China Economic Review 38 (2016) 49–59

Contents lists available at ScienceDirect

China Economic Review

Measuring media bias in China☆

Han YUAN

Department of Economics, The Chinese University of Hong Kong, Shatin, Hong Kong

ARTICLE INFO

Article history:
Received 24 April 2015
Received in revised form 19 November 2015
Accepted 19 November 2015
Available online 27 November 2015

JEL classification:
L82

Keywords:
Media bias
Chinese media
Text categorisation

ABSTRACT

Major advances in research on media bias have been achieved in recent years. However, methods used in the literature are primarily applied to American media and usually dependent on the two-party system. This paper attempts to detect and quantify the principal difference, or 'media bias', of Chinese media. We extract a document-term matrix from articles on the Eighteenth Party Congress in November 2012 from 21 Chinese newspapers from seven provinces, as well as the *People's Daily*. With this matrix, hierarchical clustering is subsequently used to divide newspapers into two groups. Using the dendrogram and intergroup dissimilarities, we can construct an index to indicate the direction and the magnitude of media bias. In our sample, newspapers from Zhejiang and Guangdong constitute one group, and the rest constitute the other group. The principal difference of Chinese media is reflected in two dimensions: central/local and political/economic.

1. Introduction

Major advances have been achieved in research on media bias in recent years,[1] including, in particular, two papers by Gentzkow and Shapiro (2010) and Groseclose and Milyo (2005). Unlike earlier studies, which specifically define media bias and uses human coders to conduct content analysis according to the given definition(Covert, Adkins, & Wasburn, 2007; Lott & Hassett, 2004; Niven, 2003; Niven, 2004; Schiffer, 2006), these two papers exploit the United States' two-party system and compare quantitative attributes of newspapers with those of members of Congress, whose political stance is usually well defined.

Groseclose and Milyo (2005) compared the number of times think tanks are referenced by each media outlet with the number of times they are cited in speeches by members of Congress cited think tanks. Gentzkow and Shapiro (2010) 'examined the set of all phrases used by members of Congress in the 2005 *Congressional Record* and identify those that are used much more frequently by one party than by another'. The two papers construct their indices based on Americans for Democratic Action (ADA) scores[2] of members of Congress and the similarity of language patterns between newspaper articles and speeches by members of Congress.

The aim of this paper is to calculate an index of media bias of Chinese media. The lack of a two-party system in China prevents us from applying to Chinese media a method similar to the one described above. From the perspective of statistical learning, the method used by Gentzkow and Shapiro (2010) and Groseclose and Milyo (2005) can be considered supervised learning, with

☆ This paper is based on my undergraduate thesis at Zhejiang University. I am grateful for comments and suggestions from Zhangyong He, Lu Wei, Bing Ye, Jianqin Li, Yefeng Chen, Zheng Song, Bei Qin, and Chun-Fang Chiang. The suggestions of two referees greatly improved the quality of this paper.

E-mail address: hanyuan@link.cuhk.edu.hk.

[1] (Groeling, 2013) is an excellent and comprehensive survey.

[2] An index of elected officials' political positions calculated by ADA. ADA selects 20 votes it considers the most important during that session and calculates a score for each member of Congress based on his/her voting records on 20 votes. See more at: http://www.adaction.org/pages/publications/voting-records.php.

http://dx.doi.org/10.1016/j.chieco.2015.11.011
1043-951X/

members of Congress being the labelled training set. Our solution to the absence of training data is to use unsupervised learning techniques, which do not require any prior information or belief regarding media bias.[3]

To measure media bias, we collected articles on the Eighteenth Party Congress in November 2012 from 21 Chinese newspapers in seven provinces and the *People's Daily*. We extracted a document-term matrix from the text. Each row in this matrix is also an object in the vector space that represents a particular newspaper. Hierarchical clustering is later used to divide newspapers into two groups. Using the dendrogram and intergroup dissimilarities, this paper successfully constructs an index to indicate the direction and the magnitude of media bias. Given the two groups, we can also extract characteristic words to see what distinguishes one group from the other.

To the extent that we employ hierarchical clustering, media bias reveals itself without our having to define it prior to analysis. Hence, our index is always conditional on the sample: Theoretically, if we increase the sample size, all 22 newspapers can fall into one group. Correspondingly, if we change the topic of study, newspapers from Zhejiang and Guangdong may fall into different groups. Using unsupervised learning also enables us to analyse media bias in a broader sense. What we find in the data may have nothing to do with political bias. Instead, the principal difference between two groups stems from characteristic word. This difference is especially relevant because all newspapers in mainland China are subject to Party control. Because the term 'media bias' has been widely used in the literature, we use 'media bias' and 'principal difference' interchangeably.

Based on the methodology described above, our sample set of newspapers can be divided into two groups, including Zhejiang and Guangdong on the one side and the rest on the other. We determine that newspapers from the same province are close on the political spectrum, which may be the result of the structure of the Chinese propaganda system. The principal difference amongst the Chinese media is reflected in two dimensions: 'central/local' and 'political/economic'. Given that newspapers form Zhejiang and Guangdong give more coverage to local and economic issues, we designate them as the LE (local/economic) group. Accordingly, the other group is the CP (central/political) group.

Although our methodology is primarily inspired by Gentzkow and Shapiro (2010) and Groseclose and Milyo (2005) (from supervised learning to unsupervised learning), we also borrow insights from other papers. Previous research, for instance, indicates that we can control for nonideological factors by focusing on specific issues such as congressional party switchers (Niven, 2003), scandals (Niven, 2004), economic events (Lott & Hassett, 2004), elections (Schiffer, 2006), and social issues (Covert et al., 2007), which is why we choose to focus on the Eighteenth Party Congress of November 2012 in this paper.

Another predecessor is a recent paper by Qin, Wu, and Strömberg (2012), which measures media bias in the context of political control and commercial motive. These researchers' analysis is driven by the theory of the function of unfree media in China. To measure the degree of political control, these authors define three types of content: party line, mass line, and bottom line. Word counting is used to measure the respective prevalence of these three types of content.

In comparison, this paper is data driven: As long as there are data, our method can be used to detect and quantify the principal difference in this sample. We do not need theories regarding media bias to use this method. Moreover, when there are competing theories, our method can be used to determine which theory has more explanative power. However, the lack of theories can cause some problems. The danger is that a poor choice of sample and data will lead to meaningless results.

The paper is organised as follows. Section 2 introduces our method with a hypothetical sample. In Section 3, we use our data to test our method. The political implications of the results are discussed in Section 4. Section 5 discusses the robustness of our method. The conclusions are presented in Section 6.

2. Methods

In this section, we present our method of measuring media bias. Our premise is that with a properly defined distance and using the hierarchical clustering technique, the ideological structure of the press can be determined. We can subsequently project this structure onto a number line. Below, we use a hypothetical example to illustrate the method. A more detailed description of the method can be found in the appendix.

A concept central to our method is the document-term matrix, which has been widely used in the field of natural language processing.[4] In a document-term matrix, columns correspond to terms (words or phrases) and rows correspond to documents. In Table 1, we have 10 documents. They are randomly generated around (3,9) (Group A) and (9, 3) (Group B). To provide a visual understanding of their relative positions, the sample is plotted in Fig. 1. If we agree that 'efficiency' is associated with the right wing and 'equality' is associated with the left wing, the data seem to indicate that Group A is the left wing and Group B is the right wing. However, in the real world, there are many terms, and there usually is no clear agreement regarding the political meanings of these terms. Hence, let us proceed as if we do not know the meaning of 'efficiency' and 'equality'.

The documents are also objects in a vector space associated with this matrix. To conduct clustering analysis, we need to define distance between objects and distance between groups. In this paper, we use cosine distance, which is used most commonly (Feldman & Sanger, 2007), and average distance, which makes our index more interpretable. To connect this matrix with media bias in the real world, we require the following assumption:

Assumption 1. All else being equal, as the political stance of a newspaper deviates from that of another newspaper, the distance between these two newspapers in vector spaces associated with relevant document-term matrices will increase.

[3] Thanks two referees for suggesting this line of reasoning.
[4] In this paper, it should be the newspaper-term matrix. However, we still use the conventional term.

Table 1
A hypothetical document-term Matrix.

Documents / Terms	Efficiency	Equality
A1	2.060319	7.654628
A2	3.275465	9.277274
A3	1.746557	11.381768
A4	5.392921	7.304436
A5	3.494262	8.879622
B1	7.557100	3.325132
B2	8.561211	2.186261
B3	9.388182	4.336717
B4	7.271802	3.893971
B5	9.293674	5.453427

Note: Entries in a document-term matrix are frequencies of corresponding term ('efficiency' or 'equality') that occur in corresponding documents (5 documents in Group A and 5 documents in Group B) and should be integers. However, decimals will not change the results.

This assumption states that the difference in political stance will be reflected in word use. The results of the studies of Gentzkow and Shapiro (2010) and Groseclose and Milyo (2005) support this assumption: Newspapers (politicians) tend to have different language patterns depending on their political stance. We have to be cautious, however, regarding the 'all else being equal' condition. This condition may not hold. For example, writing styles in newspapers can vary widely. Articles that express the same opinion in different styles can be distant from each other in the vector space and vice versa. We do not expect this to happen for a large corpus, however. The effect of writing styles can be decreased by collecting more texts and controlling for types of newspaper. Even if the impact of writing style is substantial, we do not expect it to be of serious concern because we will be able to detect it with the steps described below.

Another potential problem is that different newspapers will focus on different topics. We can solve this problem by controlling for types of newspapers. We may then find that even the same types of newspapers will focus on different topics during the same period, which is later reflected in the vector space. However, this phenomenon is exactly what we want to measure. This type of agenda-setting behaviour—deciding what is newsworthy and what is not—is of interest and will theoretically lead to suboptimal public policy decisions (Anderson & McLaren, 2012; Baron, 2006; Bernhardt, Krasa, & Polborn, 2008; Besley & Prat, 2006).

Using the matrix described above along with properly defined distance between documents and between groups, we can conduct hierarchical clustering to divide the documents into two groups. While there are different clustering algorithms, we use hierarchical clustering because it provides more information, thus facilitating the construction of the index. As with other clustering algorithms, the aim of hierarchical clustering is to obtain clusters such that objects within a cluster are similar to one another, but different from those in other clusters.

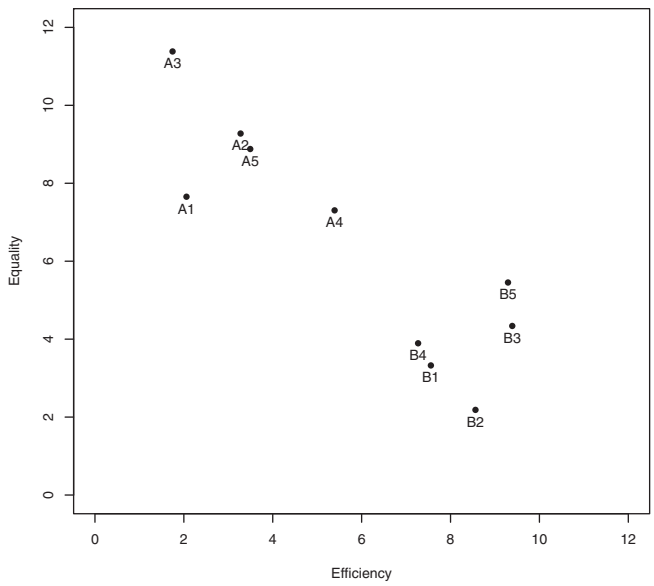

Fig. 1. Ten hypothetical documents. Note: The two dimensions are frequencies of corresponding term ('efficiency' or 'equality') that occur in corresponding documents (5 documents in Group A and 5 documents in Group B).

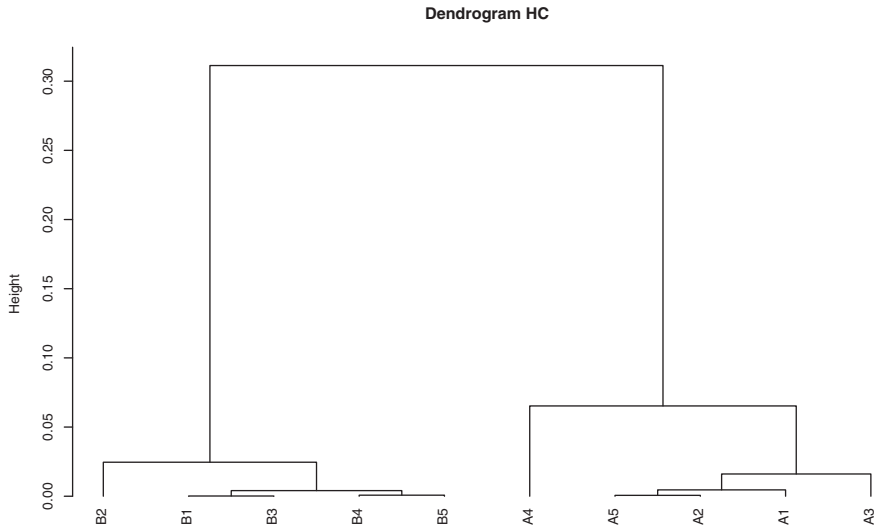

Fig. 2. The dendrogram of hierarchical clustering. Notes: (1) The nodes in this figure represent groups (the root node is the largest possible group—all the 10 documents). Each nonterminal node ('parent') has two daughter nodes, representing the two clusters merged to form the parent. (2) The height of each node is proportional to the value of the intergroup distance between its two daughters. We use cosine distance and average distance in the clustering process.

The result of hierarchical clustering for our hypothetical example is illustrated in Fig. 2, which is a binary tree or dendrogram. The nodes in a dendrogram represent groups. Each nonterminal node ('parent') has two daughter nodes, representing the two clusters merged to form the parent. If nonideological factors have been controlled for, media bias should exist between the daughters of the root node. In a dendrogram, the height of each node is proportional to the value of the intergroup distance between its two daughters. In Fig. 2, the nodes are ordered to reflect distance from each other. For instance, in Fig. 2, the distance between A4 and Group B is smaller than the distance between {A1,A2,A3,A5} and Group B. The structure of Fig. 2 is similar to that of Fig. 1 with A3 and B2 being the most 'extreme' points, and A4 and B5 being the most 'moderate' points.

When we obtain two clusters, we may want to determine how one cluster differs from the other, i.e., the principal difference between the two clusters. This difference is relatively straightforward in this hypothetical example because there are only two words and their meanings are clear: one group cares more about efficiency, and the other group cares more about equality. However, with real data, there will be thousands of terms in the document-term matrix, and the meanings of these words will sometimes be unclear.

Our aim is to find words characteristic of each cluster. The key point is that characteristic words of one group are used more frequently by this group and less frequently by the other group.[5] Hence, in this example, the word characteristic of Group A is 'equality' and that of Group B is 'efficiency'. In the appendix, we describe in more detail our algorithm to find characteristic words.

Note that characteristic words can vary from sample to sample. The principal difference detected in a sample about a party congress are likely to be notably different from that in a sample regarding other topics, such as a natural disaster, for example. Furthermore, if the effect of writing style is substantial, the characteristic words will be correlated with writing style rather than political stance. To complicate matters, there is a danger that the characteristic words may not reflect anything, which means that newspapers are very similar based upon all meaningful criteria. Distances between newspapers might result from stochastic error. This finding would suggest that the sample collection of documents provides little evidence for the existence of media bias. Note that statistical techniques have been developed by Shimodaira (2002), Shimodaira (2004), and Suzuki and Shimodaira (2004) to calculate *p*-values for hierarchical clustering via multi-scale bootstrap resampling.

Next, we calculate the index of media bias. An index will be of interest if we want to incorporate media bias into further quantitative research. An intuitive approach is to project the tree in Fig. 2 onto a number line ('the political spectrum'). This is our aim in this paper. Note that some modifications are required to do so. In the appendix, we describe how this projection works. The results are shown in Table 2.

3. Data and results

This paper investigates media bias in China. To control for irrelevant factors, we analyse news articles on the Eighteenth Party Congress that took place in November 2012. We investigate newspapers from Beijing, Shanghai, Zhejiang, Guangdong, Jilin, and Shanxi. To make this analysis more representative, we choose three newspapers (one party newspaper and two municipal newspapers) from each province based upon their circulation numbers. Because of the importance of the *People's Daily*, we also include

[5] (Gentzkow & Shapiro, 2010) also have this step. They use statistical testing to find those words that can distinguish Democrats from Republicans.

Table 2
An index of bias for the hypothetical example.

Indices / Documents	Original index	*Z* score
B2	−0.148952561	−1.0058789
B1	−0.150025103	−0.9353636
B3	−0.150004343	−0.9352392
B4	−0.149047683	−0.9295080
B5	−0.148952561	−0.9289381
A4	0.139336029	0.7981667
A5	0.169649456	0.9797710
A2	0.169688839	0.9800070
A1	0.170243222	0.9833282
A3	0.171966929	0.9936548

Note: 'Original index' is calculated based on the projection method in the appendix and distances obtained in the clustering process. *Z* score is the standardised index of 'original index'.

Table 3
The sample.

Province/municipality	Newspaper
Beijing	*Beijing Daily*, *Beijing News*, *Beijing Times*
Zhejiang	*Zhejiang Daily*, *City Express*, *Qianjiang Evening News*
Shanghai	*Liberation Daily*, *Oriental Morning Post*, *Xinmin Evening News*
Guangdong	*Southern Daily*, *Southern Metropolis Daily*, *Yangcheng Evening News*
Shanxi	*Shanxi Daily*, *Shanxi Evening News*, *Taiyuan Evening News*
Jilin	*Jilin Daily*, *New Wenhua Post*, *City Evening News*
Chongqing	*Chongqing Daily*, *Chongqing Times*, *Chongqing Morning Post*
Countrywide	*People's Daily*

Note: For the seven provinces, the first newspaper is the official newspaper of its Provincial Committee.
Sources: CNKI (cnki.net) and the China Digital Library (www.apabi.com).

it in our sample (see Table 3). Every day during the month of November 2012, one article from each newspaper is randomly selected from articles (e.g., feature reports, opinions, and editorials) that contain the word '十八大' (the abbreviation in Chinese for the 18th Communist Party of China National Congress in Chinese).[6] Our data sources are CNKI (cnki.net) and the China Digital Library (www.apabi.com).

The environment for data analysis is R. After obtaining the sample, we break all sentences into words. Next, we delete punctuation marks and stop words. Stop words are words that are frequently used but convey little information. Examples in English include 'the', 'a', and 'which'. After these steps, we obtain our document-term matrix. This matrix is usually sparse, containing tens of thousands of terms. We thus need to remove the 'unimportant' terms. In our analysis, terms that are not used by at least 90% of newspapers are removed,[7] resulting in a 22×7492 matrix.

Next, we can conduct clustering analysis. The resulting dendrogram is shown in Fig. 3. We detect two groups and an outlier, *Southern Metropolis Daily*. As mentioned in the introduction, let LE denote the cluster containing *City Express*, *Yangcheng Evening News*, *Southern Daily*, *Qianjiang Evening News*, and *Zhejiang Daily*, and CP the cluster containing the other newspapers.

Is this result statistically significant? This question is difficult to answer. Because we have only 22 objects (newspapers), it is meaningless to use a resampling method, which is the only method we have, to calculate the *p*-values of clusters. Fortunately, we can compare the words characteristic of the two groups respectively to assess whether they exhibit political differences. Following the steps described in the appendix, we obtain two frequency tables, shown in Table 4. Note that because the distance between *Southern Metropolis Daily* and LE is smaller than the distance between said paper and CP, we merge *Southern Metropolis Daily* with LE when we extract their keywords. We translated these phrases into English based upon the official translation of the *Report to the Eighteenth National Congress of the Communist Party of China*. Note that some phrases can have multiple meanings; conversely, multiple phrases can have virtually the same meaning, which is why LE's frequency table contains the term 'people' twice. The first refers to '群众' and the second to '人民'. We include the original table in Chinese in the appendix.

Based on the results in Table 4, we can infer that the LE group covers mostly local economic issues, whereas the CP group covers mostly political issues and repeatedly expresses loyalty to the central government/committee. We discuss these terms in more detail later.

[6] In our sample, we choose the first article in the search results.
[7] Another way is to remove words with low overall frequency. In that case, we have to set different thresholds for different samples. In contrast, a percentage threshold can be constant across samples. Another reason is that words with low frequency can still be important if they only appear in several documents because they signal the characteristics of these documents.

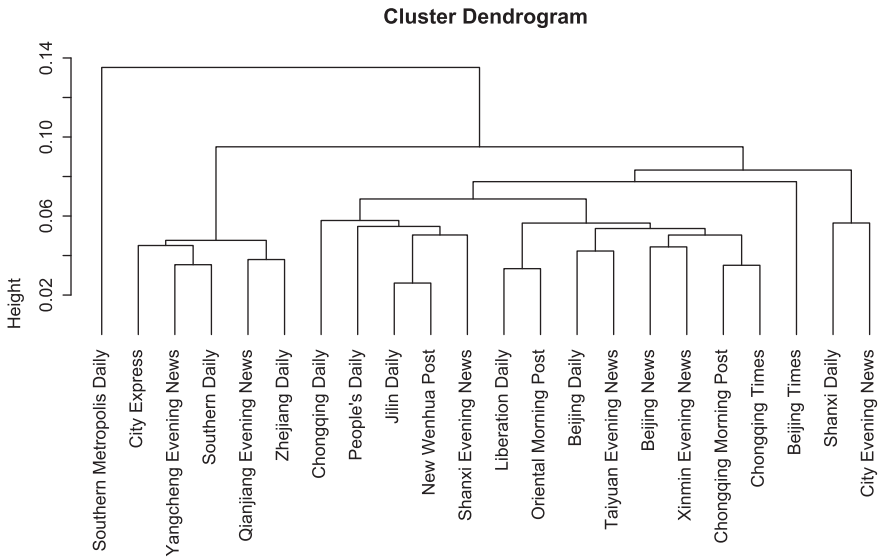

Fig. 3. The cluster dendrogram. Notes: (1) The nodes in this figure represent groups (the root node is the largest possible group—all the 22 newspapers). Each nonterminal node ('parent') has two daughter nodes, representing the two clusters merged to form the parent. (2) The height of each node is proportional to the value of the intergroup distance between its two daughters. We use cosine distance and average distance in the clustering process. Sources: CNKI (cnki.net) and the China Digital Library (www.apabi.com).

Next, we calculate the index. We obtain the results shown in Table 5. The sign of the index number indicates the direction of media bias. We address the outlier by ignoring it at first and treating {LE, CP} as the root node. After identifying LE and CP's respective positions on the number line, we calculate the position of *Southern Metropolis Daily* according to the distance between it and {LE, CP}.

Table 4
The characteristic words of CP and LE.

Terms (CP)	Frequency	Terms (CP)	Frequency	Terms (LE)	Frequency	Terms (LE)	Frequency
Publicity	535	Work hard	290	Enterprise	415	Activity	127
Thought	485	Emphasise	289	Spirit	223	People	125
Congress	467	Party members	281	Market	213	City	122
Economics	465	Continue	280	Learn	185	Have	121
Requirement	454	Educate	279	Hangzhou	180	This year	120
Apply	433	Plan	276	Socialism	179	Goal	119
Question	432	Opening up	276	Report	176	Set forth	119
Politics	428	Theory	274	Shenzhen	172	Official	116
Country	418	Further	271	Focus	163	Implement	114
Enhance	415	Great	270	All around	160	Currently	114
Comrade	386	Many	266	Policy	156	World	111
Intensive	375	Aspects	262	International	153	Environment	110
Task	371	Profound	259	Industrial	150	Shunde	108
News	366	Outlook on development	257	Progress	149	Management	107
Major	359	Serve	255	Characteristic	145	Carry out	105
Civility	358	Accelerate	249	Zhejiang	144	Project	104
Moderately prosperous	344	Guidance	247	Represent	142	Transformation	104
More	342	Living	246	Leadership	141	Adhere to	102
Cause	337	Strategy	244	Journalist	140	Promote	102
Complete	315	Hu Jintao	243	Make	137	Conference	101
Hold	308	Standard	243	Future	136	Carry out	101
Must	300	Basic	238	Science	132	Improve	101
Innovate	297			Ecology	132	Technology	99
Harder	293			People	130	Requirement	99
Central committee	293			Guangdong	128	System	99

Notes:

1. LE denotes the cluster containing *City Express*, *Yangcheng Evening News*, *Southern Daily*, *Qianjiang Evening News*, and *Zhejiang Daily*, and CP denotes the cluster containing the other newspapers. We include Southern Metropolis Daily, the outlier, in LE when we calculate the results in this table.
2. We translate these phrases into English based upon the official translation of *Report to the Eighteenth National Congress of the Communist Party of China.*

Sources: CNKI (cnki.net) and the China Digital Library (www.apabi.com).

Table 5
The index of media bias.

Indices	Original index	*Z*-score
Newspapers		
Southern Metropolis Daily	−0.135209271	−2.911730479
City Express	−0.065073794	−1.423198896
Yangcheng Evening News	−0.05601426	−1.230922418
Southern Daily	−0.051593451	−1.137096667
Qianjiang Evening News	−0.040344209	−0.898346569
Zhejiang Daily	−0.030849889	−0.69684234
Chongqing Daily	0.010932343	0.189929585
People's Daily	0.013685746	0.248366876
Jilin Daily	0.014900601	0.274150547
New Wenhua Post	0.015104302	0.278473828
Shanxi Evening News	0.015790125	0.293029503
Liberation Daily	0.019031372	0.36182063
Oriental Morning Post	0.020073887	0.383946615
Beijing Daily	0.021910612	0.422928644
Taiyuan Evening News	0.022571381	0.436952581
Beijing News	0.02335146	0.453508713
Xinmin Evening News	0.023698201	0.460867826
Chongqing Morning Post	0.024175045	0.470988202
Chongqing Times	0.024449015	0.476802848
Beijing Times	0.036375903	0.729935075
Shanxi Daily	0.06127516	1.258388463
City Evening New	0.075394259	1.558047433

Note: 'Original index' is calculated based on the projection method in the appendix and distances obtained in the clustering process. *Z* score is the standardised index of 'Original Index'.
Sources: CNKI (cnki.net) and the China Digital Library (www.apabi.com).

4. Discussion of results

4.1. Party supervises mass media

First, we observe that newspapers from the same province are close on the political spectrum. *Liberation Daily* and *Oriental Morning Post*, *Chongqing Morning Post* and *Chongqing Times*, and *Jilin Daily* and *New Wenhua Post* are next to each other on the political spectrum. Similarly, the LE group comprises six newspapers from Zhejiang and Guangdong.

To test the hypothesis that newspapers from the same province have similar political stances, we create a 22*22 dummy matrix, where 0 stands for 'from same province' and 1 for 'from different provinces'. Nest, we use the Mantel test to test the null hypothesis that this dummy matrix is uncorrelated with the distance matrix calculated from our index (rather than the original distance matrix). The significance is well below 0.01, which suggests that we can safely reject the hypothesis that the dummy matrix and the distance matrix are uncorrelated. The observed correlation, $r = 0.1665$, suggests that the matrix entries are positively correlated. Therefore, we can claim that newspapers from the same province tend to have similar political stances.

This fact can be interpreted as resulting from the principle that the party supervises the work of the mass media. Indeed, party committees (directly or indirectly) control newspapers in their jurisdictions. For example, local commercial newspapers are required to reprint politically important articles, usually editorials, of party newspapers of the same level. This requirement is likely to reduce the distance between the newspapers in the ideology space.

4.2. Central/local and political/economic

To determine the principal difference amongst newspapers in the sample, we cannot rely on any predetermined meaning of 'media bias', as it varies from country to country and time to time. Fortunately, we can observe how the LE group and the CP group differ from each other by examining Table 4. On the one hand, the CP group repeatedly mentions political/ideological issues ('publicity', 'congress', 'thought', 'comrade', and 'theory'), whereas the LE group frequently references economic issues ('enterprise', 'market', 'policy', 'international', 'industrial', 'science', 'ecology', and 'technology'). This is the political/economic dimension. At the same time, the CP group is constantly expressing loyalty to the central committee ('apply', 'enhance', 'task', 'complete', 'must', and 'work hard'), while the LE group regularly mentions toponyms of their home markets ('Zhejiang', 'Hangzhou', 'Shenzhen', 'Guangdong', and 'Shunde'). This is the central/local dimension.

Why do newspapers from Zhejiang and Guangdong cover predominantly economic and local issues? A possible explanation is that rich entrepreneurs in Zhejiang and Guangdong, where the private sector prevails, exert considerable influence over local governments. It is beyond the scope of this paper, however, to model interactions amongst governments, media and elites. The political/economic dimension, moreover, can also be perceived as a part of 'depoliticised politics' by Wang (2006).

4.3. Cherry-pick by politicians

We observe that terms in Table 4 are not antagonistic. This finding makes sense because all newspapers in China are subject to Party control. Instead of debating with each other directly, which is not allowed, newspapers choose what they want to report from official documents, leaving out what they do not like. This is best exemplified by the following statements:

(We should) plan for the long term and take all things into consideration. Distribute the pie more fairly while enlarging the pie. [Bo Xilai, the then Secretary of the Chongqing Municipal Committee of the Communist Party of China (CPC), 10 July 2011].

Unswervingly adhere to economic development as the central task, and unswervingly adhere to the Scientific Outlook on Development. This is of great pertinence... A fair division of the pie is not the emphasis of our work; it is enlarging the pie that is important—this is pertinent. These words are not new, but it is novel to emphasise them now. [Wang Yang, the then Secretary of the Guangdong Provincial Committee of CPC, 11 July 2011].

This is a rare case of public debate between CPC officials. Bo Xilai's speech was the headline story of the next day's Chongqing Daily, whereas the informal and fierce words of Wang Yang only appeared in *New Express*, a tabloid affiliated with *Yangcheng Evening News*. In fact, we can find arguments in official documents of the Central Committee to support both of them. The following statements are extracted from the *Report to the Eighteenth National Congress of the Communist Party of China*:

Taking economic development as the central task is vital to national renewal, and development still holds the key to addressing all the problems we have in China... A proper balance should be struck between efficiency and fairness in both primary and secondary distribution, with particular emphasis on fairness in secondary distribution.

The first statement can be used to support Wang Yang, and the second can be used to support Bo Xilai.

5. Robustness

As we discussed in Section 2, we should control for irrelevant factors to claim that our index is truly an index of media bias. Given that there are many factors that can influence news articles, we most likely fail to control for all of them. We maintain that our method still works. In fact, at the initial stage of this project, we conducted clustering analysis using 21 newspapers from seven provinces along with the *Report to the Eighteenth National Congress of the Communist Party of China*. It came as no surprise that the report itself constituted one group, and all of the newspapers composed the other group. The result is not *invalid* in that in this particular sample the principal difference is the type and writing style of these documents. The wording of an official report is most likely different from that of a newspaper.

Consider another example. Say that we are looking at a collection of documents containing only local newspapers from New York and Washington, and we want to identify left wing and right wing papers in the sample. We may find that the characteristic words are words or phrases related to the two cities. This is not what we want to measure. However, we can simply extract more clusters. The dendrogram of hierarchical cluster analysis depicts all possible clusters. Thus, if we divide the collection of documents into four clusters, we may be able to distinguish left from right within each city. At any rate, controlling for irrelevant factors as the first step is always the best choice.

The existence of nonpolitical differences is not a concern, as long as we obtain politically meaningful characteristic words, as shown in Table 4. In our analysis, one of our initial hypotheses was that the principal difference might be the difference between commercial and party newspapers. Had this been the case, we would have found all the party newspapers in one group and all the commercial newspapers in the other. Fig. 3, however, rebuts this scenario. While party newspapers and commercial newspapers are indeed highly different in terms of readership and what they can report, what our results suggest is that this type of difference is not the most relevant one in our sample.

In a word, the ability of our framework to detect and quantify the principal difference in 'any' sample makes it applicable to the analysis of other topics in other samples.

6. Conclusion

Although the CPC adheres to the principle that the Party supervises the work of the mass media, we observe substantial differences amongst Chinese newspapers. The principal difference observed in Chinese media is reflected in two dimensions: central/local and political/economic. This finding is different from what (Gentzkow & Shapiro, 2010) observe in the U.S. Our method can also convert the principal difference into an index, which can be incorporated into related research.

The method that we use in this paper can detect and quantify the principal difference in a sample without imposing prior beliefs. This method requires only text from newspapers and therefore has the potential to be applied in other spheres. For instance, when there is a natural disaster or an accident, party newspapers may behave differently from market-oriented newspapers. Our method can be used to detect and quantify this difference.

Our sample is limited, however, which can be problematic for two reasons. First, all of our results are conditional on this limited sample. The structure that we observe may be violated if we add more newspapers. The detected principal difference could change substantially if we change the topic. Second, only with a large sample can we use a resampling method to calculate the *p*-values of clusters. The characteristic words identified in this paper are meaningful. With more solid statistical validation, however, our results would be more convincing.

Appendix A. Characteristic words

To find out how one cluster is different from another, we use a simple algorithm.

First, we add together rows in the same cluster, yielding a matrix with just two rows. The value of each entry in this new matrix is the frequency of the corresponding term in the corresponding cluster. Next we extract the 100 (or 200, such that we can obtain meaningful results) most used words in each cluster, obtaining two frequency tables A and B. Next, we can use a set of rules to extract two subsets B_p and A_p from B and A. The rules are as follows:

1. If a term (word or phrase) only appears in B, we keep it in B_p. The same goes for A and A_p.
2. If a term appears in both B and A, and the difference of the ranking numbers of this term in B and A is more than or equal to 10 (or another meaningful number), this term is kept in the subset of the frequency table where it is ranked higher.
3. If a term appears in both B and A, but the difference of the ranking numbers of this term in B and A is less than 10, we do not keep it in either subsets.

After these steps, we obtain two new frequency tables, which may show us how one cluster is different from the other; this is the principal difference. We can also use statistical testing to find those words that can distinguish the two groups, as in (Gentzkow & Shapiro, 2010).

A.1. Projection of the structure

In this section, we discuss how to project the dendrogram onto a number line to obtain an index. First, the width of each node should be proportional to the value of the intergroup distance between its two daughters. Below, we use a hypothetical example to illustrate how to construct the index of media bias.

In Fig. 4, there are four terminal nodes: A, B, C, and D. The width and the height of each node are equal and proportional to the value of the intergroup distance between its two daughters. For example, the width of the root node is 12, which is exactly the height of the root node. Moreover, the order of terminal nodes is based on their similarity to their parent's siblings (Alon et al., 1999). Within the cluster $\{A,B\}$, A is on the left side because the distance dist$(A,\{C,D\})$ between A and $\{C,D\}$ is greater than the distance dist$(B,\{C,D\})$ between B and $\{C,D\}$. C and D are ordered in the same way.

Next, we project this dendrogram onto a number line, where the origin is the midpoint of the root node. The signed distance between a point and the origin can be used as an index of media bias. The sign of the distance represents the direction of media bias and the absolute value of the distance represents the magnitude of media bias. Dendrograms from real data are usually much more complex. In this situation, we could start from the root node and iterate this process.

This would be mistaken, however. If C is not a terminal node, then the daughter nodes of C might be on the left end of the number line, which contradicts its structural position. Let a_i denote the distance between the two daughter nodes of the ith node (a_1 is the intergroup distance between two daughters of the root node). Because we use hierarchical clustering, $a_i > a_j \forall j > i$. Following the method above, we might have $0.5\sum_{i=2}^{\infty} a_i > 0.5a_1$. This is because the sequence $\{a_i\}_{i=1}^{\infty}$ has no limits.

Thus we adopt a different method. Let b_i denote the width of the ith node. By defining $b_i = \left(\frac{1}{2}\right)^{i-1} a_i$, we have $0.5\sum_{i=2}^{\infty} b_i < 0.5b_1$. The implication of this method is that the difference between two groups is more important than differences within each group (Table 6).

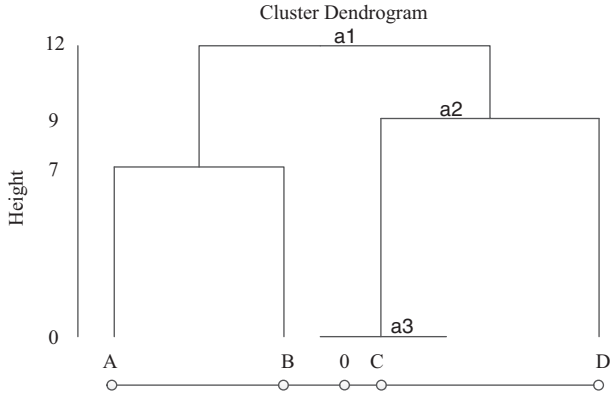

Fig. 4. A way to project the dendrogram. Notes: (1) This is a hypothetical example used to illustrate our method. The nodes in this figure represent groups. Each nonterminal node ('parent') has two daughter nodes, representing the two clusters merged to form the parent. (2) The height and the width of each node are the same and proportional to the value of the intergroup distance between its two daughters.

Table 6
The characteristic words in Chinese.
Sources: CNKI (cnki.net) and the China Digital Library (www.apabi.com).

词组(CP)	频数	词组(CP)	频数	词组(LE)	频数	词组(LE)	频数
宣传	535	努力	290	企业	415	活动	127
思想	485	强调	289	精神	223	人民	125
大会	467	党员	281	市场	213	城市	122
经济	465	不断	280	学习	185	已经	121
要求	454	教育	279	杭州	180	今年	120
落实	433	部署	276	社会主义	179	目标	119
问题	432	开放	276	报告	176	提出	119
政治	428	理论	274	深圳	172	干部	116
国家	418	进一步	271	中心	163	贯彻	114
加强	415	伟大	270	全面	160	目前	114
同志	386	各项	266	政策	156	世界	111
深入	375	方面	262	国际	153	环境	110
任务	371	深刻	259	产业	150	顺德	108
新闻	366	发展观	257	推进	149	管理	107
重大	359	服务	255	特色	145	进行	105
文明	358	加快	249	浙江	144	项目	104
小康	344	指导	247	代表	142	转型	104
更加	342	生活	246	领导	141	坚持	102
事业	337	战略	244	记者	140	推动	102
建成	315	胡锦涛	243	成为	137	会议	101
召开	308	水平	243	未来	136	开展	101
必须	300	基本	238	科学	132	提高	101
创新	297			生态	132	技术	99
认真	293			群众	130	需要	99
中央	293			广东	128	制度	99

Note: LE denotes the cluster containing *City Express*, *Yangcheng Evening News*, *Southern Daily*, *Qianjiang Evening News*, and *Zhejiang Daily*, and CP denotes the cluster containing the other newspapers. We include Southern Metropolis Daily, the outlier, in LE when we calculate the results in this table.

References

Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., & Levine, A.J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, *96*(12), 6745–6750 PMID: 10359783.

Anderson, S.P., & McLaren, J. (2012). Media mergers and media bias with rational consumers. *Journal of the European Economic Association*, *10*(4), 831–859.

Baron, D.P. (2006). Persistent media bias. *Journal of Public Economics*, *90*(1–2), 1–36.

Bernhardt, D., Krasa, S., & Polborn, M. (2008). Political polarization and the electoral effects of media bias. *Journal of Public Economics*, *92*(5–6), 1092–1104.

Besley, T., & Prat, A. (2006). Handcuffs for the grabbing hand? Media capture and government accountability. *The American Economic Review*, *96*(3), 720–736.

Covert, T.J., Adkins, & Wasburn, P.C. (2007). *Measuring media bias: A content analysis of Time and Newsweek coverage of domestic social issues 1975-2000 , 88.*

Feldman, R., & Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data.* Cambridge University Press.

Gentzkow, M., & Shapiro, J.M. (2010). What drives media slant? Evidence from U.S. Daily newspapers. *Econometrica*, *78*(1), 35–71.

Groeling, T. (2013). Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, *16*(1), 129–151.

Groseclose, T., & Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics*, *120*(4), 1191–1237.

Lott, J.R., & Hassett, K.A. (2004). Is newspaper coverage of economic events politically biased? *SSRN Electronic Journal.*

Niven, D. (2003). Objective evidence on media bias: Newspaper coverage of congressional party switchers. *Journalism and Mass Communication Quarterly*, *80*(2), 311–326.

Niven, D. (2004). A fair test of media bias: party, race, and gender in coverage of the 1992 house banking scandal. *Polity*, *36*(4), 637–649.

Qin, B., Wu, Y., & Strömberg, D. (2012). The determinants of media bias in China. *Tech. rept, 00010*, .
Schiffer, A.J. (2006). Assessing partisan bias in political news: The case(s) of local senate election coverage. *Political Communication*, *23*(1), 23–39.
Shimodaira, H. (2002). An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, *51*(3), 492–508.
Shimodaira, H. (2004). Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *The Annals of Statistics*, *32*(6), 2616–2641.
Suzuki, R., & Shimodaira, H. (2004). An application of multiscale bootstrap resampling to hierarchical clustering of microarray data: How accurate are these clusters. *The Fifteenth International Conference on Genome Informatics. 34*.
Wang, H. (2006). Depoliticized politics, multiple components of hegemony, and the eclipse of the sixties. *Inter-Asia Cultural Studies*, *7*(4), 683–700.