Causal Discovery in Bilateral and Multilateral Relations from Global Dyadic Events on the Web

Lei Jiang
lionelchange@gmail.com
Microsoft Corporation

ABSTRACT

The ever-increasing amount of web information usually leads to finding implicit relationships by applying data-driven methods. While country-country relations are not exactly new knowledge, we discover the causal links between them as a new perspective, where the causality could be illustrated by "these two countries always stand in the same line" and "how the tension between two certain countries would influence other countries' foreign policy against them". Using the global dyadic events data that record the involving parties and timestamp, we identify the causal structure as a directional acyclic graph (based on conditional independence), and also check with Granger causality that captures signals via autoregression, while unobserved confounders could be the actual reason that is not accounted for in the model.

To formalize the discovery process, we first validate commonsense country-country relations in a small-world network by finding the causal links with longer-term data without any priors, suggesting the traditional allies and rivals as we know in the real world. In particular, the causal links found are calibrated with correlation, in order to remove noise out from sparsity. Then, the short-term effects in both bilateral and multilateral relations are examined: while allies are supposed to be stable, does the level of friendship fluctuate? Experiments are performed on a variety of data granularity and sentiment analysis from actual news texts, in comparison, largely matches with the trends as causal models demonstrate. Our results show that the event time series, combined with lexical-based validation, has potential to be a means to measure and monitor the climate of complicated geopolitics.

CCS CONCEPTS

• **Information Systems**; • **World Wide Web**; • **Web Mining**; • **Web search and information discovery**;

KEYWORDS

causal discovery, global dyadic events, time-series analysis, sentimental analysis

ACM Reference Format:
Lei Jiang. 2019. Causal Discovery in Bilateral and Multilateral Relations from Global Dyadic Events on the Web. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

The evolution of web, regardless of whether the content comes from an authoritative source or an ordinary online user, leads to the increasing capability of discovering latent patterns hidden in the vast amount of data. Country-country relations are traditionally learned from news media: journalists and political scholars collect and analyze information from government releases and other sources, and then write reports as public knowledge. Undoubtedly, bilateral and multilateral relations are complicated in nature. Countries are not equal in terms of economic power and level of participation in international affairs, and the dynamics in global events means the relations could change over time as well. With the news gathering process becoming automated in the web era, it is the interest of data mining to derive such relations given the data of global dyadic events. Then, it comes to how to represent such relationships: we consider the country-country relationships show causality, rather than correlation, in ways like "these two countries always stand in the same line" (in other words, country A's action against country C would cause country B's following action) and "the tension between two certain countries would influence other countries" (due to the escalated tension between country A and B, country C's foreign policy towards them would adjust correspondingly). In this paper, we identify the causal links between countries and validate them using text-based methods.

Causal discovery has been an active topic in statistics and data analysis for decades. Inspired by its wide range of applications in econometrics [2, 4], biomedical [17, 30], social sciences [19, 24] and etc., many approaches emerge as they seek to find the links among multiple variables. Among these, depending on the specific scenario, causal inference can be made *i)* through a pair-wise statistical significance test or *ii)* by searching through a network with subtle relationships in between: e.g. a single link between corporate philanthropy and revenues is up for testing with the null hypothesis "more giving will result in more customer satisfaction, and thus bringing more future sales "; on the other hand, more often in epidemiology, a network of causal links is to be established [5, 20, 32], it is illustrated that the tremendous time complexity makes learning Bayesian networks infeasible for many cases. Meanwhile, it is important to validate the potential causal links found in order to avoid spurious causal effects, and in some cases, differentiating association from causation, as well as identifying instantaneous causality, would play a role in affecting the analytical results.

Meanwhile, discovering causal effects from web data presents several technical challenges:

- Defining causal links is not straightforward in a context with geopolitical entities.
- Low data quality or skewed distribution could result in bias and even lead to false causal relationship.
- Differentiating correlation with causality requires additional steps of validation.

We address these questions in a formal framework that utilizes the Global Database of Events, Location and Tone (GDELT) as well as crawling related Wikipedia web data. First, conceptually, we consider that a *causal link* can be established between two different geopolitical entities over another entity (e.g. the action taken by Russian on Syria would cause United States' reaction). Accordingly, we can extract structured dyadic event counts from GDELT data sets and use PC algorithm [26] as a baseline to exploit all potential pairwise causal links as well as DAGs (directed acyclic graph). Then, having the discovered structures as a set of hypothesis, we further adopt an automated text-based approach to perform the validation process: we scrape Wikipedia pages on bilateral relations to match all the words with dictionaries of sentiment/effect and the frequency-based score becomes a quantifiable measurement for checking each causal link. While the latter step of validation is considered a more reliable approach as it is based on documented words that present a ground truth, a set of hypotheses are tested and verified with regard to the causal links found . This consistency shows the validity in detecting causal relationships in multilateral relations from the dyadic event counts (as a non-text-based method). Further, we examine instantaneous Granger causality which reflects the dynamics in causal links: while Granger causality is prone to sporadic edges where they actually have the same confounder, it is complimentary in our overall methodology and we examine the causality in multiple scales of data granularity.

2 RELATED WORK

Causality, or causation, is built upon a deeper link between events and entities which reveals the intrinsic relationship. More than easy-to-calculate correlation, it takes conditional probability into account and aims to remove spurious links brought by confounding or latent factors [27]. In a canonical framework, the PC algorithm [26], as a constraint-based learning algorithm, performs statistical tests to derive a set of conditional independence and dependence statements: the skeleton, which is an undirected graph, is found by pairwise comparisons and then directions are added to graph with collision removed to ensure its acyclic nature, resulting in a Bayesian belief network. This approach has been extended for high-dimensional space [16] and nonparanormal models [14].

Data on the web naturally carry the attribute of graph structure from the content network in the background, so it contains causal relationships while it's not easy to verify from observational data. In our study, we treat the geopolitical entities in the world as a social network and the causal relationship between them can be interpreted as peer effects. In economics and statistics, this is a familiar subject [13, 31]. Peer effects can be formulated as a type of causal inference, and the treatment effects can thereby measured through randomization design, where a linear model is trained using parametric methods. Recently, the broad range of network effects continue to be actively researched [1, 6]. [25] also uses GDELT data set to model Bayesian Poisson tensors for predictive analysis as a regression problem. In this paper, we put an emphasis on finding causal links and exemplifying the relationship between geopolitical entities hidden in the data, while the causal links can actually help with regression as well.

Instead of establishing a theoretical framework, our proposed approach aims at validating the graph structure using another data source. And we treat these two steps as "non-text-based "and "text-based "respectively. Text mining and sentimental analysis has been an active area in data mining: when relations to be extracted from comparative sentences, where entities can be evaluated for rule derivation [15], a classifier works on the keywords and the categorization like non-equal gradable, equative and superlative results in features; while more entities can be involved in the context in social media text, they can be discovered and assigned by part-of-speech tagging, identifying opinion indicators and pattern matching [8]. The applications based on similar methodology can scale up with data size and derive more analytical results with business intelligence value [3, 21].

3 DATA SETS AND CONCEPTS

The Global database on events, location and tone (GDELT) [18] is a publicly accessible data set[1] that monitors the worldwide media sources in multiple forms (including print, broadcast and web), multiple languages and records context information of them. Also, it labels each such event with a standard format that encodes two involved entities (*initiator* and *reactor*), location, type and severity for each event record using CAMEO (Conflict and Mediation Event Observations) [11]. That is, the GDELT data is a complete time-dependent representation of news around the world. As event definition already has a directional setting with one single initiator and one reactor, it applies to the data extraction process of this study as well.

While the raw data has a rich set of features, we aggregate the dyadic event counts between all the pairwise entities (excluding domestic events, where the initiator and reactor are the same entity). To map the event frequencies to real-world relations, the underlined assumption is that dyadic event frequency does reflect the intensity of diplomatic activities and its trend over time does indicate the development of bilateral relations.

For an exploration, the ranking of countries by the number of initiation and reaction events are listed in Table 1. Intuitively, the larger the number is, the more actively this geopolitical entity behaves. The distribution doesn't necessarily represent the wealth of a country but show the degree of becoming a spotlight as in media reports and discussions.

Bringing it to a mutual relationship level, Figure 1-top demonstrates the relation by event frequencies between China and United States, each of which as initiator respectively: as the blue and red curves highly overlap with each other, a strong correlation is easily seen; However, United States-China and China-United States events share a high correlation coefficient (Pearson's r) 0.928, but United States-Japan and United States-South Korea can only achieve 0.208. In sense of correlation, a symmetric matrix can thereby derived for the frequency of events from United States to each country in

[1] http://gdeltproject.org/

Table 1: Number of events ranked by country (initiator and reactor respectively) (1998-2007)

	Country	#Initiations	Country	#Reactions
1	USA	1157368	USA	1115651
2	Russia (RUS)	573606	RUS	527904
3	Israel (ISR)	485240	IRQ	502551
4	GBR (UK)	371077	ISR	476792
5	China (CHN)	367154	PSE	363218
6	Iraq (IRQ)	343047	CHN	355157
7	France (FRA)	289072	GBR	341797
8	Japan (JPN)	285731	IRN	272851
9	Iran (IRN)	281072	JPN	263255
10	Palestine (PSE)	279630	EUR	251667
11	Egypt (EGY)	219457	FRA	250137
12	EUR	219225	TUR	197102

Table 2 (using data of 2000), where quite a few of them is less than 0.135.

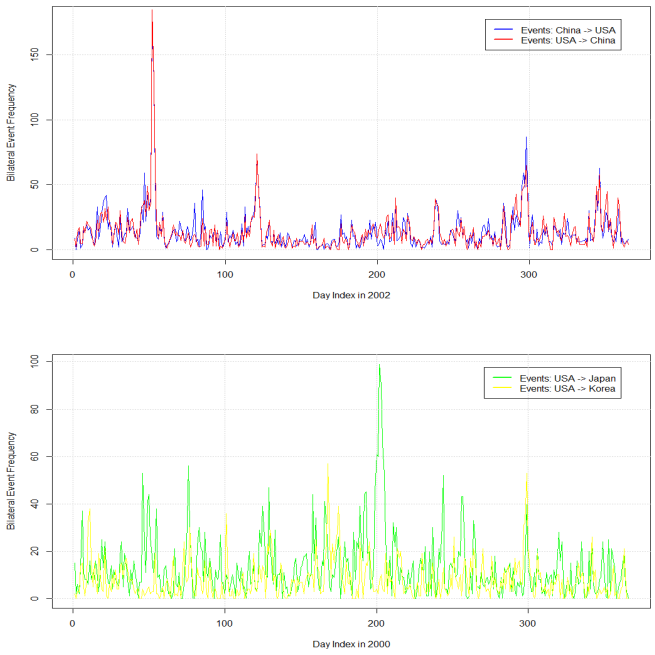

Figure 1: top: The frequency of events between China and United States (each as initiator) in 2002; bottom: Events acted by United States to Japan and South Korea respectively in 2000

With the concept of causality as well as the multi-facet nature of country-country relationship, the following definitions are derived

Definition 1 *Co-initiation link: when two dyads B and C show a causal link from B to C over a common reactor A from data sets (B ▷ A) and (C ▷ A), there is a co-initiation link from B to C associated with A, denoted as (B → C |▷ A).*

Table 2: Pearson's r for US-output event frequency to the five countries

	Brazil	China	S.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

Definition 2 *Co-reaction link: when two dyads B and C show a causal link from B to C over a common initiator A from data sets (A ▷ B) and (A ▷ C), there is a co-initiation link from B to C associated with A, denoted as (B → C |◁ A).*

4 CAUSAL LINK DISCOVERY BASED ON CONDITIONAL INDEPENDENCE

The data used for each run of causal link discovery is a subset extracted from the entire set: a target entity is selected as initiator or reactor, and the other entities' events with this target are counted. Considering the entire 211-country data set is highly sparse, only top 18 active countries (or geopolitical entities) are selected for the most part of the causal discovery, including United States, the European Union, China, Russia, Japan, South Korea, Israel and Palestine. The relations between top entities and others are involved in Section 6.

4.1 Causal discovery algorithm

The PC algorithm [26] is commonly used in the field of causal discovery for its setting of no hidden or selection variables. The steps of PC algorithm include: first constructing a complete directed graph and then test conditional independence for all the pairs to build a skeleton with all undirected links; then it identifies colliders and handles them with structural constraint, which ensures no cycle would occur throughout the process; next, the directions can be set up and the remaining undirected links can be randomly assigned as long as it's a DAG.

By selecting a different country as target, for both initiator and reactor roles, the outcome of causal discovery comes from the input data of corresponding monthly-aggregated dyadic event counts. Figure 2 shows the DAG[2] with United States as target reactor (we use $\alpha = 0.05$ in conditional independence test).

Considering how the relationship on the graph could imply the real-world situation, in Figure 2-bottom, United States is clearly in the center as this node's outdegree is 4 (the largest). However, it has an inlink from Iran, which can be roughly interpreted as "the US policy to Russia somehow depends on Iran's actions ". In other words, "Russia could use Iran to influence the relation with United

[2]In Figure 2 and 3, some edges carry two arrows and they represent undecided orientations. In a strict DAG, they would have a random orientation to ensure acyclic.

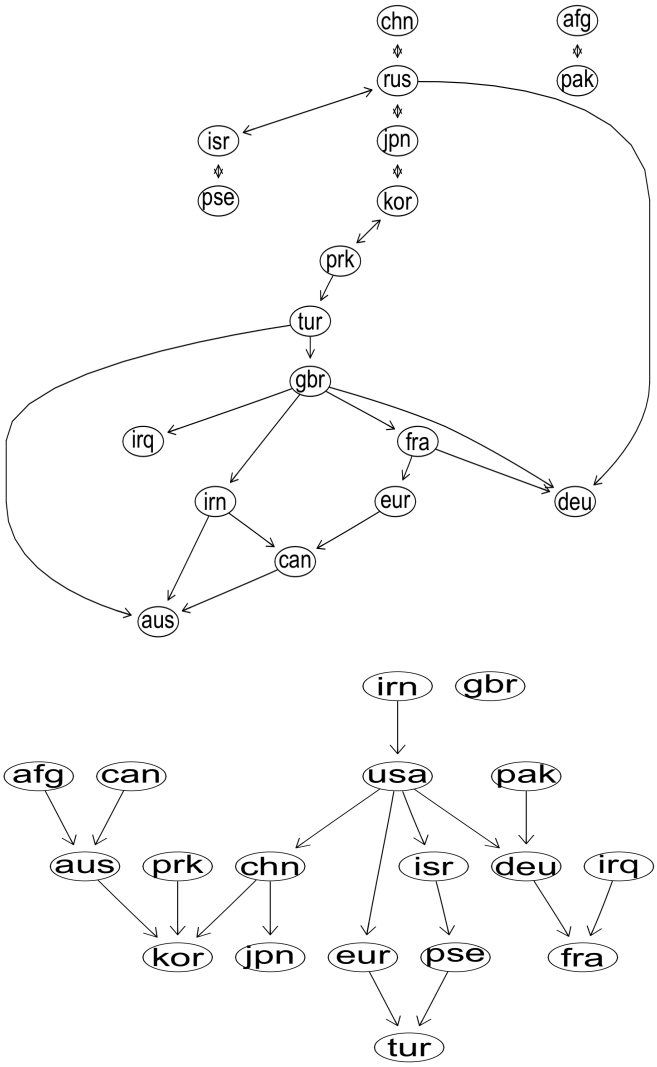

Figure 2: Causal links (co-initiation links) with United States (top) and Russia (bottom) as reactor respectively using 1998-2007 data for other entities

States ". For this individual case, the background story can be found from Wikipedia pages [3].

A quick integrity check is performed: switch the scenario of finding co-initiation links over the United States to co-initiation ones, would we get another directed graph with all orientations reversed? Figure 3 illustrates the discovery with regard to a same target being initiator versus reactor, and most of the links do remain and get reversed. One exception is that the United Kingdom is singled out in co-initiator links while it gets involved in the co-reactor structure, where the interpretation could be that "UK doesn't necessarily follow what the US do with Russia, but when Russia has some activities with Iran, it would react ".

[3]https://en.wikipedia.org/wiki/Nuclear_program_of_Iran

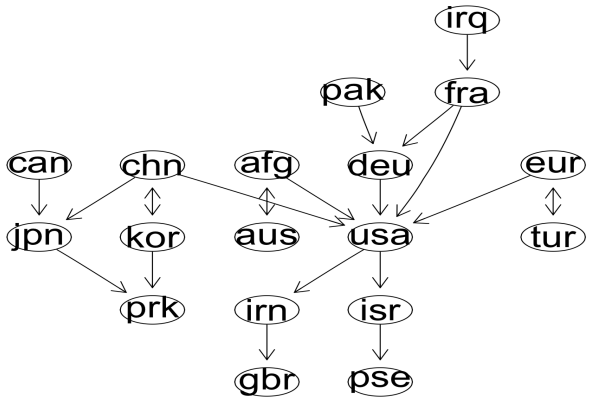

Figure 3: Co-reactor links with Russia as initiator using 1998-2007 data for other entities

Diving deeper into the links discovered, some exemplify geographic adjacency between two entities: e.g. South Korea is usually linked to North Korea or Japan; Israel is always connected with Palestine. While they are true in real-world scenarios, it illustrates the promise of this approach where no prior knowledge about country-country relationship is in the input.

In the following stage, we validate such causal links in a quantifiable way for the overall graph structure.

5 VALIDATING THE CAUSAL LINKS FOUND

The validation consists of two steps: first, a run-through over all the countries as targets provides an expanded and more complex picture for examining the nature of the causal links found; then we use text mining method to scrape related ***Wikipedia*** web pages as an independent source, where the links are scored by sentiment lexicon matching.

5.1 Scan through data as exploitation

The validation stage starts from further identifying causal links from the discovery process, attempting to get rid of potential spurious effects in a single run. As each time we run a causal discovery routine with one single target (as initiator or reactor) and the results are to be combined. With the combination, it also expands the relationship map to more entities: e.g. if we see both (B $\longrightarrow$ C $|\triangleright$ A) and (B $\longrightarrow$ C $|\triangleright$ D), the causal link $B \longrightarrow C$ appears stronger.

Algorithm 1: Exploiting and combining pairwise causal links

Input: all the pairwise time series where data a point represents the event counts at (C_i, C_j, t_k), $i, j \in 1...n$ and $i \neq j$. It is different from (C_j, C_i, t_k) when entity C_i and C_j exchanges as initiator and reactor.

Begin

for each entity C_i ($i \in 1...n$)
−− run a causal discovery routine with the subset of data represented by $(C_i, C_{1...i-1,i+1...n}, t_{\text{inf}})$
−− run a causal discovery routine with the subset of data represented by $(C_{1...i-1,i+1...n}, C_i, t_{\text{inf}})$

end for

Combine the above results by same causal link over varied targets, and remove links such that $(C_i \rightarrow C_j | \triangleright C_p)$ exists but $(C_j \rightarrow C_i | \triangleleft C_p)$ doesn't hold.

End

Output: All the causal links denoted as

$$(C_i \rightarrow C_j | \triangleright C_p 1...C_p r)$$
$$(C_i \rightarrow C_j | \triangleleft C_q 1...C_q s)$$

We define *special causal link* and based on its appearance pattern after the initial discovery process.

Definition 3 *Special causal link* is a co-initiation or co-reactor link that shows some irregularity in appearing in the graph with different targets or a target being another role: *a.* the link reverses its orientation when the target switches from initiator to reactor (and vice versa); and *b.* the link exists/doesn't exist only if the target is a certain entity.

Definition 4 *Stable causal link* is a co-initiation or co-reactor link that shows a pattern in appearing in the graph with different targets or a target being another role: *a.* the link reverses its orientation when the target switches from initiator to reactor (and vice versa); and *b.* the link persists under two or more than two targets.

This definition further probes the uni-directional attribute in the network, while eliminating some bias or numerical error by requiring it to show up in both initiator and reactor roles. Thus, after an exploitation of running PC algorithm through all the important entities being either role (set α lower to 0.005 for raising the bar of causal link). From the findings, interpretations can be made on some interesting links: the special causal link (EUR → USA |▷ RUS), meaning the European Union's actions on Russia would make the United States to follow and Russia is the only possible target for this link; (AFG → PAK) appears as a stable link for a set of entities as the target reactor: IRN, USA, JPN, GBR, DEU, FRA and EUR, and this could be related to the military actions in Afghanistan after 9/11.

Out of the total 598 links, there are 72 special causal links and 10 stable ones discovered. While the distribution follows a power law, those links with more than 1 common target indicates peer effects and the prevalence of special causal relationship, with reasonable interpretations on many cases, exemplifies the complication in world politics. In particular, 36 out of 72 involves the United States, Russia or China as a part of the causal link, while another 13 can be added with any of the three serving as a target entity. Using this approach, the influence of titans could be quantitatively measured.

5.2 Sentimental analysis as validation

Next, in the interest of finding an all-round way of further examining the causal links, some text mining against an independent data source is easily conceivable: while causal links are established through numeric data (dyadic event counts) without any priors, the texts from news are the real clue. So, we consider ***Wikipedia*** as a reliable source as it sets up a page between many major bilateral relations.

By retrieving such a page, we use a sense-level lexicon dictionary to measure the sentiment [7]. Matching related words to the content makes the level of effect quantifiable. In this scenario, we specifically probe the comparison in a triangle relationship: e.g. when there's a co-initiation link (B → C |▷ A), we would evaluate the contents of (B ▷ A) and (C ▷ A). Given that these are two distinct ways and independent sources, it's too idealistic to say that every piece of the graph structure of causal links can match the text-based approach. However, some key features do show consistency from frequencies to contents.

In applying the lexicon matching, the positive and negative effects can be scored respectively and thereby metrics can be derived as *significance score*, *positivity score* and *effect score* using the following equations.

$$E_{sig} = C_{pos} + C_{neg}$$

$$E_{pos} = C_{pos}/(C_{pos} + C_{neg})$$

$$E_{effect} = (C_{pos} + C_{neg})/C_{total}$$

where C_{pos} and C_{neg} are the counts of positive or negative effect lexicons; C_{total} is the total word count on the web page.

The hypotheses from the graph structure of causal links can be proposed and tested:

Hypothesis 1: The bilateral relations involved in the causal links are more significant than average.

By performing a scoring procedure that takes into account all words having a sentiment effect, the average significance score is 10% higher than that of the whole set of pairs available from Wikipedia. Also, as Wikipedia doesn't have all the pairwise bilateral relations page, the availability is 86.5%. With the constraint of causal links, the level of absence slightly reduces: the availability is now 90.0% out of the 130 relations extracted from causal links.

Note that we consider two pairwise relations "United States and Japan "and "Japan and South Korea "as *involved* in a causal link (USA → JPN |▹ KOR).

Hypothesis 2: The causal relationship of two parties is a stronger indication of the significance of their relation than that of "involved "entities.

The relation between United States and Japan is drawn from a causal link, while it's detected through their relations with South Korea respectively. Our analysis shows that this causal relationship, as it literally means, does represent a even larger margin in significance score (48% more versus 10% more).

Hypothesis 3: Measured by *effect score*, the bilateral relation from and within causal links represent more ties, connections or historical activities between the two entities.

The *effect score* calculates the proportion of "words/lexicons with an effect "out of the entire text. The higher the score is, it shows a more storied past and present in the bilateral relation. It doesn't simply represent how long the history is as it's a normalized score. In our result, "involved "relations stand for 1% higher effect score than the average, while relations indicated by causal link achieve a 3.5% high.

Hypothesis 4: Stable causal links have a larger significance score as well as a higher effect score.

The stable causal links, as they do represent a even stronger geopolitical connection (e.g. Iran → United States, but not the other way around), stand out from the metrics in both ways. With the calibration mechanism in place that those non-reversing causal links after an initiator-reactor exchange are discarded in verification, it would make stable links less error prone. That doesn't mean special causal links are less useful. Instead, special causal links, or even those we temporarily eliminate for non-reversing, may reflect more uni-directional nature in the global situation.

Meanwhile, the *positivity* score doesn't show an obvious change between different groups of relations. It may suggest that the Wikipedia articles are monitored by editors and they are meant to be as neutral information source.

After going through the multi-step validation process, the causal links do represent a deeper level of relations, compared to those unconnected entities. Use the attributes in graph structure as well as in a frequentist manner, it shows coherence between the non-text and text based approaches, and the hypotheses tested indicate a good likelihood that the causal links found are true.

6 GRANGER CAUSALITY WITH SHORT-TERM DYNAMICS

Granger causality [12], in contrast to PC algorithm based on conditional independence (and time-insensitive), carries time dependency in its formation and 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.

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

Table 3: The number of dyadic links (Poisson and linear kernels respectively in regression model) over the years

	2001	2002	2003	2004	2005
#(Total links with Poisson)	1268	1190	1389	1185	868
#(linear fails to find)	1017	933	1086	932	707
#(linear finds reverse)	78	78	88	93	52
#(Poisson fails to find)	18	14	23	15	11

$$y_t = b_1 y_{t-1} + b_2 y_{t-2} + \ldots + b_i y_{t-i} + b_{i+1} x_{t-m} + \ldots + b_{i+n-m+1} x_{t-n} + \sigma \tag{2}$$

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 "$'A- > B$" or "$B- > A$" below p-value in a Granger test is considered a link.

For a comparison, both Poisson [25] and linear regression models are attempted, where Poisson is the main model as it can build significantly more links after each run from 2001 through 2005.

$$log(E(y_t|y_{t-1}, y_{t-2}, ..., y_{t-m})) = a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_m y_{t-m} + \epsilon_t \tag{3}$$

$$log(E(y_t|y_{t-1}, ..., y_{t-m}, x_{t-p}, .., x_{t-q})) = \tag{4}$$
$$a_0 + a_1 y_{t-1} + a_2 y_{t-2} + \ldots + a_m y_{t-m} + b_p x_{t-p} + \ldots + b_q x_{t-q} + \epsilon_t$$

The exploitation checks all the dyadic events from the total 210 entities to those influential ones (event frequencies between non-influential countries is extremely sparse that would cause the regression to suffer from numerical problems). In either model, we use a lag of 2 days for autoregressive fit.

Table 3 shows (rows from top to down) the total number of dyadic links over a $210 * 18 * 18$ search space: using the number of total links with Poisson as baseline (as Poisson has stronger modeling capability for capturing the nonlinearity), the following rows record the numbers of "links that linear model fails to find (but Poisson finds them) ", "links that linear finds its reverse (B → A, while Poisson only finds A → B) ", "links that Poisson models can't find (but linear model gets them) ".

So, the comparison demonstrates the representational power in regression model can affect the network structure a lot. In some circumstances, a "good "model that fits non-linear data well might be prone to spurious effects. And after applying the same Algorithm 1 as in Section 4, we observe some interesting links.

For the pair of dyadic links: US → EUR and EUR → US, they attract several followers on each side. Figure **??** shows the situation of 2005. Through the years, some countries may not appear there every time (e.g. China) or change side (e.g. Italy, Ukraine), but Azerbaijan (AZE) always stays with the European Union side. The Wikipedia page [34] of Azerbaijan-European Union relations suggest that they have maintained a firm and positive relationship through the years, while the counterparts like Ukraine "has long been a difficult partner" [36]. Also, Itay-United States relations appear closer, as Italy "had a strong tendency to support American foreign policies, despite the policy divide between the U.S. and

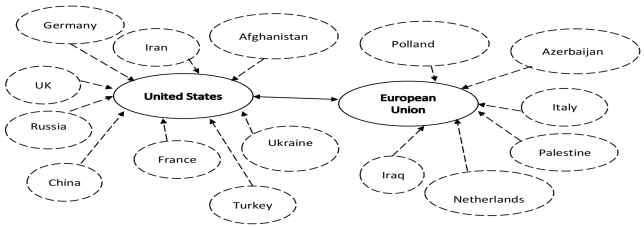

Figure 4: Dyadic links on United States vs. European Union using data of year 2005

many founding members of the European Union ... during the Bush administration" [37].

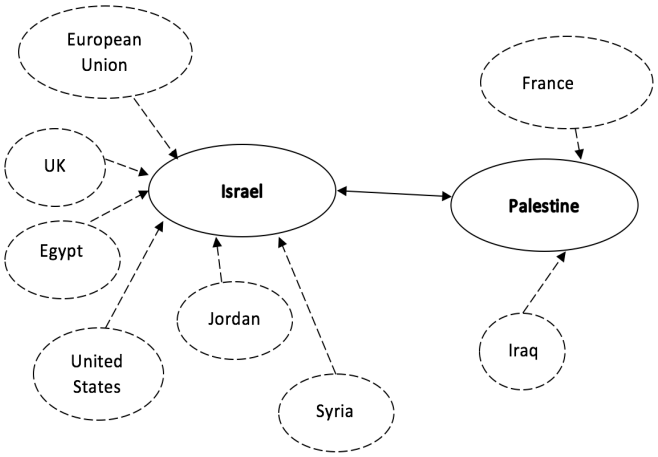

Figure 5: Dyadic links on Israel vs. Palestine using data of year 2004.

In regional spotlights, one interesting pair is ISR ↔ PSE (Israel-Palestine): as shown in Figure 5, not necessarily all the followship links mean an ally: it can also be an enemy who's closely monitoring its rival and actions can be quickly triggered. Meanwhile, France is on the side of Palestine for its softer attitude towards it, compared to many other European Union countries.

Another subject of interest is the stability-over-time of such causal links. It may indicate some links could change and some stay firm. We explore it from the original causal links to the dyadic links after reduction.

The number of causal links change over time in a steady rate. As in Table 4, 60-70% of the links over all 18 countries of interest would maintain in the next year, while such a rate varies from country to country: the United States keeps almost all the links in the next year, and China can only retain half of them. For Israel and Palestine, there's a significant drop from 2004 to 2005 and a notable event can be found as evidence: Israel completely withdrew from Gaza in 2005. It is described as a new era for the Israeli-Palestinian conflict since 2005 [35].

Specifically, stable dyadic links are extracted where the co-initiator satisfies: *a.* the dyadic link exists more than half of the times on a yearly basis; and *b.* the opposite dyadic link never appears (Given $(B \rightarrow C|A)$, there's no $(C \rightarrow B|A)$. In the results, we find 663 stable dyadic links in total. Most of them are asymmetric: if $C \rightarrow B$ is discovered as a stable dyadic link, $B \rightarrow C$ doesn't have a good chance to appear and there is not likely to have equal force on each side. The exceptions include the aforementioned (EUR → USA | AZE) and (SYR → IRQ | RUS) (while many others like United States, Turkey and Saudi Arabia choose Iraq as a cause).

Table 4: Number of causal links that stay the same in next year (2001 2005)

	2001-2002	2002-2003	2003-2004	2004-2005
All-18	599/895	621/906	690/962	502/890
US	135/136	133/136	131/134	134/135
China	26/63	27/61	39/51	34/72
Japan	16/50	17/31	31/60	24/46
Russia	90/113	91/102	78/112	51/88
France	28/54	24/50	27/50	27/49
UK	51/70	61/85	59/84	49/66
Israel	43/56	44/63	43/55	10/52
Palestine	26/38	21/44	12/31	6/34

7 DISCUSSION AND ANALYSIS

While existing work in causal discovery on real-world observational data focuses more on "factors ", where the two ends of a causal link are not sovereign entities, we identify the causality between countries, as initiators and reactors in global events, because they are, as a matter of fact, influenced in a causal manner.

The canonical theoretical framework of causal inference usually presents a randomized design [31]. The PC algorithm, as used in our work, is established on the statistical foundation and consists of steps in removing confounding effect from graph. PC algorithm doesn't have a time dimension in its modeling, instead only considering the relativity for the values of multiple entities at the same time. In general, the longer the span is, the higher confidence causal discovery result would be, as causal relationship is defined to be persistent. On the other hand, global politics may change its situation in 10 years or longer. It is understandable that the relationship between two foes may improve after some treaties or deals. Also, the economic situation would also strengthen or weaker some certain geopolitical ties. Practically, in terms of parameter tuning, we do see that a high level of significance (e.g. $\alpha = 0.005$) and a short data span like 2 years would lead to very few causal links found by PC algorithm.

In the validation phase, we systematically run the causal discovery process as exploitation and use text-based methods to provide reference scores. We propose several hypotheses with regard to the significance of causal links, and then we derive content-based scores to provide quantifiable measurements. Especially, Hypotheses 1-3 point to the overall significance and Hypothesis 4 is a further indication about graph structure. Wikipedia pages on bilateral relations is a good public data set that records the important diplomatic activities and characterize the relation using featured words. It is worth noting that country-country relationship is complicated in nature,

so getting a labeled data set as ground truth, in a black-or-white style, is nearly impossible.

On the other hand, Granger causality [12] uses short-term regression as its core and is still active in quite a few areas [22, 23, 29, 32]. Granger causality is native to time series and possess an advantage for its efficiency. Corresponding to the existing research about Granger causality [10, 27], we did find more causal links than using PC algorithm, and thereby didn't apply the above-mentioned validation on its overall structure due to its known issue. However, it can be regarded as a tool for quickly identifying featured links, as well as its capability of checking the time-dependent dynamics.

A causal link is not going to change in a day, so naturally the short-term prediction can help, with forecasting the intensity of international events in a certain area [9, 33]. In this direction, causal links would potentially help: by incorporating the causal relation into regression, which means that the other entity's event frequencies can be auxiliary variables for model training.

After all, digging into real-world bilateral and multilateral relations is eventually multidisciplinary. Without too much speculations on specific bilateral relations, we see promises from the perspective of web mining and consider that a larger-scale system on top of the results in this paper[4], combined with fresher data and more calibration and validation processes, has the potential of performing continuous causal discovery as a pipeline and detecting interesting dynamics from within.

8 CONCLUSION

In this paper, we tackle "causal discovery "in country-country relations as a web mining problem: by aggregation dyadic events between pairwise geopolitical entities, we apply a formal framework in the discovery process: *i.* Using PC algorithm to identify relatively stable links; *ii.* Applying validation using text-based methods; and *iii.* Finding Granger causality for short-term dynamics. The causal links found, through validation, are in accordance with the current world order and largely match what human can infer from news: PC algorithm is good at locating stable links, while Granger causality is a tool in quickly identifying featured relations with dynamics. Our overall methodology and results demonstrate the capability of causal modeling in this domain as an interesting application of web mining.

REFERENCES

[1] Arbour, D., Garant, D. and Jensen, D., Inferring network effects from observational data, *Proc. of ACM KDD Int'l Conf. of Data Mining and Knowledge Discovery*, 2016.
[2] Asimakopoulos, I., Ayling, D. and Mahmood, W.M., Non-linear Granger causality in the currency futures return, *Economics Letters*, Vol. 68, pp. 25-30, 2000.
[3] Castellanos et al., LCI: A social channel analysis platform for live customer intelligence, *Proc. of SIGMOD Conf.*, 2011.
[4] Cheng, B.S. and Lai, T.W., An investigation of co-integration and causality between energy consumption and economic activity in Taiwan, *Energy Economics*, Vol. 19(4), pp. 435-444, 1997.
[5] Chickering, D., Heckerman, D., and Meek, C., Large-Sample Learning of Bayesian Networks is NP-Hard, *Journal of Machine Learning Research*, Vol. 5, pp. 1287âĂŞ1330, 2004.
[6] Choi, D., Estimation of Monotone treatment effects in network experiments, *Journal of American Statistical Association*, 2016.
[7] Choi, Y. and Wiebe, J., +/-EffectWordNet: Sense-level lexicon acquisition for opinion inference, *Proc. of the Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, 2014.
[8] Ding, M., Chen, Y., and Bressler, S.L. *Granger causality: Basic theory and application to neuroscience.* In Schelter. S., Winterhalder, N., & Timmer, J. Handbook of Time Series Analysis. Wiley, Wienheim, 2006.
[9] Erikson, R., Pinto, P. and Rader, K., Dyadic analysis in international relations: a cautionary tale, *Political Analysis*, vol. 22, iss. 4, 2014.
[10] Friston, K.J., Bastos, A.M., Oswal, A., et al., Granger causality revisited, *Neuroimage*, Vol. 101, 2014.
[11] Gerner, D, Schrodt, P.,Abu-Jabr, R., and Yilmaz, O., Conflict and mediation event observations (CAMEO): A new event data framework for the analysis of foreign policy interactions. Working paper, 2012.
[12] Granger, C. W. J., Investigating causal relations by econometric models and crossspectral methods, *Econometrica*, Vol 37, pp. 424âĂŞ438, 1969.
[13] Goldsmith-Pinkham, P. and Imbens, Social networks and the identification of peer effects, *Journal of Business and Economic Statistics*, Vol. 31, 2013.
[14] Harris, N. and Drton, M., PC Algorithm for nonparanormal graphical models, *Journal of Machine Learning Research*, Vol. 14, 2013.
[15] Jindal, N. and Liu, B., Mining comparative sentences and relations, *Proc. of AAAI Conference*, 2006.
[16] Kalisch, M. and Buhlmann, P., Estimating high-dimensional directed acyclic graphs with the PC-algorithm, *Journal of Machine Learning Research*, Vol. 8, 2007.
[17] Kleinberg, S., Hripcsak, G., A review of causal inference for biomedical informatics, *Journal of Biomedical Informatics*, Vol. 44(6), pp. 1102-1112, 2011.
[18] Leetaru, K., Schrodt, P.A., GDELT: Global data on events, location and tone, 1979-2012, *ISA Annual Convention*, 2013.
[19] Lev, B., Petrovits, C. and Radhakrishnan, S., Is doing good good for you? How corporate charitable contributions enhance revenue growth, *Strategic Management Journal*, Vol. 31, pp. 182-200, 2010.
[20] Li, J, Thuc, D.L., Liu, L, Liu, J., Zhou J., Sun, B., Ma, S., From observational studies to causal rule mining, *ACM Transactions on Intelligent Systems and Technology*, Vol. 7(2), Article 14, 2015.
[21] Li, Y.-M. and Li, T.-Y., Deriving market intelligence from microblogs, *Decision Support Systems*, Vol. 55, Iss. 1, 2013.
[22] Li, Z. et al., Discovery of causal time intervals, in *Proc. of SIAM Conference of Data Mining*, 2017
[23] Nguyen, H. et al., Discovering Congestion Propagation Patterns inSpatio-Temporal Traffic Data, *IEEE Transactions on Big Data*, 2017
[24] Reiss, J., Causation in the social sciences: evidence, inference and purpose, *Philosophy of the Social Sciences*, Vol. 39(1), pp. 20-40, 2009.
[25] Schein, A., Paisley, J., Blei, D.M., Wallach, H., Bayesian Poisson tensor factorization for inferring multilateral relations from sparse dyadic event counts, *Proc. of ACM KDD Int'l Conf. of Data Mining and Knowledge Discovery*, 2015.
[26] Spirtes, P, Glymour, C, Scheines, R. Causation, Prediction, and Search. Adaptive Computation and Machine Learning, 2nd edition. MIT Press, Cambridge, 2000.
[27] Spirtes, P. and Zhang, K., Causal discovery and inference: concepts and recent methodological advances, *Applied Informatics*, Vol. 3(3), 2016.
[28] Stern, D., From correlation to Granger causality, working paper, 2011.
[29] Sun, Y., Li, J., Liu, J., Chow, C., Sun, B. and Wang, R., Using causal discovery for feature selection in multivariate numerical time series, *Machine Learning*, DOI 10.1007/s10994-014-5460-1, 2014.
[30] Tam, G.H., Chang, C., Hung, Y.S., Gene regulatory network discovery using pairwise Granger causality, *IET System Biology*, Vol. 7(5), pp. 195-204, 2013.
[31] Toulis, P. and Kao, E., Estimation of causal peer influence effects, *Proc. of Int'l Conf. of Machine Learning*, 2013.
[32] Yao, S., Yoo, S., Yu. D., Prior knowledge driven Granger causality analysis on gene regulartory network discovery, *BMC Bioinformatics*, 16:273, 2015.
[33] Yonamine, J.E., Predicting future levels of violence in Afghanistan districts using GDELT, *working paper*, 2013.
[34] Azerbaijan-European Union relations, Wikipedia page.
[35] Timeline of the Israeli-Palestinian conflict, Wikipedia page.
[36] UkraineâĂŞEuropean Union relations, Wikipedia page.
[37] Italy-United States relations, Wikipedia page.

[4]Source code available via https://github.com/lionelc/GDELT_Causal_Discovery