2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Against the Others! Detecting Moral Outrage in Social Media Networks

Wienke Strathern*, Mirco Schoenfeld†, Raji Ghawi*, and Juergen Pfeffer*

* Bavarian School of Public Policy
Technical University of Munich, Munich, Germany
{wienke.strathern, raji.ghawi, juergen.pfeffer}@tum.de

† University of Bayreuth
Bayreuth, Germany
mirco.schoenfeld@uni-bayreuth.de

***Abstract*—Online firestorms on Twitter are seemingly arbitrarily occurring outrages towards people, companies, media campaigns and politicians. Moral outrage can create an excessive collective aggressiveness against one single argument, one single word, or one action of a person resulting in hateful speech. With a collective "against the others" the negative dynamics often start. Using data from Twitter, we explored the starting points of several firestorm outbreaks. As a social media platform with hundreds of millions of users interacting in real-time on topics and events all over the world, Twitter serves as a social sensor for online discussions and is known for quick and often emotional disputes. The main question we pose in this article is whether we can detect the outbreak of a firestorm. Given 21 online firestorms on Twitter, the key questions regarding the anomaly detection are: 1) How can we detect changing points? 2) How can we distinguish the features that indicate a moral outrage? In this paper we examine these challenges developing a method to detect the point of change systematically spotting on linguistic cues of tweets. We are able to detect outbreaks of firestorms early and precisely only by applying linguistic cues. The results of our work can help detect negative dynamics and may have the potential for individuals, companies, and governments to mitigate hate in social media networks.**

***Index Terms*—Firestorms, Twitter, Change Detection**

I. Introduction

Twitter is a social media platform with millions of users exchanging ideas about daily topics [1]. Its influence on societal processes is widely discussed. Acting as social sensors for real-time discussions, users provide information about ongoing discussions for live events and media topic strategies. User interactions have provided real-time information about the success and not-success of media campaigns and public relation events. One phenomena are online firestorms [2]–[8]. Negative online dynamics can be very dangerous in real life and can do harm to people. A single statement or media outlet can trigger a collective brawl that seems to escalate uncontrollably until a certain point of exhaustion.

Research questions. Analyzing real-time communication data on Twitter can help to understand the emergence of online moral outrages, negative dynamics and collective action [9]. Hence, the key research questions of our study are, "what are the major features that indicate the outbreak of a firestorm?" and "how can we detect relevant occurrences by exploring firestorm data?"

IEEE/ACM ASONAM 2020, December 7-10, 2020
978-1-7281-1056-1/20/$31.00 © 2020 IEEE

Methods. In order to address our research question of the relationship between lexical features and firestorm participation, we use the extracted characteristics of firestorms to detect an outbreak at an early stage. Our approach provides a method from network analysis and text statistics by examining the dynamics of linguistic cues over time.

On this account, we assume that detecting change based on sentiment analysis plus the usage of pronouns is more significant in how people connect with each other to form an outrage. Combining the automated processes that is done by the LIWCTool [10] and looking for explicit lexical features, could help to answer the above posted questions. Function words are psychologically and linguistically interesting and have been studied broadly [11]. Pronouns refer to a referent, hence, tell to whom somebody is speaking [12]. In this way, we might figure out if actors in social media networks stop talking about themselves and start talking collectively against somebody emotionally and with the words they use.

Contributions. The goal is to detect sentimental and lexical changes as a signal of an underlying change in a social network. In summary, our contributions are:

- Model: We propose a novel change detection model that accounts for linguistic cues and is able to detect the outbreak of a firestorm closely and quickly.
- Algorithm: We are able to detect firestorms on streaming Twitter data by only monitoring a couple of lexical features.

II. Related Work

Online firestorms are similar to rumors to some extent, e.g. they often rely on hearsay and uncertainty, online firestorms pose new challenges due to the speed and potential global reach of social media dynamics [2]. Why do people join online firestorms? Based on the concept of moral panics the authors argue that participation behavior is driven by a moral compass and a desire for social recognition [7]. Social norm theory refers to understanding online aggression in a social-political online setting, challenging the popular assumption that online anonymity is one of the principle factors that promotes aggression [4].

With respect to firestorms on social media, the analysis of dynamics and their early detection often involves research

from the field of sentiment analysis, network analysis as well as change point detection.

Sentiment analysis. Sentiment analysis was applied to analyze the emotional shape of moral discussions in social networks [13]. It has been argued that moral-emotional language increased diffusion more strongly. Highlighting the importance of emotion in the social transmission of moral ideas, the authors demonstrate the utility of social network methods for studying morality. A different approach is to measure emotional contagion in social media and networks by evaluating the emotional valence of content the users are exposed to before posting their own tweets [14]. Modeling collective sentiment on Twitter gave helpful insights about the mathematical approach to sentiment dynamics [15]. Arguing that rational and emotional styles of communication have strong influence on conversational dynamics, sentiments were the basis to measure the frequency of cognitive and emotional language on Facebook [16]. Extracting the patterns of word choice in an online social platform reflecting on pronouns is one way to characterize how a community forms in response to adverse events such as a terrorist attack [17].

Network analysis. Social media dynamics can be described with models and methods of social networks [18]. Approaches mainly evaluating network dynamics are, for example, proposed by Snijders et al. Here, network dynamics were modeled as network panel data [19]. This study demonstrated ways in which network structure reacts to users posting and sharing content. While examining the complete dynamics of the Twitter information network, the authors showed where users post and reshare information while creating and destroying connections. Dynamics of network structure can be characterized by steady rates of change, interrupted by sudden bursts [20]. Dynamics of online firestorms where analyzed applying an agent-based computer simulation (ABS) [21]—information diffusion and opinion adoption are triggered by negative conflict messages. In other works, techniques from social network analysis were combined with those from statistical process control in order to detect when significant change occurs in longitudinal network data [22].

Change point detection. The best known approaches for change point detection include Binary Segmentation [23], [24], Segment Neighborhood [25], and Optimal Partitioning [26], all of which suffer from certain drawbacks when considering monitoring streaming data: Binary Segmentation is quite efficient in terms of computational complexity, i.e. $\mathcal{O}(n \log n)$, but it cannot guarantee to find the global minimum. Segment Neighborhood approaches suffer from computational complexity which might degenerate to $\mathcal{O}(n^3)$. A more recent approach was proposed by Killick et al. and it is based on the Optimal Partitioning in that it yields a guaranteed identification of the exact minimum while retaining a computational complexity that is linear in the number of samples n [27]. Their approach is called the Pruned Exact Linear Time (PELT) method and is based on a work by Jackson et al. [26]. Most importantly, their method has a linear computational complexity which renders it especially useful for applications on streaming data.

Mixed approaches. More recent approaches analyze online firestorms by analyzing both content and structural information. A text-mining study on online firestorms evaluates negative eWOM that demonstrates distinct impacts of high- and low-arousal emotions, structural tie strength, and linguistic style match (between sender and brand community) on firestorm potential [28]. Online Firestorms were studied to develop optimized forms of counteraction, which engage individuals to act as supporters and initiate the spread of positive word-of-mouth, helping to constrain the firestorm as much as possible [5]. By monitoring both linguistic and psychological features of anomaly in the mention networks of online firestorms, we also combine analysis of content with the focus on structural information. To be able to detect online firestorms quickly, we also employ a method of change point detection on time series of the extracted features.

III. Data

We used the same set of 21 firestorms as in [3], whose data source is an archive of the Twitter decahose, a random 10% sample of all tweets. Mention and re-tweet networks based on these samples can be considered as *random edge sampled* networks [29] since sampling and network construction is based on Tweets that constitutes the links in the network. The set of tweets of each firestorm covers the first week of the event including on average 8199.29 tweets from 6641.76 users.

We augmented this dataset by including all decahose tweets from the users that participated in the firestorms from the 7 days before and the 7 days after the starting day of the firestorm, i.e. 15 days overall with the start of the firestorm in the middle. The fraction of firestorm-related tweets is between 2% and 8% of the tweets of each event—it is important to realize at this point that even for users engaging in online firestorms, this activity is a minor part of their overall activity on the platform.

A. Mention and Retweet Networks

To get insight on the evolution of each event, we opt to split time into units of *half hours*. This allows us to perform analysis at fine granularity. The result of this splitting is a series of about 720 time slices (since the studied time-span of an event is 15 days, this period corresponds to 720 half hours). At each time point we construct *mention networks*, and *retweet networks* taking into account all the tweets during the last 12 hours. This way we obtain a moving window of tweets: with a window size of 24 slices at steps of half hours. The *mention network* of each moving window contains an edge ($user_1$, $user_2$) if a tweet (among tweets under consideration) posted by $user_1$ contains a mention to $user_2$. The *retweet network* of each moving window contains an edge ($user_1$, $user_2$) if a tweet (among tweets under consideration) posted by $user_1$ is a retweet of another (original) tweet posted by $user_2$.

IV. Analysis

A. Analysis of Mention Networks

For each event, the mention networks constructed at the different time points are directed, unweighted networks.

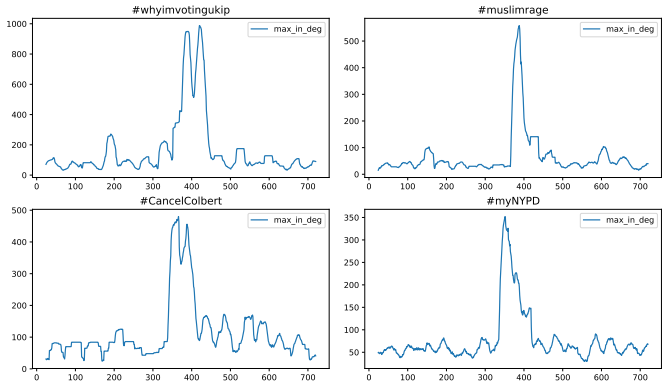


Fig. 1. Maximum in-degree in mention networks.

We performed several types of social network analysis and extracted a set of metrics, including: number of nodes N, and edges E, and density, average out-degree (which equals avg. in-degree), maximum out-degree, and maximum in-degree absolute and relative size of the largest connected component, as well as ratio of mention tweets to all tweets, mention per tweet ratio: E / nr. tweets, mention per 'mention' user ratio: E / N, tweet per 'mention' user ratio: nr. tweets / N.

Each of the aforementioned features leads to a time-series when taken over the entire time-span of the event. We find the maximum in-degree feature is one of the best features to detect this change. Figure 1 shows the time-series of maximum in-degree for the events with the largest number of tweets. The ability of this feature to detect a firestorm can be interpreted by considering that, generally speaking, a firestorm occurs when one user is being mentioned unusually high. However, the change of focus to a particular user can be the result of different (including positive) events.

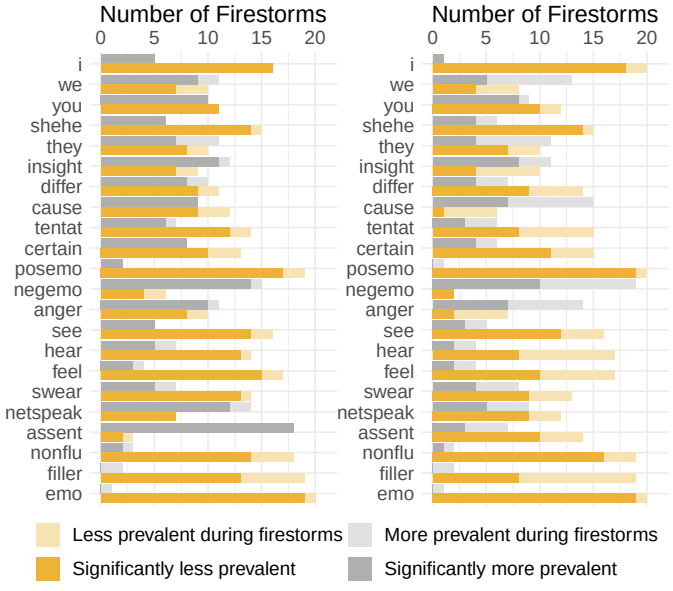
A more rigorous analysis of the change in behavior of such features is necessary in order to devise a formal method/algorithm of change detection as we will see in the next section.

B. Change of language

The first step was to uncover the linguistic peculiarities of firestorms. We classified all tweets using the LIWC classification scheme [30] and compared between firestorm tweets and non-firestorm tweets. The comparisons refer to the following categories: personal pronouns, affective processes, cognitive processes, perceptual processes, and informal language.

These categories each contain several subcategories that can be subsumed under the category names. The category of personal pronouns, for example, contains several subcategories referring to personal pronouns in numerous forms. One of these subcategories 'I', for example, includes—besides the pronoun 'I'—'me', 'mine', 'my', and special netspeak forms such as 'idk' (which means "I don't know").

Netspeak is a written and oral language, an internet-chat, which has developed mainly from the technical circumstances: the keyboard and the screen. The combination of technology



(a) All Tweets (b) Mention Networks

Fig. 2. Comparison of linguistic features between firestorm tweets and non-firestorm tweets

and language makes it possible to write the way you speak [31]. For each individual subcategory, we obtain the mean value of the respective LIWC values for the firestorm tweets and the non-firestorm tweets.

In Figure 2 the comparisons between firestorm tweets and non-firestorm tweets are shown with regard to the individual subcategories. The firestorm-tweets were compared with tweets from the week immediately before the firestorm.

Figure 2a refers to all tweets, while for Figure 2b only tweets from the mention network were considered. In both cases every subcategory was examined separately for all 21 firestorms. The grey bars represent the number of firestorms in which terms from the respective category occurred more frequently during the firestorms. The orange bars visualize the number of firestorms in which the same words occurred less frequently during the firestorms. The sum of the orange and grey bars is therefore always 21.

The light areas of the bars indicate that the mean values were different from each other. The strongly colored areas of the bars indicate that these differences were significant in terms of t-tests with $p < 0.01$. For category 'I' in Figure 2a, this means that in 5 Firestorms people used words of this category significantly more often, while in 16 Firestorms these words were used significantly less. Words of the same category were used less in 20 Firestorms considering the mention networks alone as depicted in Figure 2b. In 19 of these Firestorms, the differences were also significant.

In addition to the category 'I', the categories 'posemo' and 'assent' should also be highlighted. Words representing positive emotions like 'love', 'nice', 'sweet'—the 'posemo' category—are used clearly (and often significantly) less in almost all firestorms: positive emotions were less present in 19

out of 21 firestorms. In 17 out of 19 firestorms the differences were significant. This effect even increases when looking at mention networks. For the third remarkable category ‘assent’, which contains words like ‘agree’, ‘OK’, ‘yes’, this effect is reversed for all tweets—words in this category are used significantly more often during almost all firestorms (18 out of 21). When looking at the mention networks, however, this feature lacks accuracy. The differences are significant only in 13 firestorms.

Finally, we constructed our own category ‘emo’ by calculating the difference between positive and negative sentiments in tweets. Thus, weights of this category can be negative and should describe the overall sentiment of a tweet. There are 19 firestorms in which the ’emo’ values were significantly lower during a firestorm. At the same time, there was only one firestorm with higher values of ‘emo’ but these differences were not significant. Checking if the differences remain visible decomposing the mention networks into components comparing tweets inside the largest component to tweets outside that component, we see no effect. There are only a few firestorms, in which the use of ‘we’ is significantly larger inside the largest component of the mention network.

C. Change point detection

The goal is to identify a firestorm at an early stage with the help of linguistic features. For the detection of change points, we use an efficient method that is suitable for being applied to streaming data and that was proposed by Killick et al. [27].

We constructed individual time series of the linguistic features. For this purpose, we first split the timeline of each of the firestorm data sets into buckets of half hours and assign tweets to buckets based on their timestamp. When constructing a time series of linguistic characteristics, we describe a bucket of tweets by the mean value of the corresponding LIWC values. To detect the change points in streaming data, we simulate the arrival of new tweets every half hour. Being able to decide at any time t whether a change has occurred, we use historical data from the past 24 hours. The time series applied for change point detection thus consists of 49 values for the interval $[t-48:t]$ each representing the mean of the LIWC values of the respective tweets. By doing so, we create separate time series for each of the subcategories mentioned above—see Figure 3. We do not apply any smoothing to the time series.

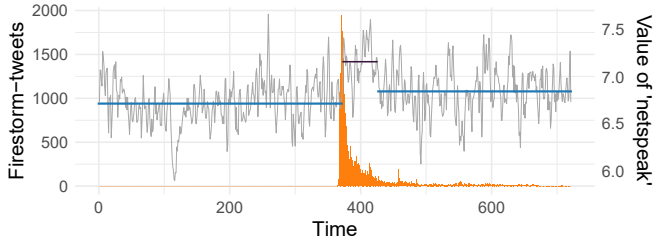


Fig. 3. Example change point detection conducted on a linguistic category. The number of Firestorm tweets is depicted in the background.

Decisive for successful change point detection is the choice of the penalty parameter. We select this parameter using elbow criterion. Therefore, we iterate penalty parameters from 2 to 10 and obtain the number of identified change points. From this data we determine the optimal penalty parameter as the configuration with the maximum absolute second derivative using the approximation of the second derivative of a point x_i as $x_{i+1}+x_{i-1}-2x_i$. Choosing a higher penalty value results in fewer change points detected and vice versa.

We were able to detect the start of the firestorms in -0.55 ∓ 2.34 hours. We have defined the start time as the first interval of half an hour at which the hashtag or @user mention of the firestorm was the most frequent hashtag or @user mention in the data set. Due to the focused data collection process, the set contains little hashtags or @user mentions that were used frequently. Hence, this definition of a starting point of firestorms is quite sensitive, i.e. a low number of tweets suffices to boost the relevant hashtag or @user mention.

Also, it takes two intervals of half an hour until the beginning of the firestorm is noticed, i.e. the minimum deviation from the start time was measured. Hence, we are able to detect a change in the linguistic behaviour of the users quickly.

In a next step, we explored change near the peak of a firestorm. Determining this peak was done in two ways. We were able to approximate the peak of network dynamics with an average of $+1.19 \mp 2.51$ hours meaning that the change point closest to the peak is on average shortly after this peak. The second, natural definition of the peak is the interval of half an hour in which most firestorm-related tweets were recorded. We were able to approximate this peak of tweet accumulations with an average of $+0.14 \mp 1.30$ hours.

With respect to the identified differences in language use that were discussed in relation to Figure 2, we further evaluated how many change points were identified on the timelines of the linguistic categories. Figure 4 depicts how often a change point could be detected from the timeline of an individual characteristic. The top three categories were ‘netspeak’, ‘I’, and ‘posemo’ which corresponds to the insights from Figure 2—these were the categories with the most significant differences just after our own category ‘emo’.

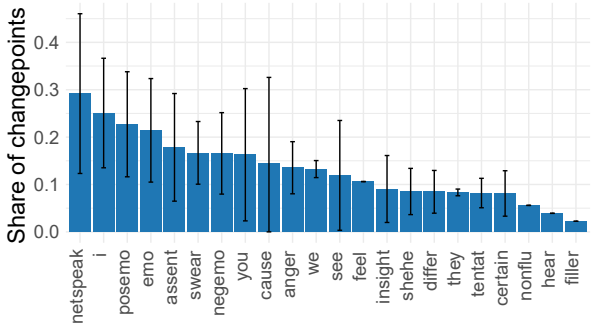


Fig. 4. How often were predictors relevant for the detection of a relevant changepoint

V. Discussion & Conclusion

From a network perspective, a firestorm occurs when one user is being mentioned unusually high—focusing on a Twitter handle or a hashtag. The maximum in-degree in mention networks is significantly deviating from comparable time periods. By evaluating lexical cues from the Tweet comments, we evaluated collective behavior manifesting in individual choices of words.

During firestorms, users talk significantly less about themselves compared to non-firestorm periods. Simultaneously, the positivity in firestorms tweets vanishes and negativity rises. The extracted lexical features were applicable to streaming data. Using lexical features to monitor change in behavior has the advantage of constant memory requirements.

By applying a straightforward change point detection, we were able to detect the starting point of the firestorms closely and quickly. We further provide insight into which linguistic categories proved to be useful for monitoring change.

According to our posed questions, combining sentiment analysis and text statistics to explore firestorm data can reveal how people connect with each other to form an outrage. The usage of vocabulary changes at a certain point when every single user stops commenting with the I-perspective and starts commenting on others. As mentioned, pronouns refer to a referent. If the 'I' diminishes, the focus changes significantly. All of a sudden people stop talking collectively about themselves positively and collectively more negatively—against the others!

Our model picks up these features and is able to detect the starting point of outrages giving insights into collective changing behavior. Further research questions regarding spreading of rumours and moral outrages might be: What causes evolving collective emotionality? Why does a community or society may at times come together and simultaneously communicate the same thought and participate in the same action? A better knowledge of individual motivations and collective action can help to better understand and detect online firestorms.

VI. Acknowledgements

The author(s) gratefully acknowledge the financial support from the Technical University of Munich - Institute for Ethics in Artificial Intelligence (IEAI). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the IEAI or its partners.

References

[1] D. Boyd, S. Golder, and G. Lotan, "Tweet, tweet, retweet: Conversational aspects of retweeting on twitter," in *IEEE/HICSS*, 2010, pp. 1–10.

[2] J. Pfeffer, T. Zorbach, and K. M. Carley, "Understanding online firestorms: Negative word-of-mouth dynamics in social media networks," *Journal of Marketing Communications*, vol. 20/1–2, pp. 117–128, 2014.

[3] H. Lamba, M. M. Malik, and J. Pfeffer, "A Tempest in a Teacup? Analyzing Firestorms on Twitter," in *2015 IEEE/ACM ASONAM*, New York, NY, USA, 2015, p. 17–24.

[4] K. Rost, L. Stahel, and B. S. Frey, "Digital Social Norm Enforcement: Online Firestorms in Social Media," *PLOS ONE*, vol. 11, no. 6, p. e0155923, 2016.

[5] A. Mochalova and A. Nanopoulos, "Restricting the spread of firestorms in social networks," *ECIS 2014 Proceedings*, 2014.

[6] B. Drasch, J. Huber, S. Panz, and F. Probst, "Detecting Online Firestorms in Social Media," *ICIS*, 2015.

[7] M. Johnen, M. Jungblut, and M. Ziegele, "The digital outcry: What incites participation behavior in an online firestorm?" *New Media & Society*, vol. 20, no. 9, pp. 3140–3160, 2018.

[8] L. Stich, G. Golla, and A. Nanopoulos, "Modelling the spread of negative word-of-mouth in online social networks," *Journal of Decision Systems*, vol. 23, no. 2, pp. 203–221, 2014.

[9] M. J. Crockett, "Moral outrage in the digital age," *Nature Human Behaviour*, vol. 1, pp. 769–771, 2017.

[10] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of Language and Social Psychology*, 2009.

[11] P. Dekker, "Pronouns in a pragmatic semantics," *Journal of Pragmatics*, vol. 34, no. 7, pp. 815–827, 2002.

[12] J. W. Pennebaker, *The secret life of pronouns: What our words say about us*, ser. The secret life of pronouns: What our words say about us. Bloomsbury Press/Bloomsbury Publishing, 2011, pages: xii, 352.

[13] W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, and J. J. V. Bavel, "Emotion shapes the diffusion of moralized content in social networks," *PNAS*, vol. 114, no. 28, pp. 7313–7318, 2017.

[14] E. Ferrara and Z. Yang, "Measuring emotional contagion in social media," *PLOS ONE*, vol. 10, no. 11, 2015.

[15] N. Charlton, C. Singleton, and D. V. Greetham, "In the mood: the dynamics of collective sentiments on Twitter," *Royal Society Open Science*, vol. 3, no. 6, 2016.

[16] C. A. Bail, T. W. Brown, and M. Mann, "Channeling Hearts and Minds: Advocacy Organizations, Cognitive-Emotional Currents, and Public Conversation," *American Sociological Review*, vol. 82, no. 6, pp. 1188–1213, 2017.

[17] S. Shaikh, L. B. Feldman, E. Barach, and Y. Marzouki, "Tweet Sentiment Analysis with Pronoun Choice Reveals Online Community Dynamics in Response to Crisis Events," in *Advances in Cross-Cultural Decision Making*. Springer International Publishing, 2017, pp. 345–356.

[18] M. Hennig, U. Brandes, J. Pfeffer, and I. Mergel, *Studying Social Networks. A Guide to Empirical Research*. Campus Verlag, 2012.

[19] T. A. Snijders, J. Koskinen, and M. Schweinberger, "Maximum likelihood estimation for social network dynamics," *The annals of applied statistics*, vol. 4, no. 2, pp. 567–588, 2010.

[20] S. A. Myers and J. Leskovec, "The bursty dynamics of the Twitter information network," in *WWW*, Seoul, Korea, 2014, pp. 913–924.

[21] F. Hauser, J. Hautz, K. Hutter, and J. Füller, "Firestorms: Modeling conflict diffusion and management strategies in online communities," *Journal of Strategic Information Systems*, vol. 26/4, pp. 285–321, 2017.

[22] I. McCulloh and K. M. Carley, "Detecting change in longitudinal social networks," *Journal of Social Structure*, vol. 12/3, pp. 1–37, 2011.

[23] A. J. Scott and M. Knott, "A cluster analysis method for grouping means in the analysis of variance," *Biometrics*, vol. 30, pp. 507–512, 1974.

[24] A. Sen and M. S. Srivastava, "On tests for detecting change in mean," *The Annals of Statistics*, vol. 3, no. 1, pp. 98–108, 1975.

[25] I. E. Auger and C. E. Lawrence, "Algorithms for the optimal identification of segment neighborhoods," *Bulletin of Mathematical Biology*, vol. 51, no. 1, pp. 39–54, Jan 1989.

[26] B. Jackson, J. D. Scargle, D. Barnes, S. Arabhi, A. Alt, P. Gioumousis, E. Gwin, P. Sangtrakulcharoen, L. Tan, and Tun Tao Tsai, "An algorithm for optimal partitioning of data on an interval," *IEEE Signal Processing Letters*, vol. 12, no. 2, pp. 105–108, 2005.

[27] R. Killick, P. Fearnhead, and I. A. Eckley, "Optimal detection of changepoints with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[28] D. Herhausen, S. Ludwig, D. Grewal, J. Wulf, and M. Schoegel, "Detecting, Preventing, and Mitigating Online Firestorms in Brand Communities," *Journal of Marketing*, vol. 83, no. 3, pp. 1–21, 2019.

[29] C. Wagner, P. Singer, F. Karimi, J. Pfeffer, and M. Strohmaier, "Sampling from social networks with attributes," in *Proceedings of the WWW Conference*, 2017, pp. 1181–1190.

[30] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The Development and Psychometric Properties of LIWC2015," The University of Texas at Austin, Tech. Rep., 2015.

[31] D. Crystal, "Language and the internet," *IEEE Transactions on Professional Communication*, pp. 142–144, 2002.