Discovering Bilateral and Multilateral Causal Events in GDELT

Lei Jiang [*] and Fan Mai

[1] Tapjoy Inc., San Francisco, CA 94104
[2] Department of Sociology, University of Virginia
Charlottesville, VA 22093 USA

Abstract. GDELT is one of the richest sources of event data and the hidden information could be potentially of tremendous value for its real-world reflections. While the majority of past research on these data sets puts emphasis on featured event series alone, such as Arab spring [2] and violence in Afghanistan [3], this paper is focused on exploiting the causality between multiple types of events in different countries. In other words, for example, the actual impact of an event would result in responses from other parties, or even an escalation that more events will come thereafter. With the causal link in real-world context, time series analysis is performed on GDELT data to discover such causal links: we start from an exploration of bilateral and multilateral events in certain countries, where a strong correlation is shown; then Granger test is systematically conducted in the large-scale data set to help to identify causality in a rigorous manner. Further, with the addition of data in related parties, a scenario of prediction task is tested to show enhanced predictability as a validation of this methodology.

Keywords: GDELT, Granger causality, bilateral events, time series analysis, prediction

1 Introduction

GDELT (Global Database of Events, Language and Tone) [1] provides up-to-date real-world events information with its richness in covering all countries globally, providing precise geographic event information along with actor categorization and its timespan in a quarter-century (1979 to present). Preprocessing of raw events, which are from media sources, has been already done by representing all the event types and actors using a standardized code set. The abstraction actually removes much domain-specific information and thereby makes a systematic analytical methodology become available. In a high level, the GDELT data appears as spatial time series: a count can be obtained as the value at each time point for different geographic units: district, city, state or country. Based on this format, many prediction tasks can be derived, and existing work already digs into this direction: Yonamine [3] uses autoregressive integrated moving average

---

model to make time series predictions for future levels of violence in Afghanistan districts;in [4], Brandt and Freeman discuss the effect of filtering GDELT time series, where the result shows filtering can't make an autoregressive fitted model better.

Considering more attributes in the data set, event and actor types are also helpful contexts for prediction tasks. In this data challenge, we illustrate a relatively new way for predictions by utilizing auxiliary information to dig out from data. More specifically, with the time series of bilateral and multilateral event frequencies extracted, the causality between events, represented by whether one time series is useful in forecasting another (Granger causality [5]), is examined between two and more parties. Using this method, the traditional time series prediction is enriched to a multivariate regression problem and the accuracy is thereby boosted.

In this paper, we first give an exploratory analysis to show intuitive results from GDELT data, illustrating the potential from its data in 2. Then, Granger causality is formally introduced to the context and described for its application to discovering bilateral and multilateral events 3. After showing some interesting measurements about causal event links, we construct predictive tasks for a validation of the methodology and demonstrate our results in 4. Lastly, we conclude the paper in 5.

2   Event Time Series, Correlation and Causality: An Exploration

The event time series is certainly an important piece of interest for researching GDELT data. While it is more about inter-relationship between time series in this paper, plotting the bilateral events, where the two actors represent two different countries, can give some intuitive insight. In Figure 2, left shows the relationship of event frequency between China and United States, each of which as primary actor respectively: as the blue and red curves highly overlap with each other, a strong correlation is easily seen; while correlation is bidirectional, the right interestingly shows events performed by the same primary actor, United States, but towards Japan and Korea: they share some trending as in time series, however, there are a larger amount of events for Japan at most of the time.

Correlation can partly describe some features: US-China and China-US events share a high correlation coefficient (Pearson's r) 0.928, but US-Japan and US-Korea can only achieve 0.208. Within the context of correlation, a symmetric matrix can thereby derived for the frequency of events from United States to each country in Table 1 (using data of 2000)

A close-to-zero correlation coefficient would not be of much help. Thus, with a question in mind: how to make use of the seemingly relationship between event frequencies and unleash the potential for predictive analysis, it brings in a unidirectional measurement: Granger causality [5].
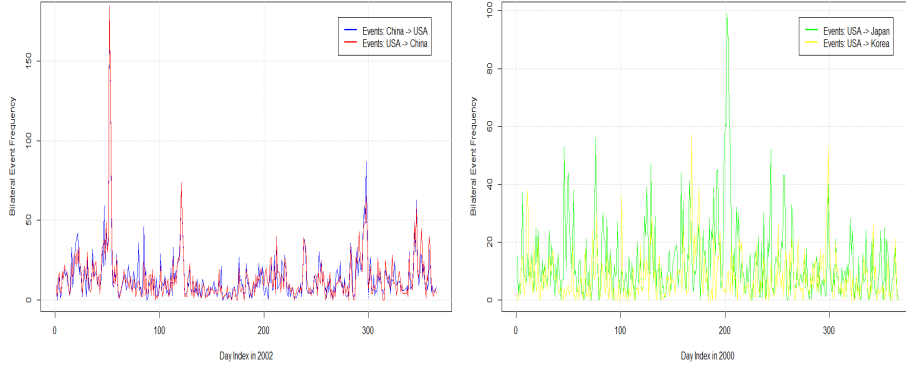
Fig. 1. left: The frequency of events between China and United States (each as primary actor) in 2002; right: Events acted by United States to Japan and Korea respectively in 2000

Table 1. Pearson's r for US-output event frequency to the five countries

	Brazil	China	Korea	Japan	Russia
Brazil	1.00	-0.0352	-0.0488	0.0346	0.0377
China	-0.0352	1.00	0.128	0.0741	0.0826
Korea	-0.0488	0.128	1.00	0.208	0.134
Japan	0.0346	0.0741	0.208	1.00	0.102
Russia	0.0377	0.0826	0.134	0.102	1.00

3   Granger Causality for Bilateral and Multilateral Events

In a real-world global network, the diplomatic behavior between countries will make an impact in the world, sometimes especially important for neighbor countries or those joining the same union. Thereby, there could be subsequent events from other countries. Such a chain reaction makes Granger causality explainable. Also, there exist both bilateral and multilateral events: in practical data mining, bilateral events can be discovered first, and then multilateral events are further derived as an optimization.

3.1   Bilateral events

Intuitively, two time series can have a link, from one to another, whenever causality can be found. In other words, one time series is helpful for forecasting the other one. Granger causality is examined towards such a statistical hypothesis: the following two equations are to be fitted and an F-test (or t-test) will go through.

$$y_t = a_1 y_{t-1} + a_2 y_{t-2} + ... + a_i y_{t-i} + \sigma \tag{1}$$

$$y_t = b_1 y_{t-1} + b_2 y_{t-2} + ... + b_i y_{t-i} + b_{i+1} x_{t-m} + ... + b_{i+n-m+1} x_{t-n} + \sigma \quad (2)$$

While x is a variable from another time series, the level of "helpfulness" of x would determine the bivariate causality based on a computed p-value from regression-based F-value. From GDELT data, we select 15 countries and conduct a search for pairwise *causal links*: with a p-value of 0.05, either "$A \to B$" or "$B \to A$" below p-value in a Granger test is considered a link (if both values are below, the smaller one is chosen). Then, a causality-graph can be constructed based upon links. In Table 2 and 3), the basic connectivity analysis is conducted for both 2000 and 2001 data: an "out-degree", for its real-world metaphor, could mean a country is an initiator for its bilateral activities, and "in-degree" means a reaction or response contingent to the cause.

Table 2. A causality-graph connectivity analysis for 2000 between 15 countries

	China	France	Russia	United Kingdom	United States
Out-degree	4	4	5	5	5
In-degree	6	5	8	4	5
	Japan	Korea	Germany	Brazil	North Korea
Out-degree	4	5	5	2	0
In-degree	5	4	5	0	5
	Afghanistan	Iran	Iraq	Saudi Arabia	Syria
Out-degree	4	8	3	4	6
In-degree	2	3	5	2	5

In general, the number of degrees could represent activeness. While we don't choose to visualize the actual graph (without misleading it to real country-country relationships from a 15-node sample network), it could show that from 2000 to 2001, the United States becomes more proactive (from (5,5) to (8,3), for a possible reason of the 911 tragic event, and other countries seem to be more responsive to that in 2001 (e.g. China, United Kingdom, Korea make the transition with more in-degrees, and that doesn't affect the pattern of Iran).

Table 3. A causality-graph connectivity analysis for 2001 between 15 countries

	China	France	Russia	United Kingdom	United States
Out-degree	5	3	5	1	8
In-degree	8	6	7	8	3
	Japan	Korea	Germany	Brazil	North Korea
Out-degree	4	2	4	6	5
In-degree	6	8	6	1	4
	Afghanistan	Iran	Iraq	Saudi Arabia	Syria
Out-degree	9	8	6	7	2
In-degree	3	3	5	3	4

3.2   Causal links and multilateral events

Following the bilateral causal links in last subsection, the discovery is extended to connecting more than two parties with causality. Practically, adding one more party to a combinational search scenario can add exponent-level computation complexity due to *the curse of dimensionality*. On the other hand, inserting more and more variables to Granger test would result in overfitting in a non-noise-free data setting. Thereby, multilateral events are considered to be an enhancement from a bivariate link, where transitive relation stands in this context.

The search algorithm is adopted in the following routine:

Step 0: Preprocessing raw GDELT data to extract bilateral events time series;

Step 1: Searching for causal links rooted from a certain country X (with an out-degree from X) by comparing all the possible pairs (A, B) such that event series "$X \to A$" and "$X \to B$' form a causal link; Then in the same way, seek for (A, B) such that event series "$A \to X$" and "$B \to X$" could be a causal link. At this stage, three-party relationship is found;

Step 2: Combining triangle relationships in pairwise causal links: e.g. from "$X \to A$" to "$X \to B$" there is a causal link, while it is also a caulsal link from "$X \to C$' to "$X \to B$"; Then it is considered that both "$X \to A$" and "$X \to C$" can help with predicting "$X \to B$" (four parties are involved).

It is worth noting that a multivariate Granger test for such multilateral events is conceivable. However, as the aforementioned reason about overfitting, the hypothesis testing here may not be performed in the same way as bivariate Granger test (p-value may not be comparable at this time. That is why multilateral causal links are detected using a heuristic approach at this stage.

Using GDELT data ranging from 2000 to 2005, we are able to exploit all the bivariate causal links that cover three parties. Figure 2 shows the change. As the extracted data over the 15 countries could be considered as a sample network, where the total number of possible links remains the same all the time, the variation over years actually indicates network *assortativity*: when more links emerge, it means there are increasing real-world cooperations being organized and produces some well-defined network structure. For an individual country, the number of causal links could also mean the level of influence or activeness: United States and China almost both follow the global trend (United States is more active with a larger number though), but that is not the pattern of Russia.

4   Predictive Analysis

From the perspective of data mining, the causal links are meant to serve predictive analysis, where the additional information from data could be utilized.
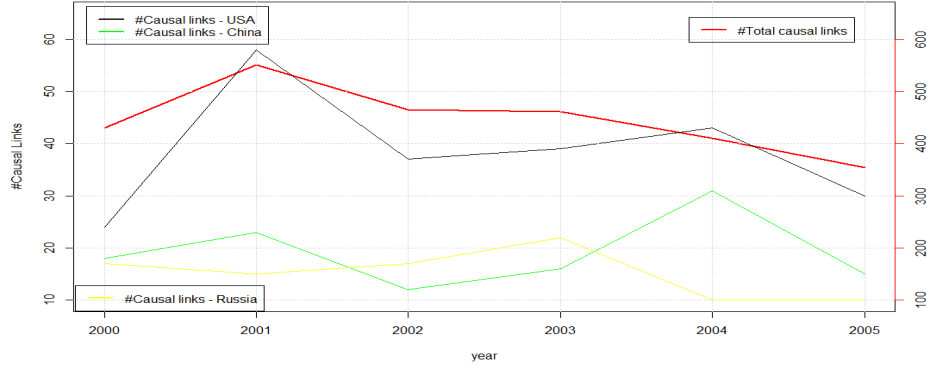
Fig. 2. Trend of number of causal links in the sample network from 2000-2005

In this section, the enhanced predictability is shown through a comparison, between traditional time series forecasting and with the help of auxiliary time series (empowered by pre-trained causal links).

When constructing regression tasks, 3-party and 4-party causal links, mentioned in the last section, are both covered in our experiment, where one and two auxiliary data sources are tested respectively.

To illustrate the advantage, the 2001 data is used, where the regression task is towards the bilateral event frequency between United States and another country (United States-Afghanistan, United States-North Korea, and United States-Saudi Arabia). A linear regression[1] is used here to fit the model: first as traditional time series prediction, it uses only the past time points (3 past time steps) of the targeted time series to predict the next; then, when auxiliary time series from other countries is brought in, the corresponding time points also participate in the fitted model. Fitted *sum of squares* (SSQ) becomes the criteria as in Table 4.

In each case , the listed two countries, as in column "One auxiliary time series 1" and "One auxiliary time series 2", are both used for the scenario of "Two auxiliary time series". Such auxiliaries, except the last one of predicting USA-Japan, all appear in multilateral causal links, identified to be helpful for the prediction task. For the last case, auxiliaries appear as the other side of causality link (in-degree), but still gain a positive effect.

---

[1] Note: the orignal regression model itself could be more sophisticated to achieve better gross accuracy. However, the claim here is that causal links can help to reduce error, no matter what actual regression model is used.

Table 4. Fitted sum of square for 2001 United States bilateral event frequency

	Time Series ($lag = 3$)	One auxiliary time series 1
USA-AFG	78369.78	77691.52/DEU
	One auxiliary time series 2	Two auxiliary time series
USA-AFG	77367.96/KOR	76883.92
	Time Series ($lag = 3$)	One auxiliary time series 1
USA-PRK	24314.52	24167.58/DEU
	One auxiliary time series 2	Two auxiliary time series
USA-PRK	24124.1/KOR	23965.08
	Time Series ($lag = 3$)	One auxiliary time series 1
USA-SAU	12890.06	12446.62/AFG
	One auxiliary time series 2	Two auxiliary time series
USA-SAU	12856.84/IRN	12424.87
	Time Series ($lag = 3$)	One auxiliary time series 1
USA-JPN	106580.1	103334.4/SAU
	One auxiliary time series 2	Two auxiliary time series
USA-JPN	103288.2/PRK	100050.2

5   Conclusion and Future Work

In this paper, we illustrate our insight for the GDELT database through a broad range of analysis, including exploratory analysis, Granger causality test, searching for pattern and predictive analysis. The current results show that the causal links hidden in data set, with a real-world metaphor of bilateral and multilateral causal events, could have a profound effect in deeply analyzing the activities and trends over different countries of the world. While the authors is new to GDELT data set via the data challenge, there are the following conceivable directions for future work: *i)* Using event type information to construct finer-level models; and *ii)* Developing larger-scale macine learning pipeline to construct rigorous model with iterative tuning. All the related source code (in R and Python) for this data challenge is available[2].

References

1. Leetaru, Kalev and Schrodt, Philip, "GDELT: Global Data on Events, Language, and Tone, 1979-2012" in *International Studies Association Annual Conference*, April 2013. San Diego, CA.
2. Z. Steinert-Threlkeld, "The Arab spring and GDELT", posted at http://blog.gdelt.org/2013/10/02/the-arab-spring-and-gdelt/ , accessed on Jan 11, 2014.
3. J.E. Yonamine, "Predicting future levels of violence in Afghanistan districts using GDELT, " published at http://gdelt.utdallas.edu/data/papers/Predicting_Future_Levels_of_Violence_in_Afghanistan _Districts_using_GDELT.pdf , accessed on Jan 10, 2014.

---

[2] https://github.com/lionelc/GDELT_SBP2014_data_challenge

4. P.T. Brandt and J.R. Freeman, "Why (Not to) Filter or Pre-Whiten the GDELT Time Series?, " published at http://www.utdallas.edu/p̃brandt/Patrick_Brandts_Website/Research_files/WhyFilterGDELT.pdf . Accessed on Jan 10, 2014.
5. C.W.J. Granger, "Investigating causal relations by econometric models and cross-spectral methods," *Econometrica*, 37(3) pp. 424-438.