BIG-2014-0072-ver9-Chen_1P
Type: research-article

Big Data
Volume 3 Number 1, 2015

DOI: 10.1089/big.2014.0072

ORIGINAL ARTICLE

Human Rights Event Detection from Heterogeneous Social Media Graphs

Feng Chen[1] and Daniel B. Neill[2,*]

Abstract

Human rights organizations are increasingly monitoring social media for identification, verification, and documentation of human rights violations. Since manual extraction of events from the massive amount of online social network data is difficult and time-consuming, we propose an approach for automated, large-scale discovery and analysis of human rights–related events. We apply our recently developed Non-Parametric Heterogeneous Graph Scan (NPHGS), which models social media data such as Twitter as a heterogeneous network (with multiple different node types, features, and relationships) and detects emerging patterns in the network, to identify and characterize human rights events. NPHGS efficiently maximizes a nonparametric scan statistic (an aggregate measure of anomalousness) over connected subgraphs of the heterogeneous network to identify the most anomalous network clusters. It summarizes each event with information such as type of event, geographical locations, time, and participants, and provides documentation such as links to videos and news reports. Building on our previous work that demonstrates the utility of NPHGS for civil unrest prediction and rare disease outbreak detection, we present an analysis of human rights events detected by NPHGS using two years of Twitter data from Mexico. NPHGS was able to accurately detect relevant clusters of human rights–related tweets prior to international news sources, and in some cases, prior to local news reports. Analysis of social media using NPHGS could enhance the information-gathering missions of human rights organizations by pinpointing specific abuses, revealing events and details that may be blocked from traditional media sources, and providing evidence of emerging patterns of human rights violations. This could lead to more timely, targeted, and effective advocacy, as well as other potential interventions.

Key words: big data analytics; data mining; machine learning; social networking

Statement of Significance

Human rights organizations, including nongovernmental organizations (NGOs) such as Amnesty International and Human Rights Watch, are increasingly monitoring and analyzing data from social media (such as Twitter and Facebook) for identification, verification, and documentation of human rights violations.[1] Such organizations' need for comprehensive verification of abuses and convincing, legally admissible documentary evidence suggests that these emerging technologies are likely to supplement rather than replace traditional, interview-based fact-finding methods. Nevertheless, social media holds great potential for monitoring emerging human rights emergencies, accessing video evidence that documents abuses, and calling attention to specific events or patterns that might otherwise escape notice. Social media helps spread information earlier and faster than traditional media: Advocates for Human Rights note that "increasingly, information and images that first came to light through social media have been used to fuel momentum for independent investigations."[2] They also identify other potential uses of social media, including detection of emerging trends in the prevalence of different types of abuses, or changes in public sentiments and perceptions, and note that it can provide detailed information that corroborates and enhances findings from other methods.[2] Social media also provides evidence of

[1]*Department of Computer Science, State University of New York at Albany, Albany, New York.*
[2]*Event and Pattern Detection Laboratory, H.J. Heinz III College, Carnegie Mellon University, Pittsburgh, Pennsylvania.*

*Address correspondence to: *Daniel B. Neill, PhD, Event and Pattern Detection Laboratory, H.J. Heinz III College, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213*, E-mail: neill@cs.cmu.edu

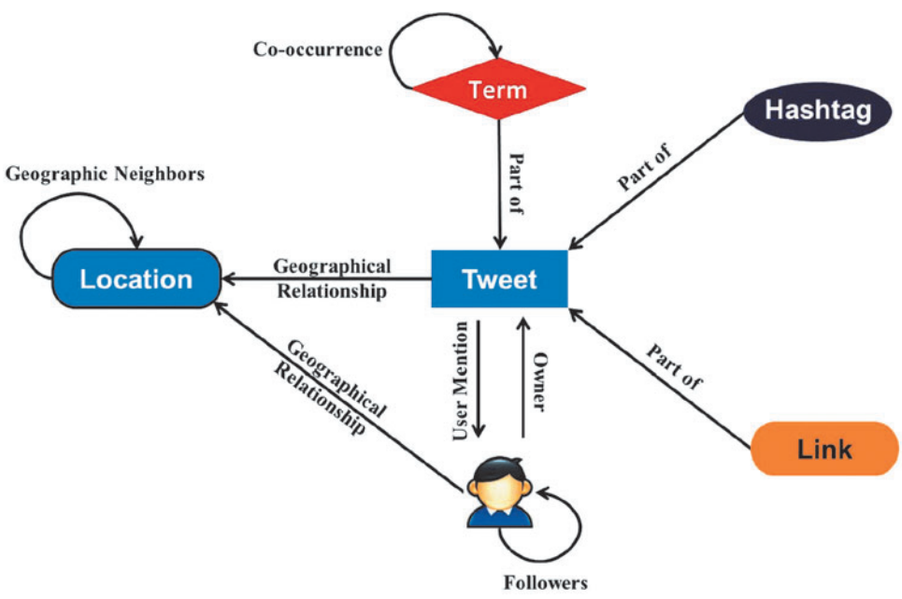
human rights abuses that may not be reported in the mainstream media, and is much more difficult for repressive governments to control or censor than traditional media sources,[3] thanks to the relative anonymity of users, the rapid spread of information, and the use of informal, dynamic language, which makes targeted censorship difficult.

However, social media monitoring poses major technical challenges for human rights NGOs, most critically the huge amount of effort that would be needed to manually identify human rights events from the massive amount of data. To address this challenge, we apply our recently developed Non-Parametric Heterogeneous Graph Scan (NPHGS) method[4] for domain-specific event detection in social media data to automatically identify human rights events using data from Twitter. This approach achieves timely and accurate event detection in other domains related to the social good, and can scale to massive amounts of data; our results below examine its potential utility and applicability for the human rights domain.

Overview of Methodological Approach

Clusters of tweets related to human rights events were identified using our recently developed approach, the NPHGS, for event detection using social media data. Complete details of the NPHGS approach are provided in Chen and Neill[4]; we summarize the essential features of this methodology in this section, and provide additional methodological details below. The key idea behind NPHGS is to consider all of the potentially useful information in social media (including, but not limited to, Twitter data) in a unified statistical framework. Previous methods tend to focus on one particular aspect of this data, such as locations, users, or tweets, but we have shown that detection performance can be improved substantially by considering the entire Twitter network as a *heterogeneous graph*.[4] In our representation of the Twitter network, each node is one of six types (User, Tweet, Location, Term, Hashtag, and Link), and the relationships between nodes can be of many different types (Fig. 1). Each node type can ◀F1 have many different features, such as the numbers of tweets and users for a given geographic location; the numbers of followers, tweets, retweets, and mentions for a given user; and the klout (a measure of influence) and sentiment score for a given tweet. An example of a portion of the Twitter network is shown in Figure 2, ◀F2 and further details are provided in Chen and Neill.[4]

Given this massive, complex, and heterogeneous graph structure, we search for clusters (connected subgraphs of the Twitter network) with anomalous activity in the recent data. As described below, we first compare each feature of each node to a reference distribution



4C▶

FIG. 1. Entity diagram for Twitter data modeling.

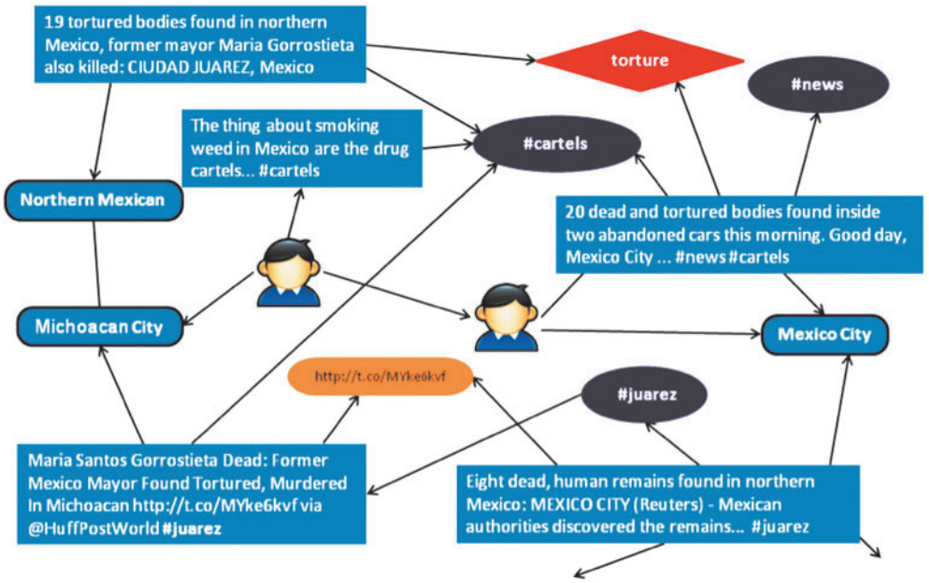


FIG. 2. A portion of the Twitter heterogeneous network.

and compute a *p*-value representing the anomalousness of that feature. For example, we may observe that a given user has been mentioned an unusually large number of times today as compared to past days. Second, we combine the multiple features into an overall measure of anomalousness (*p*-value) for each node, and third, we search for clusters where the nodes have lower (more significant) *p*-values than would be expected by chance. Such clusters can include nodes of multiple types, thus identifying the users, locations, and concepts (terms and hashtags) involved in the event, as well as providing the relevant tweets and supporting evidence (such as web links). For example, if a new human rights event occurs in a city, then we could observe anomalous patterns of activity from Twitter, including (1) certain Twitter users tweet a lot about the event; (2) the city has a large number of active users; (3) some related terms, persons, or organizations are mentioned a lot; and (4) some photos, videos, or news sources are mentioned frequently in tweets. As a result, the subgraph could be a combination of users, tweets, locations, terms, hashtags, and links.

Results

Using the Twitter event detection methodology we developed in Chen and Neill,[4] we performed an exploratory analysis of detected human rights events in Mexico from January 2013 to June 2014. We randomly collected 10% of all the raw Twitter data from Mexico from June 1, 2012, to June 30, 2014, consisting of 96 million tweets in total. Data from June 2012 to December 2012 was used to train our model, and the remainder was used to generate the experimental results described below. We first generated a small vocabulary of terms, including "torture," "murder," "rape," "disappearance," "killing," "massacre," "beating," "kidnapping," and their Spanish-language equivalents. Only the raw tweets that matched at least one term from the dictionary were preserved, and the remaining tweets were discarded. We then identified the geographic location of each tweet (at the city level) through a three-step process. We first searched for location and landmark mentions in the tweet text. If no such information was found, we searched for geotags that are available if the user enabled the geocoding function in his/her phone; otherwise, we used the location information from the user's profile. Tweets were grouped to daily windows, and for each day from January 1, 2013, to June 30, 2014, we ran our NPHGS method described below to identify the highest-scoring cluster (which could include tweets, users, hashtags, links, keywords, and locations) for that day. Our method was run prospectively; that is, for a given day's run, only data up to and including that day was used in the analysis.

This simulates the scenario in which the method is deployed in the field and used for daily analysis and early event detection and response.

Finally, the top 50 highest-scoring clusters over the entire study period were analyzed manually to identify (1) whether the cluster was human rights related, (2) the types of human rights violations, (3) the victims of the violations, and (4) the alleged perpetrators. We note that reporting 50 clusters was a somewhat arbitrary choice, and was limited by our ability to manually evaluate each cluster. In practice, the number of clusters to report should depend on the available capacity of human rights organizations to investigate these clusters. As described in Chen and Neill,[4] we could also perform a permutation test to compute the statistical significance of each detected cluster, and report only those clusters that are significant at some threshold. We now present an overall summary of our exploratory analysis of the top 50 clusters:

- 39 (78%) of the top 50 highest-scoring clusters, including 26 (87%) of the top 30 clusters, referred to human rights events. Most of the 39 clusters contained tweets referring to multiple events, but in most cases the majority of tweets referred to a single major event; all but 5 clusters referred to current as opposed to past events.
- The 11 clusters that did not refer to human rights events in Mexico included 3 clusters referring to animal rights (hunting, torture of animals, etc.), 3 clusters referring to events outside Mexico (including 2 clusters referring to the state of New Mexico, U.S.A., and 1 cluster related to the historical event of John F. Kennedy's assassination), and several other clusters containing human rights-related keywords but not actually related to human rights.
- Event types in the detected clusters included murder (31 clusters), kidnappings and disappearances (26 clusters), torture and beatings (7 clusters), rape (5 clusters), and extortion (3 clusters).
- Alleged perpetrators of the described human rights events included police and security forces (10 clusters), drug gangs (5 clusters), the army (3 clusters), and politicians or political parties (2 clusters).
- Listed victims of human rights violations included politicians or political activists (14 clusters), youths or children (14 clusters), journalists (5 clusters), and police officers (1 cluster).
- While many of the clusters expressed negative sentiments toward the described events, only two clusters referred explicitly to a planned response, such as a protest or march.

As an initial case study, we considered our 11th-highest scoring cluster, which focuses on the murder of a politician and electoral campaign coordinator Aquiles González Mayorga in the city of Zacatecas. This was believed to be a political assassination given that documents related to the upcoming election were stolen, and the event generated significant media attention. The event was first made public through a tweet by politician Jesus Ortega Martinez (@jesusortegam) on July 5th, 2013, at 6:14 pm, which translates to: "In the morning I spoke with the governor of Zacatecas to warn him about the disappearance of our political partner Aquiles. A few hours ago he was found murdered. A barbarity!!!." Local news articles about the murder appeared on July 6th, referencing tweets by @jesusortegam as confirmation, and the story appeared in international news sources on July 7th. The detected cluster on July 5th included the relevant location (Zacatecas), hashtags (#AquilesGonzalez, #Aquiles), users (@PRDMexico, @jesusortegam), and tweets (including the one mentioned above). We were able to detect this cluster using only those tweets up through July 5th (the same day as the event, and the day before the event was reported in local news). The cluster was sufficiently high-scoring that only one false-positive (not human rights related) cluster in the 18 months of data had a higher score. For two other case studies (our 2nd and 3rd highest scoring clusters), representing the murder of the secretary of tourism in Jalisco and the kidnapping and murder of journalist Alberto Lopez Bello in Oaxaca, respectively, the detected clusters occurred along with the first local news reports and included some tweets referring to those news reports, while the international news media reported the story the following day.

Discussion

The exploratory analysis and case studies described above present preliminary evidence of the potential utility of Twitter data for early detection of human rights events. Our NPHGS approach,[4] described below, was able to identify clusters of tweets related to events of interest prior to international news sources, and in some cases, prior to local news reports. A more detailed evaluation, including labeling of gold-standard events by domain experts, is necessary to more precisely quantify

the detection performance of our method, in terms of true-positive and false-positive rates, and to compare its performance to competing methods such as spatiotemporal burst detection[5] and geographic topic modeling.[6] We have conducted similar performance evaluations for detection and prediction of other event types (civil unrest and rare disease outbreaks) from Twitter data,[4] demonstrating that NPHGS provides more timely and more accurate detection as compared to the previous state of the art. We note that NPHGS was also shown to achieve accurate *prediction* of civil unrest events from 1 to 7 days in advance: organization of planned protests and marches, or widespread anger that could boil over into spontaneous demonstrations, may be visible in the Twitter data.[4] Advance prediction was not successful for the individual human rights events identified here, though it might be possible for more widespread events such as outbreaks of ethnic violence. Nevertheless, it does appear that our approach could be used to provide human rights NGOs with timely information about current events of interest, enabling them to respond quickly and appropriately.

In addition to the preliminary nature of the evaluation presented here, we note some limitations and biases of our current event detection methodology. First, we used a very simple keyword-based approach to filter tweets in the preprocessing stage, and this approach resulted in several types of false positives (e.g., animal rights and tweets related to abortion rights in New Mexico) that could easily have been filtered out by more careful preprocessing. Of course, these clusters (while considered "false positives" for the current analysis) may still be of interest to rights organizations with a different geographic or topic focus. One option going forward would be continued refinement of the keyword matching rules so that identified false positives would be excluded, while another option (which might require a substantial amount of labeled training data) would be to learn a binary classifier to distinguish human rights-related from unrelated tweets. Additionally, we chose to focus here only on keywords primarily related to physical harm of individuals, as these correspond to specific events that can be more easily localized in space and time. These constitute only a subset of human rights violations and related issues: for example, racial and ethnic discrimination, lack of adequate political representation, press censorship, and government corruption are all human rights-related issues of concern to governments and NGOs, but these might not be visible as clusters in the Twitter data unless brought to the forefront of discussion by a specific event.

We also note that the identified clusters should not be considered a representative sample of human rights events in Mexico, but are subject to selection bias in terms of which events provoke interest, discussion, outrage, and so forth, among Twitter users. For example, events where the alleged perpetrators are police officers or the military, as well as those where the victims are celebrities, government officials, or young children, may gain greater attention than the more frequent acts of violence resulting from ongoing conflicts between drug cartels, criminal gangs, and the Mexican government. Finally, we note that our method is not necessarily able to distinguish between clusters that represent true events and those that are falsely reported (e.g., false rumors of a political assassination). However, recent research[7] has shown that there are measurable differences in the way that reliable messages propagate through social networks as compared to unreliable ones. These differences can be harnessed to quickly classify messages as more or less credible with a reasonable degree of accuracy. We believe that it is important to present detected clusters to human rights organizations for further investigation, enabling them to both debunk rumors and respond to true events of interest, rather than simply assuming that described events are factual.

Methodological Details

Our NPHGS approach performs event detection in multiple steps. After preprocessing to filter out irrelevant tweets (as described above), NPHGS constructs a heterogeneous graph from the remaining tweets. It then models the network as a "sensor" network, in which each node senses its "neighborhood environment" and reports a statistical significance (p-value) measuring the current anomalousness levels of various neighborhood-related features. For a given feature, such as the number of times that a given user is mentioned in tweets by other users, we compare the daily count for that feature to a *reference distribution*: either the historical distribution of daily values of that feature for the given node, or for similar nodes if sufficient historical data for that node is not available (e.g., for a new Twitter user). In either case, we can compute the percentile rank of the current feature value as compared to the reference distribution, resulting in a (one-sided or two-sided) p-value for that feature. One key idea of NPHGS is that conversion to p-values

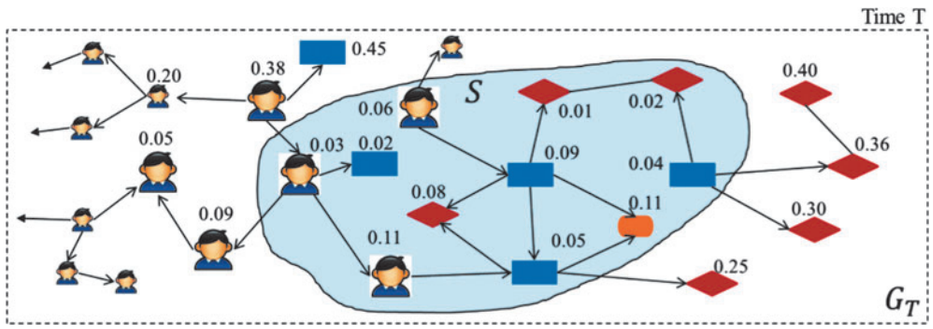


FIG. 3. Detecting anomalous subgraphs as indicators of new events. Anomalousness of each subgraph is measured by comparing the actual and expected distributions of *p*-values. A highly anomalous subgraph will have a higher than expected number of low (significant) *p*-values. We have developed fast algorithms to identify the most anomalous subgraphs in massive heterogeneous networks; more details are provided in Chen and Neill.[4]

allows the disparate node types and features to be treated on the same scale: we expect each *p*-value to be uniformly distributed from 0 to 1 if no events of interest are occurring, and lower values represent more significantly anomalous activity.

After each feature is ranked against its reference distribution to obtain a *p*-value, the multiple feature *p*-values for a given node (e.g., user) are combined into a single *p*-value representing the overall anomalousness of that node. This is done by computing the minimum *p*-value across all features for a given node, and then re-calibrating that *p*-value by comparing it to the reference distribution of minimum *p*-values for the historical data. This two-stage empirical calibration process has several benefits; for example, it does not overweight nodes that have a large number of observed features.[4]

Finally, we search for connected subgraphs of the Twitter network with a higher-than-expected number of low (significant) *p*-values. To do so, we can compute an overall measure of anomalousness, a *nonparametric scan statistic*,[8,9] for each subgraph. The nonparametric scan statistic represents a measure of divergence between the actual and expected numbers of node *p*-values that are significant at some level α; this quantity is then maximized over different α thresholds and over all connected subgraphs of the Twitter network. For the results discussed above, we used the Berk–Jones nonparametric scan statistic,[4,9] which is equal to the number of nodes in the given subgraph multiplied by the Kullback–Liebler divergence (a well-known measure of the difference between two probability distributions) between the actual and expected proportions of *p*-values that are significant at level α. An efficient heuristic approach is used to identify the most anomalous network clusters as those that maximize the nonparametric scan statistic (Fig. 3). In brief, we use an iterative subgraph expansion algorithm, where on each iteration we allow the cluster to expand to some subset of its neighbors; each such expansion can be performed efficiently by exploiting a property of the nonparametric scan statistic that allows for very fast maximization over subsets. Each cluster is returned as the indicator of an ongoing or upcoming event, and is summarized with information such as type of event, geographic locations, time, and participants. More details of this process are provided in Chen and Neill.[4]

Conclusions

These results, together with those of Chen and Neill,[4] demonstrate the potential utility of our NPHGS approach for multiple domains relevant to the social good, including civil unrest event prediction, early detection of rare disease outbreaks, and human rights. For both civil unrest and outbreak detection, NPHGS outperformed five existing methods for Twitter event detection, increasing detection power, forecasting accuracy, and forecasting lead time while reducing time to detection.[4] We plan to evaluate the approach for several additional domain-specific event detection tasks, such as traffic prediction and emergency response.

With respect to the human rights domain, our analysis suggests that NPHGS can accurately identify clusters of tweets corresponding to human rights events of interest, as well as providing information (such as time, locations, key terms and hashtags, and influential users) and documentation (such as links to web pages, pictures, or videos) for each such cluster. Thus, this approach could enhance the fact-finding missions of human rights organizations by pinpointing specific abuses and violations of human rights, in some cases before even the local news media is aware of these events. In countries such as Mexico, NPHGS could serve as an aggregator and filter of social media for human rights NGOs, reducing the effort needed to review multiple local news sources and enabling them to identify and draw connections between events. Integrating data from Twitter and local news media might help to reduce false positives and provide verification of detected events. In other countries where the news media is tightly controlled, Twitter event detection might provide essential information that is blocked from traditional media.

We believe that an important next step for this work is to assist human rights organizations with the process of drawing broader patterns, relationships, and conclusions about emerging human rights issues. Our future work will provide additional support for sensemaking and storytelling by extending NPHGS with approaches such as dynamic topic modeling,[10,11] which can provide more detailed analysis of tweet content and draw connections between related clusters over time.

Finally, it is worth noting that social media is now impacting human rights in multiple ways: organizing of local activists and political movements, sharing of videos as documentary evidence of human rights abuses ("citizen journalism"), and calling international attention to violators of human rights, both through comprehensive information gathering and advocacy by human rights NGOs, and through the so-called "hashtag activism," which attempts to inform and rally public support around a particular cause.[12] At the same time, repressive governments are beginning to monitor social media to identify dissidents, which might lead to further human rights violations. We believe that automated, large-scale discovery and analysis of human rights-related events could shed light on these emerging phenomena as well. For example, a potential use case could be identifying government crackdowns on dissidents related to some particular issue (e.g., journalists or leaders of a particular labor union) by detecting when a group of connected users in a particular geographic region simultaneously becomes inactive or disappears from social media.

Acknowledgments

This work was partially supported by National Science Foundation Grants IIS-0916345, IIS-0911032, and IIS-0953330. Additional support was provided by the John D. and Catherine T. MacArthur Foundation. The authors wish to thank Jay Aronson (Center for Human Rights Science, Carnegie Mellon University) and Eric Sears (MacArthur Foundation) for helpful comments and feedback on preliminary versions of this work.

Author Disclosure Statement

No competing financial interests exist.

References

1. Koettl C. Twitter to the rescue? How social media is transforming human rights monitoring. February 20, 2013. http://blog.amnestyusa.org/middle-east/twitter-to-the-rescue-how-social-media-is-transforming-human-rights-monitoring/
2. The Advocates for Human Rights. *Action and Advocacy: A Practitioner's Guide to Human Rights Monitoring, Documentation and Advocacy*, 2011. www.theadvocatesforhumanrights.org/uploads/final_report_3.pdf
3. Joseph S. Social media, political change, and human rights. Boston Coll Intl Comp Law Rev 2012;35:145–188.
4. Chen F, Neill DB. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. Proceedings of the 20th ACM SIGKDD Conference on Knowledge Discovery and Data Mining 2014; 1166–1175.
5. Lappas T, Viera MR, Gunopulos D, Tsotras VJ. On the spatiotemporal burstiness of terms. Proc VLDB Endowment 2012;5:836–847.
6. Yin Z, Cao L, Han J, et al. Geographical topic discovery and comparison. Proceedings of the World Wide Web Conference 2011; 247–256.
7. Castillo C, Mendoza M, Poblete B. Information credibility on Twitter. Proceedings of the World Wide Web Conference 2011; 675–684.
8. Neill DB, Lingwall J. A nonparametric scan statistic for multivariate disease surveillance. Adv Disease Surveill 2007;4:106.
9. McFowland E, Speakman S, Neill DB. Fast generalized subset scan for anomalous pattern detection. J Mach Learn Res 2013;14:1533–1561.
10. Blei D, Ng A, Jordan M. Latent Dirichlet allocation. J Mach Learn Res 2003;3:993–1022.
11. Blei DM, Lafferty JD. Dynamic topic models. Proceedings of the 23rd International Conference on Machine Learning 2006; 113–120.
12. Hilal M. Resource guide: technology for human rights monitoring, documentation, awareness raising and solidarity. December 10, 2014. www.internationalpeaceandconflict.org/profiles/blogs/resource-guide-technology-for-human-rights-monitoring-documentati

Cite this article as: Chen F, Neill DB (2015) Human rights event detection from heterogeneous social media graphs. *Big Data* 3:1, 1–7, DOI: 10.1089/big.2014.0072.

Abbreviations used

NGOs = nongovernmental organizations
NPHGS = Non-Parametric Heterogeneous Graph Scan