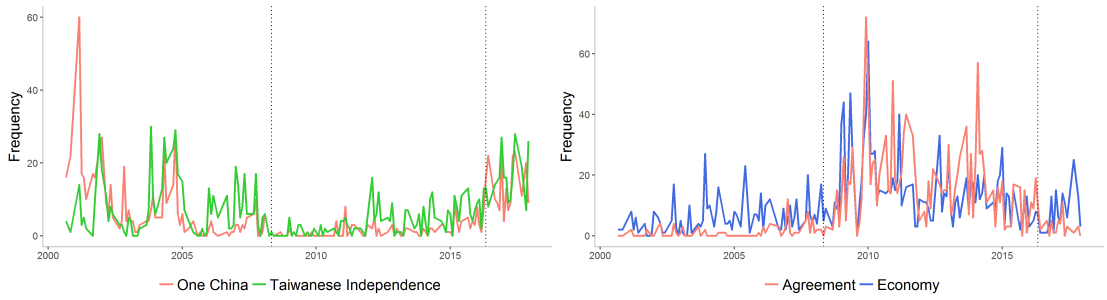
[1]The correct word is "李 維 一" (Li Weiyi), the name of the former spokesman of the Taiwan Affairs Office, but both jieba and CKIP wrongly segmented the term as "李維/一" (Li Wei / yi).

[2]Three Direct Links: shorthand for direct links of trade, mail, and air and shipping services across the Taiwan Strait.

[1]The correct word is "安 峰 山" (An Fengshan), the name of the former spokesman of the Taiwan Affairs Office, but both jieba and CKIP wrongly segmented the term as "安/峰山" (An / Fengshan).

Table 3.1: Keywords based on term frequency and CKIP word segmentation

DPP (2000.5 - 2008.4)		KMT (2008.5 - 2016.4)		DPP (2016.5 - 2017.12)	
keyword	*translation*	*keyword*	*translation*	*keyword*	*translation*
兩岸	cross-Strait	台灣	Taiwan	兩岸	cross-Strait
台灣	Taiwan	兩岸	cross-Strait	台灣	Taiwan
大陸	mainland	大陸	mainland	關係	relationship
同胞	compatriot	同胞	compatriot	大陸	mainland
中國	China	共同	common	共識	consensus
關係	relationship	合作	cooperation	政治	politics
李維 (一)[1]	Li Wei(yi)	關係	relationship	中國	China
當局	authorities	雙方	both sides	同胞	compatriot
發展	development	協議	agreement	共同	common
現在	now	海峽	strait	台獨	Taiwan independence
實現	realization	發展	development	和平	peace
台商	Taiwanese businessmen	經濟	economy	基礎	foundation
台獨	Taiwan independence	和平	peace	發展	development
包機	charter flight	企業	enterprise	當局	authorities
合作	cooperation	目前	currently	大家	everyone
原則	principle	中國	China	交流	exchange
三通[2]	Three Links	范麗青	Fan Liqing	馬曉光	Ma Xiaoguang
張銘清	Zhang Mingqing	論壇	forum	合作	cooperation
共識	consensus	交流	exchange	海峽	strait
企業	enterprise	一下	a bit	堅持	adhere to



(a) Political issues **(b)** Economic issues

Figure 3.4: The word frequency of selected words in the TAO press conference transcripts from 2000 to 2017.

Note: Figure 3.4a displays the word frequency of "台獨" (Taiwan Independence) and "一個中國" (One China), while Figure3.4b shows the frequency of "協議" (agreement) and "經濟" (economy).

Table 3.2: Keywords based on term frequency and jieba word segmentation

DPP (2000.5 - 2008.4)		KMT (2008.5 - 2016.4)		DPP (2016.5 - 2017.12)	
keyword	*translation*	*keyword*	*translation*	*keyword*	*translation*
台灣	Taiwan	台灣	Taiwan	台灣	Taiwan
大陸	mainland	大陸	mainland	兩岸關係	cross-Strait relations
交流	exchange	交流	exchange	大陸	mainland
兩岸關係	cross-Strait relations	兩岸關係	cross-Strait relations	和平	peace
中國	China	雙方	both sides	共識	consensus
台灣同胞	Taiwanese compatriot	同胞	compatriot	基礎	foundation
李維 (一)	Li Wei(yi)	和平	peace	政治	politics
台灣當局	Taiwanese authorities	協議	agreement	中國	China
發佈會	press conference	經濟	economy	同胞	compatriot
台獨	Taiwan independence	協商	negotiation	交流	exchange
台商	Taiwanese businessmen	企業	enterprise	台獨	Taiwan independence
共識	consensus	大家	everyone	大家	everyone
包機	charter flight	民眾	people	原則	principle
原則	principle	國台辦	Taiwan Affairs Office	民眾	people
三通	Three Links	論壇	forum	台灣同胞	Taiwanese compatriot
大家	everyone	海峽兩岸	cross-Strait	馬曉光	Ma Xiaoguang
同胞	compatriot	居民	resident	國台辦	Taiwan Affairs Office
企業	enterprise	共識	consensus	評價	evaluation
協商	negotiation	范麗青	Fan Liqing	(安) 峰山[1]	(An) Fengshan
國台辦	Taiwan Affairs Office	中國	China	立場	position

tion of words.[1] In the context of this paper, a document is a composition of the keywords from TAO press conference transcripts each month[2], and there are 162 documents from 2000 to 2017. Each document consists of 20 topics with certain weights, and each topic is represented by a number of keywords with certain probabilities. For example, a text of the keywords from the TAO press conference transcripts in September 2000 is a document composed of Topic 2 with weight 0.57 and Topic 8 with weight 0.43, and Topic 2 and Topic 8 are probability distributions of keywords (the top 10 keywords with the highest probabilities are listed in Table 3.3).

Figure 3.6 displays the weight of each topic in the monthly press conference transcripts over time. The colored points are topics with high weights consistently[3], and it turns out that these topics contain "台獨" (Taiwan Independence), "協議" (agreement), or "共識" (consensus) in the top 10 keywords, respectively. The keywords in these topics are listed in Table 3.3. Again, Figure 3.5 demonstrates the transition of China's key policies regarding Taiwan across presidencies in Taiwan. During the first DPP presidency, the Chinese government concentrated on attacking "台獨" (Taiwan Independence) and defending the "一個中國" (One China), whereas various forms of "協議" (agreement) became the focus when KMT was in power. After the Sunflower Student Movement took place in March 2014 and DPP regained its power in 2016, "共識" (consensus), which refers to "九二共識" (The 1992 Consensus) and represents a new form of One China Policy, has been consistently emphasized. Figure 3.6 displays the word frequency of the above keywords and exhibits a similar trend.

3.3.2.2 TF-IDF

The second algorithm for keyword extraction is TF-IDF[4], which stands for Term Frequency and Inverse Document Frequency:

[1] The number of topics in each document is chosen by the user manually.

[2] The goal is to find out the clusters of keywords, so I filter out all other words that are not classified as keywords in the TAO press conference transcripts.

[3] The selected topics contain weights higher than 0.3 for at least five months.

[4] TFIDF is implemented by a Python package called gensim (Řehůřek & Sojka, 2010).

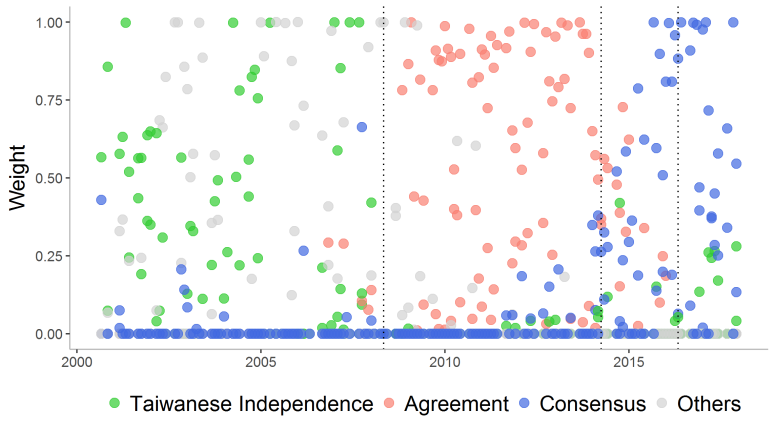


Figure 3.5: The weights of selected topics in the LDA model from 2000 to 2017.

Note: Among the 20 topics in the LDA model, only topics with weights over 0.3 for at least 5 months are presented in the graph.

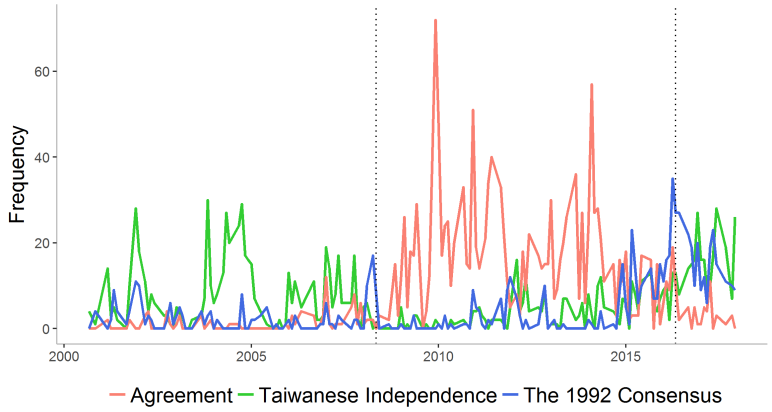


Figure 3.6: The frequency of "台獨" (Taiwan Independence), "協議" (agreement), and "九二共識" (the 1992 consensus) in the TAO press conference transcripts from 2000 to 2017.

Table 3.3: Keywords in the selected LDA topics from 2000 to 2017.

	台獨 (Taiwan Independence)			**協議** (Agreement)			**共識** (Consensus)	
topic	*keyword*	*translation*	*topic*	*keyword*	*translation*	*topic*	*keyword*	*translation*
0	台灣	Taiwan	11	兩岸	cross-Strait	8	兩岸	cross-Strait
	兩岸	cross-Strait		台灣	Taiwan		台灣	Taiwan
	大陸	mainland		大陸	mainland		關係	relationship
	同胞	compatriot		范麗青	Fan Liqing		大陸	mainland
	楊毅	Yang Yi		合作	cooperation		發展	development
	關係	relationship		**協議**	agreement		和平	peace
	合作	cooperation		企業	enterprise		安峰山	An Fengshan
	台獨	Taiwan independence		經濟	economic		同胞	compatriot
	政策	policy		台資	Taiwan-funded		交流	exchange
	居民	resident		共同	common		**共識**	consensus
2	台灣	Taiwan	12	兩岸	cross-Strait	18	兩岸	cross-Strait
	兩岸	cross-Strait		台灣	Taiwan		台灣	Taiwan
	中國	China		大陸	mainland		馬曉光	Ma Xiaoguang
	大陸	mainland		合作	cooperation		大陸	mainland
	原則	principle		楊毅	Yang Yi		關係	relationship
	張銘清	Zhang Mingqing		**協議**	agreement		發展	development
	當局	authorities		關係	relationship		同胞	compatriot
	台獨	Taiwan independence		交流	exchange		中國	China
	關係	relationship		發展	development		**共識**	consensus
	陳水扁	Chen Shui-bian		同胞	compatriots		大家	everyone
10	台灣	Taiwan						
	兩岸	cross-Strait						
	大陸	mainland						
	同胞	compatriot						
	李維 (一)	Li Wei(yi)						
	中國	China						
	發展	development						
	當局	authorities						
	台獨	Taiwan independence						
	關係	relationship						

$$\text{TFIDF}_{i,t} = \text{Frequency}_{i,t} \times log_2 \frac{D}{\text{Document Frequency}_i}$$

$\text{Frequency}_{i,t}$ is the number of occurrences of word i in document t, where a document is the text of all press conference transcripts in a month. $\text{Document Frequency}_i$ is the number of documents that contain word i, and D is the total number of documents. In total, there are 162 documents (months) with TAO press conference transcripts from 2000 to 2017.

TF-IDF uses the inverse of document frequency as a weight to each word and extracts words that are specific to a month but rare in other months. If a word appears in many months, then it is unlikely to have a high TF-IDF score despite its high frequency. For example, the average frequency of "原則" (principle) is comparable to "鮑威爾" (Powell, which refers to the 65th United States Secretary of State, Colin Luther Powell): *Frequency*(原則 (principle)) = 5.48 and *Frequency*(鮑威爾 (Powell)) = 5.50. However, because "原則" (principle) appears in most months (82.72%) and "鮑威爾" (Powell) only shows up in 2 months (1.23%), the average TF-IDF score of "原則" (principle) is much lower: *TFIDF*(原則 (principle)) = 0.0098 and *TFIDF*(鮑威爾 (Powell)) = 0.2173.

Table 3.4 shows the 20 most frequent keywords based on the TF-IDF algorithm and CKIP word segmentation in each presidency from 2000 to 2017. The TF-IDF algorithm is designed to select keywords related to popular topics in each month, such as "奧運會" (Olympic Game) and "災區" (disaster area), not topics existing for years, such as "台獨" (Taiwan Independence). There is no common word across the three presidencies in Table 3.4, whereas 40% of keywords based on the first algorithm are the same in the three presidencies.

The two algorithms are complements to each other–the first one selects keywords representing essential issues that are relatively stable over time, while the second one chooses keywords relating to issues that are very present in some months but not over longer time periods. Figure 3.7 displays the percentage of common keywords between the two algorithms from 2000 to 2017. Given the discrepancy between the two algorithms, it is not surprising that on average, only 22.3-25.9% of the keywords are identical between the two algorithms each month.

Table 3.4: Keywords based on TFIDF and CKIP word segmentation

DPP (2000.5 - 2008.4)		KMT (2008.5 - 2016.4)		DPP (2016.5 - 2017.12)	
keyword	*translation*	*keyword*	*translation*	*keyword*	*translation*
李維 (一)	Li Wei(yi)	范麗青	Fan Liqing	馬曉光	Ma Xiaoguang
包機	charter flight	楊毅	Yang Yi	安峰山	An Fengshan
張銘清	Zhang Mingqing	ECFA	ECFA	青年	youth
三通	Three Links	論壇	forum	李明哲	Lee Ming-che
春節	Chinese New Year	協議	agreement	停擺	stop
陳水扁	Chen Shui-bian	馬曉光	Ma Xiaoguang	詐騙	fraud
農業	agriculture	赴台	go to Taiwan	(蔡) 英文	(Tsai) Ing-wen
航空	aviation	峰會	summit	受害人	victim
熊貓	panda	貨幣	currency	答卷	answer
祖國	motherland	台資	Taiwan-funded	縣市	county and city
當局	authorities	平潭	Pingtan	港獨	Hong Kong independence
奧運會	Olympic Game	互設	mutual establishment	年度	annual
WTO	WTO	同比	year-on-year (yoy)	台企	Taiwanese companies
奧運	Olympics	清算	clearing	世衛	WHO
農民	farmer	青年	youth	一帶	One Belt (One Way)
汪辜	Wang-Koo	災區	disaster area	WHA	WHA
直航	direct flight	影片	video	大會	conference
公投	referendum	航班	flight	衛生	health
水果	fruit	自由行	self-guided travel	電信	telecommunications
選舉	election	商簽	business visa	巴拿馬	Panama

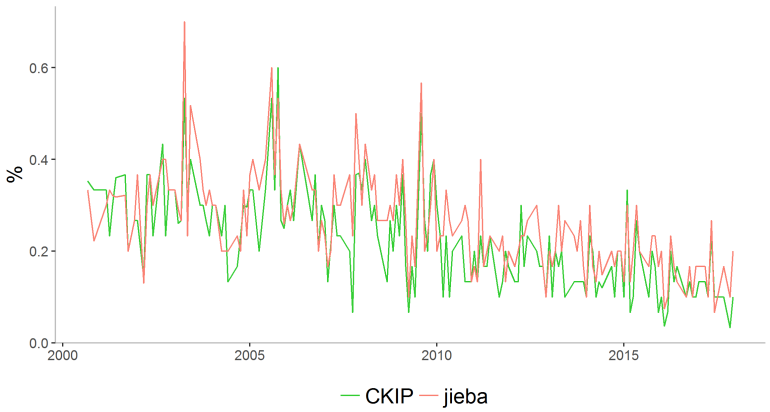


Figure 3.7: The percentage of common keywords between the algorithm based on word frequency and the algorithm based on TF-IDF from 2000 to 2017.

Figure 3.8 displays the difference in media ideology under the two algorithms for all 12 newspapers from 2000 to 2014. The media ideology based on the TF-IDF algorithm is significantly lower (p-value < 2.2e−16), but the final measures of media ideology under the two algorithms are highly correlated across newspapers over time (corr = 0.78).

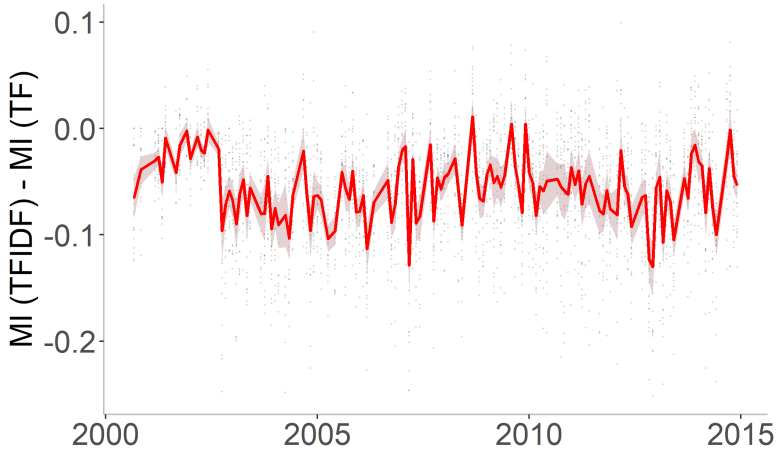


Figure 3.8: The difference in the final measures of media ideology between the algorithm based on term frequency and the algorithm based on TF-IDF from 2000 to 2014.

3.4 The Context of Keywords

To study the context of each keyword in the lexicon, I investigate the neighboring words around each keyword. A neighboring word is defined as a content word comprising at least two characters, existing within the same sentence (or the same headline) as a keyword, and appearing more than once in a month. Figure 3.9 shows the number of neighboring words per keyword from 2000 to 2017. On average, there are ca. 44.21 neighboring words for each keyword.

There are several advantages to consider the context of keywords. First, the perspective on a topic can be inferred from the context. For instance, one of the most commonly used words in the lexicon is "台獨" (Taiwan Independence). Despite its high frequency, "台獨" (Taiwan Independence) is by no means advocated by the Chinese government. Through examining the neighboring words of "台獨" (Taiwan Independence) such as "犯罪行為"

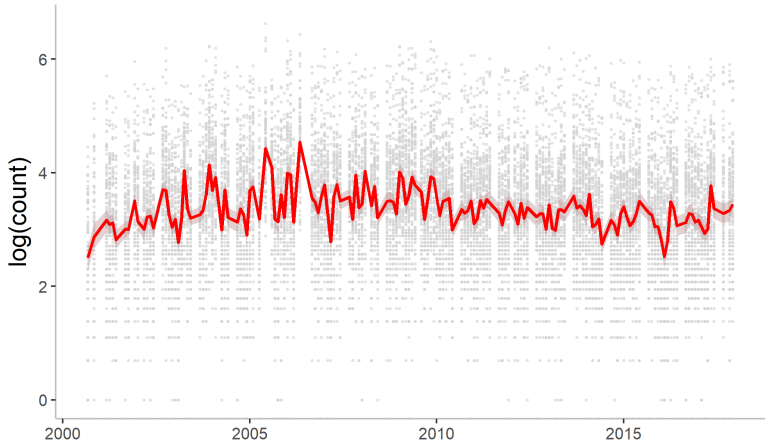


Figure 3.9: The number of neighboring words per keyword from 2000 to 2017.

(criminal behavior), "反對" (oppose), and "拋棄" (abandon), it is clear that the Chinese government holds a negative view on "台獨" (Taiwan Independence).

Figure 3.10 shows the average sentiment scores of the neighboring words for each keyword from 2000 to 2017. I first use *Google Translate*[1] to translate each neighboring word to English. Then, I use the most popular Python package in Natural Language Processing, NLTK, to assign a sentiment score to each translated word, which ranges from -1 to +1 (Hutto & Gilbert, 2014). If a word is given a higher sentiment score, then the word is considered more positive. The average sentiment score for all neighboring words across different keywords over time is 0.028. Suppose a positive sentiment score is classified as "Positive", a negative sentiment score is considered "Negative", and a zero sentiment score is labeled as "Neutral". Overall, there are 86.02% neutral words, 3.32% negative words, and 10.66% positive words. In Figure 3.10, the highlighted green dots represent the average sentiment scores of neighboring words for the keyword "台獨" (Taiwan Independence), which are mostly below the red line, the average sentiment scores of the neighboring words for all keywords. The pattern further corroborates the point that the Chinese government tends to use negative neighboring words around the keyword "台獨" (Taiwan Independence)

[1]The translation is implemented by a Python package called googletrans (https:// github.com/ ssut/ py-googletrans).

and does not support “台獨” (Taiwan Independence). Table 3.5 and Table 3.6 display top 20 keywords with the most positive and negative neighboring words, respectively.

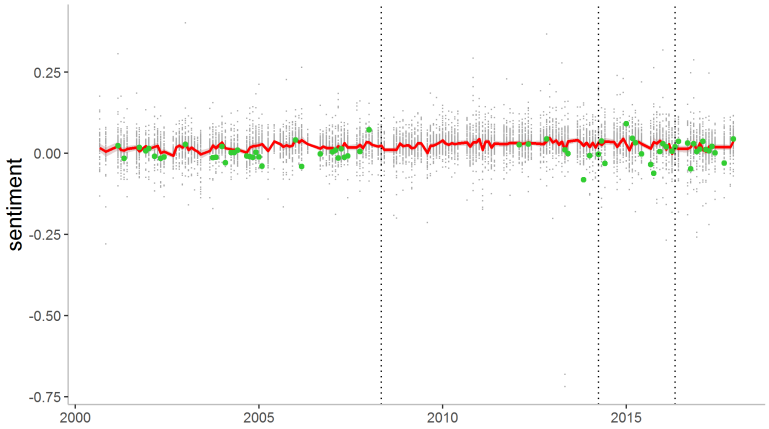


Figure 3.10: The average sentiment scores of neighboring words per keyword from 2000 to 2017.

Note: The red line denotes the average sentiment scores of neighboring words for all keywords from 2000 to 2017, while the green dots stand for the average sentiment scores of neighboring words for the keyword “台獨” (Taiwan Independence).

Secondly, analyzing context helps avoid the confusion from lexicon differences between the official language in Taiwan, Taiwanese Mandarin (國語) and the official language in China, Putonhua (普通話). Although both languages are classified as Mandarin, there are many “false friends”, i.e. words that sound or look similar but with different meanings, such as “小姐” (Miss in Taiwanese Mandarin vs. prostitute in Putonhua), and it is easier to distinguish one from another if neighboring words are considered. Thirdly, it helps alleviate the negative effects of word segmentation errors on the measure. Word segmentation is a complicated task and mistakes cannot be completely avoided. For example, “台中國小” should be segmented as “台中/國小” (Taichung Elementary School) but is wrongly interpreted as “台/中國/小” (Taiwan China small) by jieba. Since “中國” (China) is a keyword in the lexicon, the mistake may be perceived as a sign of similarity between a Taiwanese newspaper and TAO, and introduce a bias to the measure. Nevertheless, when context is taken into account, the measure should not be much affected because the neighboring words around an elementary school are very different from the ones around China. Lastly, it helps eliminate ambiguities from homographs, i.e. words with the same written form but

Table 3.5: Top 20 positive keywords by the average sentiment scores of neighboring words

keyword	translation	keyword score	neighbor score
特事	special matter	0.4019	0.4019
民共	DPP and CCP	0.4019	0.3673
三句話	three sentences	0.0516	0.306967
獎牌	medals	0	0.29295
愛國主義者	patriots	0	0.287433
習蕭會	Xi-Xiao meeting	0	0.2787
吉祥鳥	lucky bird	0.4215	0.26455
吉祥	auspicious	0	0.26455
自豪	pride	0.34	0.263225
降格	degrade	0	0.24695
依法治國	rule by law	0	0.239267
撈取	gain	0.5267	0.2294
外貌	appearance	0	0.2268
相向	face-to-face	0	0.2202
飛航點	flight point	0	0.212133
體育健兒	sportsman	0	0.208095
倍加	double	0	0.2027
健兒	athlete	0	0.190382
成績	achievement	0	0.186682
經貿關係	economic and trade relations	0	0.185022

Table 3.6: Top 20 negative keywords by the average sentiment scores of neighboring words

keyword	translation	keyword score	neighbor score
凶手	murderer	-0.6808	-0.7184
嚴懲	punish severely	-0.7184	-0.6808
愚蠢	silly	0.0258	-0.2793
晚集	evening market	0	-0.267933
註定	doomed	-0.6369	-0.243941
失敗	failure	-0.5106	-0.221386
祭典	festival	0.4939	-0.2202
困惑	confused	-0.3182	-0.20095
極端分子	extremist	0	-0.199437
最低	lowest	-0.3818	-0.191286
一己之私	selfishness	-0.4019	-0.188486
身心	body and mind	0	-0.182139
何在	where	0	-0.1806
遣送回	deportation	0	-0.172983
很小	very small	0	-0.171029
堅強意志	strong will	0.5106	-0.166767
台灣籍	Taiwanese	0	-0.166356
地步	extent	0	-0.165854
金剛	King Kong	0	-0.16495
交代	confess	0	-0.1629

different meanings, such as 易 (easy vs. change) and 長 (long vs. elder). Since homographs tend to have different contexts, examining the neighboring words around a homograph can help identify the word and solve the issue.

3.4.1 Positive Pointwise Mutual Information (PPMI)

To systematically examine the context of each keyword, I first calculate one of the most popular measures on word association, Positive Pointwise Mutual Information (PPMI):

$$PPMI_{i,j} = \max\left\{log_2 \frac{p(i,j)}{p(i)p(j)}, 0\right\}$$

$$\text{PPMI}_i = (PPMI_{i,1}, PPMI_{i,2}, \ldots, PPMI_{i,J})$$

PPMI compares $p(i,j)$, the actual joint distribution of keyword i and neighboring word j, to $p(i) \times p(j)$, the joint distribution under the assumption that keyword i and neighboring word j are independent from each other. If neighboring word j often appears with keyword i, PPMI will be positive and it is called word co-occurrence. Table 3.7 displays the top 25 neighboring words of keyword "台獨" (Taiwan Independence) along with the average probabilities and PPMI in the TAO press conference transcripts.[1] Take two neighboring words, "台灣" (Taiwan) and "決不允許" (never allow) as an example. The two neighboring words have comparable average joint probabilities in the TAO press conference transcripts, but "台灣" (Taiwan) has a much higher average marginal probability and expected joint probability with "台獨" (Taiwan Independence) than "決不允許" (never allow) does. As a result, the measure of PPMI for "台灣" (Taiwan) is much lower than the one for "決不允許" (never allow).

Using PPMI helps mitigate several potential problems. Firstly, the number of news stories included in the database is unbalanced between small newspapers and major newspapers. As a result, major newspapers have much larger text corpora than small newspapers.

[1] All the values are the average of monthly probabilities or PPMI from 2000 to 2014.

Table 3.7: Top 25 neighboring words for keyword "台獨" (Taiwan Independence) in the TAO press conference transcripts

j	translation	prob(i,j)	prob(j)	PPMI(i,j)
個別	individual	0.001881215	0.00268745	7.024973407
分裂主義	separatism	0.001698754	0.001698754	7.464341533
鐵桿	hardcore	0.001688316	0.001688316	8.045384248
分裂	split	0.001608869	0.002545965	7.383177579
主義	doctrine	0.001549587	0.00464876	6.011972642
黨員	party members	0.001147734	0.003730134	5.897535781
犯罪行為	criminal behavior	0.001132503	0.002265006	6.464341533
個別人	Individuals	0.001126071	0.001126071	7.474044311
不可動搖	unshakeable	0.001018849	0.001018849	8.616916363
活動	activity	0.001017438	0.002376222	6.741321055
言行	words and deeds	0.001011893	0.001011893	6.924268685
陳水扁	Chen Shui-bian	0.000987658	0.004620135	5.512937593
台聯黨	Taiwan Solidarity Union	0.00094436	0.002833081	6.005567734
不斷	constantly	0.000925497	0.001388246	6.405058015
民進黨	DPP	0.000898096	0.003590558	6.130260268
黨綱	party program	0.000849892	0.000849892	7.963991174
分子	activists	0.000832075	0.000901085	8.059182288
反對	oppose	0.000816797	0.001600096	7.102743516
法律依據	legal basis	0.000810701	0.000810701	7.361651405
公開	public	0.000795441	0.000795441	7.857236131
拋棄	abandon	0.000781514	0.000781514	7.575121168
本質	essence	0.000757406	0.000757406	7.452110141
依據	in accordance with	0.000755715	0.000944644	7.141051248
台灣	Taiwan	0.000754512	0.025816816	2.630092519
決不允許	never allow	0.000753012	0.000753012	8.053111336

If I use word frequency to measure media ideology, then small newspapers are deemed less pro-China simply because of fewer news stories. Figure 3.11 and Figure 3.12 display the number of news stories and text corpus size for each newspaper from 2000 to 2014, respectively. For example, 10,621.14 news stories from the *United Daily News* are included each month, which is 8.34 times larger than the number of news stories from the *Keng Sheng Daily News*. Similarly, the average text corpus size of the *Keng Sheng Daily News* is 12,341.78 words each month, which is less than 10% of the corpus size of the *United Daily News*. Using PPMI lessens the influence since it uses probabilities instead of raw frequency. Secondly, the number of news stories from a newspaper grows over time in the database. In 2000, the average number of stories is 293.55 for each newspaper in a month, but the number surges to 6050.35 in 2014. Using PPMI reduces the concerns for intertemporal comparison of media ideology.

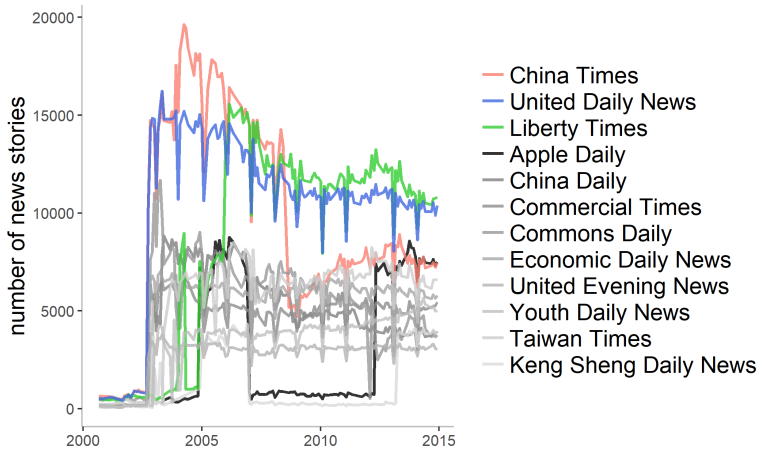


Figure 3.11: The number of news stories for each newspaper over time.

Figure 3.13 displays the average word association statistics per keyword for each newspaper over time, where the legend is based on the total number of words ranging from 377 to 214,949 words for a newspaper in a month. Figure 3.13a demonstrates that newspapers with larger total number of words tend to have higher raw frequencies, and the total number of words is positively correlated to raw frequencies across newspapers over time (corr = 0.43). However, this pattern does not exist in Figure 3.13b. Newspapers with small corpus size can have high values of PPMI, and the total number of words is not correlated to PPMI across newspapers over time (cor = 0.09).

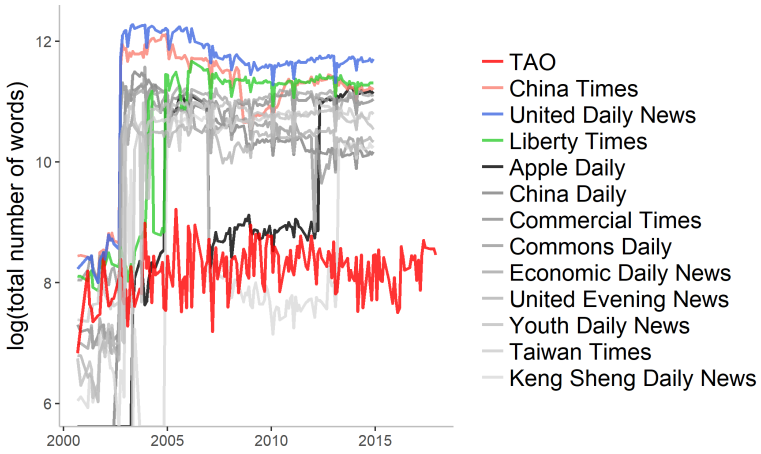


Figure 3.12: The text corpus size for each newspaper and TAO over time.

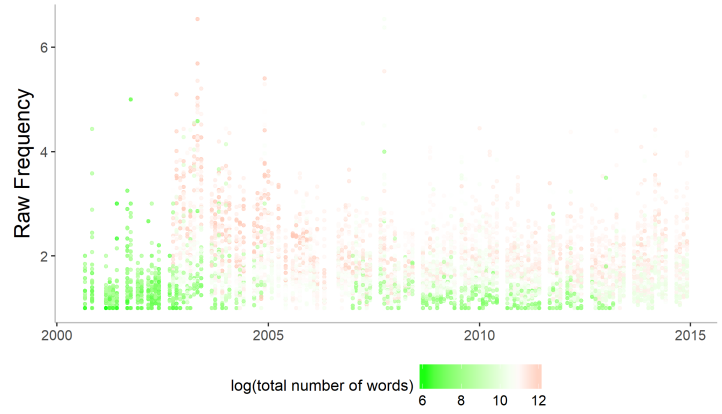
Lastly, PPMI performs better than raw frequency in terms of the comparison across different types of text, such as press conference transcripts and newspaper headlines. Press conference transcripts and newspapers headlines are utterly different types of text. Press conferences transcripts document complete conversations between a spokesperson and reporters, while newspaper headlines summarize news stories and convey information with much shorter texts. It is hard to justify the comparison between press conference transcripts and newspaper headlines, but PPMI is more comparable across different types of texts than word frequency.

3.4.2 T-TEST

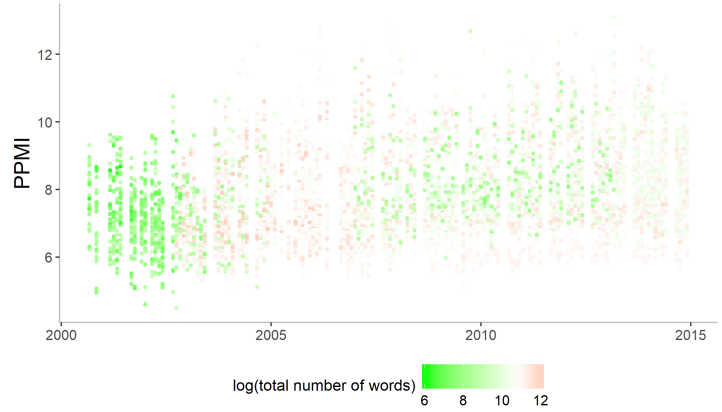
In addition to PPMI, I also calculate another popular measure, t-test, defined as follows (Curran, 2003; Jurafsky & Martin, 2000):

$$\text{t-test}_{i,j} = \frac{p(i,j) - p(i)p(j)}{\sqrt{p(i)p(j)}}$$

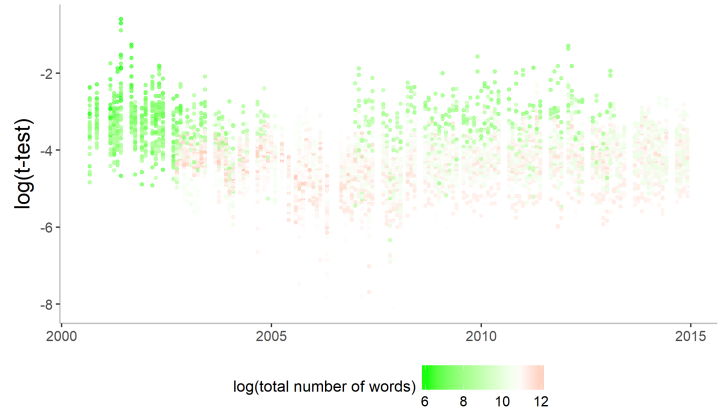
$$\text{t-test}_i = (\text{t-test}_{i,1}, \text{t-test}_{i,2}, \ldots, \text{t-test}_{i,J})$$



(a) Raw Frequency



(b) PPMI



(c) t-test

Figure 3.13: The average word association statistics per keyword across Taiwanese newspapers over time.

Similar to PPMI, the t-test statistic compares the actual joint distribution of keyword i and its neighboring word j, $p(i,j)$, to the expected joint distribution assuming keyword i and neighboring word j are independent, $p(i)p(j)$, and then normalizes the difference by the standard deviation, $\sqrt{p(i)p(j)}$. A higher t-test statistic indicates greater likelihood of word association between keyword i and neighboring word j. Given that the concept is quite comparable between PPMI and t-test, the two measures are highly correlated across neighboring words for all keywords in the lexicon among Taiwanese newspapers and TAO press conference transcripts over time (corr = 0.78). Figure 3.13c displays the average log value of t-test statistic per keyword for each newspaper over time. Likewise, the t-test statistic is based on probabilities and does not suffer from the unbalanced text corpus problem. Newspapers with relatively small text corpora can have high t-test statistics as well.

Figure 3.14 displays the difference between the final measure of media ideology based on the t-test association measure and the one based on PPMI across all Taiwanese newspapers from 2000 to 2014. On average, the measure of media ideology based on the t-test statistic is lower than the one based on PPMI. Nevertheless, the two measures are highly correlated across newspapers over time as well (corr = 0.85).

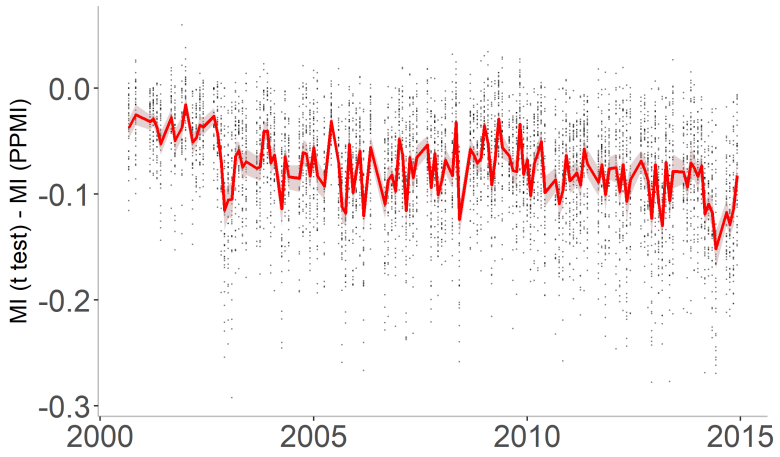


Figure 3.14: The difference in the final measures of media ideology between the two word association measures, t-test and PPMI, from 2000 to 2014.

3.4.3 Raw Frequency

As a robustness check, I calculate the word association measure, $frequency_{i,j}$, the count of keyword i and its neighboring word j at the same time. Intuitively, when a neighboring word often appears with a keyword, the two words are more likely to be associated. The word association measure based on raw frequency is moderately correlated to the one based on PPMI across neighboring words for all keywords among Taiwanese newspapers and TAO press conference transcripts over time (corr = 0.52).

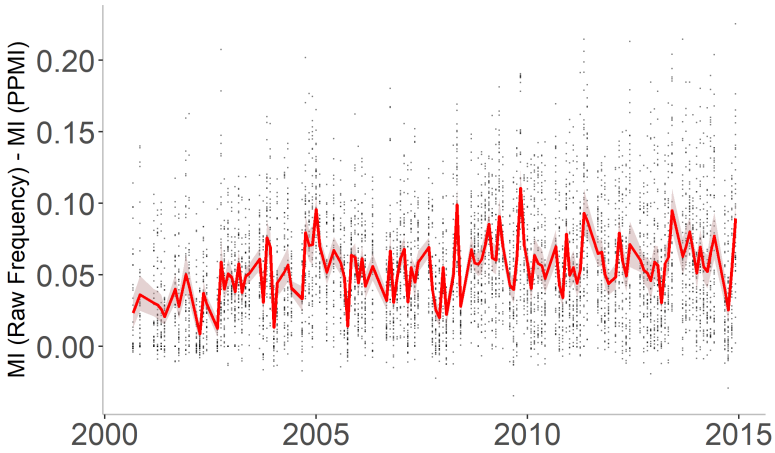


Figure 3.15: The difference between the measures of media ideology based on the two word association measures, raw frequency and PPMI, from 2000 to 2014.

Figure 3.15 displays the difference between the final measure of media ideology based on raw frequency and the one based on PPMI across all Taiwanese newspapers from 2000 to 2014. On average, the measure of media ideology based on raw frequency is higher than the one based on PPMI, but the two measures are very highly correlated across newspapers over time (corr = 0.95).

3.5 Language Comparison

3.5.1 Cosine Similarity

To compare the patterns of keywords and neighboring words between the Chinese government and each Taiwanese newspaper, I first calculate the most popular similarity metric, cosine similarity, defined as follows:

$$
\begin{aligned}
\mathrm{v}_{i,k} &\in \left\{ \mathrm{PPMI}_{i,k}, \mathrm{t-test}_{i,k}, \mathrm{frequency}_{i,k} \right\} \\
\cos\left(\vartheta_{i,k}\right) &= \frac{\mathrm{v}_{i,k}\mathrm{v}_{i,TAO}}{\|\mathrm{v}_{i,k}\|\|\mathrm{v}_{i,TAO}\|} \\
MI_k &= \frac{1}{I}\sum_{i \in I} \cos\left(\vartheta_{i,k}\right)
\end{aligned}
$$

The cosine similarity metric measures the angle $\vartheta_{i,k}$ between the vector of word association measures regarding keyword i, such as $\mathrm{PPMI}_{i,k}$, for newspaper k and the vector for the TAO. The cosine similarity compares the directions of the two vectors and does not discriminate against vectors with different levels of frequencies.

If newspaper k uses keyword i with its neighboring words in a similar pattern as the Taiwan Affairs Office does, then the cosine similarity regarding keyword i is high. Given that raw frequency and PPMI values are non-negative, the corresponding cosine similarity values always lie between 0 and 1.

Table 3.8: Three neighbors of the keyword " 台獨" (Taiwan Independence) and the corresponding average PPMI values for the TAO and four major Taiwanese newspapers.

neighbor	translation	TAO	United Daily News	China Times	Liberty Times	Apple Daily
危險性	risk	9.24	9.41	0.00	0	0
鐵桿	iron poles	8.05	7.87	7.62	0	0
老鞋	old shoes	8.66	14.30	12.85	0	0

Table 3.8 illustrates an example with the keyword "台獨" (Taiwan Independence) and three neighboring words, "危險性" (risk), "鐵桿" (iron poles), and "老鞋" (old shoes). In

addition to its literal meaning, "鐵桿" (iron poles) can be used as a metaphor to describe someone who is hardcore and stubborn. The connotation is widely used in China, but rarely used in Taiwan. "老鞋" (old shoes) is a specific metaphor that was invented by the Chinese government to criticize "台獨" (Taiwan Independence) as being old-fashioned and leading to a dead end. Despite that the two words are not common phrases in Taiwan and almost never used with "台獨", the most pro-China newspapers, the *United Daily News* (聯合報, UDN) and the *China Times* (中國時報), also use these neighboring words with "台獨" (Taiwan Independence) in the headlines and the corresponding PPMI measures are positive. On the contrary, the least pro-China newspapers, the *Apple Daily* (蘋果日報) and the *Liberty Times* (自由時報), never use these words with "台獨" (Taiwan Independence) and the PPMI measures are all zeros for the three words.

Figure 3.16: The PPMI measures related to the keyword "台獨" (Taiwan Independence) and all of its neighboring words from 2000 to 2014.

Figure 3.16 shows the PPMI measures related to the keyword "台獨" (Taiwan Independence) and all of its neighboring words for the TAO and 12 Taiwanese newspapers over time. Since the neighboring words are chosen from the TAO press conference transcripts, it is not surprising that the TAO PPMI measures are constantly higher than the measures of Taiwanese newspapers. Figure 3.16 also demonstrates that the most pro-China newspapers, the *United Daily News* (聯合報, UDN) and the *China Times* (中國時報), have relatively higher PPMI measures compared to other Taiwanese newspapers over time.

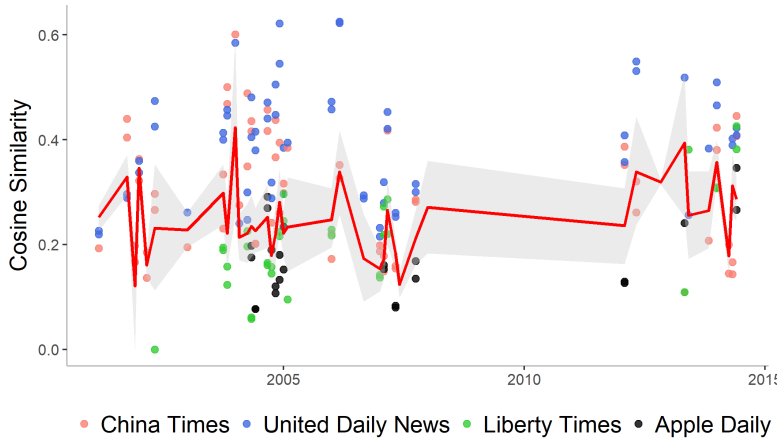


Figure 3.17: The cosine similarity measures related to the keyword "台獨" (Taiwan Independence) from 2000 to 2014.

Figure 3.17 displays the cosine similarity measures regarding the keyword "台獨" (Taiwan Independence) over time. The cosine similarity measures of the most pro-China newspapers, the *United Daily News* (聯合報, UDN) and the *China Times* (中國時報), almost always stay above the average cosine similarity measures for all Taiwanese newspapers, while the least pro-China newspapers, the *Apple Daily* (蘋果日報) and the *Liberty Times* (自由時報), have below-the-average cosine similarity measures over time.

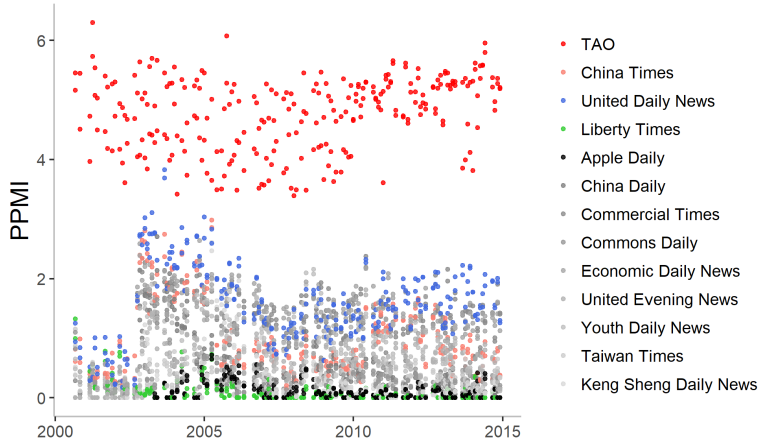


Figure 3.18: The PPMI measures related to the keyword "大陸" (mainland) and all of its neighboring words from 2000 to 2014.

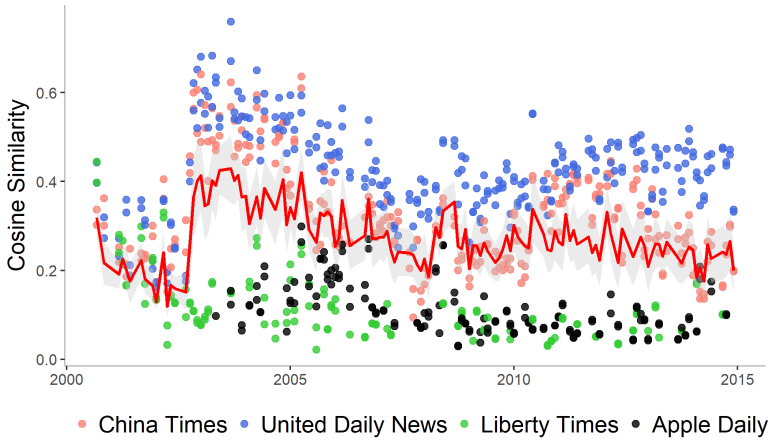


Figure 3.19: The cosine similarity measures related to the keyword "大陸" (mainland) from 2000 to 2014.

Figure 3.18 and Figure 3.19 illustrate another example of the keyword "大陸" (mainland). There are several names of the People's Republic of China (PRC) in Taiwan, such as "中國" (China), "大陸" (mainland, or continent), "內地" (inland), and "祖國" (motherland), and they contain different connotations. If one uses "中國" (China) to describe a piece of news about the PRC but not about Taiwan, this implies that the newspaper does not regard Taiwan as part of China. Alternatively, if a newspaper uses "大陸" (mainland, or continent), then it does not exclude Taiwan from China because mainland and Taiwan can both be parts of China.[1] Figure 3.18 displays the PPMI measures related to the keyword "大陸" (mainland, or continent) and its neighboring words over time. The PPMI measures for the most pro-China newspapers, the *United Daily News* (聯合報, UDN) and the *China Times* (中國時報), are always closer to the TAO than the ones for the least pro-China newspapers, the *Apple Daily* (蘋果日報) and the *Liberty Times* (自由時報). Furthermore, Figure 3.19 shows that the cosine similarity measures regarding the keyword "大陸" (mainland, or continent) for the two least pro-China newspapers are constantly below the average measures across all newspapers, whereas the two most pro-China newspapers are always above the average measures, suggesting that pro-China newspapers tend to use "大陸" (mainland, or continent) in a similar fashion as the TAO does.

[1] If a newspaper calls the PRC "內地" (inland) or "祖國" (motherland), then it assumes Taiwan is part of China. However, they are rarely used in newspaper headlines and thus not presented here.

The final measure of media ideology is the average cosine similarity metrics for all keywords in the lexicon. If the word association measure is based on PPMI or frequency (non-negative by construction), then the final measure of media ideology is between 0 and 1. A higher value indicates more similar use of the language between a Taiwanese newspaper and the Chinese government. Following the hypothesis, this implies that the Taiwanese newspaper adopts a closer stance to the Chinese government. Figure 3.20 displays the measures of media ideology based on the cosine similarity from 2000 to 2014[1], where the highlighted dots refer to the four largest newspapers, the *United Daily News* (聯合報, UDN), the *China Times* (中國時報), the *Apple Daily* (蘋果日報) and the *Liberty Times* (自由時報), and the red line indicates the average measures of media ideology of the 12 newspapers. In line with expectations, the media ideology measures of the allegedly most pro-China Taiwanese newspapers, the *United Daily News* (聯合報, UDN) and the *China Times* (中國時報), are almost always above the average red line and much higher than media ideology measures of the presumably least pro-China Taiwanese newspapers, the *Apple Daily* (蘋果日報) and the *Liberty Times* (自由時報).

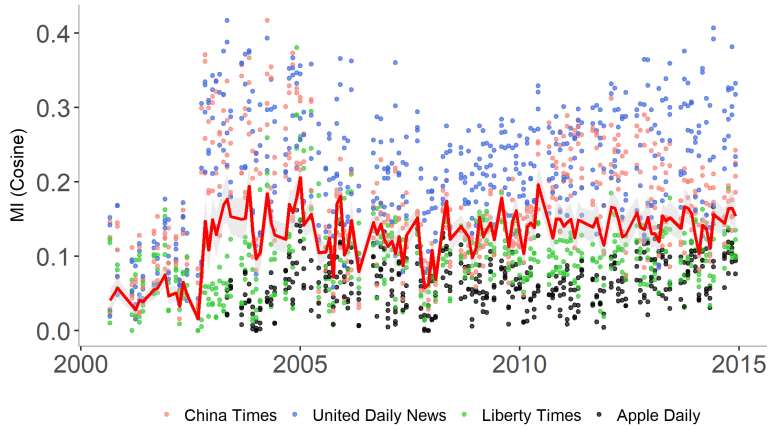


Figure 3.20: The final measures of media ideology based on the cosine similarity from 2000 to 2014.

[1]The measure is based on the PPMI word association metric across the two word segmentation methods, CKIP and jieba, and across the two keyword extraction algorithms based on word frequency and TF-IDF.

Table 3.9: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
United.Daily.News	133	0.267	0.074	0.056	0.417
Economic.Daily.News	133	0.219	0.076	0.010	0.343
Commercial.Times	133	0.216	0.072	0.017	0.343
Youth.Daily.News	133	0.203	0.078	0.018	0.330
China.Times	133	0.203	0.075	0.068	0.367
China.Daily	119	0.147	0.061	0.012	0.352
United.Evening.News	133	0.144	0.065	0.000	0.270
Commons.Daily	133	0.136	0.050	0.017	0.334
Taiwan.Times	133	0.131	0.051	0.016	0.251
Liberty.Times	133	0.121	0.042	0.030	0.295
Keng.Sheng.Daily.News	119	0.092	0.077	0.000	0.342
Apple.Daily	111	0.079	0.031	0.003	0.162

3.5.2 Jaccard Similarity

In addition to the cosine similarity, I also calculate the Jaccard similarity (Grefenstette, 1994; Jurafsky & Martin, 2000) defined as follows:

$$
\begin{aligned}
\mathrm{v}_{i,j,k} &\in \left\{ \mathrm{PPMI}_{i,j,k}, \text{t-test}_{i,j,k}, \text{frequency}_{i,j,k} \right\} \\
\text{Jaccard}_{i,k} &= \frac{\sum_{j \in J} \min \left\{ \mathrm{v}_{i,j,k}, \mathrm{v}_{i,j,TAO} \right\}}{\sum_{j \in J} \max \left\{ \mathrm{v}_{i,j,k}, \mathrm{v}_{i,j,TAO} \right\}} \\
\mathit{MI}_k &= \frac{1}{I} \sum_{i \in I} \text{Jaccard}_{i,k}
\end{aligned}
$$

The Jaccard similarity metric calculates the weighted number of shared attributes (intersection) regarding the use of keyword i between newspaper k and the TAO, where the weight is the values of word association between keyword i and its neighboring words, such as PPMI. Then, the metric normalizes the intersection by the set of all attributes (union) regarding keyword i between newspaper k and the TAO, weighted by the values of word

association regarding keyword *i*. If newspaper *k* has more similar attributes about keyword *i* to the TAO, the Jaccard similarity measure is higher and newspaper *k* is considered more pro-China regarding keyword *i*. Figure 3.21 displays the difference between the cosine similarity and the Jaccard similarity across all keywords and all newspapers over time, using the PPMI word association measure. The values of Jaccard similarity are always lower than the ones of cosine similarity, but the two are very highly correlated using the PPMI word association measure (corr = 0.96) and moderately correlated across different word association metrics (corr = 0.64).

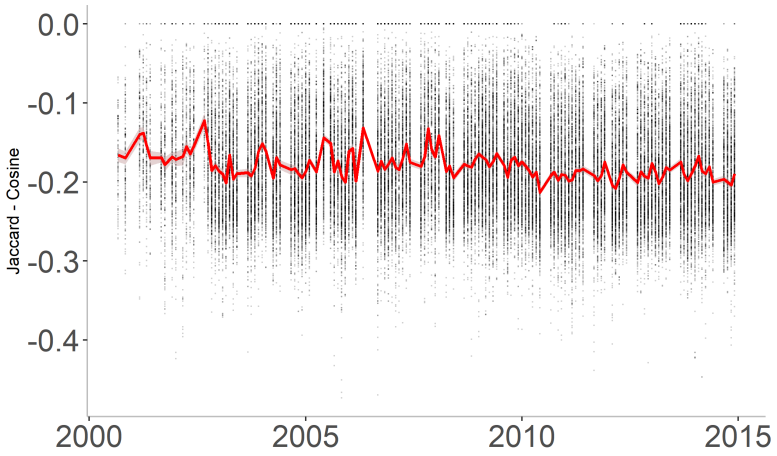


Figure 3.21: The difference between the two similarity metrics, cosine and Jaccard, across all keywords from 2000 to 2014.

The final measure of media ideology is the average Jaccard similarity metric for all keywords in the lexicon. Figure 3.22 displays the final measure of media ideology based on the Jaccard similarity over time. Consistent with the measures of media ideology based on the cosine similarity, the two most pro-China newspapers, the *United Daily News* (聯合報, UDN) and the *China Times* (中國時報), are always higher in the values of Jaccard-based media ideology than the least pro-China Taiwanese newspapers, the *Apple Daily* (蘋果日報) and the *Liberty Times* (自由時報). Figure 3.23 shows the difference between the measures of media ideology based on the Jaccard similarity and the cosine similarity over time. On average, the Jaccard-based measure of media ideology is lower than the cosine-based one, but the measures of media ideology based on the two similarity metrics are highly cor-

related across different word segmentation methods, keyword extraction algorithms, and word association metrics for all newspapers over time (corr = 0.77).

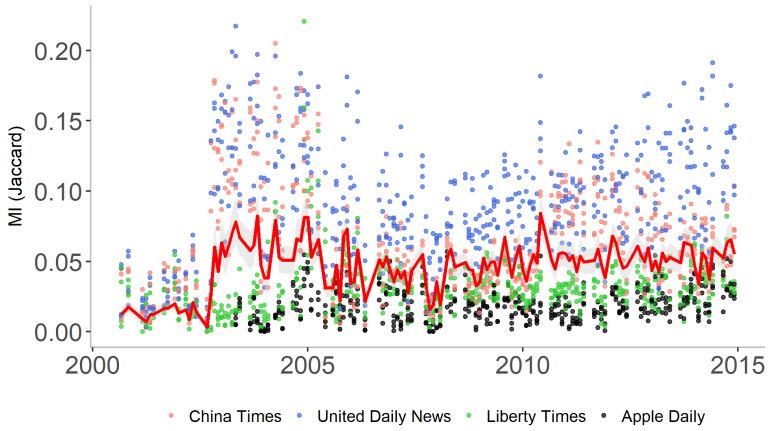


Figure 3.22: The final measures of media ideology based on the Jaccard similarity from 2000 to 2014.

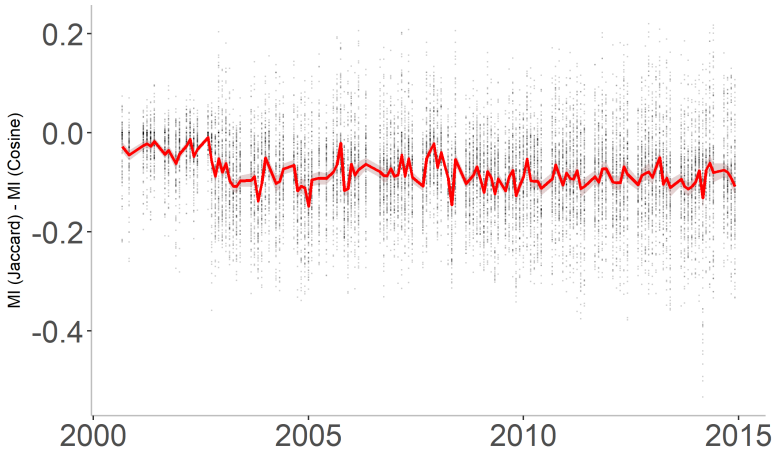


Figure 3.23: The difference between the measures of media ideology based on the two similarity metrics, cosine and Jaccard, from 2000 to 2014.

3.5.3 Dice Similarity

The Dice measure is similar to the Jaccard measure and defined as follows(Curran, 2003; Jurafsky & Martin, 2000):

$$
\begin{aligned}
\mathrm{v}_{i,j,k} &\in \left\{ \mathrm{PPMI}_{i,j,k}, \text{t-test}_{i,j,k}, \text{frequency}_{i,j,k} \right\} \\
\mathrm{Dice}_{i,k} &= \frac{\sum_{j \in J} \min \left\{ \mathrm{v}_{i,j,k}, \mathrm{v}_{i,j,TAO} \right\}}{\sum_{j \in J} \frac{1}{2} \left\{ \mathrm{v}_{i,j,k} + \mathrm{v}_{i,j,TAO} \right\}} \\
\mathit{MI}_k &= \frac{1}{I} \sum_{i \in I} \mathrm{Dice}_{i,k}
\end{aligned}
$$

The formula of the Dice measure is the same as the one of the Jaccard measure except that the normalization factor in the denominator is the average of the weighted number of the attributes regarding keyword *i* for both newspaper *k* and the TAO. The interpretation of the Dice similarity is the same as the ones of the cosine similarity and the Jaccard similarity. The Dice similarity ranges between 0 and 1 for the PPMI word association measure, and a higher value of the Dice similarity represents a closer stance to the Chinese government. Figure 3.24 displays the difference between the cosine similarity and the Dice similarity across all keywords and all newspapers over time, using the PPMI word association measure. Similar to the Jaccard similarity, the values of Dice similarity are lower than the ones of cosine similarity, but the two are very highly correlated using the PPMI word association measure (corr = 0.97). However, the Dice similarity and the cosine similarity only have a low correlation across different word association metrics (corr = 0.34). After removing 0.05% outliers, the two measures of similarity become moderately correlated (corr = 0.68).

The final measure of media ideology is the average Dice similarity metric for all keywords in the lexicon. Figure 3.25 displays the final measure of media ideology based on the Dice similarity over time. The pattern is similar to the cosine similarity and the Jaccard similarity. Figure 3.26 shows the difference between the measures of media ideology based on the Dice similarity and the cosine similarity over time. On average, the Dice-based measure of media ideology is lower than the cosine-based one, but the two similarity metrics are highly correlated across different word segmentation methods, keyword extraction algorithms, and word association metrics for all newspapers over time (corr = 0.75).

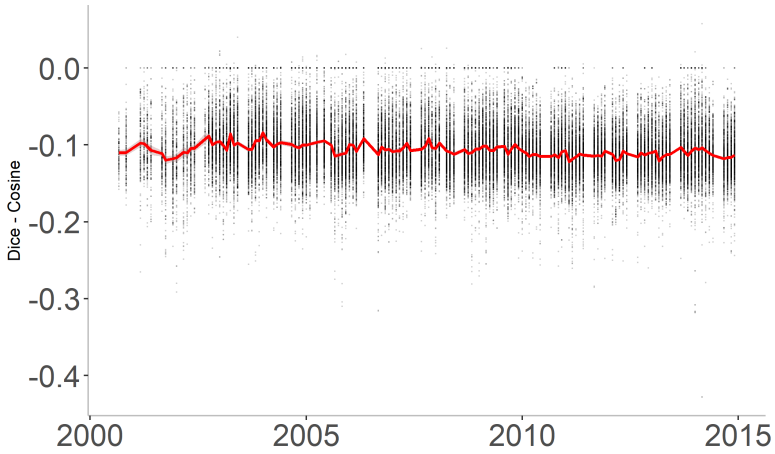


Figure 3.24: The difference between the two similarity metrics, cosine and Dice, across all keywords from 2000 to 2014.

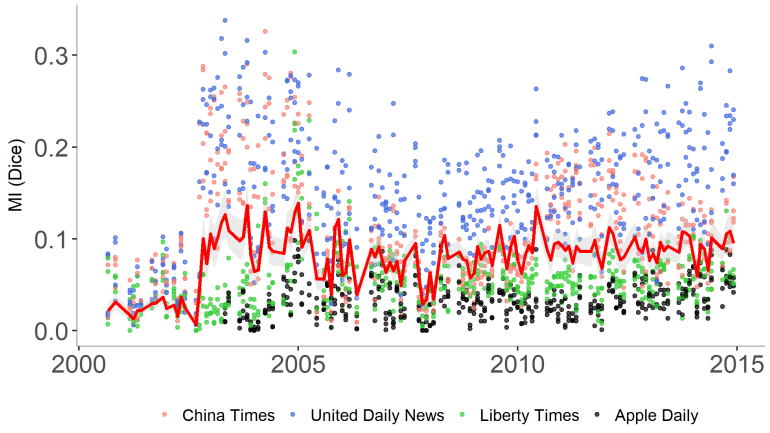


Figure 3.25: The final measure of media ideology based on the Dice similarity from 2000 to 2014.

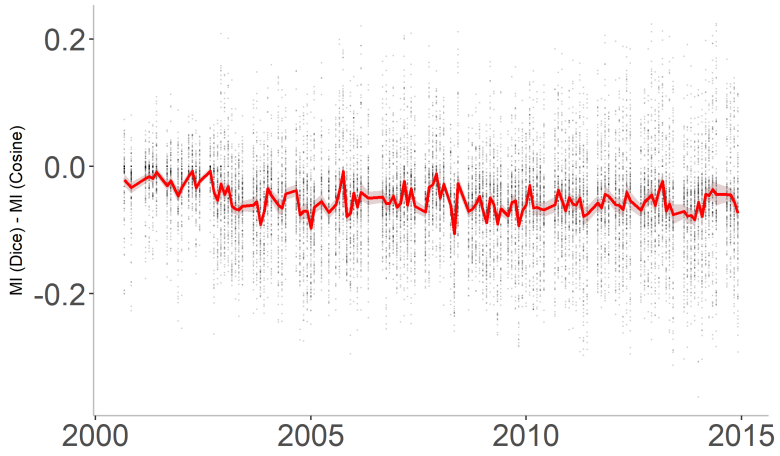


Figure 3.26: The difference between the measures of media ideology based on the two similarity metrics, cosine and Dice, from 2000 to 2014.

Note: In the graph, two outliers are removed. The dice measure of the two observations are -2.63 and -2.64, respectively. The average value of dice measure is 0.06, and the standard deviation is 0.06.

3.6 Conclusion and Discussion

This chapter discusses the construction of my media ideology measure in the first chapter, with details in word segmentation methods, keyword extraction algorithms, word association measures, and similarity metrics. Conceptually, my measure of media ideology compares the use of language in Taiwanese newspapers and in press conference transcripts of the Taiwan Affairs Office (TAO), the official Chinese government institution in charge of policies related to Taiwan. Based on the comparison, my measure evaluates how closely each Taiwanese newspaper aligns with TAO. If a Taiwanese newspaper shares similar views with TAO, then it is assigned a high score of my measure and considered more pro-China.

The first task of my measure is to find keywords to characterize the most pro-China stance represented by TAO. The task begins with word segmentation, which divides a piece of text into the smallest meaningful units–words. I adopt two different tools to implement word segmentation, and 59.5% of the keywords in the monthly lexicons are identical between the two tools. Despite moderate overlap of keywords using the two tools, the values of my media ideology measure are very highly correlated across newspapers over time

(corr = 0.95). After implementing word segmentation, I adopt two different algorithms to extract keywords from the TAO press conference transcripts. The two algorithms select keywords based on distinctive ideas: The former chooses the most frequent keywords that illustrate fundamental enduring cross-Strait issues, while the latter selects keywords about issues that are very present in some months but not over longer time periods. Nevertheless, although the two algorithms are based on quite different ideas and there are only ca. 22.3-25.9% common keywords, the final values of my media ideology measure are highly correlated across Taiwanese newspapers over time (corr = 0.78).

One important feature of my measure is the consideration of keyword context, which helps detect different views on a topic, avoids misunderstandings from the cross-Strait lexicon differences, mitigates the negative influence of word segmentation errors, and minimizes ambiguities from homographs. To understand the context of each keyword in the monthly lexicons, I examine the neighboring words around each keyword and use three different word association statistics to assess the relationship between each keyword and its neighboring words. The three statistics include PPMI, t-test, and raw frequency, and their values are higher if a neighboring word is more likely to appear with a keyword. The results from Section 3.4 show that the three word association statistics are moderately correlated across neighboring words for all keywords over time (corr = 0.52-0.78), and the corresponding values of my measure of media ideology are highly correlated as well (corr = 0.85-0.95).

After extracting keywords to characterize the most pro-China stance and studying the context of each keyword, I calculate three different similarity metrics to compare the use of language between each Taiwanese newspaper and TAO and evaluate how pro-China each Taiwanese newspaper is. The three similarity metrics are the cosine similarity, Jaccard similarity, and Dice similarity, and if the value of a similarity metric is higher, it suggests that a Taiwanese newspaper is more similar to TAO. Section 3.5 demonstrates that the values of my measure of media ideology based on different similarity metrics are highly correlated for all newspapers over time (corr = 0.75-0.77).

It is worth mentioning that my measure compares TAO press conference transcripts in simplified Chinese with Taiwanese newspapers in traditional Chinese, which generates two types of errors. First, simplified Chinese is not directly comparable to traditional Chinese

and the comparison relies on a conversion tool called Open Chinese Convert (OpenCC)[1]. Sometimes it can produce unexpected errors because many different words in traditional Chinese map to a single word in simplified Chinese (and that's why it is called "simplified"). For example, "范" in simplified Chinese can be converted to "範" (model) or "范" (Fan, a family name). "范丽青" (Fan Liqing in simplified Chinese), the TAO spokesperson, is often wrongly converted into "範麗青" (Model Liqing in traditional Chinese). I have manually dealt with some common errors, including '范丽青" (Fan Liqing), but there may be other undetected mistakes and the resulting measure of media ideology can be downward biased.

Second, there are fundamental differences in the nature between the two types of text, press conference transcripts and newspaper headlines. Even if a newspaper holds exactly the same position as the Chinese government does, my measure of its media ideology would not be equal to the maximal value of pro-China ideology. Therefore, essentially, my measure can be interpreted as the relative ideological position of a newspaper, not the absolute resemblance between each Taiwanese newspaper and the Chinese government. That is, my measure can tell if a Taiwanese newspaper shares a more similar perspective with the Chinese government compared to other Taiwanese newspapers, but it does not show the absolute ideology of a newspaper regarding issues about cross-Strait relations.

[1]https://github.com/BYVoid/OpenCC.

Appendix A

Demographics in TEDS Surveys

Age is a potential determinant of national identity because a person's nationalist feelings may change over his/her lifetime and because people of different age experience different periods of national history. Since the dataset spans over 10 years, I also include year of birth to control for generation effects. On the one hand, it may be that the longer one lives in Taiwan, the deeper s/he loves Taiwan and the more likely s/he thinks s/he is a Taiwanese. As a result, compared to the elderly, young people may have not developed emotional connections with their living place. With potentially weaker nationalist feelings, they may be inclined to change their national identity under some exogenous or endogenous shocks. For example, suppose one has to move to China because of an unexpected event. The younger when s/he moves to China, the more likely that s/he will change her/his national identity and thinks s/he is a Chinese, not a Taiwanese. On the other hand, the change of history education may lead to greater Taiwanese identification among the young over the past two decades. History education had been focused on China in the postwar period. It was not until 1997 did every middle school student in Taiwan begin learning a textbook called *Understanding Taiwan*, which covered Taiwanese history, geography and society. Because *Understanding Taiwan* was a separate textbook from other textbooks about China, students might be prone to recognize Taiwan as a different and independent identity from China. In 2002, Taiwanese government undertook an education reform and adopted an "One guide-Multiple Text (一綱多本)" textbook system. National textbooks have been abolished since then, and textbooks are now edited based on an official guide–where Taiwanese history remains isolated from Chinese history. The Taiwan-isolated history education system

may influence the national identification of students and lead to an increase of Taiwanese identity among the young. As a result, whether age contributes to or hinders a burgeoning trend of Taiwanese identity is an empirical question.

Ethnicity is a complicated issue in Taiwan. There are four main ethnic groups: Taiwanese Min-Nan, Taiwanese Hakka, Aboriginal, and Mainlander. If I assume a person's ethnicity is defined by his/her father's ethnic background, 75% of people are Taiwanese Min-Nan, 13% are Taiwanese Hakka, 1% are Aboriginal, and 10% are Mainlander in the data set. The first three are native Taiwanese, while Mainlanders consist of aliens who moved from China around 1949 and their descendants born in Taiwan. The term "provincial complex" refers to the conflicts between native Taiwanese and Mainlander, which emerged around the 228 Massacre in 1947. The 228 Massacre was an anti-government movement beginning on February 27, 1947, which was violently suppressed by the KMT government and resulted in 18,000 to 28,000 or more deaths (Research Report of the 228 Incident, 1992). The essential cause of the movement was unequal power distribution between native Taiwanese and Mainlanders–according to Encyclopedia of Taiwan edited by Council for Cultural Affairs, the government was "almost exclusively populated by Mainlanders" who represented less than 6% of Taiwanese population at that time. Therefore, the 228 Massacre can be seen as antagonism between the government (Mainlanders) and the people (native Taiwanese). Many studies argue that the 228 massacre provides an impetus for the differentiation between Taiwanese and Chinese identity (Kao & Li, 2000; F.-c. Wang, 2005). On the one hand, because the first generation of Mainlanders are mostly born in China and have more connections with China, naturally, they are more likely to consider themselves as Chinese, not Taiwanese. On the other hand, the 228 Massacre pushed native Taiwanese further away from Chinese identification. Nevertheless, as the first generation of Mainlanders fade away and the successive generations are often born from intermarriage between Mainlanders and native Taiwanese, the provincial complex lessens over time. In 1995, former President Li Teng-Hui proposed the concept of "New Taiwanese," referring to anyone who loves Taiwan and calls Taiwan home. Currently, the provincial complex is less important and the boundary between native Taiwanese and Mainlanders becomes less salient. However, since the dataset starts as early as 2001, to control for any potential lagging effects, I include ethnicity of both parents as explanatory variables.

Language plays an essential role in the formation of identity and usually constitutes a crucial part in nation-building policies (Aspachs-Bracons et al., 2008; Clots-Figueras & Masella, 2013). Taiwan is a linguistically diverse society. The ancestral languages of Taiwanese Min-Nan and Taiwanese Hakka are Taiwanese Hokkien (also called Taiwanese) and Hakka, respectively. Aboriginal consist of 14 recognized tribes and speak dozens of non-sinitic languages. Originally from diverse regions of China, most Mainlanders speak their own Chinese dialects. Nevertheless, after the KMT's implementation of Mandarin-only policy in public administration, media broadcasts and education in 1945, Mandarin has become Taiwan's lingua franca across all ethnic groups. According to Benchmark Survey 2013 administered by TEDS, 79% of people use Mandarin to conduct interview and 12% use both Mandarin and Taiwanese. Meanwhile, 42% of people usually speak Taiwanese at home, 32% speak Mandarin, and 20% speak both Taiwanese and Mandarin, suggesting that Taiwanese and Mandarin become the two most prevalent languages in Taiwan. Since Mandarin is also the official language of China, it is possible that compared to Taiwanese speakers, Mandarin speakers are more likely to identify themselves as Chinese. Dupré (2013) uses Taiwan Social Change Survey on National Identity in 2003 and finds that people speaking Taiwanese at home are associated with more Taiwanese identification. To control for any potential effect of language on national identity, I include language spoken at home as an explanatory variable.

References

Abadie, A., & Gardeazabal, J. (2003). The Economic Costs of Conflict: A Case Study of the Basque Country. *93*(1), 113–132. doi:doi:10.1257/000282803321455188

Akerlof, G. A., & Kranton, R. E. (2000). Economics and Identity*. *Quarterly Journal of Economics*, *115*(3), 715–753. doi:10.1162/003355300554881

Alatas, V., Banerjee, A., Hanna, R., Olken, B. A., Purnamasari, R., & Wai-poi, M. (2013). Does Elite Capture Matter? Local Elites and Targeted Welfare Programs in Indonesia. *NBER Working Paper*, (February), 1–41. doi:10.3386/w18798

Amnesty International. (2017). *Amnesty International Report: China 2016/2017*. Retrieved from https://www.amnesty.org/en/countries/asia-and-the-pacific/china/report-china/

Anderson, B. (1991). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. Verso.

Apple Daily. (2012). Be Careful of the Illusion of Peace. (in Chinese). Newspaper Article.

Apple Daily. (2013). How to Communicate if Press Freedom is Denied (in Chinese). Newspaper Article.

Ashworth, J., Geys, B., Heyndels, B., & Wille, F. (2014). Competition in the political arena and local government performance. *Applied Economics*, *46*(19), 2264–2276. doi:10.1080/00036846.2014.899679

Aspachs-Bracons, O., Clots-Figueras, I., Costa-Font, J., & Masella, P. (2008). Compulsory Language Educational Policies and Identity Formation. *Journal of the European Economic Association*, *6*(2-3), 434–444. doi:10.1162/JEEA.2008.6.2-3.434

Autor, D. H. (2003). Outsourcing at Will: The Contribution of Unjust Dismissal Doctrine to the Growth of Employment Outsourcing. *Journal of Labor Economics*, *21*(1), 1–42. doi:10.1086/344122

Bardhan, P. (2002). Decentralization of Governance and Development. *Journal of Economic Perspectives*, *16*(4), 185–205. doi:10.1257/089533002320951037

Berman, L. (2017). China AQI Archive (Feb 2014 - Feb 2016). doi:doi / 10 . 7910 / DVN / GHOXXO

Besley, T., Persson, T., & Sturm, D. M. (2010). Political competition, policy and growth: Theory and evidence from the US. *Review of Economic Studies*, *77*(4), 1329–1352. doi:10.1111/j.1467-937X.2010.00606.x

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research*, *3*, 993–1022. Retrieved from http://dl.acm.org/citation.cfm?id=944919.944937

Cai, Y. (2008). Power Structure and Regime Resilience: Contentious Politics in China. *British Journal of Political Science*, *38*(03), 411–432. doi:10.1017/S0007123408000215. arXiv: 0809.2258

Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics*, *90*(3), 414–427. doi:10.1162/rest.90.3.414

Celestino, M. R., & Gleditsch, K. S. (2013). Fresh carnations or all thorn, no rose? Nonviolent campaigns and transitions in autocracies. *Journal of Peace Research*, *50*(3), 385–400. doi:10.1177/0022343312469979

Chan, V. C. Y., Backstrom, J., & Mason, T. D. (2014). Patterns of Protest in the People's Republic of China: A Provincial Level Analysis. *Asian Affairs: An American Review*, *41*(February 2015), 91–107. doi:10.1080/00927678.2014.936799

Chen, T., & Kung, J. K. S. (2016). Do land revenue windfalls create a political resource curse? Evidence from China. *Journal of Development Economics*, *123*, 86–106. doi:10.1016/j.jdeveco.2016.08.005. arXiv: arXiv:1011.1669v3

Chen, X. (2017). Elitism and Exclusion in Mass Protest: Privatization, Resistance, and State Domination in China. *Comparative Political Studies*, *50*(7), 908–934. doi:10 . 1177 / 0010414016655532

Chen, Y., Li, H., & Zhou, L. A. (2005). Relative performance evaluation and the turnover of provincial leaders in China. *Economics Letters*, *88*(3), 421–425. doi:10.1016/j.econlet.2005.05.003

Chiang, C.-F., Liu, J.-T., & Wen, T.-W. (2014). *Economic Integration and National identity*.

Chou, C.-c. (2011). *When does an autocrat compromise with social forces? The political economy of labor policy reform in China, 1978-2009* (Doctoral dissertation). Retrieved from http://search.proquest.com/docview/906824215?accountid=14166

Chyi, H. I., & Huang, J. S. (2011). Demystifying the demand relationship between online and print products under one newspaper brand: the case of Taiwan and the emergence of a universal pattern. *Asian Journal of Communication*, *21*(3), 243–261. doi:10.1080/01292986.2011.559261

Clots-Figueras, I., & Masella, P. (2013). Education, language and identity. *Economic Journal*, *123*(570), F332–F357. doi:10.1111/ecoj.12051

Cragun Cragun, D., R. (2006). *Introduction to Sociology*. Seven Treasures Publications. Retrieved from http://en.wikibooks.org/wiki/Introduction%7B%5C_%7Dto%7B%5C_%7DSociology

Curran, J. (2003). From distributional to semantic similarity. *University of Edimburgh*, 177. doi:10.1.1.10.6068. arXiv: arXiv:1011.1669v3

Djankov, S., McLiesh, C., Nenova, T., & Shleifer, A. (2003). Who Owns the Media? *Journal of Law and Economics*, *46*(2), 341–381.

Dorsch, M. T., Dunz, K., & Maarek, P. (2015). Macro shocks and costly political action in non-democracies. *Public Choice*, *162*(3-4), 381–404. doi:10.1007/s11127-015-0239-x

Dupré, J.-F. (2013). In search of linguistic identities in Taiwan: An empirical study. *Journal of Multilingual and Multicultural Development*, *34*(5), 431–444. doi:10.1080/01434632.2013.783037

Durante, R., & Knight, B. (2012). PARTISAN CONTROL, MEDIA BIAS, AND VIEWER RESPONSES: EVIDENCE FROM BERLUSCONI'S ITALY. *Journal of the European Economic Association*, *10*(3), 451–481. doi:10.1111/j.1542-4774.2011.01060.x

Edin, M. (2003). State Capacity and Local Agent Control in China: CCP Cadre Management from a Township Perspective. *The China Quarterly*, *173*(4), 35–52. doi:10.1017/S0009443903000044

Environment of European Commission. (2016). Standards - Air Quality - Environment - European Commission. Retrieved October 1, 2017, from http://ec.europa.eu/environment/air/quality/standards.htm

Fell, D. (2005). Political and Media Liberalization and Political Corruption in Taiwan. *The China Quarterly*, *184*(-1), 875. doi:10.1017/S0305741005000548

Freedom House. (2015). *Freedom of the Press*. Freedom House. Retrieved from https://freedomhouse.org/report/freedom-press/2015/taiwan

Freedom House. (2017). China | Country report | Freedom of the Press | 2017. Retrieved October 4, 2017, from https://freedomhouse.org/report/freedom-press/2017/china

Gentzkow, M., & Shapiro, J. M. (2010). What Drives Media Slant? Evidence From U.S. Daily Newspapers. *Econometrica*, *78*(1), 35–71. doi:10.3982/ECTA7195. arXiv: 9809069v1 [arXiv:gr-qc]

Georgiadis, A., & Manning, A. (2013). One nation under a groove? Understanding national identity. *93*(0), 166–185. doi:http://dx.doi.org/10.1016/j.jebo.2012.10.013

Göbel, C., & Ong, L. H. (2012). *Social Unrest in China*. doi:10.1057/9781137351869_6

Gong, T., & Scott, I. (2016). *Routledge handbook of corruption in Asia*. Routledge. Retrieved from https://www.routledge.com/Routledge-Handbook-of-Corruption-in-Asia/Gong-Scott/p/book/9781138860162

Grefenstette, G. (1994). *Explorations in Automatic Thesaurus Discovery*. doi:10.1007/978-1-4615-2710-7

Groseclose, T., & Milyo, J. (2005). A Measure of Media Bias. *The Quarterly Journal of Economics*, *120*(4), 1191–1237. doi:10.1162/003355305775097542

Guibernau, M. (2004). Anthony D. Smith on nations and national identity: a critical assessment. *Nations and Nationalism*, *10*(1-2), 125–141. doi:10.1111/j.1354-5078.2004.00159.x

Ho, S. J., & Ping Sun, M. (2008). Heterogeneous multi-product competition in Taiwan's newspaper industry. *Asian Journal of Communication*, *18*(2), 102–116. doi:10.1080/01292980802021822

Hollyer, J. R., Rosendorff, B. P., & Vreeland, J. R. (2015). Transparency, Protest, and Autocratic Instability. *American Political Science Review*, *109*(04), 764–784. doi:10.1017/S0003055415000428

Hong, Y.-F. (2012). Labor costs in China surge; many module manufacturing plants return to Taiwan. (in Chinese). Newspaper Article.

Hughes, C., & Stone, R. (1999). Nation-Building and Curriculum Reform in. *The China Quarterly*, (September 1986).

Human Rights Watch. (2012). *World Report 2012: China*. Retrieved from https://www.hrw.org/world-report/2012/country-chapters/china-and-tibet%20https://www.hrw.org/world-report-2012/world-report-2012-china

Human Rights Watch. (2016). *World Report*. Retrieved from https://www.hrw.org/world-report/2016/country-chapters/china-and-tibet

Hung, C.-L. (2013). Media Control and Democratic Transition: Ongoing Threat to Press Freedom in Taiwan. *China Media Research*, *9*(2), 83–93.

Hutto, C. J., & Gilbert, E. (2014). Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Eighth International AAAI Conference on Weblogs and* ... 216–225. Retrieved from http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8109%7B%5C%%7D5Cnhttp://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

International Federation of Journalists. (2014). IFJ Condemns China's Refusal to Issue Visas to Taiwanese Journalists. Press Release. Retrieved from http://www.ifj.org/nc/news-single-view/backpid/243/category/top-news/article/ifj-condemns-chinas-refusal-to-issue-visas-to-taiwanese-journalists/

Jurafsky, D., & Martin, J. H. (2000). *Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition*. Prentice Hall. Retrieved from https://dl.acm.org/citation.cfm?id=555733

Kao, H.-H., & Li, M.-C. (2000). The Influences of Provincial Origin, Party Affiliation, and Political Ideology on Chinese vs. Taiwanese Identity in Taiwan (in Chinese). *Indigenous Psychological Research in Chinese Societies*, *13*, 231–276.

Karthigesu, R. (1988). Television as a Tool for Nation-Building in the Third World: A Post-Colonial Pattern, Using Malaysia as a Case-Study. In P. Drummond & R. Paterson (Eds.), *Television and its audience* (p. 35). London: British Film Institute.

Keller, F. B. (2015). Networks of Power : Using Social Network Analysis to understand who will rule and who is really in charge in the Chinese Communist Party.

Klandermans, B., & van Stekelenburg, J. (2013). *Social Movements and the Dynamics of Collective Action*. doi:10.1093/oxfordhb/9780199760107.013.0024

Kuo, S.-H. (2007). Language as Ideology Analyzing Quotations in Taiwanese News Discourse. *Journal of Asian Pacific Communication*, *172*(2), 28–30. doi:10.1075/japc.17.2.08kuo

Kuo, S.-H., & Nakamura, M. (2005). Translation or transformation? A case study of language and ideology in the Taiwanese press. *Discourse Society*, *16*(3), 393–417. doi:10.1177/0957926505051172

Larcinese, V., Puglisi, R., & Snyder Jr, J. M. (2011). Partisan bias in economic news: Evidence on the agenda-setting behavior of U.S. newspapers. *95*(9–10), 1178–1189. doi:http://dx.doi.org/10.1016/j.jpubeco.2011.04.006

Li, H. [Hongbin], & Zhou, L. A. (2005). Political turnover and economic performance: The incentive role of personnel control in China. *Journal of Public Economics*, *89*(9-10), 1743–1762. doi:10.1016/j.jpubeco.2004.06.009

Li, H. [Huiping], Wang, Q., & Zheng, C. (2016). Interjurisdictional Competition and Intracity Fiscal Disparity across Chinese Prefectural Cities. *Governance*, *00*(00). doi:10.1111/gove.12222

Li, Y. [Yanwei], Koppenjan, J., & Verweij, S. (2016). Governing Environmental Conflicts in China: Under What Conditions Do Local Governments Compromise? *Public Administration*, *94*(3), 806–822. doi:10.1111/padm.12263

Li, Y. [Yuan]. (2014). Downward accountability in response to collective actions. *Economics of Transition*, *22*(1), 69–103. doi:10.1111/ecot.12033

Lü, X., & Landry, P. F. (2014). Show Me the Money: Interjurisdiction Political Competition and Fiscal Extraction in China. *American Political Science Review*, *108*(03), 706–722. doi:10.1017/S0003055414000252

Lui, B. (2016). Calculate the cost of social stability maintenance. Retrieved from https://news.mingpao.com/pns/dailynews/web%7B%5C_%7Dtc/article/20160309/s00012/1457459535343

Ma, W.-Y., & Chen, K.-J. (2003). Introduction to CKIP Chinese word segmentation system for the first international Chinese Word Segmentation Bakeoff. Conference Paper. doi:10.3115/1119250.1119276

Madestam, A., Shoag, D., Veuger, S., & Yanagizawa-Drott, D. (2013). Do Political Protests Work? Evidence from the Tea Party Movemente. *The Quarterly Journal of Economics*, 1633–1685. doi:10.1093/qje/qjt021.Advance

Masella, P. (2013). National identity and ethnic diversity. *Journal of Population Economics*, *26*(2), 437–454. doi:10.1007/s00148-011-0398-0

Miguel, E. (2004). Tribe or Nation? Nation Building and Public Goods in Kenya versus Tanzania. *56*(03), 328–362. doi:doi:10.1017/S0043887100004330

Ministry of Environmental Protection (China). (2012). Ambient air quality standards. Retrieved from http://english.mep.gov.cn/Resources/standards/Air%7B%5C_

%7DEnvironment/quality%7B%5C_%7Dstandard1/201605/W020160511506615956495.pdf

Murphy, C., & Westbury, C. (2013). Expanding the Scope of Selective Exposure: An Objective Approach to Measurement of Media Ideology. *Communication Methods and Measures*, *7*(3-4), 224–246. doi:10.1080/19312458.2013.813921

Nath, A. (2014). *Political Competition and Elite Capture of Local Public Goods*.

O'Brien, K. J. (2013). Rightful resistance revisited. *Journal of Peasant Studies*, *40*(6), 1051–1062. doi:10.1080/03066150.2013.821466

O'Brien, K. J., & Li, L. (2005). Popular Contention and its Impact in Rural China. *Comparative Political Studies*, *38*(3), 235–259. doi:10.1177/0010414004272528

O'Brien, K. J., & Li, L. (2006). *Rightful Resistance in Rural China*. Cambridge Studies in Contentious Politics. doi:10.1017/CBO9780511791086. arXiv: arXiv:1011.1669v3

Ou, J., Liu, X., Li, X., Li, M., & Li, W. (2015). Evaluation of NPP-VIIRS Nighttime Light Data for Mapping Global Fossil Fuel Combustion CO2 Emissions: A Comparison with DMSP-OLS Nighttime Light Data. *PloS one*, *10*(9), e0138310. doi:10.1371/journal.pone.0138310

Panda, S. (2015). Political Connections and Elite Capture in a Poverty Alleviation Programme in India. *Journal of Development Studies*, *51*(1), 50–65. doi:10.1080/00220388.2014.947281

Persson, P., & Zhuravskaya, E. (2016). The limits of career concerns in federalism: Evidence from China. *Journal of the European Economic Association*, *14*(2), 338–374. doi:10.1111/jeea.12142

Piantadosi, S. T. (2014). Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic Bulletin & Review*, *21*(5), 1112–1130. doi:10.3758/s13423-014-0585-6

Postill, J. (2008). *Media and Nation Building: How the Iban became Malaysian*. Berghahn Books.

Prat, A., & Strömberg, D. (2013). *The political economy of mass media*.

Qiao, L., Wong, L.-t., & Mudie, L. (2016). Chinese Blogger Who Compiled Protest Data Missing, Believed Detained. Retrieved from http://www.rfa.org/english/news/china/chinese-blogger-who-compiled-protest-data-missing-believed-detained-06212016110257.html?searchterm:utf8:ustring=Lu+yuyu

Qin, B., Wu, Y., & Strömberg, D. (2016). Media Bias in China. *90089*(December), 1–54. Retrieved from http://www-bcf.usc.edu/%7B~%7Dyanhuiwu/MediaBias1.pdf

Ramzy, A. (2017). Chinese Court Sentences Activist Who Documented Protests to 4 Years in Prison. Retrieved from https://www.nytimes.com/2017/08/04/world/asia/china-blogger-lu-yuyu-prison-sentence-protests-picking-quarrels.html

Řehůřek, R., & Sojka, P. (2010). *Software Framework for Topic Modelling with Large Corpora*. University of Malta. Retrieved from http://www.fi.muni.cz/usr/sojka/presentations/lrec2010-poster-rehurek-sojka.pdf

Ren, Y., & Tian, J. (2017). Sentiment analysis of Internet performance data. In *2017 ieee 3rd information technology and mechatronics engineering conference (itoec)* (pp. 622–628). doi:10.1109/ITOEC.2017.8122373

Reporters Without Borders. (2016). 2016 RSF-TV5 Monde Press Freedom Prize: Prize awarded to Syrian and Chinese journalists, website | RSF. Retrieved June 28, 2017, from https://rsf.org/en/news/2016-rsf-tv5-monde-press-freedom-prize-prize-awarded-syrian-and-chinese-journalists-website

Reporters Without Borders. (2017). Chine | RSF. Retrieved October 4, 2017, from https://rsf.org/fr/chine

Schiller, H. I. (1985). 'Electronic Information Flows: New Basis for Global Domination' [in] Television in Transition. In R. P. (Philip Drummond (Ed.), *Television in transition* (pp. 11–20). British Film Institute. Retrieved from https://contentstore.cla.co.uk/secure/link?id=c1c11a4d-81c0-e611-80c7-005056af4099

Shao, Y., Sennrich, R., Webber, B., & Fancellu, F. (2017). Evaluating Machine Translation Performance on Chinese Idioms with a Blacklist Method. arXiv: 1711.07646. Retrieved from http://arxiv.org/abs/1711.07646

Shi, K., Huang, C., Yu, B., Yin, B., Huang, Y., & Wu, J. (2014). Evaluation of NPP-VIIRS night-time light composite data for extracting built-up urban areas. *Remote Sensing Letters*, *5*(4), 358–366. doi:10.1080/2150704X.2014.905728

Shi, K., Yu, B., Huang, Y., Hu, Y., Yin, B., Chen, Z., … Wu, J. (2014). Evaluating the ability of NPP-VIIRS nighttime light data to estimate the gross domestic product and the electric power consumption of China at multiple scales: A comparison with DMSP-OLS data. *Remote Sensing*, *6*(2), 1705–1724. doi:10.3390/rs6021705. arXiv: arXiv:1011.1669v3

Shih, V., Adolph, C., & Liu, M. (2012). Getting Ahead in the Communist Party: Explaining the Advancement of Central Committee Members in China. *American Political Science Review*, *106*(01), 166–187. doi:10.1017/S0003055411000566

Simeunovic, N. (2009). The role of media in European identity formation: understanding the complexity of today's European media landscape. *CEU Political Science Journal*, *4*(4), 501.

Slinko, I., Yakovlev, E., & Zhuravskaya, E. (2005). Laws for sale: Evidence from Russia. *American Law and Economics Review*, *7*(1), 284–318. doi:10.1093/aler/ahi010

Smith, A. D. (1991). *National Identity*. Reno: University of Nevada Press. Retrieved from http://books.google.com/books?hl=en%7B%5C&%7Dlr=%7B%5C&%7Did=bEAJbHBlXR8C%7B%5C&%7Dpgis=1

Somers, M. R. (1993). Citizenship and the Place of the Public Sphere: Law, Community, and Political Culture in the Transition to Democracy. *American Sociological Review*, *58*(5), 587. doi:10.2307/2096277

Strömberg, D. (2015). Media and Politics. *Annual Review of Economics*, *7*(1), 173–205. doi:doi:10.1146/annurev-economics-080213-041101

Sun, L. (2011). Social Order ls Currently a Severe Challenge. Retrieved from http://opinion.hexun.com/2011-02-25/127571301.html

Tong, Y., & Lei, S. (2010). Large-scale mass incidents and government responses in China. *International Journal of China Studies*, *1*(2), 487–508. doi:10.1525/as.2013.53.4.629

Ulfelder, J. (2005). Contentious Collective Action and the Breakdown of Authoritarian Regimes. *International Political Science Review*, *26*(3), 311–334. doi:10.1177/0192512105053786

US EPA. (2015). Particulate Matter (PM) Standards. Retrieved from https://www3.epa.gov/ttn/naaqs/standards/pm/s%7B%5C_%7Dpm%7B%5C_%7Dhistory.html%20ohttp://www.epa.gov/ttn/naaqs/standards/pm/s%7B%5C_%7Dpm%7B%5C_%7Dhistory.html

Wang, F.-c. (2005). From Chinese Original Domicile to Taiwanese Ethnicity: An Analysis of Census Category Transformation in Taiwan (in Chinese). *9*, 59–117.

Wang, Y. (2014). Empowering the Police: How the Chinese Communist Party Manages Its Coercive Leaders. *The China Quarterly*, *219*(August 2014), 625–648. doi:10.1017/S0305741014000769

Wasow, O. (2017). *Do Protests Matter ? Evidence from the 1960s Black Insurgency*.

Wong, L.-t., Yang, F., & Mudie, L. (2016). Chinese Citizen Journalist on Hunger Strike Over Beatings in Detention. Retrieved from http://www.rfa.org/english/news/china/china-blogger-09022016133317.html?searchterm:utf8:ustring=Lu+yuyu

Xiu, Y., Lan, M., Wu, Y., & Lang, J. (2017). Exploring semantic content to user profiling for user cluster-based collaborative point-of-interest recommender system. In *2017 international conference on asian language processing (ialp)* (pp. 268–271). doi:10.1109/IALP.2017.8300595

Xu, C. (2011). The Fundamental Institutions of China's Reforms and Development. *Economic literature*, *42*(3), 1076–1151. doi:10.2469/dig.v42.n3.11

Yu, J., Zhou, L.-A., & Zhu, G. (2016). Strategic interaction in political competition: Evidence from spatial effects across Chinese cities □. *Regional Science and Urban Economics*, *57*, 23–37. doi:10.1016/j.regsciurbeco.2015.12.003

Yuan, H. (2016). Measuring media bias in China. *China Economic Review*, *38*, 49–59. doi:10.1016/j.chieco.2015.11.011

Zhang, X., Zeng, Y., Jin, X.-B., Yan, Z.-W., & Geng, G.-G. (2017). Boosting the phishing detection performance by semantic analysis. In *2017 ieee international conference on big data (big data)* (pp. 1063–1070). doi:10.1109/BigData.2017.8258030

Zuo, C. (2015). Promoting City Leaders: The Structure of Political Incentives in China. *The China Quarterly*, *224*, 955–984. doi:10.1017/S0305741015001289

Curriculum Vitae

2006 - 2011	B.A. in Economics, double major in Finance, National Taiwan University	Taiwan
2012 - 2014	M.S. in Economics, Mannheim University	Germany
2013 - 2014	ISAP Exchange Program, Yale University	USA
2014 - 2018	Ph.D. in Economics, Mannheim University	Germany

Declaration of Authorship

I hereby declare that the paper presented is my own work and that I have not called upon the help of a third party. In addition, I affirm that neither I nor anybody else has submitted this paper or parts of it to obtain credits elsewhere before. I have clearly marked and acknowledged all quotations or references that have been taken from the works of others. All secondary literature and other sources are marked and listed in the bibliography. The same applies to all charts, diagrams and illustrations as well as to all Internet resources. Moreover, I consent to my paper being electronically stored and sent anonymously in order to be checked for plagiarism. I am aware that if this declaration is not made, the paper may not be graded.

Date, Place	Chia-Yu Tsai