Propagation of Online News: Dynamic Patterns

Youzhong Wang[1], Daniel Zeng[1,2], Xiaolong Zheng[1], Feiyue Wang[1]
[1]The Key Laboratory of Complex Systems and Intelligence Science
Institute of Automation, Chinese Academy of Sciences, Beijing, China
[2]MIS Department, University of Arizona, Tucson, Arizona, U.S.A.
wangyouzh@gmail.com,{dajun.zeng,xiaolong.zheng,feiyue.wang}@ia.ac.cn

Abstract—**A large portion of online news articles and postings are not originally created but reprinted or re-posted from other online news sources or portals. In this paper, we analyze the dynamics of online news propagation, using a large collection of Chinese online news activity data. We characterize prominent features of online news diffusion and compare them against the spreading patterns of the epidemic. Several critical factors influencing the news propagation process are identified, including the centrality and selectivity of source portals, and event variability.**

Keywords - social meida; online news propagation; complex networks

I. INTRODUCTION

Due to broad and easy accessibility, and timely and often uncensored reporting of current events and opinions, Web-based news reporting and social media have become an important channel of communication and information sharing. Studying the information diffusion process in the context of social media and online news has increasingly gained attention in social network research [1, 2]. Most current research in this area draws parallelism between epidemic spreading on complex networks and information diffusion through online news production and consumption nodes. However, information diffusion possesses many unique properties. For instance, variations of a news story are often generated in the diffusion process. In order to better understand the information ecology of social media and online news, research is critically needed to discover specific factors that influence information transmission and develop models that can help explain the mechanisms through which these factors influence the information diffusion process.

In this paper, we study news propagation through Internet-based news media and focus on empirical findings that set dynamic patterns of online news propagation apart from those studied in the existing body of literature, which are mostly based on standard epidemic spreading models. It is shown that news articles spread quickly in a star-shaped topological schema, and that the centrality or "core-ness" of source portals also has major influence over the propagation process. Our study reveals that only a limited number of news portals get "infected" even for the hottest events and that the selectivity of news portals can be clearly observed. Furthermore, we find the oscillatory effect in news events' spreading, which is related to the generation of variations of news reporting based on the original material.

Research reported in this article has been supported by the National Natural Science Foundation of China (70890084, 60875049, and 60621001), the Chinese Academy of Sciences (2F07C01and 2F08N03), and the Ministry of Science and Technology (2006AA010106).

Our empirical study is based on a home-grown large test bed of online news documents from more than one thousand Chinese online news outlets. These news documents have been collected through scanning and crawling these news outlets on a daily basis since March 2008. From this collection (and on the Web in general), many news documents are not originally created but reprinted or re-posted from other news portals. To extract the reprinting/re-posting relationship between news outlets, we have made use of a range of meta information (e.g., publishing site and time, source site, etc.) parsed from the raw HTML files collected. Reference [3] provides details about the data collection process and the construction of the News Reprinting/Re-posting Network (NRN). In an NRN, a node represents a news portal and an edge indicates the existence of a reprinting transaction between two portals. For our analysis, we have selected various types of news events, including city news, international news, economic news, and entertainment news, to study news propagation on NRNs. The focus of our research is to examine the dynamic diffusion properties of these NRNs, using epidemic spreading modeling as a baseline.

II. MODELING BACKGROUND

There is a long history of theoretical studies of infectious diseases spreading. The susceptible-infected-removed (SIR) model and the susceptible-infected-susceptible (SIS) model have been studied extensively over past decades [4]. In those models, individuals are classified into different classes according to their states (e.g., susceptible or infected). Recent years have seen active work on epidemic threshold theory and epidemic dynamics based on complex network theory.

In the SIR model, effective spreading rate λ is defined as the probability that a susceptive individual gets infected divided by the probability that an infected individual recovers. Early studies found the presence of a positive epidemic threshold λ_c on homogenous network [4]. If λ is below λ_c, infection dies out exponentially fast. Otherwise, general outbreaks can be expected. Recently, several researchers have found the absence of the epidemic threshold on scale-free networks, which means that infection will spread over and stay at a steady state if the effective spreading rate is positive [5].

Studies on epidemic dynamics on heterogeneous networks show that outbreak evolution follows certain dynamics: the most connected individuals get infected first, and then infection pervades the individuals with smaller degrees [6]. In addition, oscillation is found in growing scale-free networks, which is related to the growth, infection and immune rates [7].

III. DATASET

For our study, we use all the news documents collected from about 1086 Chinese online news outlets from September 10, 2008 to October 10, 2008. We have first constructed a News Reprinting/Re-posting Network (NRN) based on the raw news dataset. In an NRN, a node represents a news portal and there is a weighted directed edge between node/portal *u* and *v*, where the weight of the edge corresponds to the number of news documents portal *v* reprinted from portal *u*. Our previous research has shown that NRNs have several characteristics: the scale-free property, the small world effect, disassortative mixing, and the core-periphery structure [3].

In the study reported in this paper, we have selected 4 distinctive events and partitioned the related news to analyze event type-specific patterns. These events are 1) San Lu milk powder scandal (Dairy Farmer); 2) truck bomb attacks at the Marriott in Pakistan (Marriott); 3) China cutting the loan rate and deposit reserve ratio (Loan Rate); 4) Xu Deliang and Wang Wenliang quitting the famous cross-talk group Deyun She (Deyun She). The statistics about the news articles concerning these events are provided in Table 1, including the numbers of articles published, the numbers of infected sites (i.e., the sites publishing or re-posting related articles), the numbers of groups (a group capturing the source site and all the sites that have re-posted articles from the source site directly or indirectly (through other intermediary sites)) identified, as well as the life-spans of these events.

TABLE I. STATISTICS OF EVENTS

Event ID	Event	# of News	# of Sites	# of Groups	Life-Span (days)
1	Dairy Farmer	717	206	25	18
2	Marriott	2877	308	110	20
3	Loan Rate	8312	474	237	21
4	Deyun She	521	160	24	21

IV. AN EMPIRICAL ANALYSIS

A. Star-Shaped News Propagation

We have constructed the news propagation tree for each group, with the root of the tree representing the source portal in the group, and tree nodes representing those portals that have reprinted articles from the parent portal. We found that almost 80% of the portals have reprinted news directly from the source portal, and that the average height of the trees is less than 2.10. This observation indicates that many reprinting transactions can be characterized as star-shaped. Moreover, we found that the average life-span of a news group is about 50 hours, and that more than 80% of the portals reprinted news articles within 24 hours since the original news was published.

B. The Effect of Source Portals on News Propagation

Generally, major news portals publish many original news documents and these original documents are reprinted by other portals in large quantity. In an NRN network, nodes with large degrees represent major portals on the Web, and out-degree of a node represents the number of portals reprinting news from it. We analyze the effect of source portals' out-degrees on news propagation in Fig. 1. As shown in Fig. 1 (a), original news published by portals with larger out-degree tends to reprinted by more portals on the Web. It is observed that original news published by portals with out-degree falling in the interval [600,800] is reprinted by fewer portals than news published by portals with out-degree failing in the interval [500,600]. This exception can be explained by that the fact that those portals with out-degree above 600 publish a large number of original news as shown in Fig. 1 (b), but not all these news pieces are interesting enough to be picked up by other news portals.

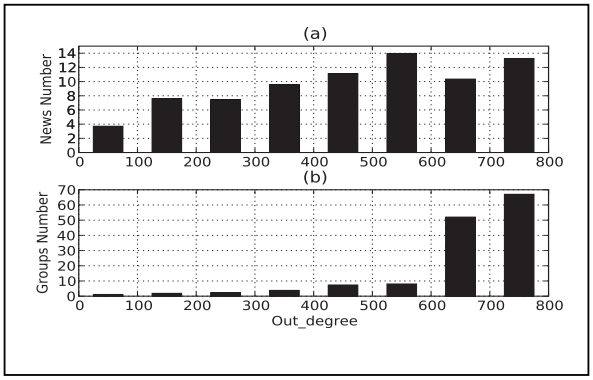

Figure 1. The effect of source portals' out-degree on online news reprintting. (a) Average # of news in the groups; and (b) Average # of groups published by one source.

V. EVENT-SPECIFIC PROPAGATION PATTERNS

The propagation of a news event can be described as follows. The source portal publishes the first news document about the event. Then other portals reprint the news documents on the Web or contribute new original reports about the event. This section summarizes our findings concerning the steady state of news event propagation and the observed oscillatory effect.

A. Steady State of News Events Propagation

From the collection of news documents from 1086 online news outlets used in our study, a small portion of the news portals have published news on the selected 4 events, as shown in Table 1. In news event propagation, a portal can get infected many times but the life-span of these events is typically around 20 days (most other events were not as hot or attention-grabbing and thus significantly shorter-lived compared to the 4 events selected for this study). The findings show major differences between news propagation and epidemic spreading on scale-free network. For instance, in the spreading of a typical computer virus, a small portion of nodes get infected but the infection can exist in a very long period [4]. The dissimilarity can be explained partially by the different spreading mechanisms between online news and epidemic. As observed above, original news documents are typically reprinted within two and a half days, and the attractiveness of news events fades away quickly (despite their "hotness"). The thread of news reporting and the attention it generates are maintained through publishing fresh, original

news documents continually. Infection quickly dies out as fresh original news (or derivative) reporting stops.

The fact that only a limited number of news portals are infected for any event is mainly due to the topic concentration or selectivity of news portals. For example, some portals are inclined to publishing international news; while others may concentrate on entertainment news. In our data set, 563 news portals got infected for at least one event; only 72 portals got infected for all four events; and 244 portals got infected by only one event. These observations strongly support that selectivity, one form of heterogeneity, of portals has a major impact on news propagation.

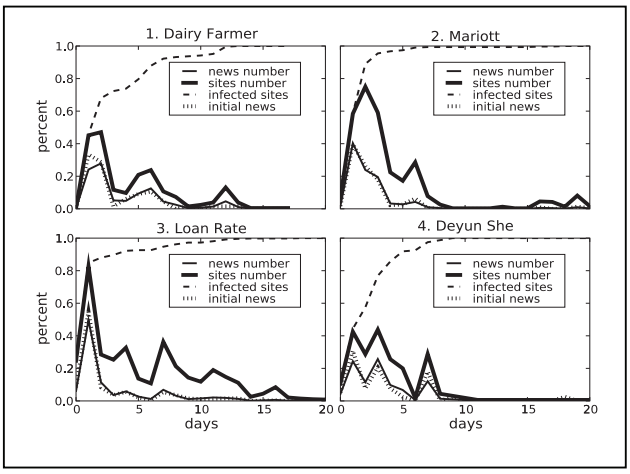

Figure 2. Information Propagation for 4 Events. The lines represent percent of news documents published in each day, percent of sites which publish news in each day, percent of sites which have published news before, and percent of initial news published in each day, respectively.

B. Oscillation in News Events Propagation

Some dynamic aspects of online news events are shown in Fig. 2. Most news portals were infected during the first three days, indicating that the events propagate at a very high speed at the beginning. Unlike most research on epidemic spreading, we found an interesting phenomenon: the rate of infection fluctuating quite a bit during a news event' life span. The number of news portals under study almost remained unchanged in our data, so the reason for this oscillation in news events propagation is not because of the network growth (e.g., in scale-free networks [7]). From Fig. 2, we observe that the fluctuation in the number of news articles published each day is strongly positively correlated to the number of variations in reports derived from the initial original material. This provides a viable explanation for the observed oscillation in news events propagation: when a change occurred during an event, various portals published follow-up reports updating the latest situation, resulting in immediate spikes in the news generation.

VI. CONCLUSION

In this paper, we have studied online news propagation from an empirical perspective. We notice significant differences as to dynamic propagation patterns between online news and epidemics and argue that modeling efforts aimed to capture these observed unique characteristics of online news propagation could lead to fruitful research results.

Our current ongoing work focuses on two extensions of the reported research. First, we are developing complex network-based mathematical model to explain the characteristics of online news propagation. Second, we are extending our data set to cover user-generated comments and feedback based on online news reporting, and are conducting related empirical and modeling studies.

REFERENCES

[1] G. Daniel, R. Guha, L.-N. David, and T. Andrew, "Information diffusion through blogspace," in Proceedings of the 13th international conference on World Wide Web, New York, NY, USA: ACM, 2004, pp. 491-501.

[2] X. Wan and J. Yang, "Learning information diffusion process on the web," in Proceedings of the 16th international conference on World Wide Web, Banff, Alberta, Canada: ACM, 2007, pp. 1173-1174.

[3] Y. Wang, D. Zeng, X. Zheng, and F. Wang, "Internet news media analysis based on complex network theory," Complex systems and complexity science, in press.

[4] X. Wang, X. Li, and G. Chen, Complex network theory and its application. Beijing: Tsinghua University Press, 2006.

[5] R. Pastor-Satorras and A. Vespignani, "Epidemic Spreading in Scale-Free Networks," Physical Review Letters, vol. 86, p. 3200, 2001.

[6] M. Barthélemy, A. Barrat, R. Pastor-Satorras, and A. Vespignani, "Dynamical patterns of epidemic outbreaks in complex heterogeneous networks," Journal of Theoretical Biology, vol. 235, pp. 275-288, 2005.

[7] Y. Hayashi, M. Minoura, and J. Matsukubo, "Oscillatory epidemic prevalence in growing scale-free networks," Physics Review E, vol. 69, p. 016112, 2004.