**Political Attention of the Party: Analyzing CCP National Congress Documents with Topic Models**

**Pei Chen, Xiaoxin Li, Chen Liu, Mingjie Xu[[1]](#footnote-1)**

**1 Introduction**

The political system in China is rightly characterized as a “party-state” in which the state wields the formal power whereas the Chinese Communist Party (CCP) exerts the actual influence. The meetings of The National Congress of the CCP are therefore viewed as crucial events by the broader Chinese society, and the documents produced by these meetings are intently studied both within and outside China as indicators for future policies. However, beyond the immediate value as policy indicators, the historical archive of these documents also reveals the evolution of the political attention of the party through various historical circumstances. A systematic understanding of these documents can therefore provide important insights for the study of Chinese politics.

Political attention is a cornerstone in the study of responsiveness of policymaking: government decision making is fundamentally an allocation of its scarce attention to its numerous tasks and an alignment of priorities (Bryan D. Jones and Baumgartner 2005). Different models have been proposed to characterize the dynamics of such allocation (e.g. Erikson, MacKuen, and Stimson 2002; Stimson, Mackuen, and Erikson 1995; B. D. Jones 2004), but measurement remains a thorny issue. An exemplary study addressing the measurement problem utilizes the Queen’s speech in the British context to construct time-series measures of political attention (Jennings and John 2009). Our study provides a first step in applying this approach to the Chinese context and uses more recent tools of automated text analysis. We build topic models upon the corpus of party congress documents and are able to present a synoptic view of the historical change of the political attention of the CCP in the 40 years of the post-reform period. We also apply this procedure to the state counterpart to the party congress in the party-state scheme—the National People’s Congress—and further illustrate the nature of CCP political attention.

To the best of our knowledge, such a systematic study of the documents produced by the party congress have been lacking in the English academic literature, with a few qualitative example (e.g. Cha 2017). Although there have been many studies of this corpus of documents in the Chinese literature, their foci are usually more specific issues covered by these documents instead of the political attention revealed by these documents as a whole (e.g. 张清敏 2009; 冉连 2017; 张凤阳 and 李智 2016). Methodologically, no study has applied topic models to the study of this data source, which is a natural tool for the study of political attention in this context. For these reasons, we believe this study makes a useful contribution to this line of research, albeit an immature one.

The rest of the article is arranged as follows. The next section details out data source and methodology choice. Our modelling results and analysis are presented in the third section. Finally we discuss the implications and limitations of our study in the fourth section.

**2 Data and Methodology**

**2.1.1 Data**

In this study we use documents from two sources: *The National Congress* of *the Chinese Communist Party* and *The National People’s Conference*. The Chinese Communist Party Congress and The National People’s Congress are two most important meetings held in the People’s Republic of China. The party congress are held for discussing issues within the party. And the people’s congress are hosted by the government, which mostly focus on the development of Chinese country.

The major materials we choose for our study are conference addresses (*Baogao*). At the conference, leaders of CCP and government shall give conclusive reports, which, as many believe, are the official announcements of the current focus of China and are thoroughly analysed by both organizations and individuals. The major conference is held in a 5-year interval. In the years between two major conferences, only the plenary session meetings will be held. One major conference and 4 plenary sessions together make up a 5-year political period.

We use data from the Eleventh Congress to the Nineteenth Congress, covering from 1977-2020. We choose the first conference right after the infamous social turmoil to avoid the obvious caveat of data. The party conference data are downloaded from *The CCP National Representative Conference Dataset*, provided by the people.com. To expand our data (for each year, we have at most one report if we only use the government address, and the maximum number of words will be less than 30,000 Chinese characters), we scraped the entire dataset and include 6 types of document: 1. Address (*Baogao*) 2. Editorial (*Shelun*) 3. Communique (*Gongbao*) 4. Decision (*Jueyi*) 5. Notice (*Tongzhi*) 6. Suggestion (*Yijian*). We exclude meaningless documents like the designation of officials. Documents lacking a decent amount of  information are also dropped due to their limited marginal contribution for big-data analysis. We end up with 160 records for 43 years. Those 160 texts are later aggregated by year to create 43 observations for our analysis.

The *People Congress* data also comes from the Internet. There exists no systematic collection of the *Government Working Report*, which is our major interest, so we manually search the key words and download the annual working reports. The original data include records from 1956-2020. To match our *Party Congress* data, we only use observations from 1977-2020, and exclude the 1985 report because there was no party conference in 1985. For the *People Conference*, the single type of document we choose is the *Government Working Report* because the *People Conference* report is more comprehensive than that of the *Party Congress*. In fact, both two datasets end up with a similar number of characters.

**2.1.2 Chinese Communist Party Congress**

According to the Chinese Communist Party Constitution, the Chinese Communist Party Congress (CCPC) convenes every five years since 1921. A series of sessions of the CPC Central Committee was held for five years to discuss China's development in the following years. Chinese Communist Party Conference plays a vital role in Chinese politics in two ways: shaping China’s political leadership by electing a new Central Committee and issuing a political report (Cha 2017). CCPC aims to evaluate previous work, assess the Party’s current situation, and provide general guidelines for the Party’s priorities, focus, and planned tasks in the following 5-year period (Cha, 2017). Through these strategies, the CCPC could convey the party’s ideological values and direction to all significant policy sectors. For instance, Hu Jintao proposed a concept of “comparatively well-off society” (小康社会) in order to create a middle-class society in the 16th Congress in 2002. (Wu 2015) demonstrated that CCPC is the “highest decision-making body” in the entire CCP’s organizational hierarchy because of the connection between the party and governmental institutions. In order words, CCPC has ultimate authoritarian power to decide the direction of development in China. After finishing each of Chinese Communist Party Conferences, different types of documents, including address, editorial, communique, decision, notice, and suggestion, would be provided to demonstrate the conferences' process and outcomes.

**2.1.3 National People’s Congress**

The National People’s Congress (NPC), is the highest organ of state power and national Legislature of the People’s Republic of China, usually convenes annually since 1954. The duration of the conference is approximately two weeks. As a national legislature, the NPC aims to enact and amend basic laws, such as civil affairs and state organs, to elect and appoint members to central State organs, to determine significant State issues. Besides, the NPC deliberates and approves economic and social development and provides a report regarding its implementation. Compared to CCPC, the NPC mainly focuses on the administrative system, governance, and legislation issues. On the other hand, CCPC provides an overall strategic plan to boost China's development in the following years. (Wu 2015) believed that the power and authority of the Party are higher than that of NPC. However, (O’Brien 1990) contended that NPC plays an active role in lawmaking, supervision, representation, and regime support, even though PRC’s political system remains illiberal and authoritarian. He believed that NPC could help rationalize the political system by routinizing and legalizing political power and considering the views of groups that were overlooked or persecuted by Communist authorities to make the government more inclusive (Van Vranken HICKEY 1992). After finishing the National People’s conference, the government working report would be offered to illustrate the meeting's outcomes.

**2.1.4 Methodology**

We briefly introduce our data cleaning and analysis procedure. We use Python to conduct most of our word and topic analysis. For *Party Conference* dataset, observations are loaded and saved as 43\*2 matrix: one column for year and one for the aggregate text. The symbol, space, blank line and indentation are removed. We use the *Jieba* package for word-cutting. After cutting, words are transformed into sparse-matrix. The frequency of each word is then counted and the most frequent words for major conferences are saved. Top words are classified as three types and the frequency for each type is summed up.

The LDA model is chosen for our topic study. We specify 8 overall topics and, similarly, divide them into 3 groups. We identify the top topic for each year and discuss the dynamic pattern of the annual top topic. Full list of topics for each year is also analyzed, with the stressing putting on the annual topic composition. We also calculate the similarity between 43 observations based on our topic model. And finally, the *People Conference* dataset is loaded for further examination. We compare the similarity for text at the same year from different conferences. We use different dictionaries to identify potential trends and compare the result of report similarity with the topic dynamic.

**3 Results and Analysis**

**3.1 Word Analysis**

Word analysis can be helpful in generating preliminary results. In fact, many analyses using official reports constantly choose word frequency as an important indicator of official attitudes. We conduct a similar process before we dive into more detailed topic study and our findings appear to be illuminating.

**Figure.1 Top Words for Major Conferences, 1977-2020**

Table

Description automatically generated

Figure 1 reports the top 8 words for each major *Party Conference* in the period 1977-2017 (we also include 2020 text to catch up the latest trend). We roughly divided words into 2 types: words related with party (*dang*), country (*guo, guojia, woguo*), communism (*shehui zhuyi, ziben zhuyi*) , which we name as the “party words”, and the rest, which we similarly name as “non-party words”. The non-party words are shaded with darker color to compare with the party words.

Result from Figure 1 shows that, at the very beginning all the top words reflecting the political attention were mainly related with party, country and communism. Frequency of non-party words gradually increased over time. After entering the 21th century, half of the top words have become non-party words. This trend shows that, though the theme of *Party Conference* mainly focuses on CCP issues, the words used in conference-related text shows their political attention are not limited to party words but instead appear to be diversified.

**Figure 2. Words Classification and Dynamic, 1977-2017**

Diagram, engineering drawing

Description automatically generated

We further split non-party words into 1. words related with economy, development, reform and 2. connected with culture, ideology and society (all these classification work are done by manual work, we host discussion about how to classify words to make sure the classification is interpretable). We aggregate the frequency of all three types of words in each year’s 30 top words and calculate their annual proportion. Results can be found in Figure 2, in which we can see a clear convergence of proportion for different words. The party words dominate at the left side, and then alternate with economy words. The share of culture words has steadily increased over years and in 2017 text it finally topped, with the word “politics” mentioned over 50 times in that year’s document. Combined with Figure 1, our word analysis gives a clear picture of the dynamic of word structure and an indication to the variation of the political attention in *Party Conference* text. Our conclusion is that, with time going by, the CCP conference has expanded their usage of words to include mentioning of non-party topics. The result is consistent with our assumption about the changing political attention of *Party Conference*.

**3.2 Party Congress Topic Analysis**

In this section we discuss our results of topic model, with the first part explaining our choice and classification of topics, the second stressing the overall topic dynamic, and the third examining the similarity between annual observations.

Figure 3 summarizes our topic choice. Eight topics are divided into 3 types, with themes similar to our classification of words: *Party Topics* (PT) , *Economy Topics* (ET) and *Culture Topics* (CT). All eight topics are: *Fight among Classes, The Leadership of Party, Consummate System, Culture and Doctrine, Civilization and Spirit, Rural and Agricultural Development, Development and Modernization and Overall Development*. We shadow words that we believe help define each topic. For instance, the first topic “Fight among Classes” includes words like “comrade”, “Chairman Mao” and “proletarianism”. Combining the historical context as well as the original text, stress of “the purity of communism” and “exclusivity of capitalism” has long been an essential part of party conferences, so we end up naming the nature of this topic as focusing on party relations. These shadowed, exclusive words, however, not only define each topic, but also create variation within groups and help us distinguish between topics labeled as the same type. For example, the PT has three sub-topics enjoying a great level of similarity, but the shadowed words help us locate the distinctive features of each sub-topic.

**Figure 3. Party Conference Overall Topics**

Table

Description automatically generated

To show the dynamic of topic, we use both annual top topic and annual full list of topics. Figure 4 and Figure 5 show our results. The LDA model calculates the topic distribution of each text, with an estimation of probability assigned for each text. We define the topic with the highest probability as the top topic of this year. For each year’s report, several topics can be identified. We collect the complete list of topics and define it as the topic list of the year.

Figure 4 plots the distribution of annual top topics. In Figure 3, each topic is given a unique number, with *Fight among Classes* coded as 1 and *Overall Development* as 8. The vertical axis thus ranges from 1-8. For most of the time, we observe PT and ET as the topic with the highest probability. The top of CT seems to be rare. But this, as Figure 5 suggests, mainly reflects the relatively limited appearance of CT as a whole. Interesting findings include the diminish of *Fight among Classes* (the bottom line of yellow circle) after the 1990s, the appearance of ET starting as the same time period, and the consistent mentioning of The Leadership of Party (the second bottom line of yellow circle). The CT rarely dominates most of the time, with two exceptions during the 1970s and 2000s. The top of CT in 2007, the year before 2008 can be justified by the stressing of “civilized society) during the pre-Olympic time. And the top in 2012 is understandable considering the catastrophic earthquake and the nation's theme of cheering and energizing the whole country.

We give further explanations and combine the top topic dynamic with important meetings in CCP history. We see a related pattern of topic dynamic around famous meetings like *the Third Plenary Session of the Eleventh Central Committee*, which introduced the famous spirit of *Jiefang sixiang, shishi qiushi* and *the Third Plenary Session of the Fourteenth Central Committee*,

**Figure 4. Yearly Top Topic, 1977-2020**

A picture containing calendar

Description automatically generated

which laid the foundation for *Chinese Socialist Market Economic System*. After the Third/Eleventh conference, we see an immediate jump from PT to CT. Similarly, before and after the Third/Fourteenth conference, ET consistently topped the list, which could be viewed as both the foundation of the system and the proceeding of it.

The complete topic list in Figure 5 can be similarly divided into three periods. It shows a clear pattern without the introduction of historical meetings. Before 1985, the type of topics was limited, often 1-2 topics for each year. Year of 1986 witnessed a sudden increase of both type and number of topics, with the emphasis of ET and PT in the following 3 decades. Things have become different in recent years with the *Party Conference* paying increasing attention on CT. The reflection of this change is that, after 2008, we observe a more even distribution of three topic types. The PT is, without doubt, the absolute center of *Party Conference*. But an increase of ET and CT should be regarded as evidence of “a party of all” instead of “a party of communist”.

**Figure 5. Yearly Full Topics, 1977-2020**

Chart

Description automatically generated

**Figure 6. Party Text Similarity, 1977-2017**

Chart, scatter chart

Description automatically generated

In Figure 6 we provide further evidence on topic similarity. The LDA model is used to calculate the similarity between each year’s aggregate document. We skip years with zero correlation and keep those with positive correlation. The similarity between years appears to be irregular. Only half observations show none-zero similarity. We increase the threshold of similarity to 0.5, 0.8 and 0.9 in the following three figures. Strong similarity is scarce, with a potential exception lies at the upper left area of the last figure.

The lack of correlation can be interpreted in many ways. Here we believe it helps explain the diversity of topics dynamics, which can also be observed in Figure 5. There has been research suggesting the CCP expanding its range of power. If the CCP has long been focusing on similar topics, the document should have higher similarity, especially for documents from neighboring years. Yet the similarity gives opposite evidence, reporting low similarity for most of our study period. Therefore, the result of Figure 6 accords with our previous discussion and support for the concept of “the big party”.

**3.3 Similarity of Party and People Conference**

We expand our work by further including text from *People Conference*. We skip previous text and topic analysis and mainly focus on comparing the similarity of each year’s reports from two different conferences. The major assumption is that each conference has its own focus and should be quite different. Similarity of conference reports can be viewed as proof of the blurring edge between government and party, the expanding focus of CCP, and a gigantic “party of all”.

Figure 7 and Figure 8 list our major findings. Figure 7 calculates the similarity using different text dictionaries. And Figure 8 compares the text similarity with the topic distribution. The reason we use different dictionaries is because we believe different conferences use different word distribution to generate their report. Word distribution for different conferences should reflect their focus. And following our previous discussion, we expect two dictionaries to focus on either party or economic development, hence, have intrinsic differences.

**Figure 7. Similarity of Party, People Conference Text, 1977-2020**

A picture containing text, device, thermometer

Description automatically generated

**Figure 8. Text Similarity and Topic Distribution**

A picture containing text, boat, several

Description automatically generated

The result using full text added up from two conferences validates our assumption. Using the full dictionary, only five years’ reports appear to be similar. And the result using the party text dictionary performs no better. Indeed, party text dictionaries have more professional words, and those words are hard to find in *the Government Working Report*, which aims to discuss economic issues. The model using people's text dictionary gives better results. The model reports not only more similar years (21 in total) but also a lasting period between 2002 and 2007, as well as higher frequency of similarity in the last decade.  We then add the topic distribution of ET into the figure. The trend of similarity accords with the mentioning of ET. We regard this correspondence as evidence of the *Party Conference’s* increasing stress on economic topics, and the CCP’s divergence from its party-nature and a more appropriate concept: the party of all.

**4. Discussion**

Our result gives preliminary evidence for CCP’s emphasis on non-party topics. However, the result suffers from problems of inaccuracy. The problems could come from several channels. In this section we talk about potential drawbacks of our data, model and interpretation.

          One significant problem about our data is its limited size. The add-up text from two datasets, after our expansion, has just about 2.5 million Chinese characters. For a text analysis study, the sample size is not sufficient. The problem is hard to resolve because of the nature of our data. The conference only holds once a year. It's impossible for the conference to leave time for a report that takes forever to read. The result could be more robust if the data size could be extended. The possible solution maybe include all conference reports from province level, which we do not take because there exist no systematic collection of provincial reports so a satisfactory data collection work can take forever.

Another problem is about our model and algorithm. The LDA algorithm uses a random sample to construct a model. The setting of random seed can have effects that are hard to neglect. Second, the word-cutting of Chinese characters is not as good as that of English words. We use the *Jieba* package for word cutting, but there are many other algorithms that could be chosen. Last, the similarity coefficient we choose here is quite simple. It only compares the similarity of topics, instead of the word distribution itself. That may be the reason for our polarized similarity result, with many being 0 and many over 0.8 and little lies in the middle. The result could be improved by using a more rigorous algorithm and more sophisticated index, which we will leave for our later work.

**References**

Cha, Chang Hoon. 2017. “China’s Search for Ideological Values in World Politics: Chinese Adaptation to Liberal Order Drawn from Political Reports of the Chinese Communist Party Congress since 1977.” *Pacific Focus* 32 (3): 416–44. https://doi.org/10.1111/pafo.12104.

Erikson, Robert S., Michael MacKuen, and James A. Stimson. 2002. *The Macro Polity*. Cambridge Studies in Political Psychology and Public Opinion. New York: Cambridge University Press.

Jennings, Will, and Peter John. 2009. “The Dynamics of Political Attention: Public Opinion and the Queen’s Speech in the United Kingdom.” *American Journal of Political Science* 53 (4): 838–54. https://doi.org/10.1111/j.1540-5907.2009.00404.x.

Jones, B. D. 2004. “A Model of Choice for Public Policy.” *Journal of Public Administration Research and Theory* 15 (3): 325–51. https://doi.org/10.1093/jopart/mui018.

Jones, Bryan D., and Frank R. Baumgartner. 2005. *The Politics of Attention: How Government Prioritizes Problems*. Nachdr. Chicago, Ill.: Univ. of Chicago Press.

O’Brien, Kevin. 1990. “Is China’s National People’s Congress a ‘Conservative’ Legislature?” *Asian Survey* 30 (8): 782–94. https://doi.org/10.2307/2644498.

Stimson, James A., Michael B. Mackuen, and Robert S. Erikson. 1995. “Dynamic Representation.” *American Political Science Review* 89 (3): 543–65. https://doi.org/10.2307/2082973.

Van Vranken HICKEY, Dennis. 1992. “Review.” *Social Science Quarterly* 73 (2): 464.

Wu, Guogang. 2015. *China’s Party Congress*. UK: Cambridge University Press. https://doi.org/10.1017/CBO9781139976800.001.

冉连. 2017. “绿色治理:变迁逻辑、政策反思与展望——基于1978—2016年政策文本分析.” 北京理工大学学报(社会科学版) 19 (06): 9–17.

张凤阳, and 李智. 2016. “从‘革命—斗争’话语到‘改革—治理’话语——以党代会报告文本为中心的政治社会学分析.” 东南大学学报(哲学社会科学版) 18 (02): 5–14.

张清敏. 2009. “六十年来新中国外交布局的发展——对党代会政治报告的文本分析.” 外交评论(外交学院学报) 26 (04): 32–42.

1. The authors contributed equally to this article. [↑](#footnote-ref-1)