Patterns of news dissemination through online news media: A case study in China

5 AUTHORS, INCLUDING:

Youzhong Wang

8 PUBLICATIONS   16 CITATIONS

SEE PROFILE

Daniel Dajun Zeng

Chinese Academy of Sciences

388 PUBLICATIONS   4,405 CITATIONS

SEE PROFILE

Zheng Xiaolong

Chinese Academy of Sciences

80 PUBLICATIONS   160 CITATIONS

SEE PROFILE

Fengyu Wang

185 PUBLICATIONS   1,420 CITATIONS

SEE PROFILE

Patterns of news dissemination through online news media: A case study in China

Youzhong Wang · Daniel Zeng · Bin Zhu · Xiaolong Zheng · Feiyue Wang

Abstract Along with the rapid development of online news production and consumption, information spreads rapidly among news portals through reprinting and re-posting. Understanding the dissemination process of online news has important implications for policy making, crisis management, and brand imaging. As the frequency of news reprinting between two news portals expresses the social relationship between them, the topological properties of complex networks of news portals indicate the dynamic patterns of online news dissemination. This research is the first study that examines patterns and organizational structure of online news networks based on historical data collected from more than one thousand Chinese news portals. Main findings reveal that information can spread widely and rapidly through online news portals. Specifically, some major news portals exchange information with each other frequently and they exchange information with small news portals directly, implying that hot-spot events, even they are first reported by a small news portal, can be re-posted by many major news portals in a short time and then spread over the Internet quickly. The dense connections between many major news portals ensure that the spread will not be influenced greatly by the refusal to report typical events of a small part of major news portals.

Keywords Online news · Information diffusion · Social network analysis · Dynamic pattern

1 Introduction

Serving as the major source of news for internet users (Journalism.org 2010), online news portals usually do not generate most of their online content by themselves. Instead, they reprint news from other sources (Wan and Yang 2007; Wang et al. 2009b) for various reasons. Through the online information reprinting, a large number of online news portals constitute a complex online news network where information is generated, transmitted and consumed (Weber and Monge 2011). Apparently, understanding the topological features of the online news reprinting network has important implications for policy making, crisis management, and brand imaging. For example, because it is always easier to disseminate information in a well-connected network than in a segmented network, spreading information in these two types of network certainly requires different communication strategies. In addition, whether network nodes form random links among them or they are connected through hubs also has great impact on methods used for information acquisition from and dissemination through the network. Finally, the communication strategy to a network varies with how the hubs form links to the rest of the network. Specific hubs need to be identified if they serve as the gate keepers for a certain part of the network of interest. While such identification is less important if hubs connect to each other densely

Y. Wang · D. Zeng (✉) · X. Zheng · F. Wang
State Key Laboratory of Management and Control for Complex
Systems, Institute of Automation, Chinese Academy of Sciences,
Beijing 100190, China
e-mail: zeng@email.arizona.edu

D. Zeng
Department of Management Information Systems, The University
of Arizona,
Tucson, AZ, USA

B. Zhu
College of Business, Oregon State University,
Corvallis, OR, USA

to form a core. And the rest of the network connects through the core. In such structure, breaking news would be spread widely through the hubs no matter which news portal reports it first, and then we can track information from any of the hubs.

However, while policy-makers and organizations demand empirical evidences to better understand the online news reprinting network, such research currently remains at a qualitative stage. Most of previous research only studies the structure patterns of some major news portals and overlooks the existence of large numbers of small news portals (Flaounas et al. 2010; Weber and Monge 2011), leading to insufficient understanding of the online news reprinting network. The main challenge is that the construction of the reprinting network requests the identification of reprinting links among thousands of news portals by reading the articles published on all of them, and it is almost impossible to do this manually.

To fill this gap, we developed a news gathering system that automatically collects news articles from more than one thousand Chinese online news portals each day. Given that hyperlinks among these news portals do not accurately reveal the information flow among them (Thelwall 2004) and a news portal usually acknowledges the source of news (Weber and Monge 2011), we developed a method to automatically identify the links between news portals through the source acknowledgements of news articles. We then construct the reprinting network, named News Reprinting/ Re-posting Network (NRN), based on the links identified. The major difference between this paper and our previous work described in (Wang et al. 2009a) is that this paper provides more in-depth analysis, using the measures such as connectivity, reachability, different centrality measures, and core-periphery coefficient that were not used before. The results presented in this paper have more significant implications for effective communication strategies to online news propagation than our previous work.

The contribution of this paper is twofold. First, the paper demonstrates how information analysis methods could be applied to collect and to process large amount of online news articles for the identification and the construction of online news reprinting network. In addition, the paper presents for the first time the topological features of such reprinting network in China, providing the first-hand empirical evidence to decision-makers interested in event identification, crisis management, branding, and interaction with society.

The rest of the paper is organized as follows. In section 2, we review related literature, followed by the description on data preparation and data preprocessing in section 3. Section 4 investigates how information is spread in NRN, while section 5 examines different roles played by different news portals. Section 6 describes how different types of

news portals interact with each other. The interpretations and the implications of results are presented in section 7, followed by the conclusion of this paper in section 8.

2 Literature review

Internet has particularly significant influences on news production and consumption. Compared with conventional newspapers, online news portals gain several advantages such as interactivity, hypertextuality, and multimediality (Paulussen 2006; Deuze 2003). The Internet not only reduces the cost of releasing and spreading digital information (Varian et al. 2004; Peng et al. 1999), but also improves users' satisfaction (Kaye and Johnson 2003; Hindman 2007). Various types of news portals thus emerged during recent two decades (Deuze 2003; Weber and Monge 2011), broadening media channels for citizens to access online news (Hindman 2007). Especially, almost all of print newspapers have built their online version, making it more convenient for people to obtain fresh information from Internet (Hindman 2009; Weldon 2008). The easy access to a large number of news portals dramatically changes the reading habits of news consumers. Online news portals have become the primary news sources for more and more users. (Lee and Carpini 2010; Journalism.org 2010; Kohut et al. 2008).

Although online news portals publish large amount of news (Herbert and Thurman 2007; Bustamante 2004; Singer 2003), they usually do not generate the content themselves. Instead, they reprint or re-post news articles developed by other news portals (Boczkowski 2002; Quandt 2008). The reprinting reduces the cost of news generation (Chyi and Sylvie 1998), providing a practical way for most news portals to publish plentiful news contents to grab users' attention (Hoffman 2006). In fact, many news portals would reprint the original news from a source in 2 days after a popular event has been occurred (Wang et al. 2009b).

Through reprinting, news portals form a connected network. The implications of understanding the topological features of news reprinting network are two folds. First, the identification of network nodes playing crucial roles, and the understanding of how those nodes form links to each other and to the rest of the network certainly has significant impact on communication strategies with the network. Secondly, understanding topological features of news reprinting network helps us predict how information spreads out over the Internet (Weber and Monge 2011). These two implications may have further impact on outbreak detection, event tracking, and crisis management (Leskovec et al. 2007; Dasgupta et al. 2009).

Therefore, we believe following topological features of an online news-reprinting network could be particularly

interesting for decision-makers seeking effective strategy to acquire and to disseminate news from/to such network.

1) *Can information spread widely and quickly in the network?*

An ideal network for spreading information widely and quickly is the one that has a strongly connected component where most pairs of nodes can reach each other through a short path (Latora and Marchiori 2001). However, in the news reprinting network, some nodes only reprint news from others and never provide original news for others to use, and some nodes only provide original news without citing other news portals. These nodes only have in-links or out-links, meaning that it is impossible for entire NRN to be a strongly connected component. Therefore, we study this question from two aspects: 1) if NRN has a strongly connected component where nodes are linked through short paths, and 2) if the nodes with only out-links are linked to the nodes with only in-links through short paths.

2) *Do all network nodes play similar roles in the network? Or if there exists network nodes that are more important than the other nodes in the same network.*

According to previous social network analysis studies, information can spread efficiently in the "scale-free" network as most nodes can reach each other through a few popular network nodes (i.e., hubs) (Cohen and Havlin 2003). Like the airport network (Barrat et al. 2004; J. Zhang et al. 2010), connecting through hubs minimizes the average travel distance between any pair of nodes in the network. Apparently, it is strategically important to know if the NRN is a scale-free network. In addition, identifying hub nodes that generate a majority of original contents or gather lots of news from many sources certainly provides strategic values to decision-makers who want to form a communication strategy to the news network.

3) *How do different types of news portals interact to each other?*

While the "scale-free" network reduces the distance between nodes, it also possesses vulnerability to the network because the removal of a hub could possibly disconnect the entire network (Crucitti et al. 2003). Such vulnerability could be addressed only when the hubs densely connect to each other to form a core. The structure with hubs forming a core in the network and the rest of the network connecting through the core is called "core-periphery" structure (Holme 2005). The interaction to a "scale-free" network with or without a core-periphery structure obviously requests different communication strategies. Therefore it is of decision-makers' best interest to know if the NRN has a core periphery structure.

In summary, a network that reveals the real re-printing/reposting relationship among online news portals needs to be constructed before the features described above can be identified. However, given a large number of online news portals, it is usually very difficult to manually collect and process information to identify the reprinting/re-posting relationship among them. As hyperlinks among these news portals do not accurately reveal the information flow among them (Thelwall 2004) and a news portal usually acknowledges the source of news, it is possible to obtain the accurate paths of information re-printing/re-posting relationships by extracting the acknowledgements from online news articles. Making use of the acknowledgements, Weber and Monge (2011) examines the information flow between 239 news portals and proposes a Source-Authority-Hub model to show the dynamic patterns of information flow (Weber and Monge 2011). The model describes that source news portals supply the main part of contents; authorities filter information from sources and then push it to hubs which collect relevant reports for a given topic. However, the study lacks the detailed investigation of the mutual relationship and organizational structure of tightly connected online news reprinting networks.

3 Data

3.1 Data preparation

This paper studies the patterns of online news reprinting between Chinese online news portals. The Chinese online news portals have been developing rapidly as more and more people have gained access to the Internet in China. As of the end of June 2011, the scale of China's Internet users reached 513 million people (CNNIC 2012). China's Internet development has contributed to the widely accessing of online news, the fourth most popular Internet service in China, occupying 71.5 % of the China's netizen (CNNIC 2012).

Online news documents in our dataset have been collected from Chinese news portals on a daily basis since March 2008. The entire process of data collection and preparation is displayed in Fig. 1, which is described in detail as below.

1) Selecting news portal seeds. We manually collect a small number of news portals as a root set of *news portals list*. These seed news portals belong to one of the two types: professional news portals such as "people.com.cn"; and popular portals such as "sina.com.cn". Professional news portals are the portals where governments release official news/announcements, providing original news contents to the Internet. And the popular portals usually have a large audience group and are commonly administrated by private internet companies. As these seed news portals provide main sources of online news
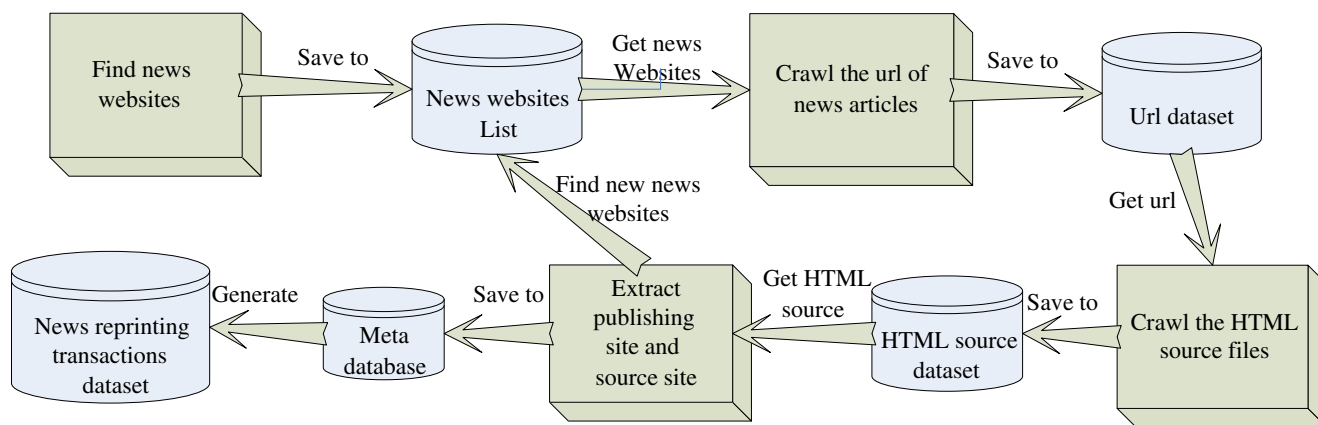
Fig. 1 Process to crawl online news articles and extract news reprinting transactions

contents and gather information from various other news portals, they provide a good entry to find other news portals through hyperlinks and source acknowledgements.

2) Identifying the patterns for source acknowledgement. Using news articles published on the seed news portals, we analyze the way in which source acknowledgement is made in news articles. This includes the position in the article and keywords using for the acknowledgement of source. For example, source acknowledgement may be declared following news title; or at the end of the article; or following keywords such as "source (来源)", "reproduced from (转载于)".

3) Collecting news articles. Software is developed to automatically download and store HTML files of news articles from collected news portals.

4) Processing news articles collected. The raw HTML files are parsed to extract meta information including news title, body, publishing news portals, publishing date/time and source news portals if available. The source portals of news articles are automatically extracted based on the patterns identified in step 2.

5) Updating news portals list. The source portals identified in step 4 are added to the news portals list if they are not on the list. We also manually check a source portal before we add it on each day.

6) The system automatically gathered news on each day from March 2008 to May 2010 by repeating the steps 3–5. Meta information of all news articles collected during this period is stored into a database.

3.2 Data preprocessing

Most news articles collected are put into one of the two categories: original reports (if the source site is the publishing site) and reprinted documents (if the source site is not the publishing site). There are also articles that do not belong to any of the two categories due to the difficulty of identifying the source information from these articles. There are three reasons for this. First, some news portals only publish news context without specifying the source site. Second, various news portals place the source information on different positions in the web pages following their specific rules, and it is nearly impossible to match all rules of thousands of news portals in our program. Third, some news documents are reprinted from the websites which don't belong to news portals, e.g. the official websites of organizations.

To improve the validity of the empirical findings this paper presents, we use two data sets for our data analysis. And we have consistent findings across these two data sets. Each data set contains one-month continuous data. The first data set, called data set A in the rest of the paper, covers the time period from September 10, 2008 to October 10, 2008. This data set collects all types of news articles, such as economic news, city news, international news, sports news, entertainment news, from 1,037 news portals in total. The second data set, called data set B in the paper, has more recent data, from March 1, 2010 to March 31, 2010. The data set B only collects economic news from 487 news portals that have released related contents before. To our knowledge, our data sets cover all primary Chinese news portals that have released large amount of news articles. Other news portals missed in our data sets post very little content, which would have little impact on the structure of the news reprinting network. In addition, despite the scope of articles and the number portals between the two data sets, the networks constructed from these two data sets show similar structure features, indicating the validity of topological features of Chinese news network found in this study.

Table 1 introduces the basic information of the two data sets. In total, data set A collects 3.3 million news articles, among which 1.3 million articles have identifiable source acknowledgements; and data set B contains 1.6 million

Table 1 The basic information of data set A and data set B

	Data set A	Data set B
Time period	2008.9.10~2008.10.10	2010.3.1~2010.3.31
# of news portals	1037	487
# of news articles	3.3 million	1.6 million
# of identifiable sources	1.3 million	0.9 million
Average number of source portals	27.66	12.96
Average number of printed news articles	980.91	849.07

news articles and 0.9 million of them contain successfully identified source acknowledgements. To respect the intellectual properties of other portals, a news portal usually acknowledges the original source of an article it posts rather than the news portal from which it copies the news article from. This indicates that news portals that are frequently acknowledged by other news portals tend to generate many original contents and thus act as main sources of other news portals. Given that reputable news portals in China usually cite the sources of articles they post, we believe that our analysis provides solid evidence to indentify major players in the network although not all of news reprinting transactions is collected. We believe that the identification of these main sources has great effects on communication strategy design for online news dissemination.

Based on the news reprinting transactions between news portals, we then construct a News Reprinting/Re-posting Network (NRN) for each data set. In the NRN, a node represents a news portal and a directed edge from node i to j indicates news portal j has reprinted news from news portal i. The weight of each edge represents the number of news documents reprinted from one news portal to another. In average, a news portal in the NRN for data set A has reprinted 980.91 pieces of news from 27.66 other news portals, while a portal in NRN B has printed 849.07 pieces of news from 12.96 other news portals at average.

One finding from the two data sets is that not every news portal has both in-links and out-links. In data set A, only 595 (57.38 %) news portals have both edges pointed in and edges pointed out, whereas 327 (31.53 %) of other news portals have only edges pointed in and 115 (11.09 %) have only edges pointed out. For data set B, 233 (47.84 %) news portals have both edges pointed in and edges pointed out, whereas 181 (37.17 %) of other news portals have only edges pointed in and 73 (14.99 %) have only edges pointed out. There are two explanations of this phenomenon. On one hand, some news portals are official announcement agents for government. They usually serve only as news sources for other news portals and do not reprint news from others. As a

result, they do not have links into them and only have links out of them. On the other hand, in China, organizations need to have the interview permissions issued by the government in order to be able to generate original news. Consequently, news portals without such permission could only reprint content from other sources and end up having no link out from them.

Although we only use one-month data in each dataset, the constructed NRN still can represent the overall patterns of online news dissemination mainly because the patterns of contents generation and transmission are usually stable over times. Our previous work has also found that most topological features of NRN remain the same over time (Wang et al. 2009a). In fact, if one news portal reprinted news from another news portal before, most likely it will do the same i in the future. This is further evidenced by the consistency between the findings from the two different data sets we use.

4 Connectivity and reachability in the NRN

Connectivity and reachability both study the social relationship between two nodes in a social network: *connectivity* examines if there is a path that connects two random nodes; and *reachability* counts the least number of edges in the path that connects these two nodes. In another word, connectivity measures whether a network is well-connected, while reachability measures how easy it is for one node to reach another node in the same network. A social network has a good connectivity if there is a path between any pair of its nodes. And the length of each path indicates the reachability. Fewer numbers of nodes the path contains, higher reachability the network has. A network is considered to be "small world" if the average length of any path that links two nodes scales logarithmically, or more slowly, with the number of nodes (Albert and Barabási 2002; Barrat et al. 2008). Connectivity and reachability thus demonstrates how wide and rapid information could diffuse in a network, respectively (Latora and Marchiori 2001; Moore and Newman 2000). Therefore it is important to know whether most news portals connect together through short paths in the NRN.

As shown in Section 3, a large fraction of news portals only have edges pointed in or edges pointed out. Thus these news portals cannot reach each other in two directions. In order to understand the range and the speed of information dissemination in the network, we first study if there is a short path between an out-link-only node and an in-link-only node. We then investigate if most nodes that have both in-links and out-links reach each other through short paths in a strongly connected component.

In the directed NRN, a *directed path* from node i to node j indicates that every node has reprinted news from the previous node in the path. For all pairs of nodes that one

only has out-links and the other only has only in-links in the directed NRN, 97.63 % of them in data set A and 85.27 % of them in data set B have a directed path from nodes with only out-links to those with only in-links. And among these reachable pairs, 79.87 % in data set A and 77.7 % in data set B are linked through at most 3 intermediate edges. The average length of the shortest directed paths that connect these pairs of nodes is 2.88 in data set A and 2.89 in data set B, while the maximal length of that is 8 in data set A and 7 in data set B. The results show that information can transmit from source news portals to contents gatherers through a few reprinting transactions in most times.

We the examined the presence of strongly connected component of the subgraph that composed by nodes with both in-links and out-links. We measured the reachability with the average and maximal distance between any two nodes in the giant component. The strongly connected component is a connected component that any two nodes in it can reach each other through a directed path. Among 595 news portals in data set A and 233 news portals in data set B that have both in and out links, 95.97 % of them in data set A and 89.27 % of them in data set B are in the strongly connected component. In the strongly connected component of data set A, 90.26 % of all pairs of nodes are connected by directed paths that have at most 3 intermediate directed edges. The average and maximal distance of all pairs of nodes is 2.54 and 7, respectively. Similarly, in the strongly connected component of data set B, 79.46 % of all pairs of nodes are connected by directed paths that have at most 3 intermediate directed edges. The average and maximal distance of all pairs of nodes is 2.74 and 8, respectively. The similar results obtained from data set A and data set B implies that no matter which news portal first reported a fresh news event, the information of this event would spread

to other news portals quickly. Figure 2(a) and (b) show that most of pairs of nodes reach each other through 2 or 3 edges.

It is possible that the path between two portals is longer than what we have found because some portals actually don't copy news from sources but from other intermediate portals. But adding more nodes to a path probably will not reduce the speed of dissemination in NRN too much. In fact, we have found that 80 % of the portals posted a piece of news within 24 h since the original news was published (Wang et al. 2009b).

5 Importance of news portals in the NRN

We identify the roles of a news portal in NRN by the type and number of connections it has to other portals in the same network. For a news portal i, we define the in-degree $k_{in}(i)$ and out-degree $k_{out}(i)$ as the number of news portals i has reprinted news from or to, respectively; and the in-strength $s_{in}(i)$ and out-strength $s_{out}(i)$ as the number of news articles i has reprinted from or by other news portals, respectively. The formal definitions of these four degree centrality metrics are as below,

$$k_{out}(i) = \sum_{j \neq i \in V} a_{ij}, k_{in}(i) = \sum_{j \neq i \in V} a_{ji} \qquad (1)$$

$$s_{out}(i) = \sum_{j \neq i \in V} w_{ij}, s_{in}(i) = \sum_{j \neq i \in V} w_{ji} \qquad (2)$$

where V denotes the set of all news portals in the NRN; a_{ij} is 1 if news portal j has reprinted news from news portal i, and it equals to 0 if otherwise; and w_{ij} denotes the number of news articles that j has reprinted from i in total.
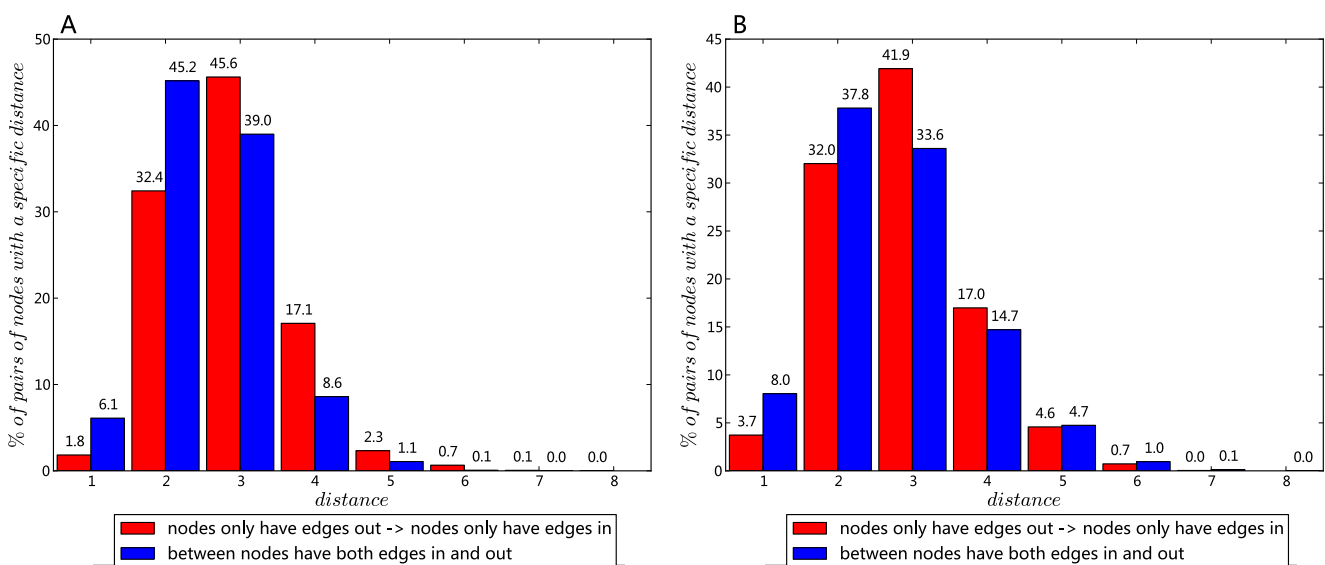

Fig. 2 The percentage of pairs of nodes that are connected by directed paths with different distances

These concepts clearly indicate the roles that news portals play in news dissemination from different aspects. Specifically, high k_{in} or s_{in} means a node collects news widely from the network, whereas high k_{out} or s_{out} value means a node is the source that many other nodes reprint news from. Obviously nodes with high values in those variables have high power and have more strategic values to decision makers. For example, the most effective way to find current popular topic is to monitor the nodes with high values in k_{in} or s_{in}, while the best way to prevent undesired spread of certain news is to control the nodes with high values in k_{out} or s_{out}.

It is interesting to study how such degree centrality values distribute in the network. Especially, if the distributions of these four centrality metrics obey power-law functions, this indicates that the network has a scale-free network structure. The "scale-free" structure means that a few hub nodes in a network possess extremely large degree centrality values, while the degree centrality values of the others are very small (Barabasi and Albert 1999). It has been shown that various types of social networks possess the "scale-free" structure (Zheng et al. 2008; Zhang et al. 2010; Panzarasa et al. 2009). Comparing to a random network, the existing of hub nodes in a scale-free network accelerates the speed of information spreading (Barthélemy et al. 2005). In addition, attacking on a randomly selected node in a scale-free network would not impact the information spreading greatly, but the structure of the scale-free network would be vulnerable if the attack focuses on hub nodes (Ni et al. 2011; Crucitti et al. 2003). Obviously decision-makers should design different communication strategies for random and for scale-free networks.

However, the limited scale (such as number of nodes) of many real social systems causes that the centrality distribution often follows power law distribution with exponential cut-off ($p(x) \propto e^{-\beta x} x^{-\alpha}$) rather than pure power-law distribution ($p(x) \propto x^{-\alpha}$) (Newman 2001). As the actual number of news portals is limited, it would be normal that centrality distribution in the NRN cannot be fitted by pure power-law functions. As the result, we only consider whether the centrality distribution follows the power-law distribution with exponential cut-off.

Figure 3(a-d) and (e-h) report the complementary cumulative distributions of these centrality metrics for data set A and data set B, respectively. We fit these distributions with power-law function, exponential function ($p(x) \propto e^{-\beta x}$), and power-law function with exponential cut-off, respectively. These fitting functions have great statistical significances as the p-values are all below 0.001. The scaling parameters γ and β and the coefficient of determination r^2 of these fitting functions are reported in Table 2 (A) for data set A and Table 2 (B) for data set B, respectively. According to high values of coefficient of determination, all distributions can be fitted very well with power-law functions with exponential cut-off. This result

indicates that a small number of news portals have very high centrality values, while most other news portals possess very small centrality values. Besides, according to the value of exponential scaling parameter β in the fitting functions of power-law distributions with exponential cut-off, distributions of s_{in}, s_{out} have much smaller effects of exponential cut-off than distributions of k_{in} and k_{out}. The range of value of k_{in} and k_{out} is limited as the amount of news portals we crawl is finite, making their distributions have big exponential cut-offs; but the value of s_{in}, s_{out} keeps increasing as time goes and thus their distributions are more likely to obey power-law functions with small exponential cut-offs.

As described before, nodes with high s_{in} values collect large amount of information from the network and present it to audience, while those with high s_{out} serve as the information source for the network. For example, the top s_{out} portal in both data set A and in data set B is the www.xin-huan.com, a news portal backed by the most authoritative news agency in China, while the top s_{in} portals, www.chi-na.com.cn in data set A and www.163.com in data set B, are both the most popular online portals where people go seeking news.

As displayed in Table 3 (A) and (B), online news portals providing original news to the network are usually not the ones that collect information from the network. There is only one overlap in data set A and two overlaps in data set B between the top five s_{in} portals and top 5 s_{out} portals. In fact, only thirty-five nodes belong to both the top 100 s_{in} list and top 100 s_{out} list in data set A, and thirty-three overlaps between top100 s_{in} list and top 100 s_{out} list in data set B. The reason for this is that in China, only a limited number of online portals have the right to interview and release first-hand reports (Harwit and Clark 2001; Zhang 2007). Out-degrees of these news portals are very large because their reports would be reprinted by many news portals directly. On the other hand, major news portals with large audience group gather news from many news media and possess large value of in-degree; but they have small number of original news and thus would not be cited as news source.

The lack of overlapping between top s_{in} list and top s_{out} list in both data sets is very different from what have been found in many other social network studies (Panzarasa et al. 2009; Newman et al. 2002), where one person who likes to send message to others tends to attract attention and receives many responds from others as well. This suggests that the NRN in china is very different from a typical social network people are familiar with.

6 Interaction patterns of different types of news portals

In some real-world social networks, it has been found that a small number of members form a densely connected core

Fig. 3 The complementary cumulative distribution of (**a**) k_{in}; (**b**) k_{out}; (**c**) s_{in}; and (**d**) s_{out} for data set A, and the complementary cumulative distribution of (**e**) k_{in}; (**f**) k_{out}; (**g**) s_{in}; and (**h**) s_{out} for data set B. The dash lines fit the distributions with power-law function with exponential cut-off
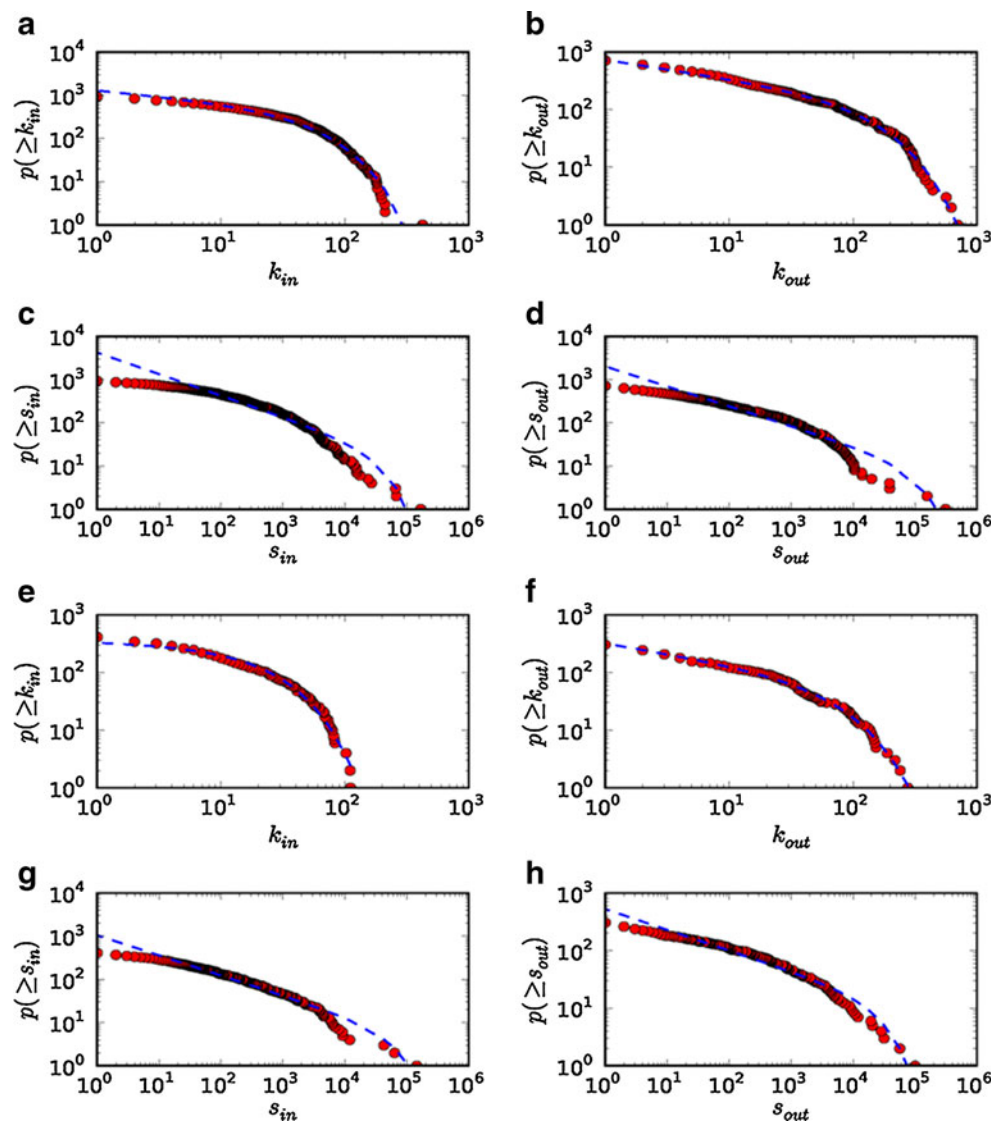

Table 2 Test of heavy tailed property of different centrality metrics

Centrality metric	Power-law		Exponential		Power-law with cut-off		
	γ	r^2	β	r^2	γ	β	r^2
A Data set A:							
k_{in}	−1.18	0.76	−1/45.31	0.95	−0.27	−1/54.69	0.96
k_{out}	−0.88	0.83	−1/111.13	0.96	−0.32	−1/154.12	0.99
s_{in}	−0.55	0.87	−1/15349	0.32	−0.50	−1/39968	0.91
s_{out}	−0.49	0.89	−1/40273	0.23	−0.46	−1/120351	0.91
B Data set B:							
k_{in}	−1.01	0.78	−1/22.72	0.98	−0.06	−1/23.73	0.98
k_{out}	−0.93	0.86	−1/52.04	0.95	−0.36	−1/75.23	0.99
s_{in}	−0.50	0.93	−1/20521	0.29	−0.47	−1/72569	0.95
s_{out}	−0.43	0.90	−1/14302	0.50	−0.36	−1/34831	0.96

For each centrality metric, we fit the complementary cumulative distribution with power-law distribution $\left(p(\geq x) \propto x^{-y}\infty\right)$, exponential distribution $\left(p(\geq x) \propto e^{-\beta x}\right)$ and power-law distribution with exponential cut-off $\left(p(\geq x) \propto x^{-y}e^{-\beta x}\right)$, respectively. The scaling parameters γ and β and the coefficient of determination r^2 is reported for each fit. The statistical p-values for all fits are less than 0.001, so we don't display the p-value in the table

Table 3 Two lists of top 5 news portals with the largest s_{in} and s_{out}, respectively

Order	5 news portals with the largest s_{in}		5 news portals with the largest s_{out}	
	Chinese name	Website	Chinese name	Website
A Data set A:				
1	中国网	www.china.com.cn	新华网	www.xinhuanet.com
2	**腾讯网**	**www.qq.com**	中国新闻网	www.chinanews.com
3	网易	www.163.com	新浪	www.sina.com.cn
4	中国经济网	www.ce.cn	人民网	www.people.com.cn
5	搜狐	www.sohu.com	**腾讯网**	**www.qq.com**
B Data set B:				
1	**网易**	**www.163.com**	新华网	www.xinhuanet.com
2	华奥星空网	www.sports.cn	中国新闻网	www.chinanews.com
3	**搜狐**	**www.sohu.com**	**搜狐**	**www.sohu.com**
4	新浪	www.sina.com.cn	人民网	www.people.com.cn
5	每日甘肃网	www.gansudaily.com.cn	**网易**	**www.163.com**

News portals in bold format are in both two lists

and other nodes reach each other through the nodes in the core. This structure is called "*core-periphery*" structure. Borgatti and Everett (2000) defines a series of models to express the "core-periphery" structure in social networks (Borgatti and Everett 2000). These models assume that in ideal situations, individuals in the core are fully connected, while members in the periphery don't connect to each other at all and they only possess links to the core. The "core-periphery" structure reveals that a few key members dominate information diffusion in the network. Meanwhile, the densely connected core reduces the vulnerability of a network. Information can still spread out if several nodes in the core are removed.

However, it could be a great challenge to decide if a real-world social network has a "core-periphery" structure (Borgatti and Everett 2000). The method to find if a social network has "core-periphery" structure consists of two steps, with the first step being core identification and the second step being the confirming that the interaction between the core and the rest network is consistent with the definition of "core-periphery" structure.

We seek to give a quantitative answer to whether the NRN is organized as a core-periphery structure. As the first step, we try to identify if there exists a core in the NRN. The identification of the core could be computational expensive. Holme (2005) presents a simple heuristic method to extract the core with the set of k-core (Holme 2005). A k-core is a subgraph in the network where the degree centrality of each node is higher or equal to k (Dorogovtsev et al. 2006). In another word, k-score is a subgraph that contains nodes linking to at least k other nodes in the subgraph. As nodes would be connected more densely in k-core when k is bigger, we identify the biggest k value such that k-core exists but $(k+1)$-core does not exist in the NRN network.

Then, we need judge whether the k-core is indeed a core of the network and whether the network has core-periphery

structure. According to the definition provided in (Borgatti and Everett 2000), a network with core-peripheral structure not only has a core that consists of a group of densely inter-connected nodes, but also nodes in the periphery with few connections among themselves And those periphery nodes connect with each other mainly through the core. Therefore, a social network has a "core periphery" structure if there is a large difference in network density between the core and the periphery. Density here denotes to the ratio of the number of links between a certain number of nodes to the maximum possible links that could have for the same number of nodes (Scott 2007). For comparison, we create a random network that has the same degree sequence as NRN (i.e., nodes in the random network have the same degree centrality values as nodes in the NRN) as bench mark. It is apparent that the density difference between the core and the peripheral in NRN should be larger than that in the random network. Using the random networks generated by the model provided in (Newman 2003), we calculate the core-periphery efficient (c_{cp}) as follows,

$$c_{cp} = \left(\frac{\rho_{core}(G)}{\rho_{periphery}(G)} \right) \Big/ \left\langle \frac{\rho_{core}(G_{rand})}{\rho_{periphery}(G_{rand})} \right\rangle, \quad (3)$$

where $\rho_{core}(G)$ and $\rho_{periphery}(G)$ denotes the density of a sub-graph of the core and periphery, G_{rand} is a random network which possesses the same degree sequence as G. If the c_{cp} is bigger than 1, the network G has a core-periphery structure.

Our measure for whether a network has core-periphery structure is different with the measure used in (Holme 2005), which detects the reachability between nodes in the core and nodes in the periphery. Because in most real world social networks the distance between any two nodes is short (Watts and Strogatz 1998; Barrat et al. 2008), it is practically

unimportant to identify a core that has a group of network nodes with the shortest distance to each other. In addition, we believe that our definition of core-periphery efficient can better represent the ideal model presented in (Borgatti and Everett 2000), which requires that nodes in the core are densely connected and the periphery has low connectivity among nodes.

We applied core-peripheral analysis to understand how news portals form links between each other. We first construct an undirected NRN by adding an undirected edge between two news portals i and j if i has reprinted news from j or j has reprinted news from i. The weight of edges (i-j) is counted by the total number of news articles that they have reprinted between them. For a news portal i, its degree $k(i)$ indicates the number of news portals i has reprinting relationship with, and its strength $s(i)$ means the number of news articles i reprints from or by other news portals. Notably, the value of $k(i)$ doesn't equal to the sum of $k_{in}(i)$ and $k_{out}(i)$.

The core of the undirected NRN is represented by a maximum k-core and the periphery of the NRN is formed by other news portals that are not in the k-core. The biggest k which satisfies the definition of k-core is 68 in data set A and the biggest k value is 35 in data set B. This means that data set A contains an inter-connected core with all nodes in the core having degree values equal to or larger than 68, and the core in data set B has nodes with degree value equals to or larger than 35. There are 98 nodes and 4,062 edges in the 68-core in data set A and 83 nodes with 2,102 edges in the 33-core of data set B. In the periphery part of the network, the data set A has 939 nodes linked by 6,456 edges, while the data set B has 404 nodes with 786 links among them.

The results from both data sets suggest that the core of NRN provides most of news contents to public and generates most of the original news to the network. In data set A, the core contains less than 10 % of the nodes but is responsible for 51.74 % of the contents delivered to the public and 55.13 % of the original contents cited by other news portals. Similarly in data set B, 75.32 % of the news articles brought to readers and 63.78 % of the original news contents are generated by the core that contains only 17 % of the nodes in the network.

The cores in both data sets have much higher density than their corresponding network and peripheries. In data set A, the core has the density of 0.8546, much higher than that in the entire NRN (0.0534) and that in the periphery (0.0147). In data set B, the core also has much higher density (0.6174) than both the entire NRN (0.0568) and the periphery (0.0097). The value for the proposed core-periphery measurement (formula (3)) is 1.49 for data set A and 1.81 for data set B, indicating that the NRN constructed from each dataset has a core-periphery structure.

Figure 4(a) and (b) present a graphical representation of the core-periphery structures in NRNs for data set A and data set B, respectively, further validating the existence of the core-periphery structure in NRN. As suggested in Fig. 4, the amount of information exchanged among core nodes and that exchanged between the core and the periphery are much higher than that exchanged among periphery nodes.

7 Discussion and implication

The network of online news reprinting reveals how information is spread among news portals in China. Our study indicates that information could traverse the network quickly and reach the majority of the news portals in the NRN. The study also indicates that there are two types of crucial roles a news portal could play in the network, providing original news to the network and collecting information from the network to attract audience. Very few important news portals play both types of roles. More interestingly, we find those popular news portals form a core among them, exchanging large amount of information frequently. In addition, the rest of the network largely depends on the core to connect with each other. This means, if a news portal is not in the core, regardless its roles, it is more likely to provide news to or receive news from the core than to interact with other periphery portals.

Our results have significant implications to policy makers, brand managers, or any people seeking to communicate to the news network effectively. First of all, it is apparent that not every news portal should be treated equally. The core-periphery organization of NRN indicates that a small number of news portals constitute a tight core, which dominates the information production and transmission. These news portals provide more than a half of the contents to public and contribute to most of news reprinting transactions between online news media. Information is exchanged frequently in the core, linking nodes in the core together extremely tightly. The results suggest that we need to monitor only the news media in the core to follow most of latest news events. Due to the agenda-setting theory, the public may be influenced strongly by these major news portals. From a more practical point of view, the core of NRN is a double-edged sword for the society and people. On one hand, a shocking event released on these major media would grab the public attention immediately and trigger a big outbreak. On the other hand, if these media post rumors, the society would be affected very widely and deeply.

In addition, decision-makers certainly should form their communication strategies based not only on the popularity of a news portal but also on the nature of links the portal has. For example, our results suggest that the fastest way to inform public is to feed messages to news portals with high

Fig. 4 The number of news articles that reprinted from one group to another group for (**a**) data set A and (**b**) data set B; the numbers on the arrows represent the numbers of news articles transmitted from one group to another group
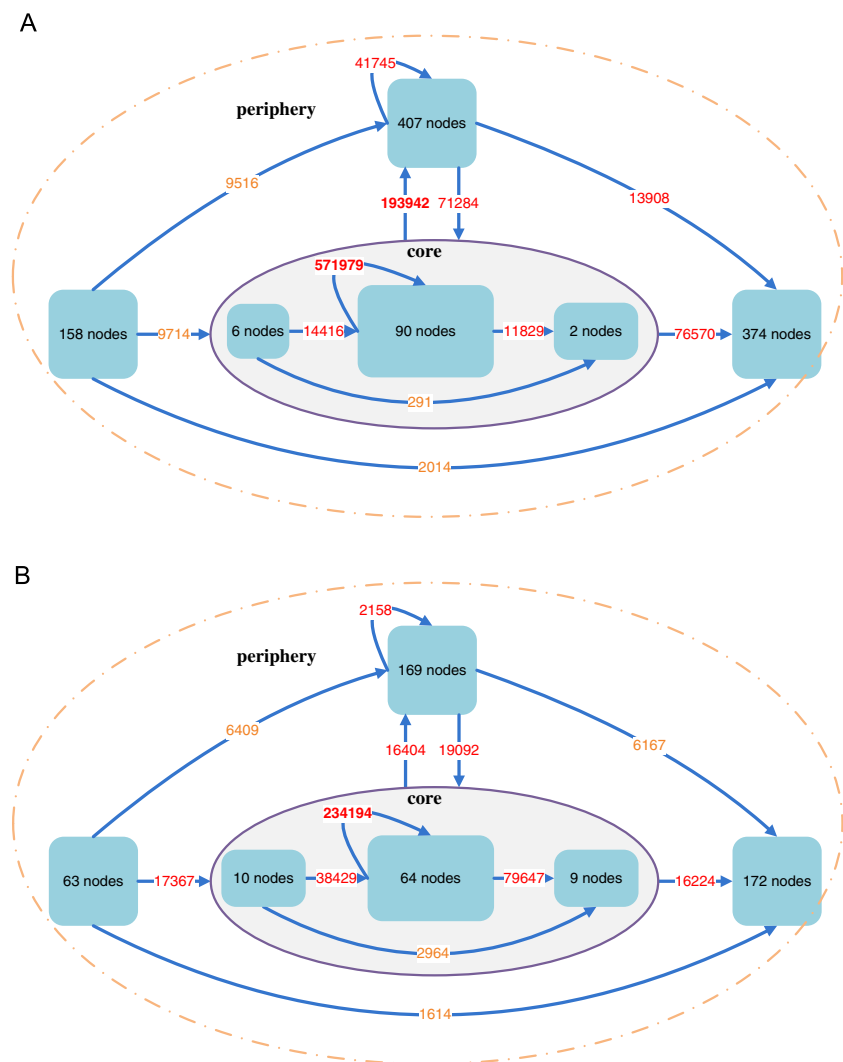
values of in-degree as they are where people go to read news. This is especially important for crisis management when important messages need to reach the public in a very fast fashion. On the other hand, while the portals with high out-degree values serve as the source of original news to the network, monitoring the frequency of news cited from those sources certainly help the understanding of the popular topics in the society, whereas the most efficient way to spread a message widely in the network is to feed the message to these sources. More importantly, the type of news cited by a portal could also help the identification of psychographics about the readers of that portal. This could largely help organizations reach their potential customers.

Finally, the core-periphery organization of NRN helps the information spread quickly in the Internet from two aspects. On one hand, news reports spread more conveniently from major news portals to small news portals. On the other hand, when a small news portal releases a breaking news event, the core can acquire the event instantly and then it transmits the event widely to other news media and to the public. As a result, this structure can explain why rumors are spread over the Internet extensively

in a short time. Besides, it would be hard to eliminate the transmission of rumors among online news portals as the core is composed by dozens of news portals. More specifically, when it comes to managing undesired dissemination of information, our results suggest that it would be too late to prevent the spreading of a message if it had already appeared on the news portals with only in-links, because those portals are usually the final step for the news dissemination in the network. New content control should start with portals with high out-degree values. They are the start point for dissemination. Furthermore, it could be more efficient to enforce the regulations about the content of news at high out-degree portals rather than at those with high in-degree values.

8 Conclusion

In this research, we examine the patterns of news dissemination through online news network. We develop software for the automatic news collection from online news portals

in China. The citation links are identified through automatic processing the acknowledgement statements in news articles that indicate the source of the news. We then construct the reprinting network existed between Chinese news portals. What we have found from network analysis is listed as follows:

1) Information can spread out quickly within the news reprinting network. A piece of message, regardless its entering point of the network, can reach most of the news portals in a short period of time.

2) Important news portals in the network are those that either serve as the source of original news or those that bring news to public. The former type of news portals have high values in out-degree because the rest of the network cites news from them, while the second type of news portals are high in in-degree values because they collect information from all possible sources within the network and they are the places public go for news. In addition, the majority of network nodes connect each other through a few intermediate nodes with high values either in in-degree or in out-degree. The online news reprinting network in China thus is a scale-free network

3) The news reprinting network not only is a scale-free network but also has a core-periphery structure. The important nodes form a densely inter-connected core, while the periphery part of the network connects through the core. The core generates most of the original news that the network has and it has the most popular portals where most people visit for news.

The implication of studies lies in that not every news portals should be treated equally when communicating to the network. A small number of news portals form a densely connected core, which play important roles both in information distribution and information acquisition. Specifically, the fastest way to reach public is to post messages on portals with high values in in-degree because those portals are the places where public go, while the effective way to spread out news to the entire network is to feed the messages to portals that serve as the major sources of original news because most other news portals copy news from them. In addition, monitoring the frequency that a certain piece of news is cited from a high out-degree source portal could be a great way to understand the popular topic in society. And the core-periphery structure of NRN indicates that it would be hard to prevent rumors spreading due to. We suggest that such prevention should quickly start with portals with high out-degree values before the information is posted by news portals with only in-links.

The limitation of this study lies into the fact that all the data comes from China and it remains unclear whether the empirical findings obtained from this study could apply to the news network in other countries. However, we believe

that our study demonstrates a systematic way to understand news network in other regions of the world. And our future studies will apply the method described in this paper to news network in other countries to identify how the demographic, cultural and political differences impact the topology of news reprinting network. We are also interested in expanding our study to how information spreads out in social media. Given that it are editors who decide what the readers read from news portals, social media, where readers themselves decide what other readers read, probably presents very different information demission patterns such as revealed in (Gonçalves et al. 2011; Fan and Yeung 2011). Finally, we will further explore content analysis to all news articles collected, investigating the factors that contribute to the popularity of a piece of news.

References

Albert, R., & Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics, 74*(1), 47–97.

Barabasi, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509.

Barrat, A., Barthelemy, M., Pastor-Satorras, R., & Vespignani, A. (2004). The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences, 101*(11), 3747–3752.

Barrat, A., Barthlemy, M., & Vespignani, A. (2008). *Dynamical processes on complex networks*. New York, NY: Cambridge University Press.

Barthélemy, M., Barrat, A., Pastor-Satorras, R., & Vespignani, A. (2005). Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *Journal of Theoretical Biology, 235*(2), 275–288.

Boczkowski, P. J. (2002). The development and use of online newspapers: What research tells us and what we might want to know. In L. Lievrouw and S. Livingstone (Eds.), *The handbook of new media* (pp. 270–286). London: Sage.

Borgatti, S. P., & Everett, M. G. (2000). Models of core/periphery structures. *Social Networks, 21*(4), 375–395.

Bustamante, E. (2004). Cultural industries in the digital age: some provisional conclusions. *Media, Culture & Society, 26*(6), 803.

Chyi, H., & Sylvie, G. (1998). Competing with whom? Where? And how? A structural analysis of the electronic newspaper market. *Journal of Media Economics, 11*(2), 1–18.

CNNIC. (2012). the 29th China Internet Development Status Survey Report. http://www.cnnic.cn/research/bgxz/tjbg/201201/P020120116330880247967.pdf. Accessed 1.18 2012.

Cohen, R., & Havlin, S. (2003). Scale-free networks are ultrasmall. *Physical Review Letters, 90*(5), 58701.

Crucitti, P., Latora, V., Marchiori, M., & Rapisarda, A. (2003). Efficiency of scale-free networks: error and attack tolerance. *Physica A: Statistical Mechanics and its Applications, 320*, 622–642. doi:10.1016/s0378-4371(02)01545-5.

Dasgupta, N., Mandl, K. D., & Brownstein, J. S. (2009). Breaking the news or fueling the epidemic? Temporal association between

news media report volume and opioid-related mortality. *PLoS ONE, 4*(11), e7758. doi:10.1371/journal.pone.0007758.

Deuze, M. (2003). The web and its journalisms: considering the consequences of different types of newsmedia online. *New Media & Society, 5*(2), 203.

Dorogovtsev, S. N., Goltsev, A. V., & Mendes, J. F. F. (2006). k-core organization of complex networks. *Physical Review Letters, 96* (4), 040601.

Fan, W., & Yeung, K. H. (2011). Online social networks—Paradise of computer viruses. *Physica A: Statistical Mechanics and its Applications, 390*, 189–197. doi:10.1016/j.physa.2010.09.034.

Flaounas, I., Turchi, M., Ali, O., Fyson, N., De Bie, T., Mosdell, N., et al. (2010). The structure of the EU mediasphere. *PLoS One, 5*(12), e14243. doi:10.1371/journal.pone.0014243.

Gonçalves, B., Perra, N., & Vespignani, A. (2011). Modeling users' activity on twitter networks: validation of Dunbar's number. *PLoS One, 6*(8), e22656. doi:10.1371/journal.pone.0022656.

Harwit, E., & Clark, D. (2001). Shaping the internet in China: evolution of political control over network infrastructure and content. *Asian Survey, 41*(3), 377–408.

Herbert, J., & Thurman, N. (2007). Paid content strategies for news websites. *Journalism Practice, 1*(2), 208–226.

Hindman, M. (2007). A mile wide and an inch deep: measuring media diversity online and offline. *Media diversity and localism: Meaning and metrics, 40*(2), 327–348.

Hindman, M. (2009). *The myth of digital democracy.* Princeton, New Jersey: Princeton University Press.

Hoffman, L. (2006). Is Internet content different after all? A content analysis of mobilizing information in online and print newspapers. *Journalism and Mass Communication Quarterly, 83*(1), 58.

Holme, P. (2005). Core-periphery organization of complex networks. *Physical Review E, 72*, 046111.

Journalism.org (2010). The state of the news media 2010: Online: Audience behavior. http://www.stateofthemedia.org/2010/online_audience.php. Accessed 2010.06.10.

Kaye, B. K., & Johnson, T. J. (2003). From here to obscurity?: media substitution theory and traditional media in an on-line world. *Journal of the American Society for Information Science and Technology, 54*(3), 260–273. doi:10.1002/Asi.10212.

Kohut, A., Doherty, D., & Dimock, M. (2008). *Key news audiences now blend online and traditional sources.* Washington, D.C.: The Pew Research Center for the People and the Press.

Latora, V., & Marchiori, M. (2001). Efficient behavior of small-world networks. *Physical Review Letters, 87*(19), 198701.

Lee, A., & Carpini, M. (2010). News consumption revisited: Examining the power of habits in the 21st century.

Leskovec, J., Krause, A., Guestrin, C., Faloutsos, C., VanBriesen, J., & Glance, N. (2007). *Cost-effective outbreak detection in networks.* Paper presented at the Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, San Jose, California, USA.

Moore, C., & Newman, M. E. J. (2000). Epidemics and percolation in small-world networks. *Physical Review E, 61*(5), 5678.

Newman, M. E. (2001). The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America, 98*(2), 404–409.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Review, 45*(2). doi:10.1137/s003614450342480.

Newman, M. E. J., Forrest, S., & Balthrop, J. (2002). Email networks and the spread of computer viruses. *Physical Review E, 66*(3), 035101.

Ni, S., Weng, W., & Zhang, H. (2011). Modeling the effects of social impact on epidemic spreading in complex networks. *Physica A: Statistical Mechanics and its Applications, 390*(23–24), 4528–4534. doi:10.1016/j.physa.2011.07.042.

Panzarasa, P., Opsahl, T., & Carley, K. (2009). Patterns and dynamics of users' behavior and interaction: network analysis of an online community. *Journal of the American Society for Information Science and Technology, 60*(5), 911–932.

Paulussen, S. (2006). Online news production in Flanders: how Flemish online journalists perceive and explore the internet's potential. *Journal of Computer-Mediated Communication, 9* (4).

Peng, F., Tham, N., & Xiaoming, H. (1999). Trends in online newspapers: a look at the US web. *Newspaper Research Journal, 20* (2).

Quandt, T. (2008). (NO) News on the world wide web? *Journalism Studies, 9*(5), 717–738.

Scott, J. (2007). *Social network analysis: A handbook.* London, UK: Sage.

Singer, J. (2003). Who are these guys?: the online challenge to the notion of journalistic professionalism. *Journalism, 4*(2), 139.

Thelwall, M. (2004). *Link analysis: An information science approach.* Bingley, UK: Emerald Group Pub Ltd.

Varian, H. R., Farrell, J., & Shapiro, C. (2004). *The economics of information technology.* Boston: Cambridge University Press.

Wan, X., & Yang, J. (2007). *Learning information diffusion process on the web.* Paper presented at the Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada.

Wang, Y., Zeng, D., Zheng, X., & Wang, F. (2009a). Internet news media analysis based on complex network theory. *Complex Systems and Complexity Science, 6*(3), 11–21.

Wang, Y., Zeng, D., Zheng, X., & Wang, F. (2009b). *Propagation of online news: dynamic patterns.* Paper presented at the Proceedings of the 2009 IEEE international conference on Intelligence and security informatics, Richardson, Texas, USA.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440–442.

Weber, M., & Monge, P. (2011). The flow of digital news in a network of sources, authorities, and hubs. *Journal of Communication, 61* (6), 1062–1081.

Weldon, M. (2008). *Everyman news: The changing American front page.* Columbia, MO: University of Missouri Press.

Zhang, X. (2007). Breaking news, media coverage and 'citizen's right to know'in China. *Journal of Contemporary China, 16*(53), 535–545.

Zhang, J., Cao, X.-B., Du, W.-B., & Cai, K.-Q. (2010). Evolution of Chinese airport network. *Physica A: Statistical Mechanics and its Applications, 389*(18), 3922–3931. doi:10.1016/j.physa.2010.05.042.

Zheng, X., Zeng, D., Li, H., & Wang, F. (2008). Analyzing open-source software systems as complex networks. *Physica A: Statistical Mechanics and its Applications, 387*(24), 6190–6200. doi:10.1016/j.physa.2008.06.050.

Youzhong Wang is a doctoral candidate in the State Key Laboratory of Management and Control for Complex Systems at the Institute of Automation of the Chinese Academy of Sciences. His current research interests include social dynamics, epidemic dynamics, complex networks, and social media.

Daniel Zeng is a professor in the State Key Laboratory of Management and Control for Complex Systems at the Institute of Automation of the Chinese Academy of Sciences and Department of Information Systems at University of Arizona. His research interests include multi-agent systems, collaborative information and knowledge management, recommender systems, and automated negotiations and auctions.

Bin Zhu is an Assistant Professor of Business Information Systems in the College of Business. Prior to OSU she was an assistant professor at Boston University. She earned her Ph.D. in Management Information Systems from University of Arizona. Her current research interests include business intelligence, information analysis, social network, human-computer interaction, information visualization, computer-mediated communication, and knowledge management systems. She has been a lead author for papers that have appeared in Information Systems Research, Decision Support Systems, Journal of the American Society for Information Science and Technology, IEEE Transaction on Image Processing, and D-Lib Magazine. Her research also received an IBM faculty award.

Xiaolong Zheng is an assistant professor in the State Key Laboratory of Management and Control for Complex Systems at the Institute of Automation of the Chinese Academy of Sciences. His research interests include complex networks, social computing, and data mining. Specifically, he is interested in social dynamics, social influence, spread dynamics, and opinion and behavior mining in social media.

Feiyue Wang is a professor in the State Key Laboratory of Management and Control for Complex Systems at the Institute of Automation of the Chinese Academy of Sciences. His current research interests include social computing, parallel control and management, web and services science, and agent-based intelligent systems.