Detecting and Forecasting Domestic Political Crises: A Graph-based Approach

Yaser Keneshloo
Department of Computer Science
Virginia Tech
Blacksburg, VA
yaserkl@cs.vt.edu

Jose Cadena
Virginia Bioinformatics Institute
Virginia Tech
Blacksburg, VA
jcadena@vbi.vt.edu

Gizem Korkmaz
Virginia Bioinformatics Institute
Virginia Tech
Blacksburg, VA
gkorkmaz@vbi.vt.edu

Naren Ramakrishnan
Department of Computer Science
Virginia Tech
Blacksburg, VA
naren@cs.vt.edu

ABSTRACT

Forecasting a domestic political crisis (DPC) in a country of interest is a very useful tool for social scientists and policy makers. A wealth of event data is now available for historical as well as prospective analysis. Using the publicly available GDELT dataset, we illustrate the use of frequent subgraph mining to identify signatures preceding DPCs, and the predictive utility of these signatures through both qualitative and quantitative results.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications–Data Mining

Keywords

GDELT, event forecasting, graph mining, domestic political crises.

1. INTRODUCTION

Predicting and monitoring political events is known to be an important and challenging task in social science research [2]. Of particular interest is forecasting domestic political crises (DPCs), which refer to significant opposition against the government usually triggered by an election or legalizing an unfavorable law [22, 15]. As discussed in [1], forecasting a DPC is an arduous tasks compared to prediction of other types of events such as rebellion and international crises.

Recent times have significantly increased the wealth of resources available to the computational social scientist. Resources such as ICEWS [9] and GDELT [12] span most of the countries of the world and have been used to develop prediction models for a range of events such as international and domestic crises, insurgency, rebellion, and ethic and religion violence [3, 4]. Methods used include discriminant analysis [17], HMMs [16], Bayesian time series forecasting [14, 5, 19, 11], and vector auto regression (VAR) methods [10, 6]. For a survey on these predictive models, we point the readers to [18].

We take a *contrast data mining* approach wherein we seek patterns in interaction graphs that are frequent in situations with DPCs but infrequent in situations without DPCs, and thus discriminative. To the best of our knowledge, our work is the first graph mining analysis of intra-country events from GDELT. Using the features (interaction patterns) extracted, we demonstrate how they are both explanatory for the underlying crises and predictive of future DPCs.

WebSci'14, June 23–26, 2014, Bloomington, IN, USA.

http://dx.doi.org/10.1145/2615569.2615698 .

2. PRELIMINARIES

2.1 The GDELT Dataset

The Global Database of Events, Language, and Tone (GDELT) is a new CAMEO-coded dataset containing geolocated events with global coverage from 1979 to the present [12]. The data are collected from news reports throughout the world. Currently, this dataset provides daily coverage on the events found in news coverage published on that day. The event types in CAMEO taxonomy are divided into four primary classifications: verbal cooperation and material cooperation, which are represented by numbers 1 to 10, and verbal conflict and material conflict, which are represented by numbers 11 to 20. Moreover, there are 32 different roles for the actors in each event, e.g., Police Forces, Government, and Military.

In GDELT, each record captures information pertaining to a specific event. To generate our models, we use the following attributes from an event: MonthYear, Actor1Type, Actor2Type, RootEventCode, AvgTone, and GoldsteinScale, where Actor1Type and Actor2Type store the role of the actors participating in the event, RootEventCode $\in \{1, \ldots, 20\}$ identifies whether this event is cooperative or conflicting, AvgTone is a subtle measure of the importance of an event and plays as a proxy for the impact of that event, and the

Primary Role Codes	Description
COP	Police forces, officers
GOV	Government: the executive, governing parties, coalitions partners
JUD	Judiciary: judges, courts
MIL	Military: troops, soldiers, all state-military personnel
OPP	Political opposition: opposition parties, individuals, activists
REB	Rebels: armed and violent (non-state) groups, individuals
SPY	State intelligence, secret service
Secondary Role Codes	**Description**
BUS	Business: businessmen, companies, etc.
CVL	Civilian individual or group
EDU	Education: educators, schools, students
ELI	Elites: former government officials or celebrities
LAB	Labor: workers, unions
LEG	Legislature: parliaments, assemblies, "lawmakers"
MED	Media: journalists, newspapers, television stations,etc.
REF	Refugees
Tertiary Role Codes	**Description**
MOD	Moderate: "moderate," "mainstream," etc.
RAD	Radical: "radical," "extremist," "fundamentalist," etc.
UAF	Armed forces that cannot be identified as MIL, COP, or REB

Figure 1: Actors defined in the CAMEO codebook.

GoldsteinScale captures the impact of this event on stability of a country.

2.2 Modeling Domestic Interactions

We define an *interaction graph* of CAMEO event-types involving CAMEO actors. Let $G = (V, E, l, w)$ denote an undirected, labeled multigraph. The set of nodes V represent CAMEO actors; each node is given a distinct actor code (label) by the *node label function* l. The set of edges E represent the CAMEO interactions (events) between actors. The *edge label function* w assigns a label to an edge e corresponding to the type of an interaction.

We construct the interaction graph from a collection of entries in the GDELT dataset. For example, a record indicating a "Demand" interaction (type 10) between government (GOV) and refugees (REF) is represented in the interaction graph as an edge $e = (u, v)$ between the nodes with labels $l(u) =$ GOV and $l(v) =$ REF; the label of this edge is $w(e) = 10$. In this paper, we use *monthly* interactions graphs; for a given country, we compile the entries in GDELT for one month and construct the graph as described above.

We now present several definitions that will be used in the detection and forecasting tasks of Section 3:

Support. Let G_s and G be two graphs, $D = \{G_1, \ldots, G_N\}$ be a collection of graphs, and let $G_s \sqsubseteq G$ denote that G_s is a subgraph of G. We define the support of G_s in dataset D, denoted as $supp(G_s, D)$, as the number of graphs $G \in D$ for which $G_s \sqsubseteq G$. In other words, $supp(G_s, D) = |\{G \in D \mid G_s \sqsubseteq G\}|$

Frequent Subgraph. Given a collection of graphs $D = \{G_1, \ldots, G_N\}$ and a threshold value $\theta \in (0, 1]$, a graph G_s is *frequent* if it is a subgraph of at least $\theta \times N$ graphs in D, or, equivalently, $supp(G_s, D) >= \theta N$.

Subgraph Matching. If a graph S is isomorphic to at least one subgraph G_s of G, then G_s is a *match* of S in G.

3. PROPOSED METHODS

We hypothesize that the interactions between specific actors of a country are important indicators of a DPC in that country. For instance, if there are a high number of conflicts between the government and civilians, the country is likely to experience a DPC imminently or in the near future. Furthermore, interactions between actors during a DPC should be different from interactions in periods of peace. The methods presented below are motivated by this idea.

3.1 Classifying and Detecting DPCs

We pose the problem of detecting DPCs as a classification task. Given the interaction graph G of a country for some period of time t, we use a subset of the subgraphs of G to classify t as *DPC* or *non-DPC*. Formally, let $X = \{x_1, x_2, \ldots, x_n\}$ be a set of multigraphs; the nodes of a graph in X are a subset of the CAMEO actor codes, and its edges represent CAMEO interactions. We refer to X as the *feature set*. Let G^t be an interaction graph corresponding to a period of time t, and let G^t_X be a vector of length n, where the the i^{th} entry in G^t_X is 1 if $x_i \in X$ is a subgraph of G^t, and 0 otherwise. We call G^t_X the *feature vector* of G^t. Our task is to find a *detection function* f that indicates whether a feature vector corresponds to a period of DPC or non-DPC; that is $f : G^t_X \rightarrow \{\text{DPC}, \text{nonDPC}\}$.

A key step in finding a good detection function is to find features, i.e. subgraphs, that appear frequently in interaction graphs and, at the same time, are discerning enough to separate DPC graphs from non-DPC ones. We separate the monthly interaction graphs in our dataset into two groups, D_+ and D_-, representing DPC and non-DPC graphs, respectively. We then find the frequent subgraphs in each dataset, F_+ and F_- using the gSpan algorithm [23]. Since we are interested in finding the most discriminative features for the classification task, we ignore all the subgraphs that are common between F_+ and F_-, thus obtaining a discriminative feature set $DFS = \{F_+ \cup F_-\} - \{F_+ \cap F_-\}$, which we use for classification. As explained in Section 4, at this point in our process, we find that, in most cases, the intersection between F_+ and F_- is very small. This means that actor interactions are very different on months with and without DPC, and the set of frequent subgraphs in the two groups are promising discriminator features for classification.

After obtaining the DFS, we can compute the feature vector of a graph using a subgraph matching algorithm [24, 25, 20]. We use a simplified version of the TreeSpan algorithm [25] to find the exact matching in this paper. Once we have the feature vectors for all graphs in our dataset, we train different classification algorithms to obtain the detection function.

3.2 Forecasting a Domestic Crisis

We now turn to the task of predicting DPCs using the interaction graph. To this end, we develop various regression models that estimate the probability of a DPC occurring in the near future. We employ the LASSO methodology (Least Absolute Shrinkage and Selection Operator) [21]. Like a regular linear regression, LASSO minimizes the sum of squared errors, but with an added constraint on the sum of the absolute values of the coefficients. We use LASSO over a standard linear regression in order to encourage a sparse representation; that is, we are interested in reducing the original feature set, as some of the initial graph properties are expected to be redundant. We develop three types of LASSO-based logistic regression models that use (i) event counts (M_Event), (ii) graph properties (M_Graph), and (iii) features from both M_Event and M_Graph (hybrid model designated as M_Event_Graph).

Event Counts: In this baseline model, we use the monthly counts of each event type in each country as explanatory variables. Moreover, we include the average AvgTone and the average GoldsteinScale associated with these events. Formally, the regression model estimates DPC_t, the probability of a DPC at time t, as

$$DPC_t = \sum_{i=1}^{20}(\alpha_i E_{it-1} + \beta_i T_{it-1} + \gamma_i G_{it-1}) + DPC_{t-1} \quad (1)$$

In the above equation, E_i is the counts of events of type i, T_i is the average AvgTone, G_i is the average GoldsteinScale for the event type i and DPC_t is the dependent variable. We also use the lagged value of DPC in all of the regression models, since DPCs can persist over consecutive months.

Graph Properties: In this regression model, we use graph-based features. We compute structural properties of the interaction graphs: Total Number of Edges (*TotEdge*), Average Weighted Degree (*AvgWDeg*), Diameter (*Diam*), Number of connected components (*Comp*). In addition, we calculate the weighted degree of each actor and their centrality, based on different measures, such as betweenness (*betwCen*), closeness (*closCen*), and degree (*degCen*). The model estimates DPC_t as

$$\begin{gathered} DPC_t = \textstyle\sum_{i\in V} \alpha_i degCen_{it-1} + \beta_i betwCen_{it-1} + \\ \gamma_i closCen_{it-1} + \omega_1 AvgWDeg_{t-1} + \omega_2 Diam_{t-1} + \\ \omega_2 Comp_{t-1} + \omega_3 TotEdge_{t-1} + DPC_{t-1} \end{gathered} \quad (2)$$

The summation is over all the nodes (actors) of the interaction graph.

Hybrid model: In this model, we combine the features of the event count model and the graph-based model. The performance and predictive power of each model are evaluated in Section 4.

4. EXPERIMENTS

Our experiments are designed to address the following questions:

- Are interactions in a country different during "normal" times and during a DPC? Can we capture this difference and use it to detect DPCs? (Section 4.2)
- How adept are graph-based properties at forecasting DPCs in a country? (Section 4.3)
- Are graph-based models for DPC detection and forecasting better than a history-based approach or a vanilla event count approach? Is there value in combining features from different models? (Sections 4.2 and 4.3)

4.1 Data

We evaluate our methods using GDELT interaction graphs from five countries: Brazil, Colombia, Mexico, Argentina, and Venezuela. The data are collected from January 2003 to the December 2013. Thus, for each country, we have 132 monthly interaction graphs. Table 4.1 shows the number of DPCs in each country for the period mentioned above. GDELT does not include information about DPCs. As ground truth for our experiments, we used the similarly motivated ICEWS dataset, which includes information on whether or not there was a DPC in a country in a given month. We note that our proposed methods exhibit quantifiably good detection and predictive power in spite of using features from one dataset and supervisory labels from another.

Table 1: Number of DPC months in the 132 months of our experiment on different countries

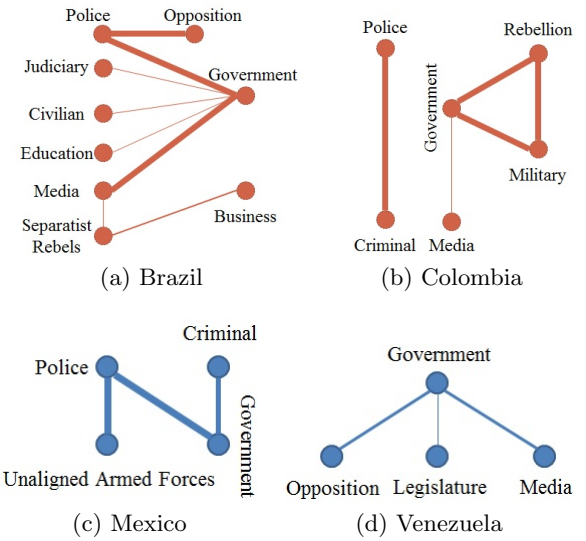
Country	# of Months with DPC (out of 132)
Brazil	6
Argentina	76
Mexico	10
Venezuela	36
Colombia	3

4.2 Classifying DPC events

We use the gSpan algorithm to find frequent subgraphs in F_+ and F_-. For our experiments, we set the threshold parameter in gSpan to obtain a number of features ranging from 500 to 1,000. Figure 2 represents the top frequent subgraphs associated with DPCs in Brazil, Colombia, Mexico, and Venezuela. The thick edges represent more adversarial interactions (event types 11-20) whereas the thin edges represent cooperative interactions (event types 01-10). The figure shows that, in Colombia, conflicts between Rebellion, Military, and Government are frequent during domestic crises. This graph corresponds to the real-life political tension between the Colombian government and the guerrilla groups in the country. On the other hand, during political crises in Brazil, the actors involved in conflict are Government, Police, Media, and the Opposition. Actors involved in a DPC and the respective interactions vary across different countries. For instance, Goverment and Media seem to engage in more conflict in Brazil than they do in Colombia or Venezuela. Understanding the role of each actor during a DPC requires a thorough analysis of the social and political aspects of each country and is beyond the scope of this paper.

For classification, we use algorithms from the LibSVM [7] and LogitBoost [8] libraries. We compare the performance of the frequent subgraph approach to a baseline model that does not take into account event types (ignoreEventType). In the baseline approach, all graphs are unlabeled which gives us a different set of frequent subgraphs. Our model and the baseline are evaluated using the Area Under Curve (AUC) metric and the Matthews Correlation Coefficient (MCC) measure. MCC is a quality metric for binary classification in unbalanced datasets; its range is $[-1, 1]$, where 1 indicates perfect classification, -1 indicates inverse classification, and 0 represents a random classifier.

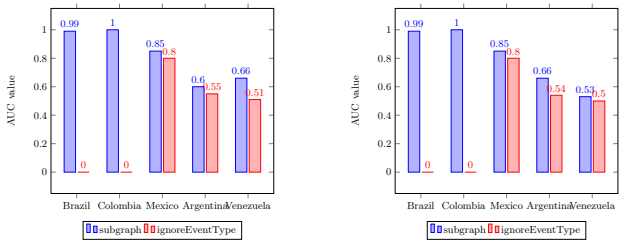
Figures 3 and 4 compare our proposed method to the baseline based on the AUC and MCC metrics. We note that, for Brazil and Colombia, the gSpan algorithm could not find the appropriate number of subgraphs for the baseline method, and it runs out of memory; this is the case even if we run a parallel version of the algorithm described in [13]. In every case in the figures, the frequent subgraph approach beats the baseline; the difference is more noticeable when we compare both methods using the MCC metric. For instance, we can see that although we have only three months of DPC for Colombia, the proposed method is able to classify all the graphs in dataset. Moreover, this illustrates the importance



(a) Brazil (b) Colombia

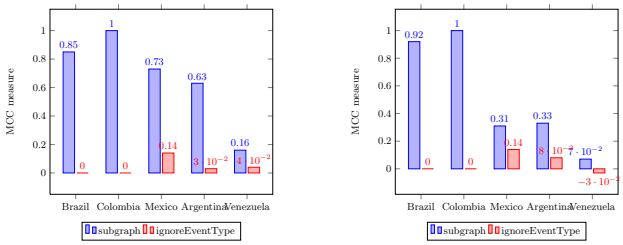
(c) Mexico (d) Venezuela

Figure 2: Frequent subgraphs for (a) Brazil, (b) Colombia, (c) Mexico, and (d) Venezuela during DPCs. Thicker edges represent more adversarial interactions.



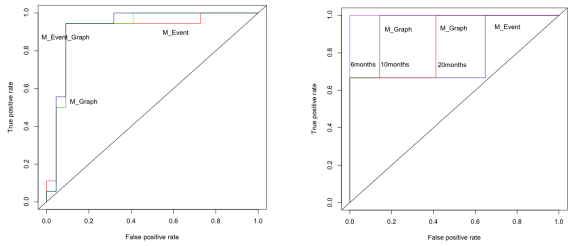
(a) LogitBoost (b) LibSVM

Figure 3: Classifying DPCs (AUC values).

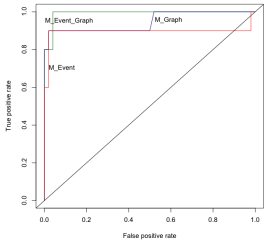


(a) LogitBoost (b) LibSVM

Figure 4: Classification results (MCC measure).



(a) Argentina (b) Mexico



(c) Venezuela

Figure 5: ROC curves for the DPC prediction models. The combination of graph features and event counts outperforms each individual model.

of features that are selected for the classification and solidifies our claim about the effect of actors' interactions on the instability of a country.

4.3 Forecasting DPC events

We evaluate the three regression models described in section 4.2. We focus on Argentina, Mexico, and Venezuela because these countries have a sufficient number of DPCs to train the models. When the distribution of DPC to non-DPC samples is not even, LASSO puts too much weight on the non-DPC months, resulting in a low-performance model. In order to deal with this issue, we focus on the months in which the countries suffered DPC and take the preceding and subsequent months as the training set. (A more systematic approach would be to adopt a classifier specifically meant for imbalanced classes, but our goal here is to explore the utility of graph features using basic machine learning methods.) In the training period (90 months), Argentina experienced 44 months of DPC, compared to 18 in the test period (42 months). For Venezuela, we use a smaller training period —70 months— in order to have a balanced number of DPCs in training and testing data (11 and 10 events, respectively). Finally, Mexico had only 10 months of DPC in 130 months, 3 of which are in the last 3 months of the dataset. Therefore, we only use 45 months to train the model. In this case, we run experiments with 20 months, 10 months and 6 months as the test set to evaluate the performance of the model over different periods. Figure 4.3 illustrates the performance of the baseline, graph-based and hybrid models for Argentina, Mexico and Venezuela. We observe that the hybrid model outperforms the individual models for all countries. The hybrid model for Argentina and Venezuela result in very high precision (around 0.89 for both), and accuracy reaches 0.92 for Argentina and 0.95 for Venezuela. As we have discussed above, we try different test periods for Mexico and we observe a significant improvement as the number of test days decrease. The precision gets as high as 1, and the accuracy 0.95. When the test period is the last 6 months in our dataset (i.e. July 2013 to December 2013), the model predicts the DPC of the last 3 months and the absence of DPC in the previous months perfectly.

5. DISCUSSION

We have introduced the problem of forecasting DPCs in a given country using graph features. Future work will focus on three aspects. First, we aim to situate our approach

in a temporal context so that frequent discriminative subgraphs can be viewed in terms of their evolution over time. Second, we seek to create subgraph features using compositions of basic CAMEO codes, to improve the expressiveness of discovered patterns. Finding such patterns without overwhelming computational complexity is a key issue here. Finally, we aim to develop a maximum entropy modeling of interaction graph evolution so that we can aim to model not just crises but surprising geopolitical developments in general.

Acknowledgements

Supported by the following grants: DTRA Grant HDTRA1-11-1-0016, DTRA CNIMS Contract HDTRA1-11-D-0016-0010, NSF ICES CCF-1216000 and NSF NETSE Grant CNS-1011769. Also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center (DoI/NBC) contract number D12PC000337, the US Government is authorized to reproduce and distribute reprints 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.

6. REFERENCES

[1] B. Arva, J. Beieler, B. Fisher, G. Lara, P. A. Schrodt, W. Song, M. Sowell, and S. Stehle. Improving forecasts of international events of interest. In *European Political Science Association*, volume 78, 2013.

[2] E. E. Azar. The conflict and peace data bank (COPDAB) project. *Journal of Conflict Resolution*, 24:143–152, 1980.

[3] B. E. Bagozzi. Forecasting civil conflict with zero-inflated count models. *Manuscript, Pennsylvania State University*, 2011.

[4] B. E. Bagozzi. Modeling Two Types of Peace The Zero-inflated Ordered Probit (ZiOP) Model in Conict Research. *Journal of Conflict Resolution*, 58, 2014.

[5] P. T. Brandt and J. R. Freeman. Advances in Bayesian time series modeling and the study of politics: Theory testing, forecasting, and policy analysis. *Political Analysis*, 14(1):1–36, 2006.

[6] P. T. Brandt, J. R. Freeman, and P. A. Schrodt. Real time, time series forecasting of inter-and intra-state political conflict. *Conflict Management and Peace Science*, 28(1):41–64, 2011.

[7] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[8] J. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The Annals of Statistics*, 28(2):337–407, 2000.

[9] D. J. Gerner, P. A. Schrodt, R. A. Francisco, and J. L. Weddle. Machine coding of event data using regional and international sources. *International Studies Quarterly*, pages 91–119, 1994.

[10] J. S. Goldstein. A conflict-cooperation scale for WEIS events data. *Journal of Conflict Resolution*, 36(2):369–385, 1992.

[11] G. R. Harris. Regime switching vector autoregressions: a Bayesian Markov chain Monte Carlo approach. *Conflict Management and Peace Science*, 28(1):41–64, 2011.

[12] K. Leetaru and P. Schrodt. GDELT: Global Data on Events, Language, and Tone, 1979-2012. In *International Studies Association Annual Conference*, 2013.

[13] T. Meinl, I. Fischer, and M. Philippsen. Parallel Mining for Frequent Fragments on a Shared-Memory Multiprocessor-Results and Java-Obstacles. *Lernen, Wissen, Adaption*, 2005.

[14] J. M. Montgomery, F. M. Hollenbach, and M. D. Ward. Improving predictions using ensemble Bayesian model averaging. *Political Analysis*, 20(3):271–291, 2012.

[15] S. P. O'brien. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12(1):87–104, 2010.

[16] P. A. Schrodt. Pattern recognition of international crises using hidden markov models. *Political complexity: Nonlinear models of politics*, pages 296–328, 2000.

[17] P. A. Schrodt and D. J. Gerner. Empirical indicators of crisis phase in the Middle East, 1979-1995. *Journal of Conflict Resolution*, 41(4):529–552, 1997.

[18] P. A. Schrodt, J. Yonamine, and B. E. Bagozzi. Data-based computational approaches to forecasting political violence. In *Handbook of Computational Approaches to Counterterrorism*, pages 129–162. Springer, 2013.

[19] S. M. Shellman. Time series intervals and statistical inference: The effects of temporal aggregation on event data analysis. *Political Analysis*, 12(1):97–104, 2004.

[20] Y. Tian, R. C. Mceachin, C. Santos, J. M. Patel, et al. SAGA: a subgraph matching tool for biological graphs. *Bioinformatics*, 23(2):232–239, 2007.

[21] R. Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

[22] M. D. Ward, N. W. Metternich, C. Carrington, C. Dorff, M. Gallop, F. M. Hollenbach, A. Schultz, and S. Weschle. Geographical Models of Crises: Evidence from ICEWS. 2012.

[23] X. Yan and J. Han. gSpan: Graph-based substructure pattern mining. In *Proceedings of IEEE International Conference on Data Mining*, pages 721–724, 2002.

[24] S. Zhang, S. Li, and J. Yang. GADDI: distance index based subgraph matching in biological networks. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 192–203. ACM, 2009.

[25] G. Zhu, X. Lin, K. Zhu, W. Zhang, and J. X. Yu. Treespan: efficiently computing similarity all-matching. In *Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data*, pages 529–540. ACM, 2012.