DISCOVERING BILATERAL AND MULTILATERAL CAUSAL EVENTS IN GDELT

Lei Jiang (Tapjoy Inc.)
Fan Mai (University of Virginia)

04-02-2014

Guideline

- Introduction
- Exploratory analysis on GDELT event (spatial) time series
- Causal discovery: bilateral and multilateral events as causal links
- Predictive analysis

Introduction

- A snapshot of GDELT (Global Database of Events, Languages and Tone) data:
  - **<u>210 spatial time series of events between 15 countries (2000-2005)</u>**

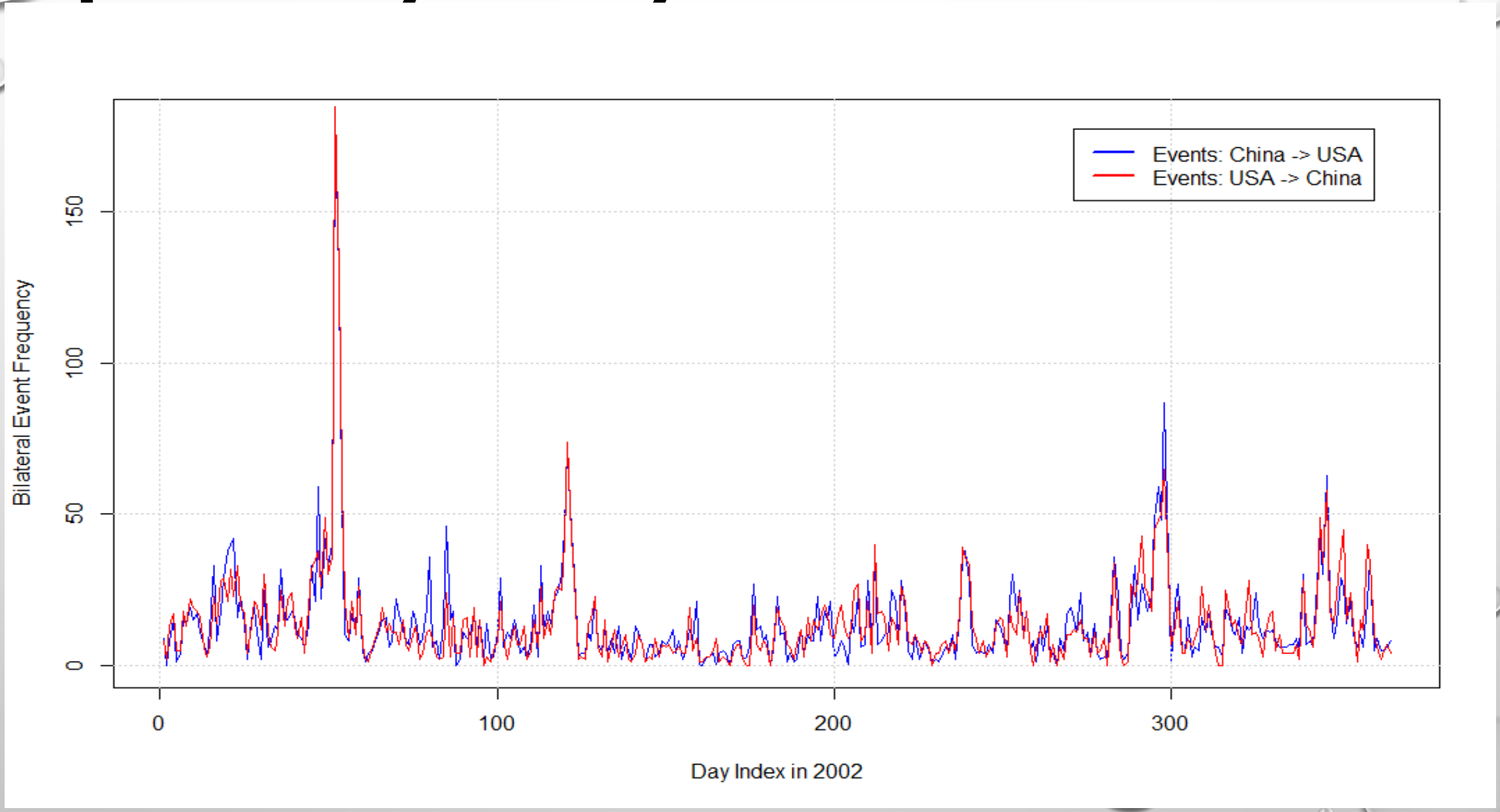
- **Default assumption:**
  - bilateral and multilateral events would be reported by each involved country, and/or other countries

  - The importance of events can be reflected by the frequency of news reports

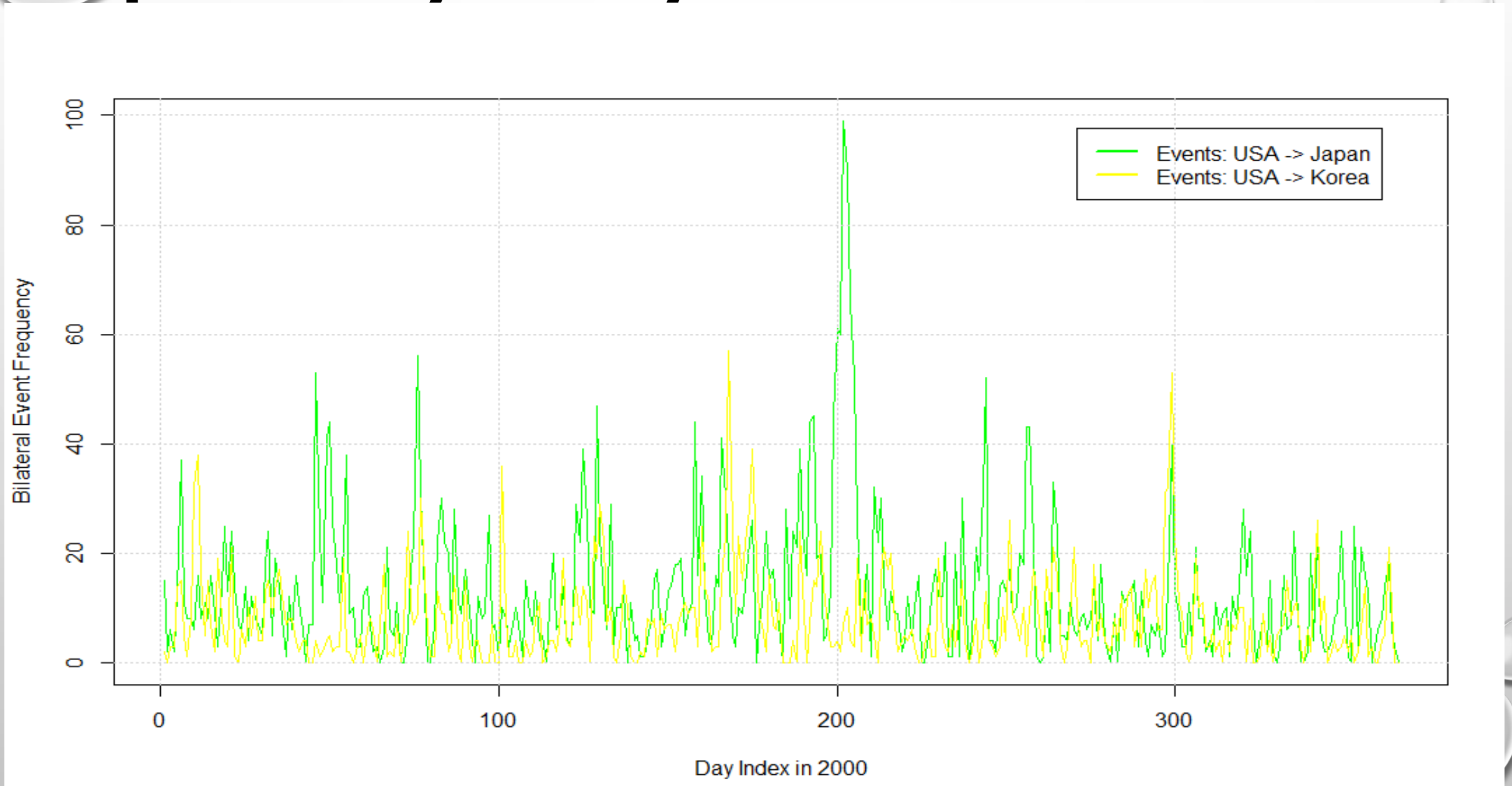
Introduction-2

- Time series of cross-country events could be correlated, which somehow reveals the relationship between countries
- The relationship could be even more complex when considering more than two countries
- It could eventually form a network --- with dynamics

Exploratory analysis - 1



Exploratory analysis - 2



Exploratory analysis - 3

- Correlation coefficient doesn't seem to be a good indicator for linking multiple countries

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

Causal links - 0

- Granger causality (Granger, 1969): examine whether a time series is useful for forecasting another one

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

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

Causal links - 1

- **Unidirectional relationship is revealed:** assuming there is one single "initiator" in country-country relationship

- By Granger testing with events series:
  - With p-value of 0.05, causal links are established
  - If p-value < 0.05 for both directions, link with smaller p-value is chosen
  - A causality-graph is thereby constructed

Causal links - 2

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

Causal links - 3

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

Causal links – 4 (Multilateral events)

- Multivariate Granger test is not preferred:
  - Search space increases exponentially (curse of dimensionality)
  - Overfitting is possible to affect accuracy

- Instead,
  - A heuristic framework is derived
  - Multilateral events are considered as optimization to bivariate links

Causal links – 5 (Multilateral events)

- Assuming transitive relation: A→B + B→C ~ A→C
- Step 1: Find bivariate links A→ X1…Xi and Y1…Yj → A
- Step 2: Between X1…Xi and Y1…Yj, find any X → Y or Y→X
- Reduce search complexity from $O(n^3)$ to $O(n^2)$
  - And prove to be effective in predictive analysis

Causal links – 6 (3-party Causal links over time)

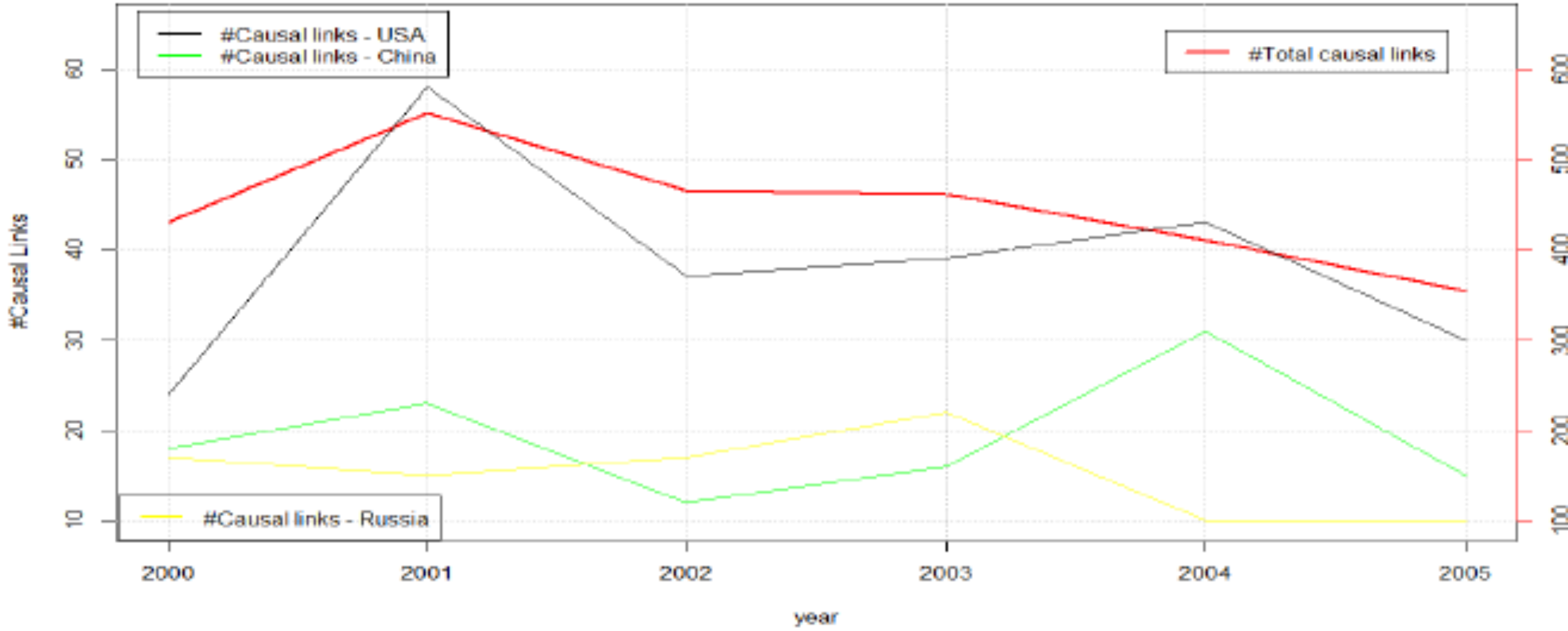


Fig. 2. Trend of number of causal links in the sample network from 2000-2005

Causal links – 7 (3-party Causal links over time)

- **Insights:**
  - Network assortativity increases when more causal links emerge in total
  - Influence/activeness can be reflected for an individual country
  - US/China follow global trending, Russia doesn't

Predictive Analysis – 1

- **Problem definition**
  - Find a third/fourth country that can help with predicting #bilateral events (regression task)
  - Additional data source as auxiliary time series
  - Use Sum of Squares (SSQ) as criteria

Predictive Analysis – 2

	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 ($lag = 1 - 3$)	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

Conclusion and possible future work

- Causal links are good indicator for revealing country-country relationship with a real-world metaphor
- More information in GDELT can be used for building finer models: e.g. event type
- Larger-scale analysis and visualization can lead to even more interesting results

THANK YOU !

QUESTIONS …