Information Network Mining: a Case For Emergency Scenarios

Anh Le
University of Pittsburgh
atl13@pitt.edu

Yu-Ru Lin
University of Pittsburgh
yurulin@pitt.edu

Konstantinos Pelechrinis
University of Pittsburgh
kpele@pitt.edu

ABSTRACT

The proliferation of available datasets from digital social media and networks has lead to a surge of analytical studies that aim into understanding topics such as the way people create ties, or the way that they influence each other. Despite this large body of literature, little - if any - attention has been given in the network dynamics of the *context* associated with the user-generated content (e.g., tweets). In particular, the context of a tweet can be captured from the hashtags used by the user who generated it. Existing studies have focused on various characteristics of hashtags individually. In this paper, however, we examine to what extent the coupling of hashtags help capture additional contextual information. We define and thoroughly study the properties of a network structure between hashtags, which we call the *co-tweet hashtag network*. Vertices in this network are hashtags and an edge between hashtags exists if they have appeared in the same tweet. After analyzing the properties of this structure, we turn our attention on its possible applications. In particular, we examine the hypothesis that accounting for the context of the generated content can further improve applications such as event detection, which have primarily exploited the volume of the content. Our preliminary results in this work further support this hypothesis.

1. INTRODUCTION

Online social media have provided us with an unprecedented amount of information on various aspects of human life and behavior. People create "social" ties and generate content. While doing so they leave behind digital traces that can be used to further understand subtle topics such as friendship formation [1], peer influence and social selection [7, 2, 26] etc. These "digital breadcrumbs", as they are commonly referred to as, have also found novel applications in stock market prediction [4], natural disaster detection [10, 20] and flu epidemic tracking [3] to name just a few.

While there is a significant body of work that utilizes the immediate network structures defined by the connections of the users, or the actual generated content itself, there is very little attention given to the context of the latter. With the term context we refer to the broader setting in which the content was produced. For example, consider the tweet "Such a bad day...". While it is clear that the user who tweeted this was upset for some reason, we do not know the exact setting. It could be a bad day at the office, a bad day because his favorite team lost and so on. Automatic extraction of the context is still an open problem in natural language processing. However, online social media platforms have introduced features that allow users to give a brief description of the context of their content. The most prevalent one, and the one that we will use in our study, is that of *hashtags*. Hashtags are short keywords that act as metadata for the content generated. Hence, they essentially provide information for the context of the content. For instance, the above tweet using the hashtag `#rainyday`, could be "Such a bad `#rainyday`...", which would make clear the broader setting of the tweet.

Studies that analyze the properties of hashtags have appeared in the literature but they are mainly focused on properties of individual hashtags as we further elaborate on in Section 5. Some very recent work [5, 12] examines the co-occurrence of hashtags in tweets. In our work, we are interested in studying the network dynamics underlying these co-occurrences and examining potential applications. In particular, users are not limited to one hashtag per "generated content"[1] but can have multiple tags. This implicitly defines a network structure of the context of the information present in the users' content/tweets. In brief, we quantitatively analyze the spatio-temporal changes in the co-tweet hashtag network, where hashtags are the vertices and an edge between two hashtags exists if the two hashtags have appeared in the same tweet.

Using a large corpus of geo-tagged tweets we analyze the properties of this structure. While it exhibits trends observed in other network structures, such as right-skewed degree distribution, we identify some differences. In particular, while a giant component exists, this covers much smaller fraction of the vertices (≈ 0.50) as compared to other networks in the literature (e.g., ≈ 0.9). Moreover, it does not exhibit significant transitivity.

Furthermore, we are interested in applications of this network structure. Our research hypothesis is that information context and its dynamics can provide additional evidence that can be exploited for a variety of applications such as the extraction of emergency events from social media data. To examine the hypothesis, in this preliminary work, we focus on (abnormal) event detection and we examine the benefits we can obtain when monitoring structural changes of the co-tweet hashtag network over time on top of the activity volume. We would like to emphasize here that event detection is not the focus of our study, but we are using it to study the above hypothesis.

The main contributions of our work can be summarized in the following: **(a)** quantitative, in-depth analysis of the dynamics of the network structure defined between co-tweeted hashtags and, **(b)** provision of concrete evidence for the applicability of the co-tweet hashtag network in one of possible applications.

Roadmap: In Section 2 we describe the dataset we used for our study and we provide definitions and notations. Section 3 presents the analysis of the co-tweet hashtag network, while Section 4 presents some preliminary findings when we explore the applicability of this network structure to event detection. Finally, Section 5 discusses related to our study literature, while Section 6 forms our conclusions.

KDD-LESI 2014, August 24, 2014, New York City, USA.

[1]While the content can refer to image, video, or text, for the rest of the paper we will focus on text and in particular on tweets.

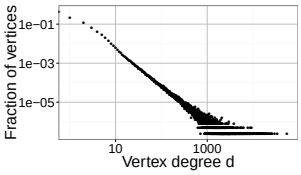

Figure 1: The degree distribution exhibits the typical heavy tail.

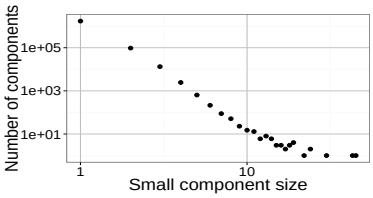

Figure 2: Many *very* small components and a few *larger*.

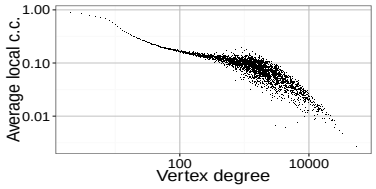

Figure 3: The local clustering coefficient drops (on average) with an increase in the vertex degree.

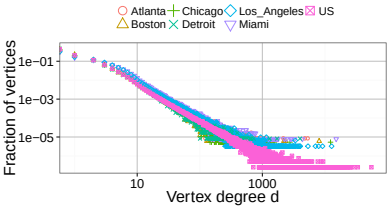

Figure 4: The degree distribution is heavy-tailed for individual cities as well.

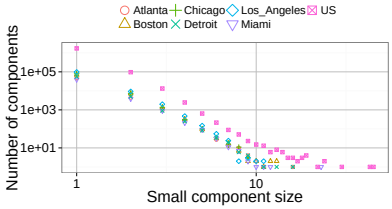

Figure 5: The small component size distribution is similar across cities.

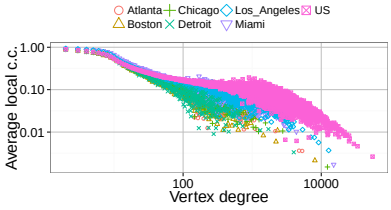

Figure 6: Local clustering coefficient drops faster in individual cities compared to aggregate.

City	Users	Tweets	Unique hashtags
Atlanta	34,783	278,320	117,963
Boston	35,766	398,118	164,356
Chicago	47,616	476,132	187,680
Detroit	23,194	301,325	121,657
Los Angeles	80,678	718,407	306,685
Miami	34,027	273,789	127,715

Table 1: Number of users, tweets and unique hashtags observed in three months in each city.

2. PRELIMINARIES

Dataset and Analysis Set-Up: Our dataset of $20,472,557$ geotagged tweets was generated from $1,311,181$ users and collected in the the period from 3/1/2013 to 5/31/2013. All the tweets in our dataset have been generated within the United States and contain at least one hashtag. We extract the hashtags of each tweet and our final dataset has the following tuple-format `<tweetId, userId, time, longitude, latitude, hashtags>`. The total number of unique hashtags in our dataset is $3,953,890$.

We further obtain smaller datasets, based on the city that each tweet was generated, that we will use to study the spatial dynamics of the co-tweet hashtag network (to be formally introduced later in this section). In particular, we focus on 6 big cities from different parts of the US, namely Los Angeles, Detroit, Miami, Atlanta, Chicago and Boston. Consequently for each of the above datasets we obtain all tweets with coordinates falling within 50 km from the corresponding city centers. Table 1 summarizes the data in each city.

Definitions and Notations: We will now formally define the co-tweet hashtag network $\mathcal{G} = (V, E)$. The set of vertices V is the set of observed hashtags, that is, $v \in V$ is a hashtag observed in at least one tweet. An undirected edge e_{ij} between two hashtags, $v_i, v_j \in V$ exists iff v_i and v_j have appeared in common to at least one tweet. We further define a weighted version of the above network, by annotating each edge $e_{ij} \in E$ with the number of tweets in which v_i and v_j appear together.

3. HASHTAG NETWORK ANALYSIS

We initially perform a static analysis, considering all tweets generated during the three months period of our dataset. Then we focus on specific cities and we further analyze the corresponding network structures, essentially capturing their spatial properties.

Static analysis: When considering all the tweets in aggregate, both in space and time, the corresponding network consists of $3,953,890$ vertices and $12,858,285$ edges. The empirical PDF of the degree distribution p_k is shown in Figure 1 in log-log scale. As we can see, the distribution is heavily right-skewed and can be approximated by a power law with exponent $\alpha = 1.9$. Furthermore, there is one large connected component containing $1,990,651$ vertices. The fraction S of vertices in this component is approximately 0.5. Note that this is significantly less than the typical cases of social, information, technological and biological networks (Table 8.1 [17]); information context appears to be less "connected". The small component sizes range from 1 (singleton vertices) to 45. The number of singleton vertices is $1,714,479$, that is, $p_0 = 0.43$. The large number of singleton vertices can be explained from the fact that many hashtags are used/created by only a few users and they do not become popular (e.g., `#SheWalksTooFast`) or they are never co-tweeted with any other hashtag, at least within our dataset (e.g., `#votemeup`). This also leads to lower value of S as mentioned above. The distribution of the small component sizes is shown in Figure 2.

Next we examine the transitivity of the network. In particular we calculate the global clustering coefficient of the network c_g, which is 0.03. However, this number alone is not informative. We need to compare it with the expected clustering coefficient of an appropriate random graph model. When we control for the degree distribution (i.e., the configuration model), the expected clustering coefficient of the random network with n vertices is given by [17]:

$$\frac{1}{n}\frac{[< k^2 > - < k >]^2}{< k >^3} \quad (1)$$

where $< k^m >$ is the m-th moment of the degree distribution of the graph. In our case the above expression gives 0.15, which is higher than c_g. This means the co-tweet hashtag network is less transitive compared to a random network with the same degree distribution! The possible explanation is that hashtags can be co-tweeted only if they can form a linked context, so the two facts that hashtag A was co-tweeted with hashtag B and hashtag A was also co-tweeted with hashtag C may not affect the probability of hashtags B and C being co-tweeted. For example, the two pairs `#love #arsenal` and `#love #seafood` are popular, but `#arsenal` and `#seafood` are very rarely co-tweeted.

We further compute the local clustering coefficient of every vertex in the network. Figure 3 shows the relationship between the average local clustering coefficient and vertex degree. This relationship follows the typical decreasing trend, that is, high degree vertices have on average lower clustering coefficient.

Given the length limit of tweets, shorter hashtags tend to be more popular. This was confirmed in [8]. We further want to test whether there is a correlation between the length of co-tweeted hashtags and the number of hashtags being co-tweeted with. We annotate every vertex/hashtag with an attribute that captures the length l (in alphanumeric characters) of the hashtag. We then calculate the correlation coefficient $r_{l,k}$ between the length l and the degree k of the hashtags. The obtained correlation of -0.04, ($p\text{-}value < 10^{-15}$), even though significant, does not indicate any strong correlation.

We also examine the assortativity mixing in the network with respect to the hashtag degree (r_k) and length (r_l). Degree assortativity and hashtag length assortativity are -0.09 and 0.10 respectively.

Hashtag networks in different cities: Since the co-tweeted hashtag network captures the inter-related contexts behind the content generated by users, one question can be asked is that will the hashtag network generated by users in different cities exhibit different characteristics? In other words, will this kind of network structure reflect any city-specific trend of user generated content, or it will remain stable across cities? In effort to answer that question, we further analyze the co-tweet hashtag networks that we obtain when we focus on specific cities by computing the same statistics as above. Table 2 summarizes the results, which indicate that despite the differences in the absolute values of the statistics, the networks obtained in different cities exhibit similar characteristics. In particular, they all appear to (i) be sparse (small edge density ρ), (ii) exhibit degree distribution with heavy tail, (iii) exhibit little or no transitivity and (iv) not have significant mixing with respect to the hashtag degree or length. More specific the degree distributions of the city-level networks are presented in Figure 4. Comparing them to the degree distribution of the whole US network (shown in pink color), the maximum degrees in city networks are orders of magnitude smaller as one might have expected. The giant component in these networks is also at the same level as compared to the giant component of the aggregate network. Figure 5 further shows that the distributions of small component sizes for the city-level networks are similar to each other. Finally, Figure 6 depicts the average local clustering coefficient for vertices with the same degree for all the networks. Despite the discrepancies observed between the city networks and the aggregate one, we still observe the typical decreasing trend.

In summary, we observed that the overall characteristics of the hashtag networks as captured by vertex degree distribution, connected component sizes and local clustering coefficients are stable across cities and are also aligned with the characteristics of the country-level hashtag network.

4. EVENT DETECTION WITH CO-TWEET HASHTAG NETWORK

In this section we focus on our hypothesis that the network dynamics of the information context can provide additional features for the design of a variety of applications. To support this hypothesis we focus on the applicability of the hashtag network on event detection.

4.1 Hashtag network during events

In Section 3 we analyzed the network by considering the aggregate time. However, for an application such as event detection the temporal evolution of the network is of crucial importance. We focus our preliminary work on the Boston marathon bombing event [24] that is part of our dataset. We begin by examining the number of hashtags that are included in tweets that were generated between April 15 and April 17, 2013. Figure 7 presents the results, where we have further used a sliding window of 4 hours and focused on tweets in Boston. The time-series for tweets with more than 2 hashtags exhibit three peaks at 4pm (GMT time) every day. After inspecting the corresponding tweets, we found that almost all of them (especially those with more than three hashtags) are advertisements. For example, *"#Banking #Job alert: Asst Branch Manager | Citizens | #Providence , RI http://t.co/xB5kjozuSC #cfgjobs!"*. Interestingly, most of the tweets with two hashtags are not related to advertisement. Instead, people were tweeting about the bombing event (e.g. *"This country, this WORLD, is one constant tragedy after another. #prayforboston #prayforlife"*) Especially for the first hour after the bombing above 80% of the tweets containing 2 hashtags were related to the event. The fact that tweets generated by spammers tend to contain more hashtags than tweets from normal users was confirmed by McCord *et al.* [16]. However we will try to further delve into the details as part of our future investigations to fully understand the reasons behind our observation.

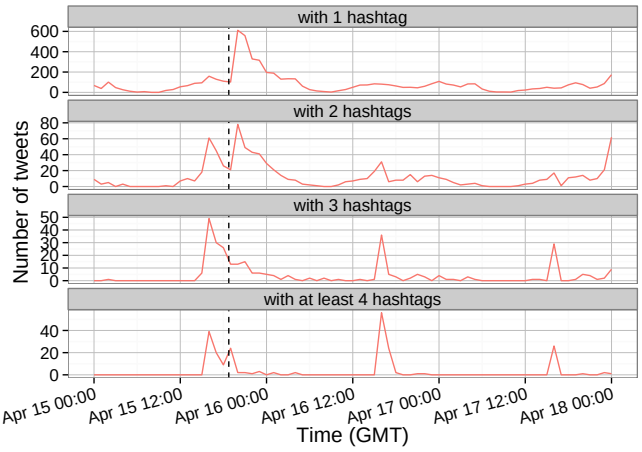

Figure 7: The number of tweets with two hashtags increases in Boston after bombing (marked by vertical dashed line). Tweets with three or more hashtags are often advertisements and appear with temporal periodicity.

Hashtag co-existence is the building block of the co-tweet hashtag network. Hence, based on the above result one might expect changes in the network structure during the event. Figure 8 visualizes the network in Boston at different times; three hours before event, one hour after event and three hours after event. To make it more visible, only vertices corresponding to the 15 most tweeted hashtags (based on the number of users) are labeled. We can observe that before the event happened, the hashtag network had only a few edges. However, after the event the hashtag network quickly became denser. Furthermore, a few hashtags became *central hubs* connecting with many other hashtags. We would like to emphasize here that central hubs appear also with daily periodicity (i.e., without the presence of an event). However, as we can see these periodic patterns are not as pronounced as when there is an event. Furthermore, and most importantly, we can detect these periodic hubs and avoid misclassifying them as events (see Section 4.2).

To quantify the changes in the hashtag network under events, we study a variety of network metrics, such as edge density, giant component fraction, maximum/average degree etc. We find that the Gini coefficient of the degree distribution (denoted as g_k) exhibits the best reflection of events. Visually during events, the hashtag network roughly consists of "hub-and-spoke" sub topologies; there are common hashtags describing the event (e.g., #bostonmarathon) and "satellites" (e.g., #explosion), which serve as sub-topics. This leads

City	$\|V\|$	$\|E\|$	ρ	α	p_0	S	c_g	random c_g	$r_{l,k}$	r_k	r_l
Atlanta	117,963	250,241	3.60×10^{-5}	2.23	0.39	0.49	0.04	0.07	-0.10	-0.04	0.14
Boston	164,356	286,166	2.12×10^{-5}	2.19	0.42	0.45	0.03	0.09	-0.09	-0.03	0.14
Chicago	187,680	372,077	2.11×10^{-5}	2.14	0.42	0.46	0.02	0.14	-0.08	-0.03	0.13
Detroit	121,657	187,831	2.54×10^{-5}	2.23	0.46	0.42	0.05	0.04	-0.12	-0.03	0.17
Los Angeles	306,685	1,108,213	2.36×10^{-5}	2.01	0.32	0.59	0.04	0.12	-0.09	-0.07	0.11
Miami	127,715	465,643	5.71×10^{-5}	2.09	0.31	0.60	0.03	0.26	-0.08	-0.05	0.13

Table 2: Statistics from hashtag networks in different cities.

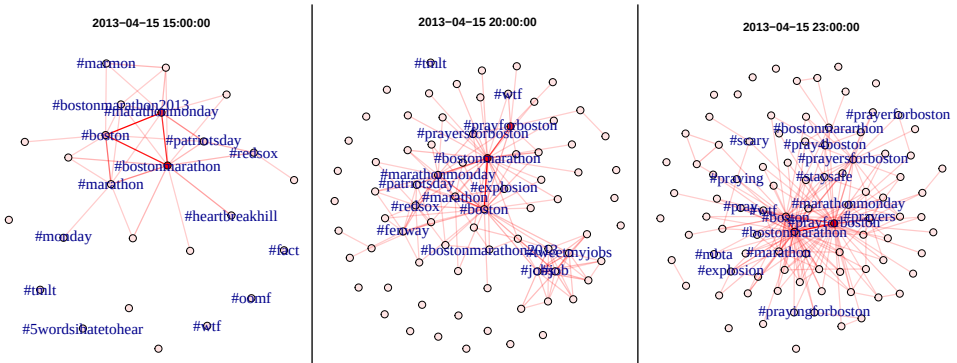

Figure 8: Temporal evolution of the hashtag network in Boston during event day. The bombing happened at 18:49 GMT on 4/15/13. After the event a few hashtags become *hubs* connecting to many other hashtags.

to an imbalance of the vertex degree distribution, which can be quantified by the Gini coefficient [25]. The latter is a measure of the statistical dispersion of a distribution. It ranges from 0 to 1, where 0 represents a *fair* distribution and 1 represents a distribution of maximal inequality (e.g., a highly skewed one). In our case, g_k represents how the degree distribution among non-isolated vertices is imbalanced.

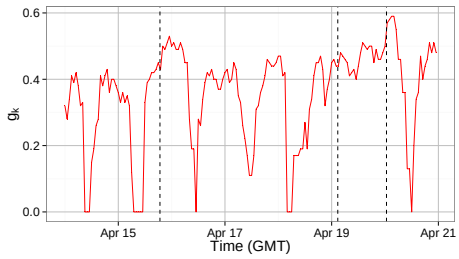

Figure 9: After an event, g_k appears to quickly and noticeably increase deviating from its daily periodicity. The dashed lines mark the Boston bombing event and the two subsequent events: shooting at MIT and manhunt in Watertown.

Based on our previous results we expect during and after the event to observe a shift of g_k to higher values, since there are central hubs that emerge in the network and concentrate the majority of edges. Figure 9 depicts g_k over time for Boston. The vertical dashed lines mark three important moments related with the event: when the first bomb exploded, when the two suspects shot a policeman at MIT and when the firefight and tracing began. As one can observe the Boston hashtag network changed quickly and g_k followed quickly and increased. To reiterate, we can also observe the daily periodicity of g_k, which essentially follows the sleep-wake patterns of the population.

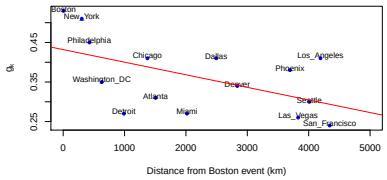

Figure 10: There is a significant trend between g_k and the distance from the event's location.

We further want to delve into the details of the the spatial dependency of g_k. Figure 10 depicts g_k three hours after the event for different cities as a function of distance from the location of the event. (Note that we expand to even more cities than in Table 1.) As expected, the effect of the event on the co-tweet hashtag network fades out as we move farther away from event center. Furthermore, this trend is both strong (correlation coefficient is -0.54) and significant ($p\text{-}value = 0.03$).

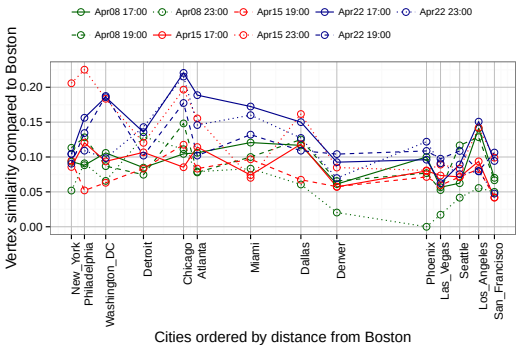

Figure 11: Jaccard similarity between vertices in other cities' co-tweet hashtag networks and Boston's network. There is an increase shortly after the event and in cities near the event location.

The above observations show that the observed effect of the event on the structure of hashtag network (as captured by g_k) changes over time and space. We then ask how the set of contexts captured by the hashtag network changes over time and space. So we examine the vertex similarity, $\sigma(\mathcal{G}_{Boston}, \mathcal{G}_c)$, between the vertices in the network in Boston and in another city c using the Jaccard similarity measure (i.e., fraction of common vertices over total vertices). Figure 11 depicts the results for different cities c at different times (5pm, 7pm, 11pm GMT on the day of event and one week before/after event). We can see that the line corresponding to 11pm on April 15 is higher than other red lines (for the same day but different hours), which shows that right after the event, users in other cities also adopt the same hashtags as users in Boston. That dotted red line also exhibits a decreasing trend as we move further from Boston, in alignment with

the results in Figure 10. Note that the solid blue has a peak at Chicago which corresponds to another event - Earth day.

4.2 Event detection

Based on the above analysis we have built a preliminary event detection algorithm that operates on the live tweet stream using a sliding window of width W and step Δ. We currently use $W = 4hours$ and $\Delta = 1hours$.

Building a baseline: The main idea in our algorithm is to track and detect abnormality in the value of g_k. The first step of our preliminary algorithm aims into building a baseline θ_{g_k} to compare with our current value of g_k. It is computed as the maximum value of g_k observed during the last two weeks (if we have detected an event during that period we ignore the corresponding data).

We further identify *regular* hashtags, that is, hashtags which appear periodically in the network and hence, are not associated with specific events. To do that we use the tweets in the last two weeks and compute the auto-correlation of the occurrence time-series of every hashtag h. The time-bin for this time-series is 5 minutes (i.e., every time-series reading provides us with the number of occurrences of h during the last 5 minutes). If we identify significant daily or weekly periodicity we label h as a regular hashtag. Examples of identified regular hashtags are #np, #ff, #oomf, etc.

Real-time analysis: Given the set of regular hashtags $\mathcal{R}$, at each time t we analyze the tweet stream in $[t - W, t]$. If $\mathcal{H}_t$ is the set of hashtags observed during this period, we build the co-tweet hashtag network $\mathcal{G}$, using hashtags in $\mathcal{H}_t$. We calculate the Gini coefficient of the degree distribution of the observed network, g_{k_t} and we compare it with θ_{g_k}. If $g_{k_t} > \theta_{g_k}$, an event is detected.

Every time an event is detected, our algorithm further reports the top-3 hashtags which do not belong to $\mathcal{R}$. Those hashtags should capture the nature of the event.

Table 3 shows the results for April in Boston. Note that among the identified events is also the NCAA men's basketball final-4 tournament. This event did not take place in Boston (tournament was held in Atlanta, GA and the champion was from Louisville, KY), neither any of the teams participating was from a Boston-based university, but due to the high popularity of the game the event was still detectable in Boston. Also, the last two lines from Table 3 show the two hashtags #eb2013 and #earthday which are related to the Experimental Biology Conference and Earthday events respectively. These could have been big events of the day but given the huge impact of the bombing event, they now became inferior.

While our prototype is by no means a full-fledged event detection system, it showcases the applicability of co-tweet hashtag network dynamics to such problems. In the future we plan to further explore and design a full-fledged event detection system based on the network dynamics of information context.

5. RELATED WORK

Twitter hashtag analysis: The creation and use of hashtags have been studied in [8] with a linguistic approach. The authors show the similarity between forming a new hashtag and forming a new term in normal languages. For example, one of the findings is that popular hashtags are often shorter. Also looking at the formation of hashtags, Tsur *et al.* [21] found a relationship between the appearance (e.g., capitalization) and the status (e.g., number of the tweet author's followers) of a hashtag with its popularity. Recently, Kamath *et al.* [14, 13] studied the spatio-temporal dynamics of hashtags and used them to predict when and where a hashtag will be popular. Furthermore, the diffusion mechanisms of hashtags is studied in [19] where it is found that repeated exposure increase a tag's adoption. Hashtags were also exploited in predicting user's location [6, 11], tweet's topics [15] and tweet's sentiment [9]. Note here that our study is complementary to the above literature. The latter examines individual hashtags while we are focused on studying and exploiting their network dynamics as obtained through their co-appearances in tweets. Some recent work also examined the co-occurrence of hashtags. Carley *et al.* [5] incorporated the hashtag-to-hashtag network into their set of tools to help analysts follow the changes in tweet stream and extract important topic being discussed. Jussila *et al.* [12] used the network of hashtags in analyzing tweets related to organizing of a conference. By visualizing the network, the author was able to detect the inconsistent use of hashtags to refer to the same paper presentation. Our work is quite different in that we do not rely on human visual examination but use analytical tool to automatically process the hashtag network. This allows us to analyze not only a static snapshot but also keep track of the spatio-temporal dynamics of that network.

Event detection using content of tweets: Studies in this line of research can be divided into two types. The first one focuses on specific classes of events (e.g., sports), while the other deals with all events. For example, using a dataset of tweets about soccer games, Nichols *et al.* [18] provide a scheme to extract the events of the game and the later's summary. Similarly, by monitoring the number of tweets containing words related to earthquake, Earle *et al.* [10] built a simple system to detect earthquakes. On the same problem, Sakaki *et al.* [20] use supervised learning and probabilistic models to classify tweets as earthquake related or not and detect the time and the location of an earthquake. In a similar way, Aramaki *et al.* [3] also use supervised learning and NLP to detect influenza outbreak events. The second type of work is more general, without any assumption on event class.For example, Weng *et al.* [23] do not focus on specific types of events and apply wavelet transform to detect peaks of words in a tweet stream. Once such words are identified they use spectral clustering to identify related words that can represent events. Taking a different approach, Vallkanas [22] first extract the mood from the tweets, and then detect peaks in one of moods as indicators for events. To reiterate, our work is not particularly focused on event detection. We use the latter as an example application to study the applicability and importance of the co-tweet hashtag network.

Time (GMT)	Top hashtags	Event
April 09 5am	#louisville #michigan #goblue	NCAA tournament
April 09 7am	#louisville #marchmadness #ncaachamp	NCAA tournament
April 15 4pm	#bostonmarathon #marathonmonday #boston	Boston marathon
April 15 8pm	#bostonmarathon #prayforboston #marathonmonday	Boston marathon
April 15 9pm	#prayforboston #bostonmarathon #boston	Boston bombing
April 19 4pm	#bostonstrong #boston #prayforboston	Boston bombing
April 19 5pm	#bostonstrong #boston #watertown	Boston bombing
April 19 9pm	#boston #bostonstrong #lockdown	Boston bombing
April 20 2pm	#bostonstrong #boston #eb2013	Boston bombing
April 22 4pm	#bostonstrong #boston #earthday	Boston bombing

Table 3: Example of detected event-related hashtags.

6. CONCLUSION

In this paper we focus on the network structure defined between hashtags present in tweets. This structure captures in essence the dynamics of information context in user generated data in social media. After analyzing the static properties of the co-tweet hashtag network, we further study the research hypothesis that such structures can facilitate a variety of applications. We propose using Gini coefficient as one of the metric to analyze the dynamics of such network structure. In particular, we explore its temporal evolution under abnormal events and we build a preliminary event detection system that exploits its dynamics. In the future, we plan on developing a full-fledged event detection scheme utilizing this network structure.

7. REFERENCES

[1] M. Allamanis, S. Scellato, and C. Mascolo. Evolution of a location-based online social network: Analysis and models. In *Proceedings of ACM IMC*, 2012.

[2] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *Proceedings of ACM KDD*, 2008.

[3] E. Aramaki, S. Maskawa, and M. Morita. Twitter catches the flu: Detecting influenza epidemics using twitter. In *Proceedings of EMNLP*, 2011.

[4] J. Bollen, H. Mao, and X. Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[5] K. Carley, J. Pfeffer, H. Liu, F. Morstatter, and R. Goolsby. Near real time assessment of social media using geo-temporal network analytics. In *Proceedings of IEEE/ACM ASONAM*, 2013.

[6] Z. Cheng, J. Caverlee, and K. Lee. You are where you tweet: A content-based approach to geo-locating twitter users. In *Proceedings of ACM CIKM*, 2010.

[7] D. Crandall, D. Cosley, D. Huttenlocher, J. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *Proceedings of ACM KDD*, 2008.

[8] E. Cunha, G. Magno, G. Comarela, V. Almeida, M. A. Gonçalves, and F. Benevenuto. Analyzing the dynamic evolution of hashtags on twitter: A language-based approach. In *Proceedings of ACM LASM*, 2011.

[9] D. Davidov, O. Tsur, and A. Rappoport. Enhanced sentiment learning using twitter hashtags and smileys. In *Proceedings of ACM COLING*, 2010.

[10] P. S. Earle, D. C. Bowden, and M. Guy. Twitter earthquake detection: Earthquake monitoring in a social world. *Annals of Geophysics*, 54(6):708–715, 2012.

[11] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. A latent variable model for geographic lexical variation. In *Proceedings of ACM EMNLP*, 2010.

[12] J. Jussila, J. Huhtamäki, H. Kärkkäinen, and K. Still. Information visualization of twitter data for co-organizing conferences. In *Proceedings of ACM MindTrek*, 2013.

[13] K. Y. Kamath and J. Caverlee. Spatio-temporal meme prediction: Learning what hashtags will be popular where. In *Proceedings of ACM CIKM*, 2013.

[14] K. Y. Kamath, J. Caverlee, K. Lee, and Z. Cheng. Spatio-temporal dynamics of online memes: A study of geo-tagged tweets. In *Proceedings of ACM WWW*, 2013.

[15] J. Lin, R. Snow, and W. Morgan. Smoothing techniques for adaptive online language models: Topic tracking in tweet streams. In *Proceedings of ACM SIGKDD*, 2011.

[16] M. McCord and M. Chuah. Spam detection on twitter using traditional classifiers. In *Proceedings of the ACM ATC*, 2011.

[17] M. Newman. *Networks: An Introduction*. Oxford University Press, 2010.

[18] J. Nichols, J. Mahmud, and C. Drews. Summarizing sporting events using twitter. In *Proceedings of ACM IUI*, 2012.

[19] D. M. Romero, B. Meeder, and J. Kleinberg. Differences in the mechanics of information diffusion across topics: Idioms, political hashtags, and complex contagion on twitter. In *Proceedings of ACM WWW*, 2011.

[20] T. Sakaki, M. Okazaki, and Y. Matsuo. Earthquake shakes twitter users: Real-time event detection by social sensors. In *Proceedings of ACM WWW*, 2010.

[21] O. Tsur and A. Rappoport. What's in a hashtag?: Content based prediction of the spread of ideas in microblogging communities. In *Proceedings of ACM WSDM*, 2012.

[22] G. Valkanas and D. Gunopulos. Event detection from social media data. *IEEE Data Engineering Bulletin*, 1:3–5, 2013.

[23] J. Weng and B.-S. Lee. Event detection in twitter. In *Proceedings of ACM ICWSM*, 2011.

[24] Wikipedia. Boston marathon bombing. `http://en.wikipedia.org/wiki/Boston_Marathon_bombings`.

[25] S. Yitzhaki. Relative deprivation and the gini coefficient. *The Quarterly Journal of Economics*, pages 321–324, 1979.

[26] K. Zhang and K. Pelechrinis. Understanding spatial homophily: The case of peer influence and social selection. In *Proceedings of ACM WWW*, 2014.