Network Agenda Setting and Social Cognition Construction of the Dengue Fever Epidemic Event based on Social Media Big Data

Completed Research Paper

Yuejiao Wang **Zhidong Cao**

Abstract

This article collected dengue-related information in China from News, Apps, Forums, Microblog and WeChat. We extract the 100 highest frequency of agenda settings from News, and built agenda setting networks of the five platforms respectively. Social network analysis methods are used to reveal the characteristics of these five agenda setting networks and the QAP is used to test the correlation between them. The robustness analysis of the results is applied. We found that: the information dissemination in dengue can be roughly divided into three categories: epidemic situation and change, prevention and control measures, and epidemic outbreak area; The context of News has the highest similarity with Apps (correlation coefficient=0.9561, $p<0.01$), the second similarity with WeChat (correlation coefficient=0.8291, $p<0.01$), the third similarity with Forums (correlation coefficient=0.7916, $p<0.01$), the lowest similarity with Microblog (correlation coefficient=0. 4280, $p<0.01$); the QAP is robust, which can be widely used in social media big data research.

Keywords: Network agenda-setting, QAP, big data, dengue

Introduction

Public health is a key area related to national security and people's lives and property. If we do not conduct effective public opinion dissemination and control, it will cause the spread of panic and the deterioration of the infectious disease situation, as well as greater economic losses. The dengue epidemic in southern China is a serious public health problem, and all previous outbreaks pose threats to social security and stability. Once the outbreak of dengue fever occurs, national and local news media will focus on the epidemic-related information reports, and these reports will spread widely through news websites, WeChat, apps, microblog, forums and other Internet platforms, thus influencing and mobilizing the public. Therefore, it is necessary to publicize public health in news and media to deepen people's awareness of prevention, and do survey feedback on social awareness. Previously, researchers often used statistics and questionnaires to get feedback about the social

cognition on public health, which takes a lot of time and efforts. We need an efficient and innovative method to investigate social cognition on health events and measure the relationships between different media platforms.

In the era of big data and Internet, with data mining methods, scholars can do many empirical study on theoretical innovation (Zhian et al. 2017): Chris J. et al. (2014) find support for agenda melding and validates the network agenda-setting (NAS) model through computer science methods with large datasets from Twitter during the 2012 U.S. presidential election; J, K., et al. (2016) use the big data from Twitter to analysis the public emotional change after the Fukushima nuclear accident; Park J, et al. (2014) use Twitter to study nonverbal cues in emoticons. We can have many amazing discoveries combining big data with traditional theory of journalism (Anbin et al. 2017). And in the era of media, data-driven communication research is the frontier of journalism and communication (Anbin et al. 2014). As for the challenges, it is the phenomena of post-truth: the 2016 U.S. presidential election (Lee et al. 2018) reflect the decline of traditional news media and the popularity of social media. The recommended system changes the social media into an echo chamber to some extents (Colleoni et al. 2014). Truth is no longer the principle of news, but some people's emotion and standpoints (Yiqing et al. 2018). So, scholars should pay attention to the agenda-setting model of social media and social cognition.

Therefore, drawing on the social cognitive analysis methods of other scholars in the political and cultural fields, we innovatively introduce the NAS theory and other social network analysis theories into the field of public health security. We take dengue, a mosquito-borne infectious disease, as the research object.

This paper is organized as follows. We first introduce the NAS theory and quadratic assignment procedure (QAP) method. Then, we describe the data analysis technologies in detail: data filter, keywords extraction, data visualization methods and so on. Next, we present our results in the form of chart and figure, and explain our findings. A test about the robustness analysis of the results is also given in this section. We conclude the paper with a discussion on our contributions and the future directions of this research project.

Theoretical Background

The Development of Agenda-Setting Theory

It is the information on news media that construct the world in our mind (McCombs 2002). Before the popularization of Internet, by headlines or emphases, newspapers and television news tell us what are important and set the agendas for public. In other words, people are more likely to pay attention to what the media emphasizes on. So, McCombs and Shaw (1972) proposed the first level of agenda-setting model that the salience of issues emphasized by the news media could be transferred to the public's mind.

As the research further developed, McCombs et al. (1997) found that each of the objects has numerous attributes which describe the object, and news media can structure the public's knowledge of what the most salient attributes of these topics are. This is the second level of agenda-setting model.

However, social media take the initiative in the last 10 years and people can receive information from various sources. It is the agenda networks that form our mind. To adapt to the situation, Lei Guo and McCombs (2012) proposed an expanded model on agenda-setting effects: the network agenda-setting (NAS) model. They did empirical tests and found the conclusion: news media can bundle sets of objects or attributes and make these bundles of elements salient in the public's mind, which forms the third level of agenda-setting model.

But why agenda-setting effects occur? The concept of need for orientation and dual psychological paths is explained by psychology research on agenda-setting effects (McCombs et al. 2014).

The agenda-setting theory shows great vitality in the past 50 years since it was proposed. McCombs and Shaw, et al. (2014) summarized seven distinct facets in the theory and three of the seven facets: need for orientation, network agenda setting, and agenda melding, are the particularly active

theoretical arenas in the research. Especially, the NAS model demonstrates a key social role for the news media in citizen's participation in public affairs (McCombs 2005). Vu et al. (2014) seek to expand the NSA model's scope by testing five years (2007-2011) of aggregated data from national news media and polls. They verified the theory and demonstrated strong network correlations of issue salience among different types of news media.

Network Agenda-Setting Theory

The NAS theory is the third level of agenda-setting theory, which is pretty well adapted to the mixed mass media on Internet. Messages on Internet have the feature of diversification, so the ways information flows and obtained have changed into a reticular pattern. The public are not only the consumers of news, but also the producers of news. Besides, the study on human cognitive structure also tells us that our cognitive structure is not linear, but nearly networked. In the networked structure, different nodes link together into a cognitive map.

The core idea of NAS theory is that agenda-setting networks of news which have the feature of topics and attributes can affect the topics and attributes in agenda-setting networks of the public. NAS theory does not focus on the relationship of single topic or attribute, but focuses on the relationship of agenda-setting networks structured with topics and attributes.

To conform the theory, Lei Guo and McCombs (2012) used the data of Texas gubernatorial election and senate election in the year of 2002. These data were once used in their prior paper to study the attribute agenda-setting model, the second level of agenda-setting theory. In that paper, they chose the top ten personality characteristics words with the highest frequency that appeared in mess media, for example, leadership, experience, ability. Then they collected the reports and descriptions about candidates using these ten personality characteristics words from both the local media and the public and found that the characteristics of a candidate shaped by the media are highly consistent with the public impression of the candidate.

But when they used these data to validate the NAS theory, what they focused on is the co-occurrence of these 10 words. Lei Guo and McCombs first set a square of ten order, and the rows and columns labels are all the 10 personality characteristics words. To find the relationship of the 10 words, they counted the frequency of the words appearing together in the text and used the co-occurrence frequency as the element value of the matrix. The more the two words appearing together, the stronger the correlation they have. So, researchers constructed the co-occurrence matrix of the news agenda as well as the co-occurrence matrix of the public agenda. And the text used to calculate the co-occurrence frequency in each matrix are local news and citizen questionnaires.

After applying the quadratic assignment procedure (QAP) test, it is found that these two co-occurrence matrices are highly positive correlated. So, the news agenda-setting network significantly affect the people's agenda-setting network. In the end, the cognitive networks of the public about these candidates are constructed in a visual way using weighted undirected graph. The nodes in graphs are the ten words, and the links are weighted by frequency.

What we need to pay attention is the concept of "co-occurrence". if two words often appear together in text, people will naturally think that these two words are related. However, this relationship can be deliberate structured by news media, causing the deviation between the public cognitive networks and the reality. In the era of social media, this bias has been further amplified and spread. That is what we want to research in this paper: how to construct the agenda-setting networks of news and the public? What is the relationship between them?

So, co-occurrence matrix can both show the agenda-setting networks of news and the cognitive structure of people. Matrix and weighted undirected graphs are two equivalent representations. And researchers use keywords to approximate agendas in the networks, which is acceptable.

Quadratic Assignment Procedure

We focus on the correlation of news' agenda-setting network and the social media's agenda-setting network in this paper. Are they positively correlated or negatively correlated? What's more, how do we test for the correlation significance? Because we use co-occurrence matrices to represent agenda-setting networks, the key problem is the correlation significance test of two co-occurrence matrices. This type of matrix is relational data, showing the co-occurrence relation between keywords. Quantitatively, since relational data themselves are data about associations, the principle of avoiding collinearity is directly violated. So, many conventional statistical techniques, such as OLS, can't be simply applied to the analysis of relational data and some specific methods are required, and quadratic assignment procedure (QAP) is an important method that can test for the correlation significance of relational data (KRACKHARDT 1987).

To explain how QAP works, we will give an example. As shown in Figure 1 in graph form, these two networks measure six persons' friendship structure and debtor-creditor relationship. Through observation, the two networks are so similar that it can be almost considered that the relationship between friendship and debtor-creditor is positively correlated.

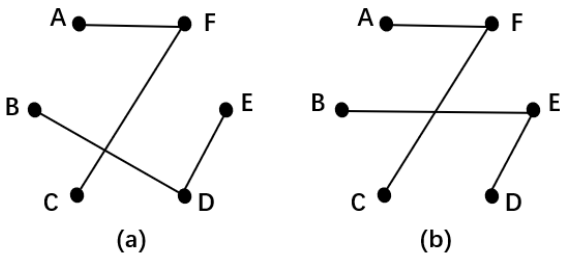

Figure 1. Structure of Friendship (a) and Debtor-creditor Relationship (b)

Figure 1 are two undirected graphs, and we represent the graphs in matrices in Figure 2. The Pearson correlation coefficient between the two original matrices is 0.659. To calculate the Pearson correlation coefficient between two matrices, we rearrange the matrices into long vectors in row order. The formula of the Pearson correlation coefficient is shown blow:

$$r = \frac{1}{n\text{-}1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right) \tag{1}$$

in Formula 1, n is the length of vector, x_i and y_i are elements of vectors, and σ_x, σ_y are standard deviations.

	A	B	C	D	E	F
A	0					1
B		0		1		
C			0			1
D		1		0	1	
E				1	0	
F	1		1			0

(a)

	A	B	C	D	E	F
A	0					1
B		0			1	
C			0			1
D				0	1	
E		1		1	0	
F	1		1			0

(b)

Figure 2. Matrices of Friendship (a) and Debtor-creditor Relationship (b)

We randomly permute the order of the six labels, and get the probability distribution function of correlation coefficients, as shown in Figure 3. It can be found from the probability density distribution diagram that among all the random permutation correlation coefficients, 98.5% of them are less than

the true coefficient 0.659. In other words, the correlation coefficient of 0.659 can't be generated randomly. It passes the confidence test with a significance of 0.05 obviously.

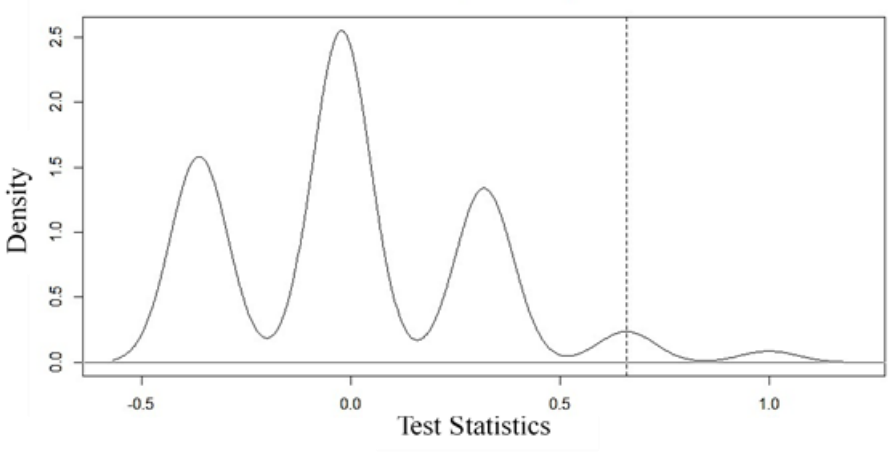

Figure 3. QAP Test Result

The above case focuses on the label rearrangement of data. If the network is larger, the times of substitutions will increase dramatically. In practical applications, QAP permute the rows and columns of a matrix simultaneously. Here are three steps. First, reshape the original two matrices into long vectors and calculate the Pearson's correlation coefficient. Then, perform random permutation on the rows and the corresponding columns of one of the matrices. One should pay attention that we can't just permutate rows or columns. And we should only permutate one matrix, the other matrix keeps the same. Last, calculate correlation coefficient between the permutated matrix and the other matrix. Repeat the last two steps hundreds or even thousands of times and we will get a distribution of the correlation coefficients. Compared the distribution with the true correlation coefficient obtained in the first step, to see whether the true correlation coefficient falls into the rejection domain or the acceptance domain, and make a judgment. If the ratio is lower than 0.05 (assuming the significance level determined by the researcher is 0.05), it indicates that there is a strong relationship between the two matrices, and the correlation coefficient between them is unlikely to be random.

Research Method

According to NAS theory, we use keywords co-occurrence matrices and weighted undirected graph to represent agenda setting networks. Then, applying QAP test on agenda setting networks, we can come to the conclusion on the relation of news media and social cognition and calculate the correlation coefficients of different media platforms. Since opinions on mass media can show social cognition, opinions on mass media is equivalent to social cognition in this paper, and what follows in the passage, we use "data of mass media" to represent "social cognition". The data processing flow chart is in Figure 4.

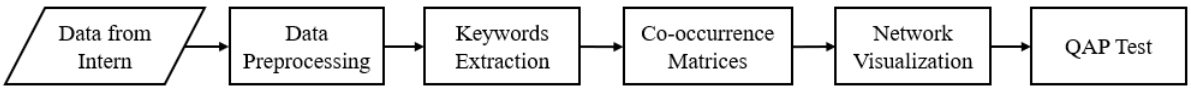

Figure 4. Data Processing Flow Chart

Data Sets and Preprocessing

Introduction of Data Sets

We get all the data about news media and mass media from Internet data system of our research institute. This data system grabs dengue-related news from different platforms and websites. Here are five data sets obtained from five kinds of platforms. The time range is from January 5, 2016 to August 13, 2018.

"News" data set contains authoritative news from newspapers' main websites and other news media websites, which represent the position of relevant departments and experts to the dengue fever. Here are 31529 pieces of data in the "News" data set and every piece of data has five labels: title, site name, URL, published time, content.

"Apps" data set contains epidemic analysis and reports from different news apps. Reports in apps may be the republications of news or other informal news from invisible advertisers. So, some news in apps may be deceptive and commercially beneficial. Here are 13183 pieces of data in the "Apps" data set and every piece of data has five labels: title, site name, URL, published time, content.

"Forum" data set contains people's comments about dengue and reports republications. Some of the posts reflect people's personal experiences and feelings, and the public's awareness to dengue is better reflected in the "Forums" data set. Here are 9521 pieces of data in the "Forum" data set and every piece of data has six labels: title, site name, URL, published time, blogger, content.

"Microblog" data set is the opinions collection of Sina microblogs. There are limits on how much text can be posted on microblog, so the opinions of bloggers are much shorter than news. Similar to the "Forums" data set, public opinion matters more in "Microblogs" data set, including different attentions to dengue fever from local communities, hospitals and Internet celebrities. Here are 9368 pieces of data in the "Microblog" data set and every piece of data has five labels: website name, URL, published time, blogger, content.

"WeChat" data set contains opinions about dengue from WeChat public accounts. WeChat public account is an application account applied by developers or merchants on the WeChat public platform. Through the public account, merchants can realize all-round communication and interaction with text. Institutions and individuals can spread information and values through their public accounts. Therefore, the "WeChat" data set mixed with truth and business interests. Here are 31191 pieces of data in the "WeChat" data set and every piece of data has five labels: title, URL, published time, blogger, content.

As we can see, in these five data sets, "News" data set can represent the voice of institutions and government departments, and the rest four data sets can represent the views of the people and non-governmental organizations, that is, social cognition, even they contain some rumors. Once the keywords are extracted and the co-occurrence matrices are constructed, the correlation of the five platforms can be obtained. And the relationship between news and social cognition is naturally clear.

Data Preprocessing

Big data often contains noise, and data preprocessing can reduce data redundancy and improve data processing accuracy. Limited by data capture technology, here are some duplicate data and irrelevant data with dengue.

We consider two pieces of data with the same "title" and "published time" to be duplicated, and message republished by the same website in different time is considered as non-repeated message. Using built-in data deduplication functions in Excel, the duplicated data can be removed easily. We do the same deduplication options for the five data sets.

Next, some data has nothing to do with dengue because the condition setting of regular expression is not rigorous enough in data acquisition stage. Through custom Excel functions, we think the data that do not have the keyword "dengue" in both "title" and "content" fields to be irrelevant with dengue and remove them. We do the same operation for the five data sets.

After data preprocessing, the sizes of the five datasets have decreased into 16986, 7190, 1459, 9342, 17239 respectively.

Agenda Keywords Extraction and Co-occurrence Matrices

One of the difficulties of this project is the construction of agenda setting networks, which are essentially the weighted undirected graphs. Agenda keywords are node labels in the graph, and

keywords co-occurrence matrices are equivalent representation of graphs. We use some text analysis techniques to extract agenda keywords and form their co-occurrence matrix.

Agenda Keywords Extraction

People's comments on a news event mainly focus on the elements and attributes of the event, and in communications, we call it an agenda. In NAS theory, researchers use keyword extraction to obtain news agendas and mass agendas. Keyword extraction is the focus of natural language processing and data mining. Methods for keyword extraction have evolved over decades. Algorithms based on statistical characteristics, frequency, are the simplest. And there are many other supervised and unsupervised learning algorithms based on word location and part-of-speech.

The keyword extraction method based on word frequency and part-of-speech can meet the needs of this study. The agenda keywords extraction method of the paper can be divided into five steps (Chang et al. 2018):

Tokenizing. Based on Chinese words segmentation dictionary and HMM model, we divide the text string into words for later operations.

Removing stop words. Stop words are words that need to be filtered out to save space and improve efficiency. They often have no meaning or acting as a conjunction, and the stop words lists are usually designed by human selection.

Part-of-speech tagging and filtering. After tokenizing and stop words filtering, we classify the rest of the words according to their part-of-speech. Then filter out adjectives, conjunctions, differentials and other words that have little semantic relevance.

Word frequency statistics and sorting. Conduct frequency statistics on the remaining words and take the top 200 keywords of frequency as candidate keywords.

Manual selection. To select agenda keywords related to dengue, we make a manual selection based on expert experience from the candidate keywords. The remaining top 100 keywords can better define the agendas of dengue.

Benefit from open source code, all of the above steps can be manipulated using python with some open source functions. We import the Chinese word segmentation kit named “jieba” and conduct data processing operations in the python3.7 environment. Because we focus on how news influences the social cognition, all the keywords extraction operations are based on the content of “News” data set.

Keywords Co-occurrence Matrices

Now we obtain top 100 keywords and their frequencies, but word frequency represents only one dimension of the agenda. In NAS theory, there is a network of connections between agendas, and people's perception of events. This agenda setting network is a weighted undirected graph in two dimensions. Its nodes are agendas keywords and the weight of its edges are the frequencies with which two keywords appear together in the same news report. To facilitate subsequent mathematical operations, we use the co-occurrence matrix as the equivalent representation of the graph.

It should be emphasized that the text set for keywords extraction is the “news” dataset, but the five co-occurrence matrices are obtained from the five datasets respectively. The matrix presents the co-occurrence of the 100 agenda keywords in each dataset. It is a 100 - dimensional square matrix which row and column tags correspond to 100 keywords and the element values in the matrix are the occurrences frequency of the corresponding row’s and column’s keywords. Similarly, the computation and storage of co-occurrence matrix can be realized by using python.

Visualization of Agenda Setting Networks

Numbers and data are often not intuitively understood by the brain. The information contained in them is boring and easily to be missed, and data often cannot be accepted by the general public. Data visualization technology makes the hierarchical information displayed in front of us vividly.

For social media and big data networks in our paper, we can use programmatic data processing languages, for example, Matlab, R, for data visualization, or we can use integrated software development platforms and its built-in functions to process the data. Gephi, the open source software we used, can easily do data sorting, statistics, segmentation, filtering and embedded advanced algorithm operations. It meets our network visualization needs and the visual network after processing will be shown in the following chapter.

QAP test

We have five keywords co-occurrence matrices for the five datasets. We need to calculate the Pearson correlation coefficient between every two matrices and then conduct significance test using QAP. So, there are totally 10 sets of operations and tests.

Especially in the significance test using QAP: first, we get the correlation coefficient between matrices A and B (A and B are co-occurrence matrices with zero diagonal elements); second, keep A the same and replace the rows and columns of B simultaneously, then calculate the correlation coefficient once again; third, repeat the second step for hundreds or even thousands of time and get the probability distribution of the correlation coefficients; lastly, check out whether the correlation coefficients in the first step can fall into the confidence interval of the probability distribution function. The final calculation results and the probability density distribution of the correlation coefficients will be shown in the next chapter.

Numerical Stability Test

Numerical stability is a measure of whether the results of an algorithm is reliable. Especially in the field of big data mining, the huge amount of data poses great challenges to both hardware and algorithm time consumption. We want to inspect that whether the result will change if we reduce the amount of data in this project. If the result has the feature of numerical stability, we can use less data to analyze public health incidents with less time and resources.

To do this numerical stability test, when it comes to the step for calculating the keywords co-occurrence matrix, the data in the “news” dataset are randomly selected in proportion. Incrementing by 10%, we get nine sub-datasets and each sub-dataset accounts from 10% to 90% of the total “news” data set. Still using the 100 keywords extracted from the original “news” dataset, we obtain nine keywords co-occurrence matrices from the nine sub-datasets. The correlation coefficients are calculated with the matrices of the other four data sets in turn.

Empirical Results

Agenda Setting Extraction

We get 100 agenda keywords from “News” dataset after filter and frequency statistics, and these words represent the agendas. Considering the length of the paper, Table 1 only lists the top 20 keywords with word frequency and their corresponding English translation.

Because the research background of this project is China and all the data collected are the news about the dengue fever, the keywords extracted are all in Chinese and some of them are medical terms. In order to make this paper easily to be understood, we added corresponding English translation for the Chinese keywords. However, we know that the meaning of Chinese words will change slightly after being translated, but it is acceptable.

The extracted keywords reflect many dimensions of dengue fever: keywords in the pathology level, for example, mosquito and virus, depicts the pathogen and vector of dengue fever; keywords in the prevention and control level, such as prevention and control, reflect measures taken by relevant departments to deal with dengue; keywords in the patient level, such as case and patient, can reflect the public's views to dengue fever. We believe that these keywords can well support our subsequent research and can be used as the expressions of the agendas.

Table 1. Keywords List

Order	Keyword (Chinese)	Keyword (English)	Word frequency	Order	Keyword (Chinese)	Keyword (English)	Word frequency
1	登革热	dengue	63366	11	症状	symptom	16411
2	蚊子	mosquito	43474	12	积水	hydrops	15652
3	传染病	infectious disease	37810	13	患者	patient	13464
4	疫情	epidemic	31566	14	防控	Prevention and control	13410
5	病例	case	23490	15	伊蚊	aedes	12717
6	病毒	virus	20899	16	防蚊	Mosquito prevention	12559
7	传播	spread	20334	17	措施	measures	10693
8	蚊虫	nyamuk	19620	18	卫生	sanitation	9337
9	灭蚊	Anti-mosquito	19190	19	密度	density	8382
10	报告	report	17715	20	监测	surveillance	7859

The Agenda-Setting Networks

We count the frequency of keywords co-occurrence in the dataset and get the five keywords co-occurrence matrices. Indeed, the five matrices are equivalent forms of five agenda setting networks for each of the datasets. We used Gephi for graph visualization to show more details of the agenda setting network. Figure 5 to Figure 9 are the agenda setting networks processed by Gephi.

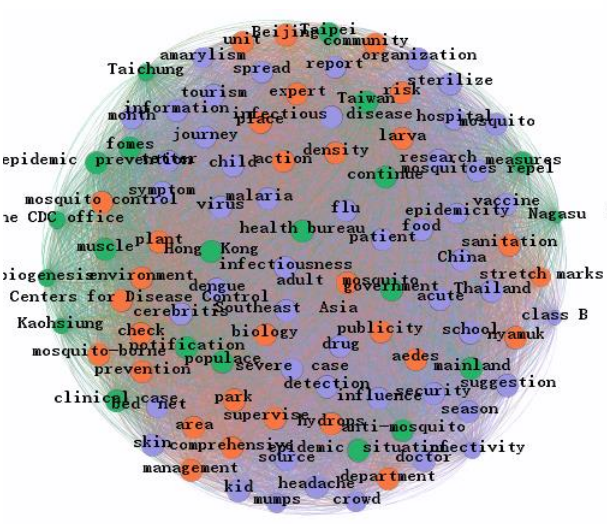

Figure 5. News Agenda Setting Network

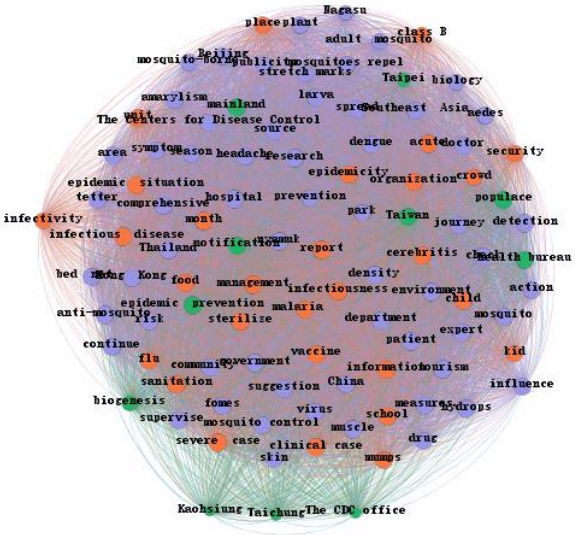

Figure 6. Apps Agenda Setting Network

We import the co-occurrence matrix extracted from the “News” database into Gephi. As we can see in Figure 5, through the built-in modular algorithm, Gephi divides 100 interrelated agenda keyword nodes into three communities, and the nodes of different community categories are stained with different colors. Using Yifan Hu's multilevel algorithm, nodes are arranged into a graph with the contour of a circle. The blue nodes mainly reflect the pathologic and genetic characteristics of dengue

fever; the orange nodes mainly reflect the prevention and control actions of local governments; the green nodes are mainly the locations for concentrated outbreaks of dengue.

The result of module partition in Figure 6 is very similar to that in Figure 5, but in different colors. The orange nodes mainly reflect the pathologic and genetic characteristics of dengue fever; the blue nodes mainly reflect the prevention and control actions of local governments; the green nodes are mainly the locations for concentrated outbreaks of dengue.

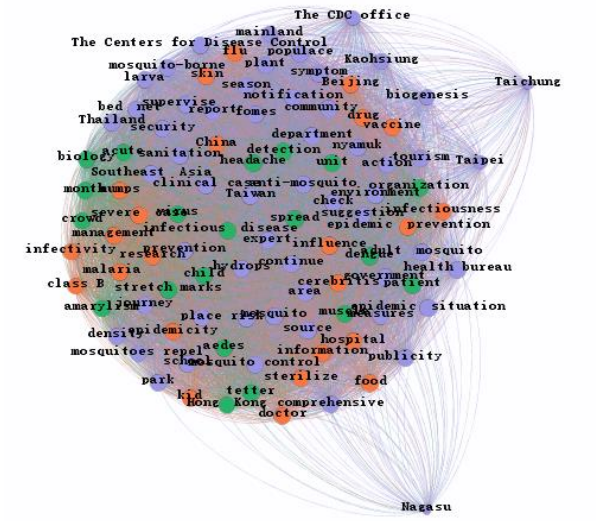

Figure 7. Forum Agenda Setting Network

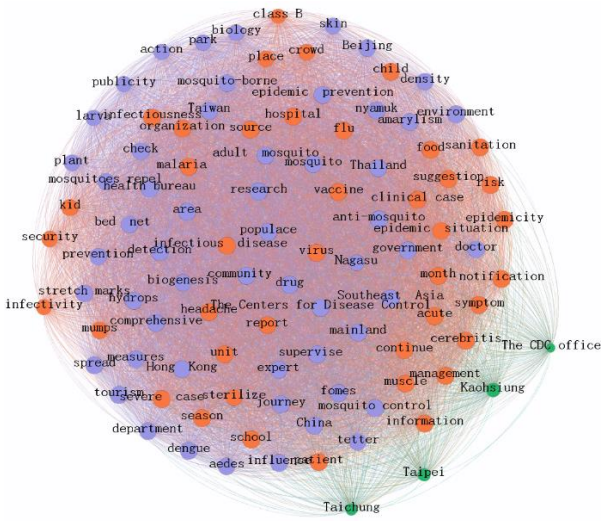

Figure 8. WeChat Agenda Setting Network

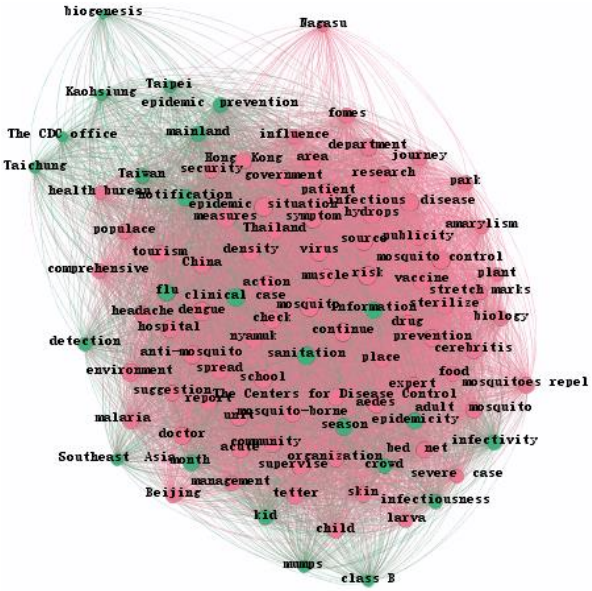

Figure 9. Microblog Agenda Setting Network

However, after careful observation of Figure 7 and Figure 8, the keywords in each module starts to change, which is different from “News” and “Apps” agenda setting networks. This means the similarity of the keyword agenda setting networks began to decline. Compared with “News”, the agenda setting networks of "Forum" and "WeChat" show a drift. We can think that the difference between social cognition and authoritative news reports began to appear. Especially Figure 9, the module division result of the agenda setting network of "Microblog" is more obvious on the drift when comparing with the previous four networks. Using the same module partition parameters, Microblog agent setting network only divide into two classes. Given that the data set size of “Microblog” is more than 9,000 pieces, which is not less than the “Apps”, we believe that it is the difference in social cognition among the masses that leads to the change in the module partition of the agenda network, rather than the data scale.

The QAP test of Agenda-Setting Networks

Using the social network analysis function package "statnet" in R language, we calculate the Pearson correlation coefficients between the five keywords co-occurrence matrices. The result is shown as Table 2. Moreover, using QAP, these results in Table 2 all pass the hypothesis test with a significance level of 0.01, and Figure 10 shows the probability density distributions of the correlation coefficients after 3000 permutations.

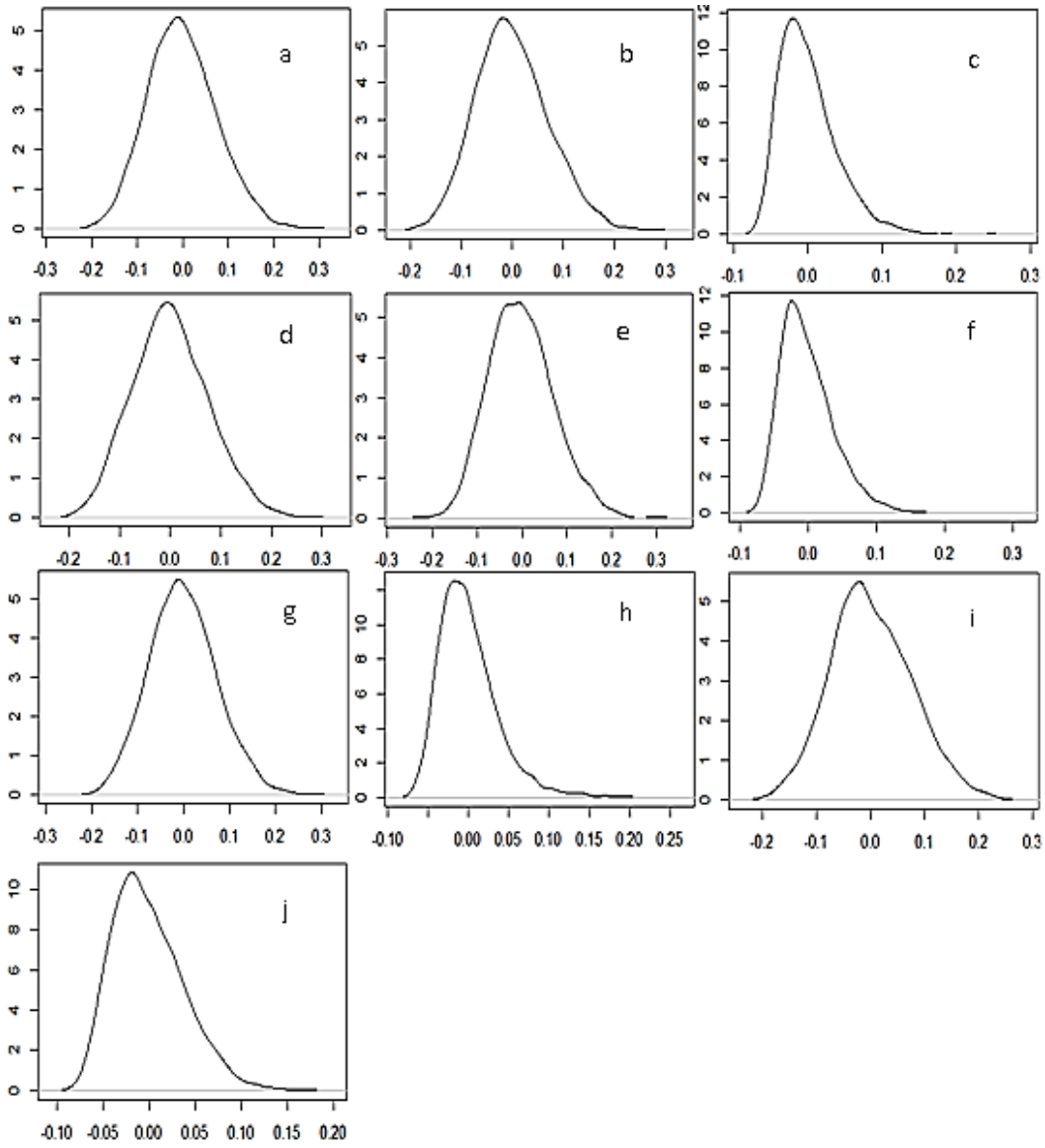

Figure 10. Probability Density Distributions of QAP (a, news & apps; b, news & forum; c, news & microblog; d, news & WeChat; e, apps & forum; f, apps & microblog; g, apps & WeChat; h, forum & microblog; i, forum & WeChat; j, microblog & WeChat)

Taking "News" data set as reference, the agenda setting networks of the remaining four data sets are different. "Apps" and "news" have the highest Pearson correlation coefficient while "Microblog" and "news" has the lowest correlation coefficient. The calculated results of correlation coefficients are consistent with the results of network visualization and module division in the previous section.

We regard the "news" dataset as the authoritative information for dengue, representing the recommendations of relevant government departments and experts, and the rest of the datasets as the reflection of different organizations and the public's awareness of dengue, namely, the social cognition. The correlation differences between the five media platforms show the differences and drift

between official reports and social cognitive. People's understanding of the dengue epidemic is affected by news reports, but there are some cognitive errors and changes in cognitive focus.

Table 2. QAP Correlation Coefficient Matrix between Apps, Forums, Microblog and WeChat

	News	Apps	Forums	Microblog	WeChat
News	---	0.96	0.79	0.43	0.83
Apps	---	---	0.80	0.39	0.91
Forums	---	---	---	0.34	0.76
Microblog	---	---	---	---	0.31

Robustness Analysis of QAP Test

To test the robustness of the algorithm results, we use sub-dataset of "news" dataset to extract keyword co-occurrence matrices, and the nine matrices are named as "sampled-news" co-occurrence matrices. These nine sub-datasets represent 10% to 90% of the original data set, respectively. Table 3 are the QAP correlation coefficients between sampled-news and other datasets.

Table 3. QAP Correlation Coefficients between Sampled-news and Other Datasets

	News	Apps	Forums	Microblog	WeChat
sampled-news (10%)	0.9999	0.9561	0.7916	0.4280	0.8291
sampled-news (20%)	0.9999	0.9561	0.7916	0.4280	0.8291
sampled-news (30%)	0.9999	0.9561	0.7916	0.4280	0.8291
sampled-news (40%)	0.9999	0.9561	0.7916	0.4280	0.8291
sampled-news (50%)	0.9999	0.9561	0.7916	0.4280	0.8291
sampled-news (60%)	0.9999	0.9561	0.7916	0.4280	0.8291
sampled-news (70%)	0.9999	0.9561	0.7916	0.4280	0.8291
sampled-news (80%)	0.9999	0.9561	0.7916	0.4280	0.8291
sampled-news (90%)	0.9999	0.9561	0.7916	0.4280	0.8291

Results in Table 3 indicate the sample from 10% to 90% of data does not affect the QAP test results. the QAP correlation coefficients between sampled-news and other datasets are nearly the same with Table 2. When the data volume reaches a certain scale, the data volume growth has little impact on the QAP test. So, the QAP test in this paper has strong robustness.

Conclusion and Discussion

Taking dengue fever as research object, the paper aims to find a proper way to construct social cognition and analyze the correlation between news report and social cognition. Once reaching the target, we can detect social cognition and evaluate how the news from government and experts influence our social cognition. If the method is proved to be efficient, it can be used to serve the prevention and control works of public health events in the aspect of Internet big data analysis.

We collect news and opinions about dengue from kinds of websites and form five data sets: "News", "Apps", "Forum", "Microblog" and "WeChat". The "News" dataset represents the voice of government, while the others represent social cognition in varying degrees. Draw on the experience of NAS theory, we extract keywords from "News" dataset to describe agendas and construct keywords co-occurrence matrix. Indeed, the matrix is equivalent to the agenda setting network, which is a weighted undirected graph. We extract matrices from the other four datasets in the same way. Using

co-occurrence matrices and agenda networks, we can describe the government's attitude as well as social cognition in both numerical aspect and graph aspect.

We imply module division on graphs and calculate the Pearson's correlation coefficient among the matrices. And There is a consistent result: news reports and social cognition have a strong correlation. The agenda setting networks of "News" and "Apps" have a correlation coefficient of 0.957, while the networks of "News" and "Microblog" have a correlation coefficient of 0.427. So, news reports on dengue fever strongly affect social cognition, but there are differences and drift in people's cognition. In order to verify the statistical significance of correlation coefficient, we conduct QAP test. QAP uses random permutation to obtain the probability distribution of the correlation coefficients and to evaluate the significance level of the original results. It shows that all the correlation coefficients passed the significance test.

To conclusion, NAS theory is a good way to construct social cognition, and news from government strongly influence the agenda setting of people. It proves our coverage of public health is effective, and the news guide people to correctly understand dengue fever. However, in addition to strong correlations, the data also reveal that social cognition still has a drift to the authoritative information. In other words, the meaning of public health news maybe changed as they spread across different media platforms. Or there are slight changes in public's focus on the health events.

There are several areas of this research that deserve further attention. Firstly, whether it is reasonable to use keywords to represent the agenda of an event. Secondly, since we notice the drift in social cognition, do we have any other way to characterize this cognitive error? Maybe we can put social cognition analysis into a richer dimension and analyze it with time or locations. By giving play to the multi-dimensional advantages of online big data, we can better serve the work of social public health or other emergency safety events with data analysis technology.

Acknowledgements

This study was funded by National key research and development program (Nos. 2016YFC1200702, 2016QY02D0305) and National Natural Science Foundation of China (Nos. 91546112, 71621002).

References

Zhian Zhang, Yanhui Cao. 2017. "Big Data and News Communication Research: Hotspots and Reflections," Chinese Publishing (10), pp. 3-11. (in Chinese)

Vargo C. J., Lei G., McCombs M., Shaw D. L. 2014. "Network Issue Agendas on Twitter during the 2012 US Presidential Election," Journal of Communication (64:2), pp. 296-316.

Kim J., Brossard D., Scheufele D. A., Xenos M. 2016. "Shared Information in the Age of Big Data: Exploring Sentiment Expression Related to Nuclear Energy on Twitter," Journalism & Mass Communication Quarterly (93:2), pp. 430-445.

Park J., Beak Y. M., Cha M. 2014. "Cross-Cultural Comparison of Nonverbal Cues in Emoticons on Twitter: Evidence from Big Data Analysis," Journal of Communication (64:2), pp. 333-354.

Anbin Shi, Peinan Wang. 2017. "Agenda Setting Theory and Research for 50 years: Traceability, Evolution, and Prospects," News and Communication Research (10), pp. 13-28. (in Chinese)

Anbin Shi, Dieer Liao. 2014. "The Development Path and Prospect of Data Journalism," News and Writing (2), pp. 17-20. (in Chinese)

Jayeon L., Weiai X. 2018. "The More Attacks, the More Retweets: Trump's and Clinton's Agenda Setting on Twitter," Public Relations Review (44), pp. 201-213.

Colleoni E., Rozza A., Arvidsson A. 2014. "Echo Chamber or Public Sphere? Predicting Political Orientation and Measuring Political Homophily in Twitter using Big Data," Journal of Communication (64:2), pp. 317-332.

Yiqing Hu, Jingyan Zhang. 2018. "40 years of Chinese Communication Studies: Reflections on the Process of Disciplinaryization," The International Press (40:1), pp. 72-89. (in Chinese)

McCombs, M. 2002. The Agenda-Setting Role of the Mass Media in the Shaping of Public Opinion. In Mass Media Economics 2002 Conference: London School of Economics.

McCombs M E, Shaw D L. 1972. "The Agenda-Setting Function of Mass Media," Public opinion quarterly (36:2), pp. 176-187.

McCombs M., Llamas J. P., Lopez-Escobar E., Rey F. 1997. "Candidate Images in Spanish Elections: Second-level Agenda-setting Effects," Journalism & Mass Communication Quarterly (74:4), pp. 703-717.

Guo L., Vu H. T., McCombs M. 2012. "An Expanded Perspective on Agenda-setting Effects: Exploring the Third Level of Agenda Setting," Revista de Comunicación (11), pp. 51-68.

McCombs M., Stroud N. J. 2014. "Psychology of Agenda-setting Effects: Mapping the Paths of Information Processing," Review of Communication Research (2), pp. 68-93.

McCombs M. E., Shaw D. L., Weaver D. H. 2014. "New Directions in Agenda-setting Theory and Research," Mass communication and society (17:6), pp. 781-802.

McCombs M. 2005. "A Look at Agenda-setting: Past, Present and Future," Journalism studies (6:4), pp. 543-557.

Vu H. T., Guo L., McCombs M. E. 2014. "Exploring "the World Outside and the Pictures in our Heads" A Network Agenda-setting Study," Journalism & Mass Communication Quarterly (91:4), pp. 669-686.

Chang YC, Zhang YX, Wang H, Wan HY, Xiao CJ. 2018. "Features Oriented Survey of State-of-the-art Keyphrase Extraction Algorithms," Journal of Software (29:7), pp. 2046-2070.