Forecasting Civil Strife: An Emerging Methodology

Dipak K. Gupta*, Sathappan Muthiah‡, David Mares† and Naren Ramakrishnan‡
*Professor Emeritus
Department of Political Science
San Diego State University
†Chair for Inter-American Affairs
Institute of the Americas
University of California, San Diego
‡Discovery Analytics Center
Department of Computer Science
Virginia Tech, Arlington, VA

***Abstract*—From the earliest time of recorded scholarship, forecasting civil strife has been the Holy Grail to political theorists. Yet, without actual data and ability to conduct empirical analyses, until the 1960's such analyses were no more than speculation. The advent of high-speed computing along with collection of data on civil unrest allowed political scientists to empirically test their hypotheses. Yet, these analyses did not result in short term prediction due to the lack of real time data. Today the rise of social media has witnessed a radically different methodology in how we can understand, monitor, and forecast incidents of social strife in real time. This emerging methodology, however, requires a multi-disciplinary effort no one even contemplated until recently. This paper presents results of forecasting events of politically motivated violence based on monitoring open source information (Twitter, blogs, newspaper articles) in 10 Latin American countries by a multi-university, multi-disciplinary team of academics, supported by a grant from the Intelligence Advanced Research Projects Activity (IARPA).**

***Keywords*–Forecasting, Civil Strife, Social Media, Multidisciplinary Methodology**

I. Introduction

Generally speaking, social scientists are reluctant to forecast incidents of civil unrest. The open system, within which these outcomes are derived, has been considered far too complex to forecast anything more than the direction of trends or the increased likelihood of an event occurring. In fact, the noted social scientist Paul Collier [1] flatly admits concerning his own model: “More fundamentally, our model (of political conflict) cannot be used for prediction, it cannot tell you whether Sierra Leone will have another civil war next year. That depends on a myriad of short-term events.” As a result, the efforts of social scientists have largely been confined to the understanding of the determinant variables using various types of causal models or by using time-series projections. From a public policy standpoint, the importance of forecasting imminent political upheavals is undeniable. The urgency of developing such predictive capabilities has been further precipitated by the spread of social media and the Internet. The prime example of the potential of social media to mobilize civil strife was amply demonstrated during the sudden uprising, dubbed, the “Arab Spring.” Rise of social media has been a double-edged sword; along with its ability to mobilize the disgruntled, it has accorded us a new avenue through which we can understand society in a way that was impossible in the past. Our ability to understand and forecast civil strife has been vastly expanded because of two primary reasons. First, social media, especially in democratic nations, allows individuals to express their opinions freely. As a result, when monitored properly, social media allows us to comprehend societal trends in a way that even the best-constructed opinion surveys are not able to capture. Second, thanks to the recent advancements in a number of related academic disciplines as well as computational capabilities, we are witnessing a revolutionary change in how we understand and then predict various societal events [2] [3], such as spreading chaos in the financial market [4] [5], spread of infectious diseases [6]–[9], etc. These efforts require a multi-disciplinary focus that was until recently largely absent.

Our multidisciplinary research effort, which brings in expertise from wide-ranging disciplines across social sciences, linguistics, geographic information systems, and computer science, concentrates on “Civil Strife“, by which we mean mass movements, such as protest movements and riots and other acts of collective rebellion and not on terrorism or the so-called “lone wolf” attacks, plotted and carried out by a small group or an isolated individual.

For our study, we chose 10 large Latin American countries. These countries offer a unique combination of mostly functioning democracies with a long vibrant tradition of public discourse of political issues along with excellent penetration of Internet communication.

Our article is divided in four broad parts. The first part introduces the problems of forecasting from a social science perspective and provides a theoretical basis of human motivation for participating in mass movements (Section I and Section II). The second part provides the foundation of our subsequent analyses (sections III and IV). The third part (Section V) explains the index of accuracy of our research effort. The final part (Section VI) of the paper discusses ways of moving our research forward.

II. Theoretical Basis of Forecasting Through Social Media

Social media is a noisy medium, where individuals and groups participate using messages that are confusing in language, terminologies, and expressions, made even more complicated by their temporal malleability. Therefore, before we begin to make sense of this vast and chaotic linguistic landscape, we need to put our analysis within a broad theoretical framework.

As scholars looked deeper into the process by which individual participants are moved into creating collective actions, Olson [10] raised a theoretical concern. When it comes to volunteering for a political cause, there is an inevitable inertia of what is known in social sciences as the "collective action" problem, where a rational actor asks the inevitable question, "why me?" The answer to this question may be found in the work of social psychologists, starting with the path breaking work of Tajfel and his associates [11] [12]. Their work, widely accepted as "social identity" theory argues that, contrary to the economic assumption of human behavior that equates self-interest with human rationality [13], individuals are also motivated to voluntarily participate in collective actions out of their community concerns. Group identity as a political force, however, does not develop spontaneously [14]. For that, the society needs to have leaders or "political entrepreneurs." These leaders need to clearly define the perimeters of the "us" and "them," by instilling in the minds of the adherents, the most primal emotion of all: fear [15].

We can forecast future actions in social media by searching for markers of extant grievances ("injustice", "deprivation", "inequality", etc.), the boundaries of the in-group (viz., "students", "workers", "farmers", "indigenous people", etc. along with the names of the "heroes" of the community) the out-group (viz., "politicians", "thieves", "oppressors", "police" etc., along with names of the "villains" of the movement). Finally, we can also track the actions suggested by the leaders (viz., "protest", "march", "destroy", etc.). Based on this theoretical framework, we developed a library of keywords, reflecting existing grievance, the leaders' framing of collective identity, and prodding of actions against the offending party. The results of tracking such keywords in social media lead us to our forecasts of impending civil strife. For our forecasting, we found a dichotomy between planned and unplanned events to be useful. We search for "planned events" by searching for news or announcements by various organizations for upcoming protest events. Figure 1 illustrates the theoretical framework, where political entrepreneurs take the prevailing grievances and frame those in terms of "us" and "them," provide an action plan to bring about events of political demonstrations or riots. By tracking this process, we predict incidents of civil strife in a country.

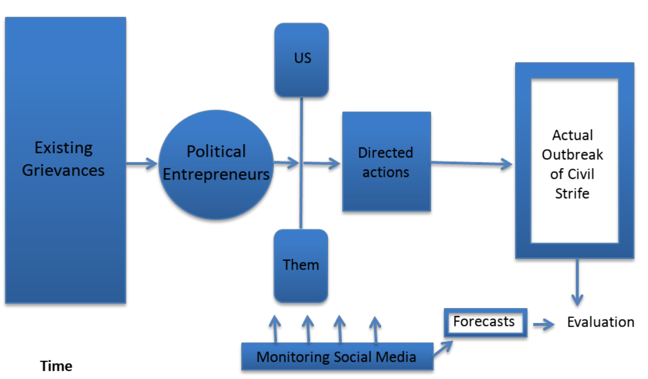

Figure 1. Theoretical Framework of Forecasting Civil Strife

III. Features of Our Study

Our study offers a number of important features:

1) On the Internet there is information that is public ("blasts" to world) and private (communication between individuals or groups, otherwise protected by law). We strictly use only public information obtained from open sources, such as newspapers, blogs, and public areas of social media such as Facebook and Twitter. In other words, none of our data are private or classified.
2) The events of civil strife were classified by strict definitions of who participates (general population, labor unions ...), issues (political, environmental ...), geographic location, and time of event. The *who* part of an event is described by the categories defined in the "social" sector of the TABARI [16] event coding system. The issues or the *why* part of an event is classified into 7 categories. Apart from encoding the *who* and *why* part of an event, our study also strives to identify if a civil strife will turn violent or not. An event is deemed violent if there is significant damage to property or if there is any act of violence directly associated with the event that results in injuries to anyone involved. The geographic location entry of an event description uses a 3-level typology with the country at the top, province or state (admin level 1) in the middle and city at the last level. The geonames gazetteer [17] is used as the reference standard for the location specification of an event. Finally, the time of an event is encoded with date level accuracy.
3) The actual post-facto events were recorded from the local, national, and international newspapers and codified by a third party (in this case, MITRE Corporation) and were published in IARPA's Gold Standard Report (GSR). The list of newspapers were identified using the rankings as provided by 4 International Media & Newspapers [18] and with subject matter expert input.
4) Forecasts had to be submitted in real time and at the end of every month would be matched against the GSR by MITRE Corporation and rated for accuracy as well as lead time. False positives and false negatives were also tracked as were the confidence in the forecasts.
5) Ultimately, the delivery of warnings had to be fully automated, without a "human in the loop". Humans may be involved in the development and training phases of modeling, but the final warnings must be fully machine generated and automatically submitted without any interference (in form of filtering or guiding) from subject matter experts.For illustrative purposes, we have also presented a sample of a hypothetical warning in Figure 2.
6) Every forecast should have an audit trail for IARPA to trace it back to the causal factors.

IV. Algorithms for Forecasting

In this section, we provide details of how open source data is harvested and enriched by our system and finally how this

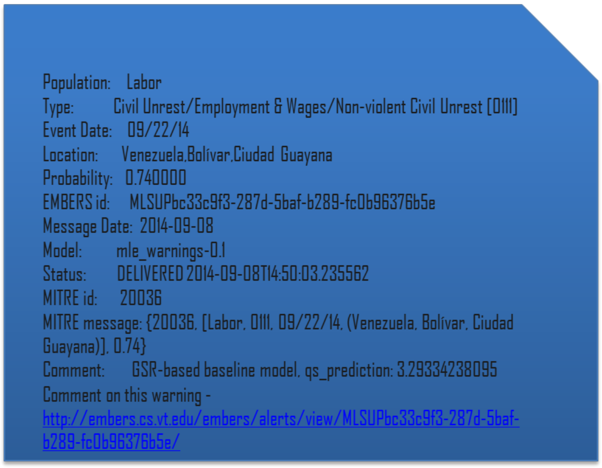

Figure 2. An example of computer generated warning

enriched data is used in various models to produce forecasts or alerts.

A. Data Collection and Enrichment

Our work begins with searching the vast open source (public) areas of the Internet by following the keywords seeded by social scientists (the subject matter experts). Figure 3 explains our system architecture. These searches scoop up a significant quantity of linguistic information, much of which are unrelated to our objective. These data are stored in a "vat" and "enriched" by a Natural Language Processing(NLP) pipeline (includes language identification, tokenization, lemmatization, named entity recognition and Part-of-Speech tagging) to make the necessary connections between words and their sought meaning. The Basis technologies RLP suite [19] is used for all the NLP tasks. The textual data is then passed through a temporal normalization system like Heideltime [20] which converts any relative temporal expressions such as *today, tomorrow, a week ago* to absolute time based on the article/tweets publish time. It is not enough, however, to understand the underlying meaning of the collected keywords; for forecasting, we need to also know the geographic points of origin. A geographic information system allows us to geo-locate these conversations onto maps. We use different geocoding systems for different types of data sources. For example, twitter geocoding is achieved by looking at different parts of a tweet in an orderly fashion starting from the least available source, viz. geotags, followed by Twitter places field, the text fields contained in user profile (location, description) and finally the tweet text itself to find mentions of relevant locations. For news articles and blog posts, we develop a probabilistic reasoning engine using Probabilistic Soft Logic (PSL) [21] to identify which among the multiple location names mentioned within the text(such as the location of reporting, the incident location, etc.) is the main geographic focus of the article. These data points are now ready for statistical analyses, yielding forecasts.

B. Data Modeling

In this section we introduce briefly some of the main prediction models that work on different real-time datasets to produce alerts and a fusion engine that combines these alerts to achieve high quality and performance.

Figure 3. System Architecture

Protests in many countries are regular occurrences (for example Mexico has an average of 245.95 events per month in the GSR during 2013-2014). A baseline model that produces forecasts using only historical occurrences of events and no real-time data performs reasonably well and can be beat only in situations where there is an abnormal spike (or dip) in number of events The baseline model is trained using the ground truth event data from the previous three months to make forecasts for the current month. The baseline model estimates the expected number of events for the current month, for a given location, reason (or issue) and population (the *who* part) as the average of the number of such events in the training period. Finally, the time information (expected date of occurrence of event) is obtained by performing a random draw from a uniform distribution. We have conducted a maximum entropy analysis wherein we evaluated the baseline model's ability to forecast surprising events (i.e., events whose frequency falls outside the norms of historical data) and found that EMBERS consistently beats the baseline model [22]. EMBERS uses a suite of six forecasting algorithms that posit different approaches to modeling civil unrest.

The planned protest model aims to identify incidents of organized and preannounced protests from news and Twitter using language-processing techniques [23]. It searches for occurrences of phrases denoting "calls for protest" (like "planar protesta", "Announce Strike", "Manhã de mobilização",.etc) in news and Twitter along with the mention of a future date to produce alerts. The phrase list is developed in a semi-automatic manner using template based matching techniques on about 1 year worth of data (twitter, news articles and blog posts) and has mostly remained static barring a few changes for the entire duration of our project.

A second model uses spatial scan statistics to identify geo-located clusters of tweets enriched with a defined vo-

 ISBN: 978-1-61208-578-4

cabulary of 726 keywords [24].The goal of the model is to identify anomalous spatial regions based on Poisson mixtures. A fast linear time subset scan [25] is applied to identify the anomalous regions and each such region is scored using p-values computed by Monte Carlo simulations. The keyword filtered twitter stream is split into day level chunks and the fast subset scan is run separately for each day. A spatial cluster on any day is considered to be an continuation of a previous day cluster if the two clusters have a jaccard similarity (between the keywords present in each cluster) is greater than a threshold. In this manner the growth of a cluster both in terms of size and density is tracked over time and these characteristics are then used to issue a forecast for a given spatial region. The spatial scan model is trained using 6 months of twitter data. In comparison to the planned protest model, this model can track the development of a protest from its birth.However this models performance is limited by the amount of coverage/popularity the issue of a protest achieves on social media as not all protests are covered by social media like twitter. Also since social media is more prone to rumors it can lead to alerts being generated falsely.

The cascade regression model recognizes situations where social media, such as Twitter, is utilized as the staging ground for galvanizing support for protests via online recruitment to the underlying causes [26]. The model studies activity cascades in twitter to understand spread of influence and information and tries to forecast date of event based on this information spread besides identifying a critical subset of users responsible for the formation and survival of the activity cascade. The model analyzes over 353 million tweets over a 1.5 year period. Each tweet contains at least 3 keywords from a dictionary of over 900 keywords in 3 languages related to civil unrest. This ensures the activity cascades are relevant to our topic of interest. Next, for making forecasts a regression model (LASSO) using a feature set based on structural properties of cascades like size of cascade, number of participants, duration of cascades, change in the number of participants and tweets, average growth rate of tweets ,.etc, is used to predict the probability of a civil unrest event in a given day.

The dynamic query expansion model is similar to the spatial scan model but aims to learn new emerging keywords, unlike the static vocabulary used by the spatial scan model [27]. In 2013, for instance, there were a series of protests in Venezuela due to a shortage of toilet paper, a novel circumstance that was uncovered using this model. The dynamic query expansion model starts with a very small set of seed keywords (like protest, march, demonstration etc.) and iterates through the data identifying semantically similar and co-occurring terms. The model repetitively sweeps through the data, learning a larger set of relevant keywords at each iteration, unless it converges. This dynamic learning of new relevant keywords helps the model identify novel/unusual circumstances of protest.

Both the cascade regression model and dynamic query expansion suffer from the same disadvantages as the spatial scan model and is trained on at least 6 months worth of historical data.

The volume-based LASSO model uses every possible data source in our study (news, tweets, blogs, economic indicators, TOR, and smiles) to forecast the imminence of protests in the next day or two [28]. The volume-based LASSO model provides insights about the underlying social dynamics in different countries by identifying predictive features that are tied to unrest. The model also is capable of identifying the value of different data source in predicting an unrest and is more robust to changes in individual data characteristics as compared to the previously mentioned models. The main disadvantages of such a model is low recall and lead-time. Also, this model works better at country and state levels as opposed to city level (not all datasets are available at city level granularity)

Finally, the MLE(Maximum Likelihood Estimate) model aims to identify regularities in the GSR and provides a baseline performance level [29]. Each of these models is tuned for high precision, and their fusion aims to achieve high recall.

All the above-mentioned models produce alerts independent of each other with each model tuned for high precision. The fusion model then aims at combining the different predictions from these individual models to achieve high quality forecasts with tunable precision and recall. The goal of the fusion model is to (i)identify duplicate alerts (alerts with same location, date, type and population) as models share data sources and thus their hypothesis space overlap, (ii) fill-in missing values if any in an alert (for example certain forecasts of the volume based model are only at country level and it is the fusion engines responsibility to add city information) , (iii) re-write warning fields if necessary as it is possible for a model to issue alert for an improbable combination of <*date,location,type,population*> due to noisy data and finally, (iv) balance recall and quality. The recall-quality trade-off is achieved by first building a random forest regression model to predict the expected quality score of an alert and then a threshold set on this expected quality score can be used to tune the precision and recall (with a lower bound on quality) of our overall system.

V. ACCURACY OF FORECASTS

The forecasts for a given month are evaluated by MITRE at the end of every month using the Gold Standard Report (GSR). The forecasts are matched against the GSR using the Hungarian [30] bipartite matching algorithm. At the end of the bipartite matching set of alert-event match pairs are obtained along with the list of unmatched events and unmatched alerts. Each alert-event match pair is assigned a quality score out of 4. The problem of measuring forecasting accuracy is that each forecast is a multi-dimensional phenomenon. Thus, a specific forecast must match the time and geographic place where it was supposed to have taken place and it should also correspond to the type of protest and its specific cause. Thus, the quality score is a composite measure of four components. These are: location score, date score, event-type score and event-population score. Precision refers to the fraction of alerts that got matched to a true event, whereas recall refers the fraction of true events (GSR) that got matched to an alert (i.e., were accurately forecasted). The average lead-time of the system of was 9.76 days i.e., on an average our system produced an alert for a civil unrest event 9.76 days in advance of the first report of the event.

Our ability to forecast events in Latin America was put to test by a sudden explosion of public anger in Brazil during mid-2013 and in Venezuela in early 2014. Table I provides the performance metrics for the two countries during these

 ISBN: 978-1-61208-578-4

TABLE I. COMPARISON OF ACCURACY FOR BRAZIL AND VENEZUELA

Country	*Period*	*Quality*	*Precision*	*Recall*	*Lead-Time*
Brazil	May'13-Aug'13	3.44	0.53	0.69	7.04
Venezuela	Feb'14 - March'14	3.66	0.91	0.50	2.25

periods. Similar to the so-called "Arab Spring," where three years earlier the self-immolation by an obscure fruit vendor in Tunisia's capital city touched off a cascading wave of protests engulfing nearly the entire Arab world, a sudden spread of protests inundated Brazil. Clearly, these were not protests that were planned weeks in advance. Therefore, our models had to quickly adapt to the rapidly changing political landscape with government forces interacting with the protesters, sometime quelling, other times adding fuel to the fire through their highhanded reactions. We have presented our forecasts in Figure 4. The demonstrations originally started in protest of rising bus prices in June. We have also presented our results of Venezuelan protests of February 2014 in Figure 5. The map of geolocation for the Venezuelan protests is shown in Figure 6.

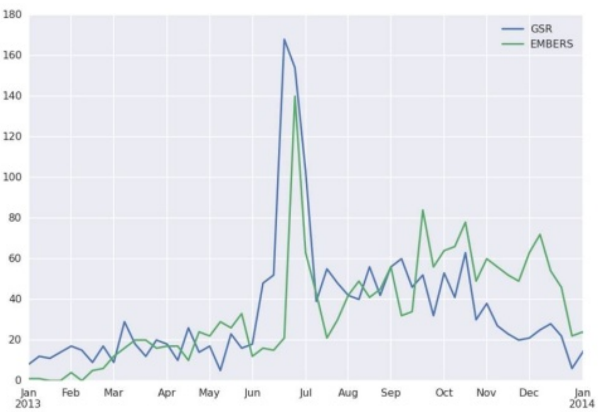

Figure 4. Forecasting performance during Brazil Spring

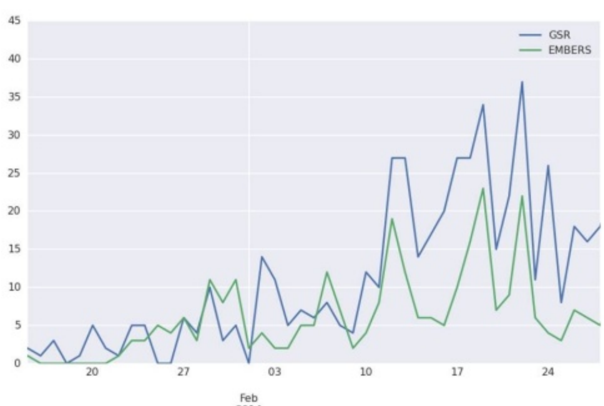

Figure 5. Forecasting performance in Venezuela 2014

Figure 6. Geolocation of Venezuelan Protests

VI. DISCUSSION: A BRAVE NEW WORLD OF FORECASTING AND CONTROL?

The rise of social media has allowed us to break down the barriers of geographic space and to create virtual communities of like-minded people. This trend has simultaneously united and divided people all over the world. Our current effort aims at detecting these trends to forecast incidents of civil strife in real time. Similar to any new innovation, this emerging methodology has its obvious downside; it carries the risk of being a tool of state oppression. Authoritarian regimes all over the world are trying to find ways to control their citizens by manipulating Internet conversations, particularly in the aftermath of the "Arab Spring." Given the newness of the technology, the ethical implications of all of these issues are still evolving. However, there are some important considerations related to forecasting acts of civil strife. First, in light of recent events there is a heightened awareness of privacy issues surrounding developing surveillance capabilities, even when they involve publicly available information.Yet, our project does not use any information that is not publicly available. Our project does not use any information that is not publicly available. Our project has developed a completely automated system based on 10 Latin American nations that begins with monitoring social media and ends with generating warnings in one smooth loop.Second, the genie is out of the bottle. As new modes of social media spring up, their use spread throughout the world and algorithms are perfected to monitor the conversations-whether for public policies or for private commercial gains - there is simply no effective way of stopping these monitoring efforts.

There is little doubt that we are witnessing a brand new world of information processing and trans-discipline inquiries. Although we can clearly see that we have developed the capability for real-time forecasting of incidents of civil strife, we must realize that our efforts at forecasting events is only in the short term. As for the long-term course of a mass movement, we must agree with Collier [1] ; it is impossible to conceive of any algorithm that would be able to do an effective job. Finally, our effort was strictly focused on forecasting and not about finding causal relationships that create political instability. We must use this methodology to gain a deeper insight into this aspect of academic inquiry.

ACKNOWLEDGEMENT

Our project (EMBERS: Early Model-Based Event Recognition using Surrogates) is supported by the Intelligence Advanced Research Projects Activity (IARPA) via DoI/NBC contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints of this work for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the US Government.

References

[1] P. Collier, "Bottom billion," The Wiley-Blackwell Encyclopedia of Globalization, 2007.

[2] L. An, M.-H. Tsou, S. E. Crook, Y. Chun, B. Spitzberg, J. M. Gawron, and D. K. Gupta, "Space–time analysis: Concepts, quantitative methods, and future directions," Annals of the Association of American Geographers, vol. 105, no. 5, 2015, pp. 891–914.

[3] L. An, M.-H. Tsou, B. H. Spitzberg, D. K. Gupta, and J. M. Gawron, "Latent trajectory models for space-time analysis: An application in deciphering spatial panel data," Geographical Analysis, vol. 48, no. 3, 2016, pp. 314–336.

[4] H. Mao, S. Counts, and J. Bollen, "Predicting financial markets: Comparing survey, news, twitter and search engine data," arXiv preprint arXiv:1112.1051, 2011.

[5] E. J. Ruiz, V. Hristidis, C. Castillo, A. Gionis, and A. Jaimes, "Correlating financial time series with micro-blogging activity," in Proceedings of the fifth ACM international conference on Web search and data mining. ACM, 2012, pp. 513–522.

[6] K. Lee, A. Agrawal, and A. Choudhary, "Real-time disease surveillance using twitter data: demonstration on flu and cancer," in Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2013, pp. 1474–1477.

[7] H. Achrekar, A. Gandhe, R. Lazarus, S.-H. Yu, and B. Liu, "Predicting flu trends using twitter data," in Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on. IEEE, 2011, pp. 702–707.

[8] A. A. Aslam et al., "The reliability of tweets as a supplementary method of seasonal influenza surveillance," Journal of medical Internet research, vol. 16, no. 11, 2014.

[9] A. C. Nagel et al., "The complex relationship of realspace events and messages in cyberspace: case study of influenza and pertussis using tweets," Journal of medical Internet research, vol. 15, no. 10, 2013.

[10] M. Olson, The logic of collective action. Harvard University Press, 2009, vol. 124.

[11] H. Tajfel and J. Turner, "The social identity theory of inter group behavior in s worchel & wg austin (eds) psychology of intergroup relations," Chicago: Nelson, 1986.

[12] H. Tajfel, "Differentiations between social groups: Studies in intergroup relations," 1978.

[13] A. K. Sen, "Rational fools: A critique of the behavioral foundations of economic theory," Philosophy & Public Affairs, 1977, pp. 317–344.

[14] D. K. Gupta, Understanding terrorism and political violence: The life cycle of birth, growth, transformation, and demise. Routledge, 2008.

[15] J. M. Gawron, D. Gupta, K. Stephens, M.-H. Tsou, B. H. Spitzberg, and L. An, "Using group membership markers for group identification." in ICWSM, 2012.

[16] P. A. Schrodt, "Automated production of high-volume, near-real-time political event data," in American Political Science Association meetings, 2010.

[17] M. Wick and B. Vatant, "The geonames geographical database," Available from World Wide Web: http://geonames. org, 2012.

[18] http://www.4imn.com/, [Online; accessed 2012].

[19] "Basis Technology Rosette Linguistics Platform," http://www.basistech.com/text-analytics/rosette/.

[20] J. Strötgen and M. Gertz, "Heideltime: High quality rule-based extraction and normalization of temporal expressions," in Proceedings of the 5th International Workshop on Semantic Evaluation, ser. SemEval '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 321–324. [Online]. Available: http://dl.acm.org/citation.cfm?id=1859664.1859735

[21] A. Kimmig, S. Bach, M. Broecheler, B. Huang, and L. Getoor, "A short introduction to probabilistic soft logic," in Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications, 2012, pp. 1–4.

[22] S. Muthiah et al., "Embers at 4 years: Experiences operating an open source indicators forecasting system," in Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ser. KDD '16. New York, NY, USA: ACM, 2016, pp. 205–214. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939709

[23] S. Muthiah, B. Huang, J. Arredondo, D. Mares, L. Getoor, G. Katz, and N. Ramakrishnan, "Planned protest modeling in news and social media." in AAAI, 2015, pp. 3920–3927.

[24] F. Chen, J. Arredondo, R. P. Khandpur, C.-T. Lu, D. Mares, D. Gupta, and N. Ramakrishnan, "Spatial surrogates to forecast social mobilization and civil unrests," in Position Paper in CCC Workshop on "From GPS and Virtual Globes to Spatial Computing-2012, 2012.

[25] D. B. Neill, "Fast subset scan for spatial pattern detection," Journal of the Royal Statistical Society: Series B (Statistical Methodology), vol. 74, no. 2, 2012, pp. 337–360.

[26] J. Cadena, G. Korkmaz, C. J. Kuhlman, A. Marathe, N. Ramakrishnan, and A. Vullikanti, "Forecasting social unrest using activity cascades," PloS one, vol. 10, no. 6, 2015, p. e0128879.

[27] L. Zhao, F. Chen, J. Dai, T. Hua, C.-T. Lu, and N. Ramakrishnan, "Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling," PloS one, vol. 9, no. 10, 2014, p. e110206.

[28] G. Korkmaz, J. Cadena, C. J. Kuhlman, A. Marathe, A. Vullikanti, and N. Ramakrishnan, "Combining heterogeneous data sources for civil unrest forecasting," in Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on. IEEE, 2015, pp. 258–265.

[29] N. Ramakrishnan et al., "'Beating the news' with EMBERS: forecasting civil unrest using open source indicators," in Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2014, pp. 1799–1808.

[30] "Hungarian Algorithm," https://en.wikipedia.org/wiki/Hungarian_algorithm, [Online; accessed 01 July, 2017].