On Burst Detection and Prediction in Retweeting Sequence

Zhilin Luo[1], Yue Wang[2], Xintao Wu[3(✉)], Wandong Cai[4], and Ting Chen[5]

[1] Shanghai Future Exchange, Shanghai, China
luo.zhilin@shfe.com.cn
[2] University of North Carolina at Charlotte, Charlotte, USA
ywang91@uncc.edu
[3] University of Arkansas, Fayetteville, USA
xintaowu@uark.edu
[4] Northwestern Polytechnical University, Xi'an, China
caiwd@nwp.edu.cn
[5] Northeastern University, Boston, USA
tingchen@ccs.neu.edu

Abstract. Message propagation via retweet chain can be regarded as a social contagion process. In this paper, we examine burst patterns in retweet activities. A burst is a large number of retweets of a particular tweet occurring within a certain short time window. The occurring of a burst indicates the original tweet receives abnormally high attentions during the burst period. It will be imperative to characterize burst patterns and develop algorithms to detect and predict bursts. We propose the use of the Cantelli's inequality to identify bursts from retweet sequence data. We conduct a comprehensive empirical analysis of a large microblogging dataset collected from the Sina Weibo and report our observations of burst patterns. Based on our empirical findings, we extract various features from users' profiles, followship topology, and message topics and investigate whether and how accurate we can predict bursts using classifiers based on the extracted features. Our empirical study of the Sina Weibo data shows the feasibility of burst prediction using appropriately extracted features and classic classifiers.

1 Introduction

Microblogging, such as Twitter and Sina Weibo, has attracted a huge number of users and becomes increasingly popular. In Twitter, a user can tweet any message within 140-character limit or share pictures, follow any interesting users, and comment or retweet messages that she received from her followees. A tweet can reach the immediate followers of the owner user and can further reach other users when retweeted by some followers. Hence the retweeting mechanism empowers users to spread their ideas beyond the research of the original tweet's followers. Message propagation via retweet chain can be regarded as a social contagion process.

T. Cao et al. (Eds.): PAKDD 2015, Part I, LNAI 9077, pp. 96–107, 2015.
DOI: 10.1007/978-3-319-18038-0_8

In this paper, we examine burst patterns in retweet activities. A burst is a large number of retweets of a particular tweet occurring within a certain short time window. We can consider a burst as a spike and its duration is often short as compared to the surrounding non-burst durations. The occurring of a burst indicates the original tweet receives abnormally high attentions during the burst period. As a result, it will be imperative to characterize burst patterns and develop algorithms to detect and predict bursts.

Many tweets receive little interests in their life cycle and have no burst at all. The propagation of those no-burst tweets often experiences only two stages: low growth and long extinction. However, for tweets that receive significant attention and spread widely in microblogging sites, their propagation often experience eruption, continuance and extinction. Some tweets have a distribution with single-burst whereas other tweets have a distribution with multi-burst. The single-burst indicates that the original tweet receives wide intensive attention in its short period of eruption and then fades away gradually without raising any further significant attention. On the contrary, some tweets may receive intensive attention in several different periods of times during their life cycle due to some triggering event. As a result, they have a distribution with multi-burst. The multi-burst is often characterized by slowly alternating phases of near steady state behavior and rapid spikes. The propagation of multi-burst tweets has five cyclic stages: eruption, continuance, decay, dormant and reflourish, within their (often long) life durations.

In this paper, we propose the use of the Cantelli's inequality to identify bursts from retweet sequence data. We treat bursts as outliers (i.e., significantly different from the average) in retweet sequence data. We then conduct a comprehensive empirical study of burst pattern using the Sina Weibo data and examine various factors, including tweet users and topics, that may have effects on burst. We extract various features from users' profiles, followship topology, and message topics and investigate whether and how accurate we can predict bursts using classifiers based on the extracted features.

2 Burst Characterization

We define the life duration of a particular tweet as the time period from when it was originally posted to when it was lastly retweeted. We convert the retweet frequency information of a given tweet into a time series where each value indicates the number of occurrence of retweets during the time window. The size of the time window could be minutes, hours, or even days dependent on the application. Formally, denote t_i the i^{th} time window after the original tweet is posted, and x_{t_i} the number of retweets in the i^{th} time window. The retweet time series is defined by $X = x_{t_1}, x_{t_2}, \ldots, x_{t_n}$. A burst is a large number of retweets occurring within certain time windows of the tweet's life duration. We define the **burst duration** of the tweet as the total number of time windows for which all bursts last.

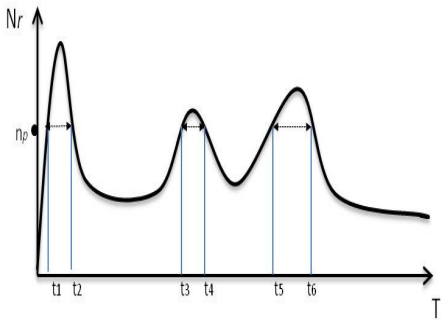


Fig. 1. The retweeting time series of a tweet

Figure 1 shows an example of a retweet series with three bursts: t_1 to t_2 for the first burst, t_3 to t_4 for the second burst, and t_5 to t_6 for the third burst. The retweeting sequence often has a much longer non-burst duration which includes the period before the the first burst, the period between two consecutive bursts, and the period after the last burst. We propose the use of the Cantelli's inequality to identify those bursts.

Theorem 1. *(Cantelli's inequality) Let X be a random variable with finite expected value μ and finite non-zero variance σ^2. Then for any real number λ*

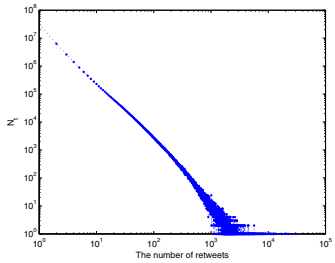
$$Pr(X - \mu \geq \lambda) \begin{cases} \leq \frac{\sigma^2}{\sigma^2 + \lambda^2} & if \quad \lambda > 0, \\ \geq 1 - \frac{\sigma^2}{\sigma^2 + \lambda^2} & if \quad \lambda < 0. \end{cases} \tag{1}$$

The Cantelli's inequality is a generalization of Chebyshev's inequality in the case of a single tail. When $\lambda > 0$, we have $Pr(X - \mu \geq \lambda\sigma) \leq \frac{1}{1+\lambda^2}$. We treat as outliers those values that are more than λ standard deviations σ away from the mean μ. The number of outliers are no more than $1/(1 + \lambda^2)$ of the distribution values. Those outliers form the bursts in the retweeting sequence. In our paper, we set the λ value as 2.

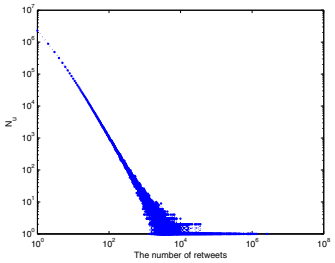
3 Empirical Evaluation of Burst Patterns

We conduct an empirical study using the WISE 2012 Challenge Data [1]. The WISE 2012 Challenge is based on a dataset collected from the Sina Weibo, one of the most popular Microblogging service in China. In the data, content of tweets are removed and some tweets are annotated with events. For each event, the terms that are used to identify the event are given. Each tweet includes the basic information such as time, user ID, message ID, mentions (user IDs appearing in tweets), retweet paths, and whether containing links. The followship network is also provided. The data set contains 5,636,858 users with 46,584,914 original tweets being retweeted by 190,920,026 times.

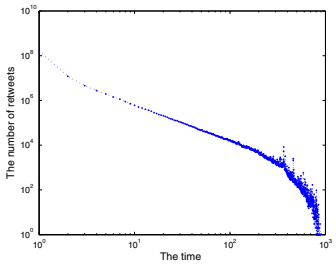
[1] http://www.wise2012.cs.ucy.ac.cy/challenge.html



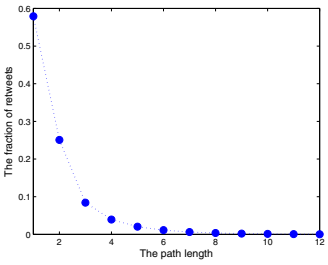
(a) Retweet-Tweet distribution



(b) User-Retweet distribution



(c) Retweeting time (days) from original post)



(d) Path length

Fig. 2. Retweet distribution from WISE 12 Challenge data (200M tweets)

3.1 An Overview of Retweet Patterns

Our preliminary findings are summarized below.

- Figure 2(a) shows the distribution of the number of retweets that each tweet receives. We observe from the figure that most of the original tweets receive less than 10 retweets in their lifetime, while a small number of tweets receive hundreds or even thousands retweets, e.g., the largest number of retweets from a single tweet reaches 34,096 in the data set.
- Figure 2(b) shows the distribution of the number of retweets that each user receives. Tweets authored by a small number of influential users (e.g., celebrities, actors, stars) are very popular and receive most retweets. For example, the top 100 most influential users receive 46,094,478 retweets in total, about 24.2% of all retweets; while the top 1000 most influential users receive 86,501,021 retweets, about 45.3% of all retweets.
- Figure 2(c) shows the distribution of retweeting time of each retweet. We can observe that most retweets occur in a very short period of time after the tweet's posting. For example, 81.8% of the tweets would not be retweeted any more after the first day and only 6.28% of the tweets would last for more than two days. However, a small number of tweets would still be retweeted even after 100 days.

– Figure 2(d) shows the path length distribution of each retweet. The path length of a retweet is defined as the number of hops of the retweeting user away from the original user who posts the tweet. For example, given a retweeting sequence of $A \rightarrow B \rightarrow C \rightarrow D$, user A's tweet is retweeted by user B, then retweeted by user C through user B, and finally retweeted by user D through user C. The path length of the retweet by user C is 2 while the path length of the last retweet by user D is 3. We can observe from the figure that 57.9% of the retweets have one single hop and 98.6% of the retweets are within five hops, matching the concept of *six degrees of separation* in social networks.

3.2 Burst Pattern

In the previous section, we found that some popular tweets are widely retweeted and their retweets last a long time after their posting. On the contrary, a majority of tweets would not be retweeted any more shortly after their posting. In this section, we focus on those tweets that have been retweeted more than 100 times in the data set. We extract 207,259 such tweets.

For those 207,259 popular tweets (each receiving more than 100 retweets), the majority (68.71%) include only one single burst, which often occurs in the first day when the original tweet is posted. 12.84% tweets have no burst and 18.45% of tweets have multi-burst. Tweets with multi-burst often have longer path lengths and longer active duration time than tweets with single or no burst, as shown in Table 1. Among the tweets with multi-burst, there are 31,300 tweets with two bursts and 2,782 tweets with more than 4 bursts. The maximum number of bursts is 17 in the data set.

Table 1. The burst distribution and the average path length of 207,259 original tweets (each receiving more than 100 retweets)

Burst	Number of tweets	Ratio(%)	Avg. of path length	Avg. of life duration (days)
No	26620	12.84	2.03	4.78
Single	142406	68.71	2.09	10.27
Multi	38233	18.45	2.32	18.87

Different Bursts. We examine whether the average path length of retweets occurred in each burst period is different. Our conjecture is that for a tweet authored by user A, its retweets occurred in the first burst are more from user A's immediate followers and retweets occurred in later bursts are more from A's indirect followers. Our findings show that the average path length of retweets in the first burst is shorter than that in following bursts. Specifically, the average path lengths for the first four bursts in our data set are 2.08, 2.29, 2.92, and 2.99 respectively, which validates our conjecture.

We further examine the path length distribution of retweets occurred in each burst. Each curve in Figure 3 shows the fraction of retweets for each path length value from 1 to 10. For retweets occurred in the first burst, 45.6% of retweets

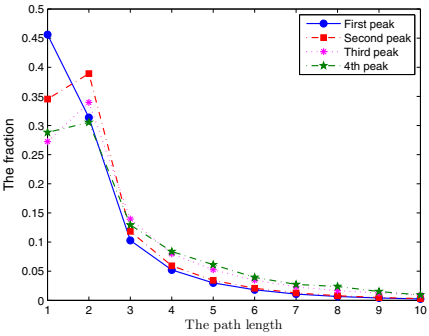


Fig. 3. Path length distribution of retweets occurred in each peak

have the path length of 1 and 30.6% of retweets have the path length of 2. However, for retweets occurred in the second burst, 39.3% of retweets have the path length of 2, which are more than the number of retweets (35.1%) with the path length of 1. This shows that the second burst is mainly caused by non-immediate followers of original users who post the tweets. We have the similar phenomenon for the third burst and the fourth burst.

Burst Pattern vs. Topics. We examine whether topics of original tweets have effects on burst patterns. We extract four hot topics, i.e., house price, xiao mi release, family violence of Li Yang, and case of running fast car in Hei Bei university. We denote them as *House*, *Xiaomi*, *Li Yang*, and *He Bei*, respectively. For those tweets with no assigned topic, we group them in the *Unknown* category.

Table 2. The comparison of different topics

Topic	Avg. of path length.	Avg. burst duration(days)	Avg. life duration(days)
Unknown	1.90	2.95	15.85
House	2.66	3.05	22.75
Xiao Mi	2.62	3.39	16.32
Li Yang	2.89	3.71	21.20
He Bei	2.97	3.64	28.15

Table 2 shows the general comparison of different categories in terms of path length, burst duration, and life duration. We can see that the tweets from the *Unknown* category have shorter path length, shorter burst duration, and shorter life duration time than the tweets with known topics. This indicates tweets with some particular hot topics are often widely propagated in the microblogging site.

Figure 4 shows the path length distribution for each topic category under study. The curve of *No Topic* (aka, *Unknown*) is significantly different from other curves corresponding to known topic categories. We can observe that the proportions of retweets from *Unknown* category with path length 1 and 2 are 45% and 38% respectively, which are much higher than the corresponding proportions for retweets with known topics.

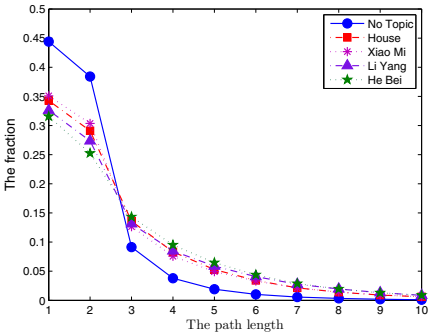


Fig. 4. Path length distribution of retweets of each topic

Table 3. The information of different types of users

Users	Avg. of path length.	Avg. burst duration(days)	Avg. life duration(days)
Top 100	1.86	2.73	10.63
Top 100-1000	2.17	3.05	18.34
Normal	3.04	3.77	28.46

Burst Pattern vs. Users. We examine whether different types of users who post tweets have effects on burst pattern. Table 3 shows the general comparison of three types of users: top 100, top 100-1000, and normal users. We define the top 100 users as those who rank among the top 100 in terms of the total number of retweets each user receives. We can see there are significant differences in terms of path length, peak time, and duration time among three types of users. Tweets from the top 100 most influential users have much shorter path lengths, burst duration, and tweet life duration than the top 100-1000 and normal users.

Figure 5 shows the path length distribution for each type of user under study. We can observe that the proportions of retweets of those tweets authored by the

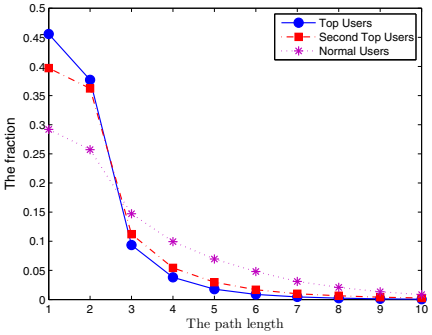


Fig. 5. Path length distribution of retweets from three types of users

top 100 users with path length 1 and 2 are 46% and 39%, respectively, which are much higher than the corresponding proportions for tweets from the top 100-1000 and normal users. This phenomena shows that the top 100 most influential users can propagate their messages more quickly in the microblogging site than other users.

4 Burst Prediction

We are interested in the following prediction problem: given a tweet with known information about its content, its user profile, the followship topology, and the observed retweet sequence in the first 12 hours, can we predict whether the tweet will have multi-burst in the future of its life cycle.

One challenge here is what kind of features we can extract from the known information and how useful they are for burst prediction. In our study, we extract 178 features from the a-priori known information of a tweet (i.e., its topics, user profile, followship topology, and its observed retweet sequence in the first 12 hours). The extracted features can be roughly grouped into two main classes: user-related and tweet-related.

In the user-related class, we extract features from the profile of the user who posts the original tweet. For example, we extract the number of his immediate followees, the number of his two-hop followees, the number of tweets the user has authored, the average number of retweets received in the first 12 hours for all his tweets, and the numbers of tweets with no, single, and multiple bursts.

In the tweet-related class, we extract the features such as the tweet's post time, first retweeting time, the presence/absence of hot topics in the tweet, the presence/absence of hot topics in its retweets, the presence/absence of @users in the tweet, the presence/absence of @users in its retweets, the number of retweets containing @users and the number of @users in its retweets, etc. For each tweet, we also build a retweet tree from its observed retweet sequence in the first 12 hours and extract features such as the maximum width, the maximum height, the number of retweet users, and the average path length.

In our experiment, we exclude from the Sina Weibo dataset those records in which the original tweets' user ID could not be found in the followship network. Finally, we build a training data set with 30,084 tweets with no multi-burst and 30,030 tweets with multi-burst.

We run a suite of 7 classifiers: Logistic Regression (LR), Random Forest(RF), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and k-Nearest Neighbor (kNN). We take the 10 fold cross-validation for each classifier. The accuracy result is shown in Figure 6. We can observe that Random Forest, Decision Tree, k-Nearest Neighbor, and Logistic Regression achieve good prediction results in terms of accuracy (higher than 72%).

We then analyze the effect of each feature on prediction. We take the logistic regression coefficient as the effect. The regression coefficients represent the change in the logit for each unit change in the feature. The larger the absolute value of the coefficient is, the more effect the feature takes. Formally, we can

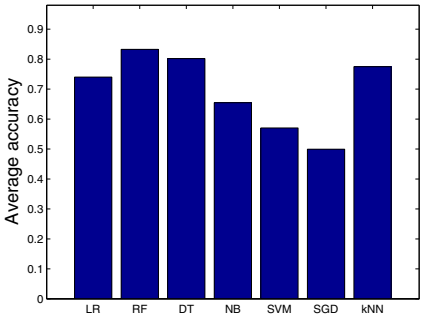


Fig. 6. Accuracy of Classifiers: Logistic Regression (LR), Random Forest(RF), Decision Tree (DT), Naive Bayes (NB), Support Vector Machine (SVM), Stochastic Gradient Descent (SGD), and k-Nearest Neighbor (kNN)

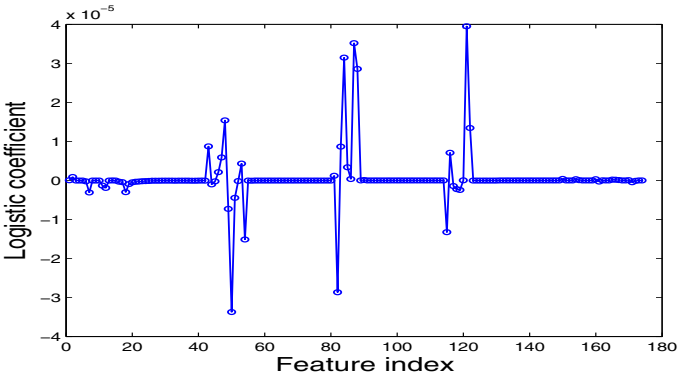


Fig. 7. Logistic Coefficient of Features

use the likelihood ratio test or the Wald statistic to assess the significance of an individual feature. Our results show that there are only 20 features with relatively large coefficient values. Figure 7 plots the logistic regression coefficient for each feature where X-axis represents different features and Y-axis shows each feature's coefficient value. We list top 5 most significant features in Table 4. We can see that the average number of retweets with path length 1 of the user's all

Table 4. Top 5 most significant features (PL1 denotes path length 1)

Index	Meaning	Coefficient
121	Avg no of PL1 retweets of user's all tweets	3.95E-05
87	Avg no of PL1 retweets (first 12h) of user's no-burst tweets	3.51E-05
50	Avg no of retweets (first 12h) of user's multi-burst tweets	-3.37E-05
84	Avg no of retweets (first 12h) of user's no-burst tweets	3.14E-05
82	Avg no of retweets of user's no-burst retweets	-2.86E-05

tweets is the most significant feature with the coefficient value 3.95E-05. In our future work, we will conduct detailed correlation analysis and examine prediction performance after removing those redundant features.

5 Related Work

Examining retweet behavior has been an active research area recently [7–9,12,13]. For example, the authors in [7] studies the coverage prediction of retweets, i.e., what is the number of times that a particular message posted by a user will be retweeted. In [13], the authors examine various factors such as user, message, and time and propose a factor graph model to predict whether a user will retweet a message. The authors in [9] study why people retweet and examine the anti-homophily phenomena. In [8], the authors examine the use of log-linear modeling to identify multi-way interactions between retweet and various features such as power ratio, link structure and users' profile information. In [12], the authors analyze the ways in which hashtags spread on twitter and show widely-used hashtags on different topics spread significantly different.

Change detection models [1,4] provide a standard approach to detecting deviations from baseline. Usually we assume the mean and variance of a distribution representing normal behavior and the mean and variance of another distribution representing behavior that is abnormal. We can measure deviations from normal using the generalized likelihood ratio. For example, in [4], the authors assume both distributions are Gaussian with the same variance and the change is reflected in the mean of the observations. In this context, they apply the generalized likelihood ratio to score changes from baseline.

Techniques for finding burst patterns in data streams have also been presented in [6,11,15,16]. In [6], the authors examine bursty structure in temporal text streams (e.g., emails or blogs). They examine how frequency words change over time. The burstiness of words is defined as those words with significantly higher frequency than others. They propose to model the stream using an infinite-state automaton, in which bursts appear naturally as state transitions. In [16], the authors examine point monitoring and aggregate monitoring in time series data streams and design a new structure, called the Shifted Wavelet Tree, for elastic burst monitoring. In [15], the authors propose a family of data structures based on the Shifted Binary Tree for elastic burst detection and develop a heuristic search algorithm to find an efficient structure given the input. In [11], the authors study how to detect, characterize and classify bursts in user query logs of large scale e-commerce systems. The authors build several models that continually detect newer bursts with minimal computation and provide a mechanism to rank the identified bursts based on a number of factors such as burst concentration, burst intensity and burst interestingness. They also propose several quantities to rank bursts including duration of burst, mass of burst, arrival rate for burst, span ratio, momentum of burst, and concentration of burst, and apply unsupervise learning techniques to classify the bursts based on their patterns. Although extensive work has been done in related fields for mining various

temporal patterns, we notice that very little work has been done to detect and predict interesting burst patterns from large-scale retweet sequence data.

Message propagation can be regarded as a social contagion process. There has been research on rumor propagation [5,10,14]. In [14], the authors study the dynamics of an epidemic-like model for the spread of a rumor on a small-world network. In [10], the authors study the dynamics of a generic rumor model on complex scale-free topologies and investigate the impact of the interaction rules on the efficiency and reliability of the rumor process. In [5], the authors apply the susceptible-infectious-recovered and susceptible-infectious-susceptible models to study the spreading process in complex networks. However, we notice that very little work has been done to detect and predict burst patterns.

6 Conclusion

In this paper, we have proposed the use of the Cantelli's inequality to identify bursts from retweet sequence data. With the use of the Cantelli's inequality, we do not need to assume the distribution of the retweet sequence data and can still identify bursts efficiently. We conducted a complete empirical study of burst pattern using Sina Weibo data and examined what factors would affect burst. We extracted various features from users' profiles, followship topology, and message topics and investigated whether and how accurate we can predict bursts using various classifiers based on the extracted features. Our empirical evaluation results show the burst prediction is feasible with appropriately extracted features and classifiers.

In our future work, we will investigate various regression analysis methods [3] on extracted features to predict when a tweet produces its first burst as well as following bursts. We will analyze the bursts to see what their causality was by matching external events that might have caused the bursts. In our future work, we will also study how to classify bursts based upon their shapes, durations, and derived burst characteristics. We will examine various burst characteristics such as burst concentration, burst intensity and burst interestingness. We will study how the window size affects burst detection and categorization. Finally, we will study the use of topic modeling [2] to analyze tweet content and automatically identify the topics of every tweet.

Acknowledgments. The authors would like to thank anonymous reviewers for their valuable comments and suggestions. This work was supported in part by U.S. National Science Foundation (CCF-1047621), U.S. National Institute of Health (1R01GM103309), and the Chancellor's Special Fund from UNC Charlotte.

References

1. Basseville, M., Nikiforov, I., et al.: Detection of abrupt changes: theory and application, vol. 104. Prentice Hall, Englewood Cliffs (1993)
2. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. JMLR **3**, 993–1022 (2003)

3. Cohen, J., Cohen, P.: Applied multiple regression/correlation analysis for the behavioral sciences. Lawrence Erlbaum (1975)
4. Curry, C., Grossman, R., Locke, D., Vejcik, S., Bugajski, J.: Detecting changes in large data sets of payment card data: a case study. In: KDD, pp. 1018–1022. ACM (2007)
5. Kitsak, M., Gallos, L., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H., Makse, H.: Identification of influential spreaders in complex networks. Nature Physics **6**(11), 888–893 (2010)
6. Kleinberg, J.: Bursty and hierarchical structure in streams. Data Mining and Knowledge Discovery **7**(4), 373–397 (2003)
7. Luo, Z., Wang, Y., Wu, X.: Predicting retweeting behavior based on autoregressive moving average model. In: Wang, X.S., Cruz, I., Delis, A., Huang, G. (eds.) WISE 2012. LNCS, vol. 7651, pp. 777–782. Springer, Heidelberg (2012)
8. Luo, Z., Wu, X., Cai, W., Peng, D.: Examining multi-factor interactions in microblogging based on log-linear modeling. In: ASONAM (2012)
9. Macskassy, S.A., Michelson, M.: Why do people retweet? anti-homophily wins the day! In: ICWSM (2011)
10. Moreno, Y., Nekovee, M., Pacheco, A.: Dynamics of rumor spreading in complex networks. Physical Review E **69**(6), 066130 (2004)
11. Parikh, N., Sundaresan, N.: Scalable and near real-time burst detection from ecommerce queries. In: KDD, pp. 972–980. ACM (2008)
12. Romero, D.M., Meeder, B., Kleinberg, J.: Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In: WWW, pp. 695–704. ACM (2011)
13. Yang, Z., Guo, J., Cai, K., Tang, J., Li, J., Zhang, L., Su, Z.: Understanding retweeting behaviors in social networks. In: CIKM, pp. 1633–1636. ACM (2010)
14. Zanette, D.: Dynamics of rumor propagation on small-world networks. Physical Review E **65**(4), 041908 (2002)
15. Zhang, X., Shasha, D.: Better burst detection. In: ICDE, pp. 146–146. IEEE (2006)
16. Zhu, Y., Shasha, D.: Efficient elastic burst detection in data streams. In: KDD, pp. 336–345. ACM (2003)