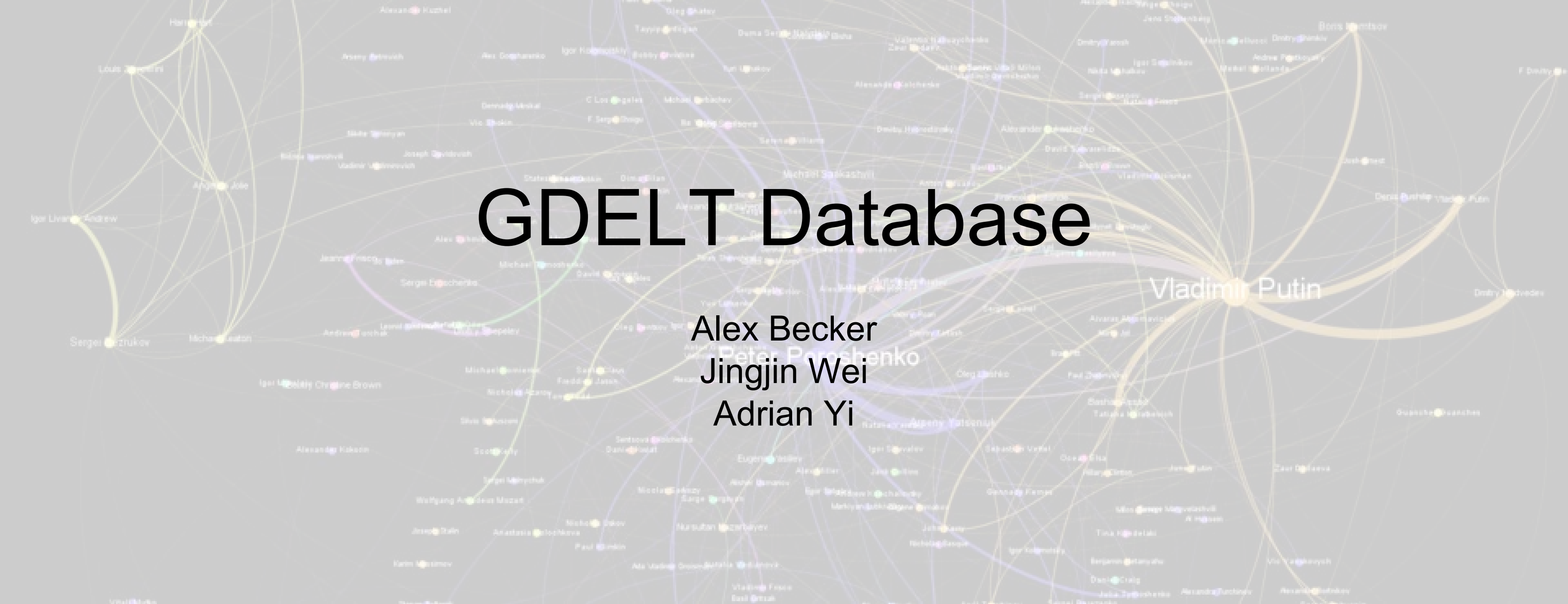


GDELT Database

Alex Becker
Jingjin Wei
Adrian Yi

GDELT Database

- “The vision of the GDELT Project is to codify the entire planet into a computable format using all available open information sources that provides a new platform for understanding the global world.”
- GDELT Project consists of over a quarter-billion event records in over 300 categories covering the entire world from 1979 to present.
- GDELT relies on tens of thousands of broadcast, print, and online news sources from every corner of the globe in 15 languages → Big Data!

News Coverage around the World

News coverage around the world

- Global Knowledge Graph
  - “Expands GDELT’s ability to quantify global human society beyond cataloging physical occurrences towards actually representing all of the latent dimensions, geography, and network structure of the global news.”
  - Timeline of news coverage
  - Normalized with respect to growth of database
- Given a set of keywords or themes, how often and how much is a country mentioned in this respect?

News coverage around the world: Themes

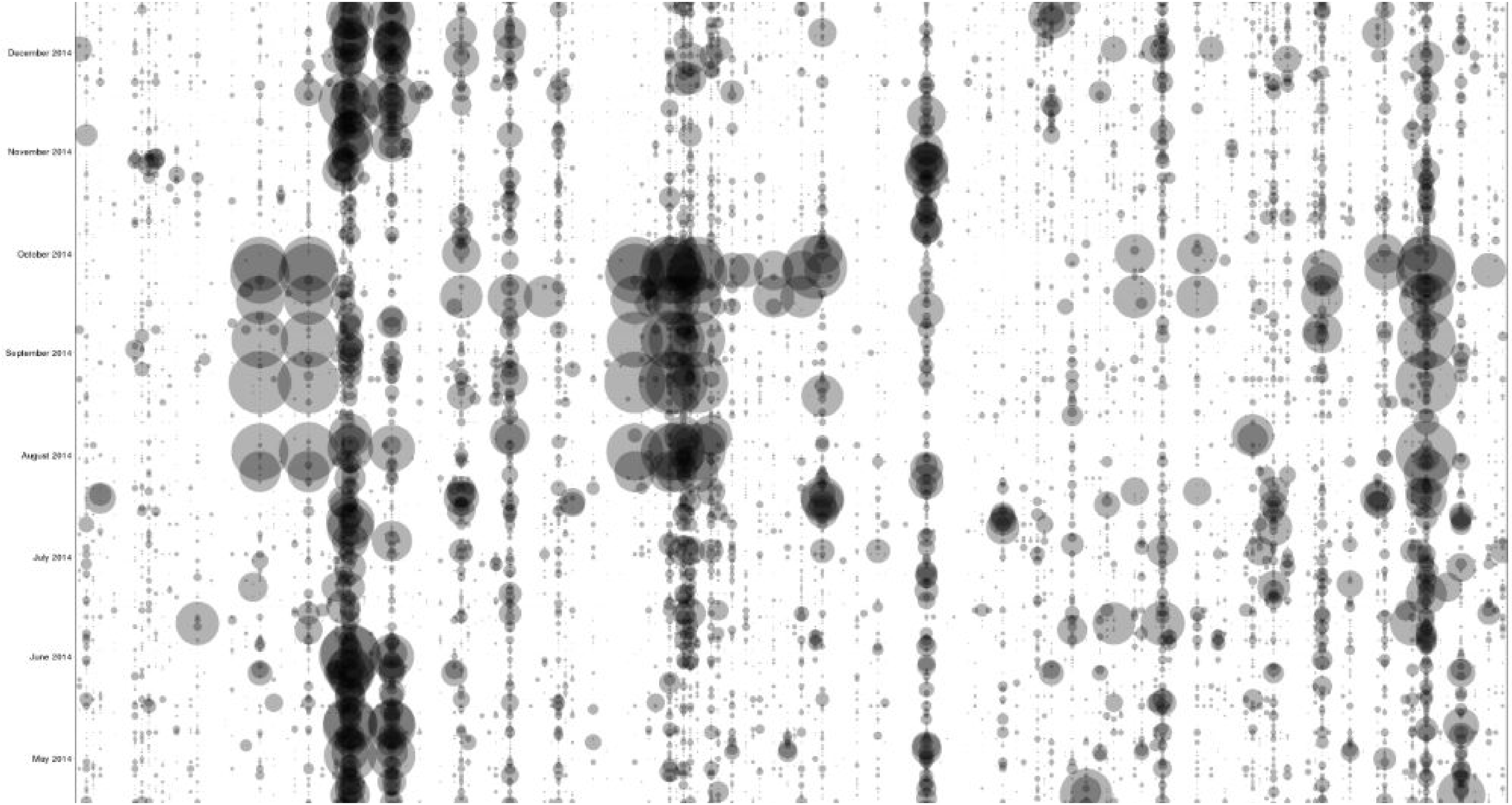
- Keywords and themes we analyzed:
    - Revolution
    - Sovereignty
    - Terrorism

- Auxiliary terms like “fighting”, “borders”, “independence” to narrow search and focus on political conflicts.

News coverage around the world: Data Preparation

- For each theme we get time series for all the countries that are mentioned with respect to the keyword.
- Filter data to only include countries which go on record at least 70% of the days of the time period from 01/01/2014 to 10/31/2015.
- GDELT measures the volume of news coverage regarding the keyword and a particular country on any given day and quantifies numerically.
- If no news and sentiments are reported on a given day, we fill in 0.

News coverage around the world: Data Preparation



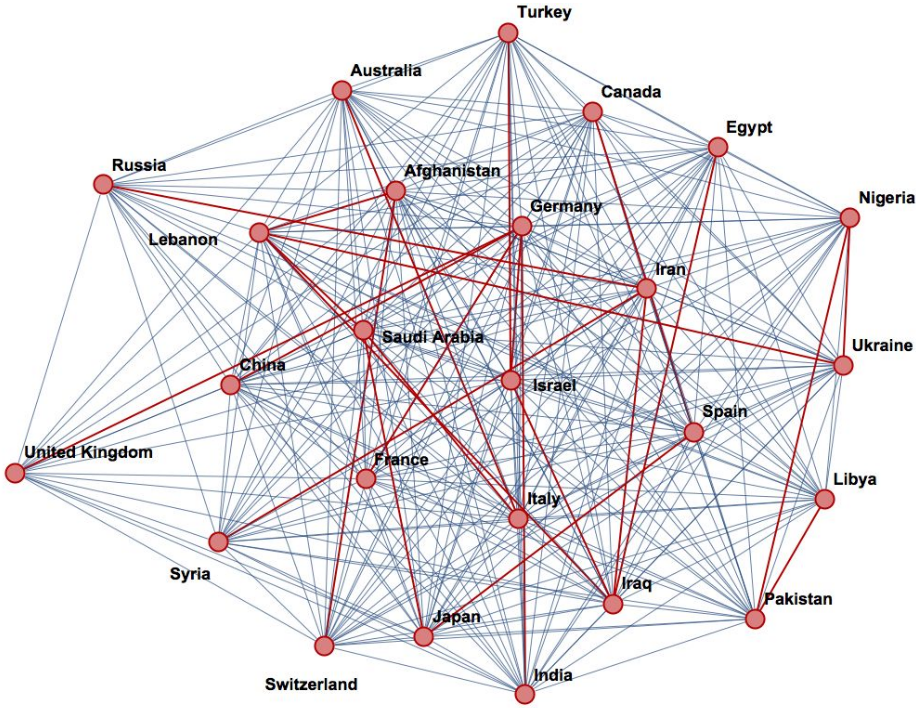
News coverage around the world: Data Analysis

- Find Correlation between time series
  - Use Spearman Rank Correlation measure: Is the date of the largest value in X also the date of the largest value in Y, etc?
    It is the Pearson Correlation for the ranking of the variables.
  - +1 for positive monotonic relationship, -1 for negative monotonic relationship
  - In our project: Do news burst come up at the same time for two different countries?

News coverage around the world: Filtering

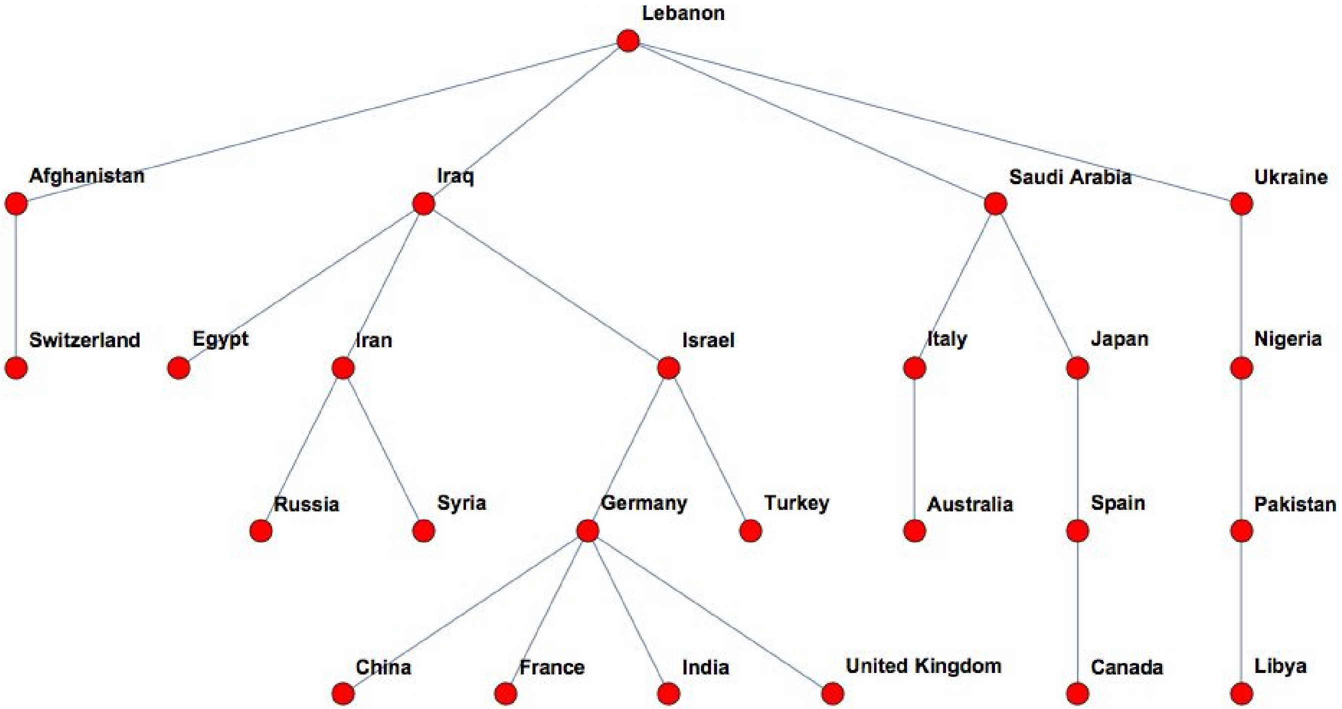
- Every time series is correlated to some extent; by noise, information etc.
- Draw adjacency matrix with weights: $\surd 2(1-\varrho)$ where ϱ is the correlation.
- How do we filter?
    - Threshold: Only correlations above a certain threshold cause links.
    - Minimum Spanning Tree: Connect all nodes such that all the overall weights are minimized.
    - Planar Maximally Filtered Graph: Connect all nodes starting with the smallest weights such that the graph stays 2-dimensional.

Analysis for “Revolution”

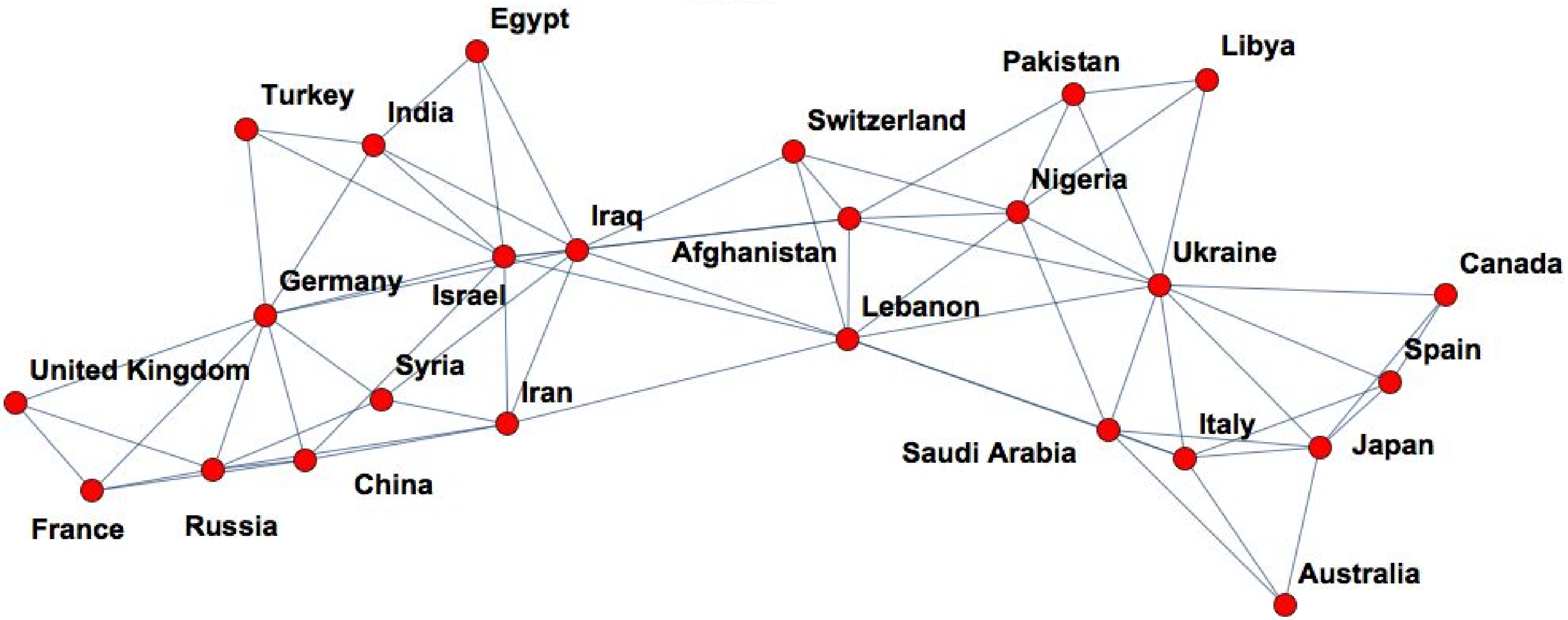


- Fairly high correlations among the different countries
- Use MST and PMFG to visualize information a little better.

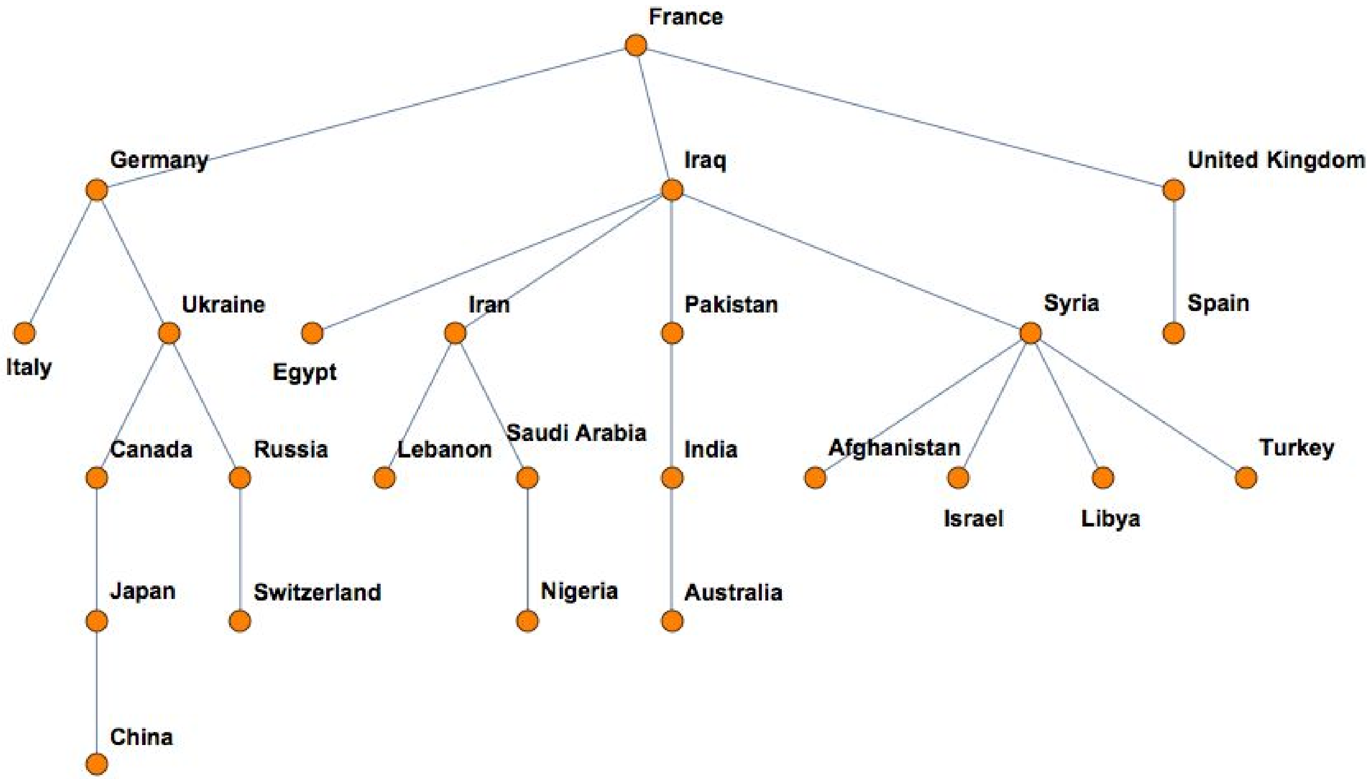
Analysis for “Revolution”



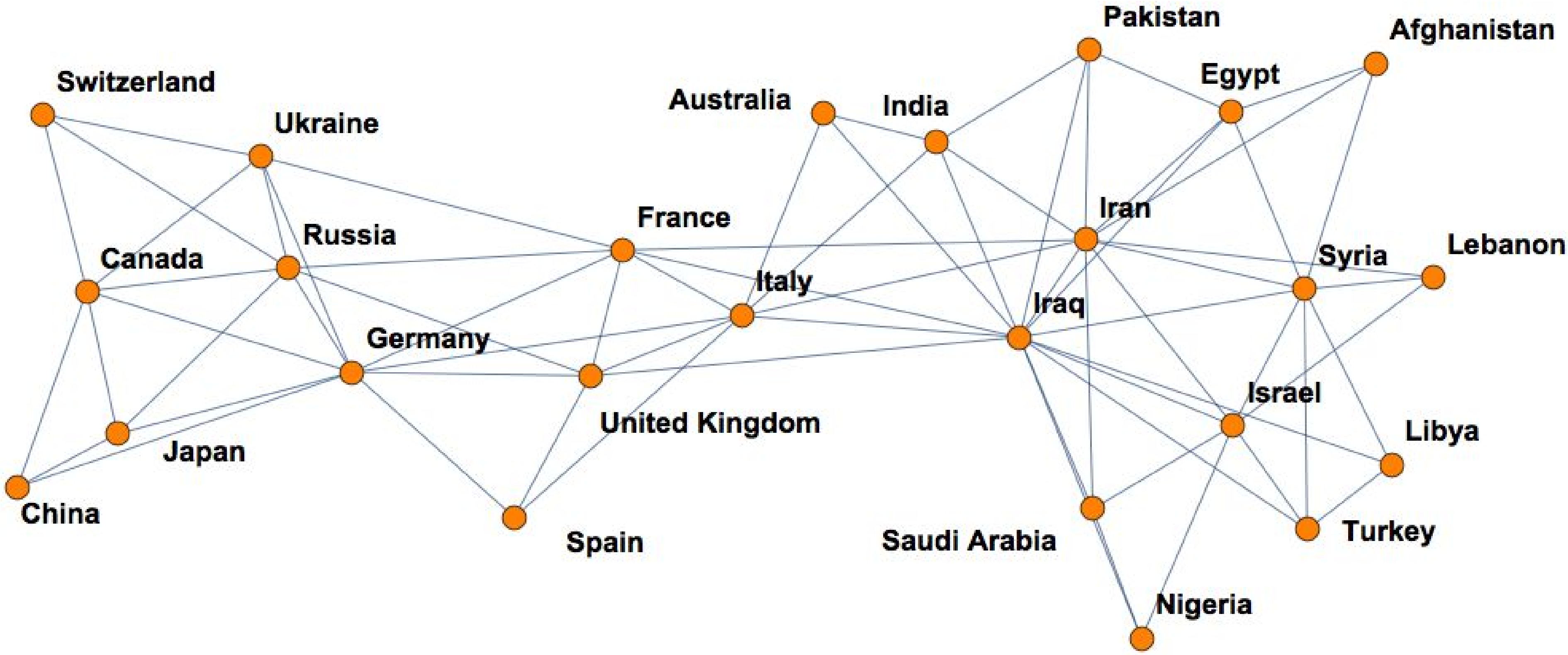
Analysis for "Revolution"



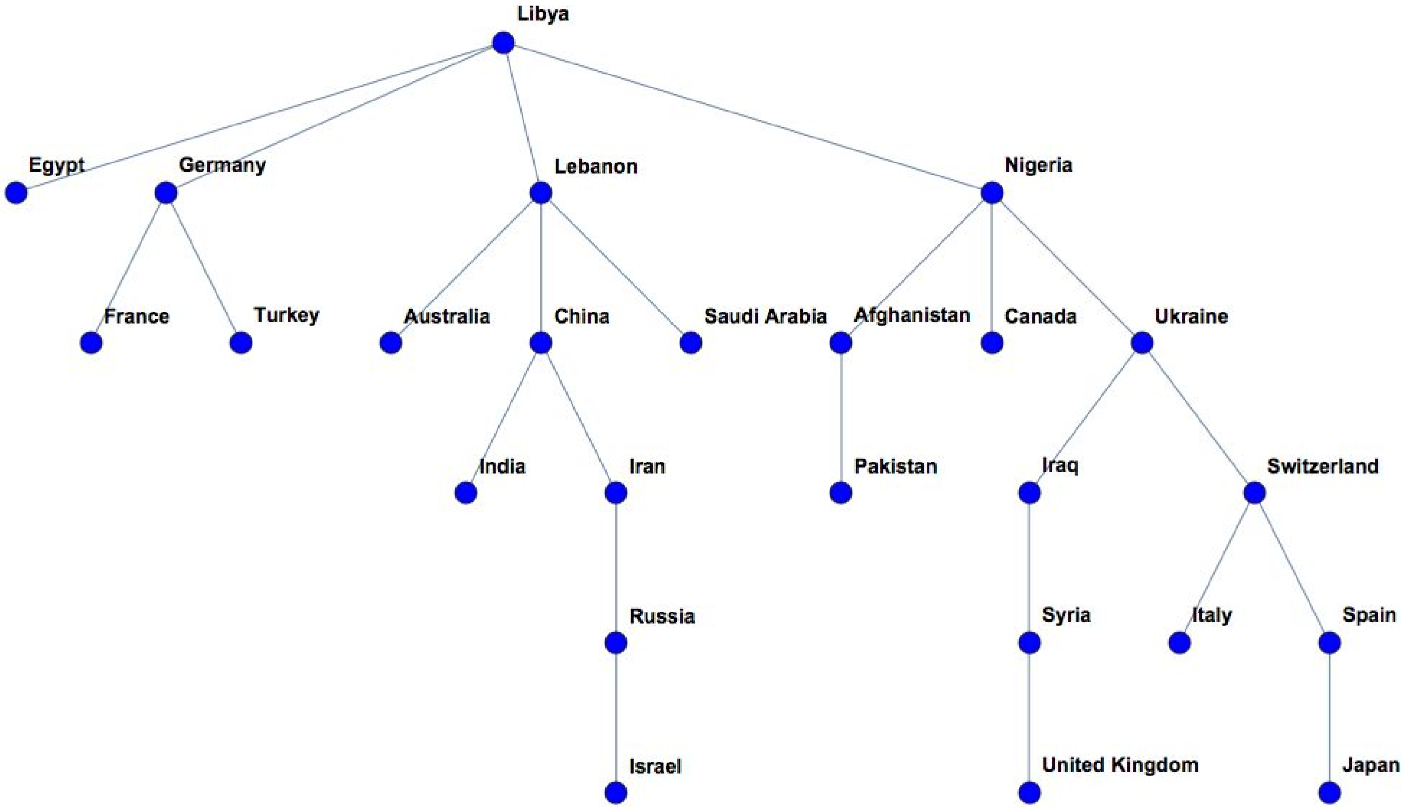
Analysis for “Sovereignty”



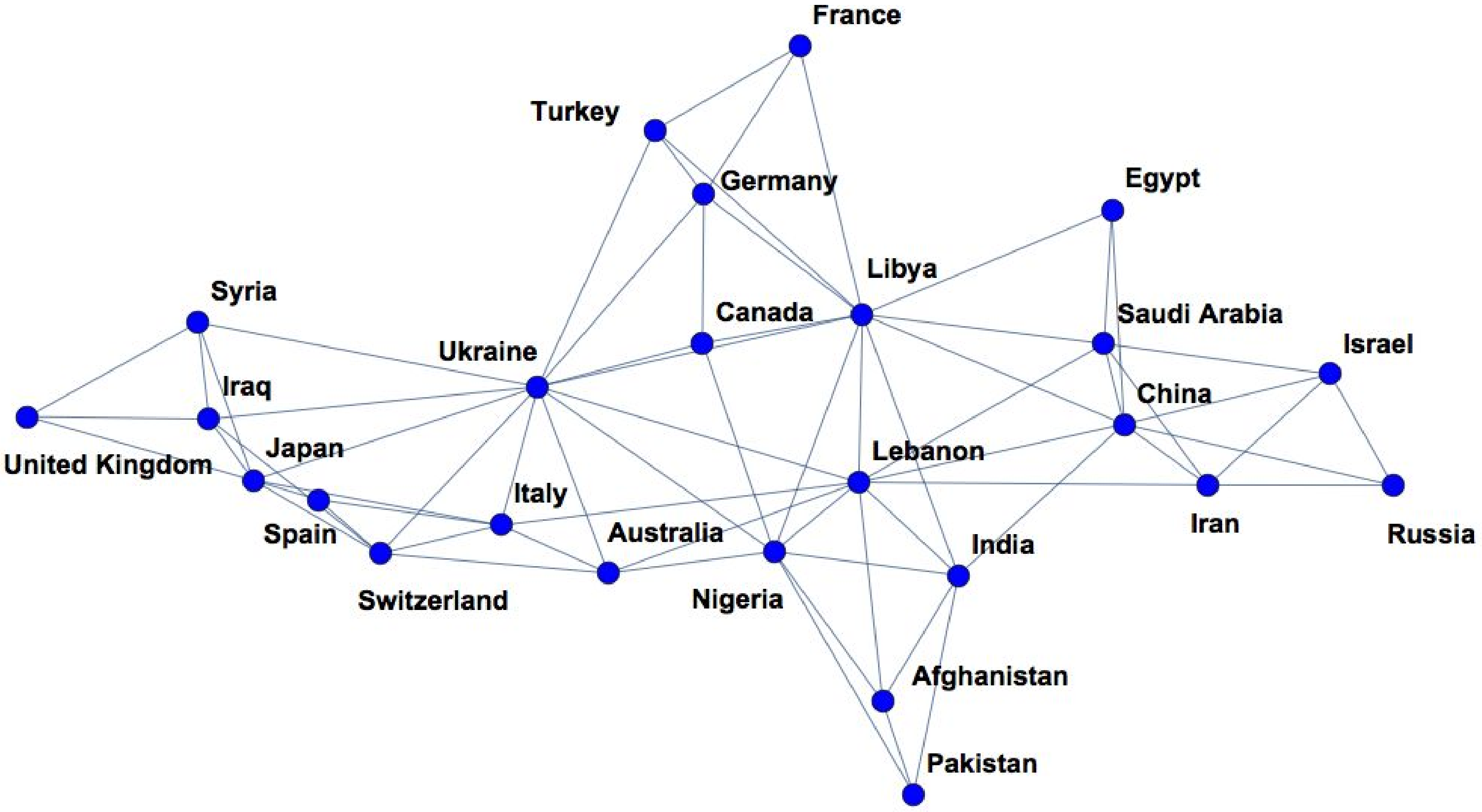
Analysis for “Sovereignty”



Analysis for "Terrorism"



Analysis for “Terrorism”



News Coverage around the world: Outlook

- Knowledge Database also offers weighed news count instead of just the number
- Lagged correlations: news about which country in what context trigger events or reports somewhere else
- Include sentiment in the analysis

Tone Analysis for Stocks

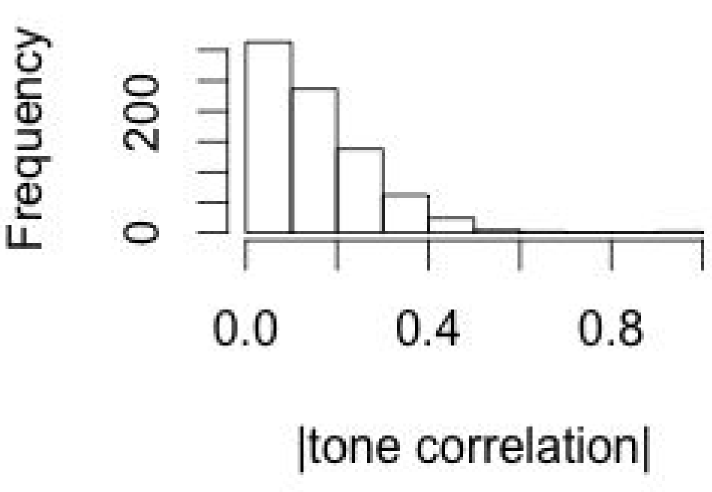
GDELT: Tone

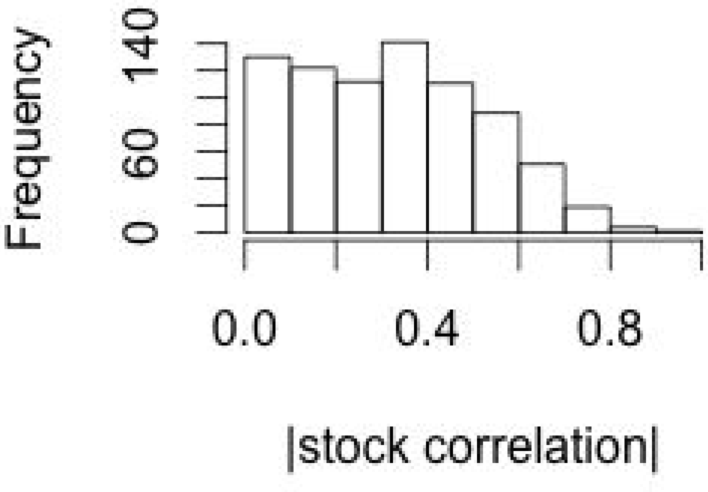
- Organizations analyzed:
  - Top 50 Fortune 500 companies (Walmart, ExxonMobil, Chevron, Berkshire Hathaway, Apple, etc.)
- Tone
  - Percentage of words that are positive - percentage of words that are negative
  - Common values range between -10 and +10
- Question
  - Correlation between tones for different companies
  - Correlation between tone and company health (stock price)

GDELT: Tone - Data preparation

- For each company (40 companies)
    - Weekly average (Mon-Sun) tone was downloaded from GDELT for date range 3/1 - 10/31
        - Luckily no missing values
    - Weekly stock percentage change was downloaded from Yahoo for date range 3/1 - 10/31
- Pearson correlation
    - Between each pair of companies (within tone data and stock data)
    - Between tone and stock data for each company

GDELT: Tone - Initial visualization

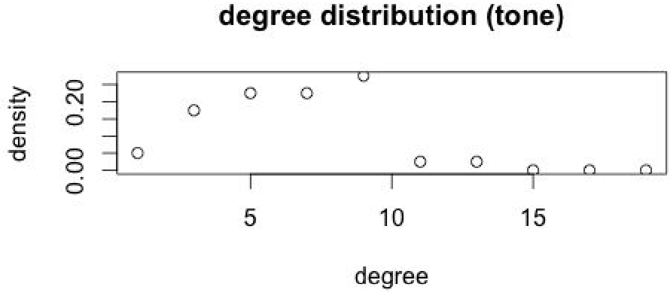


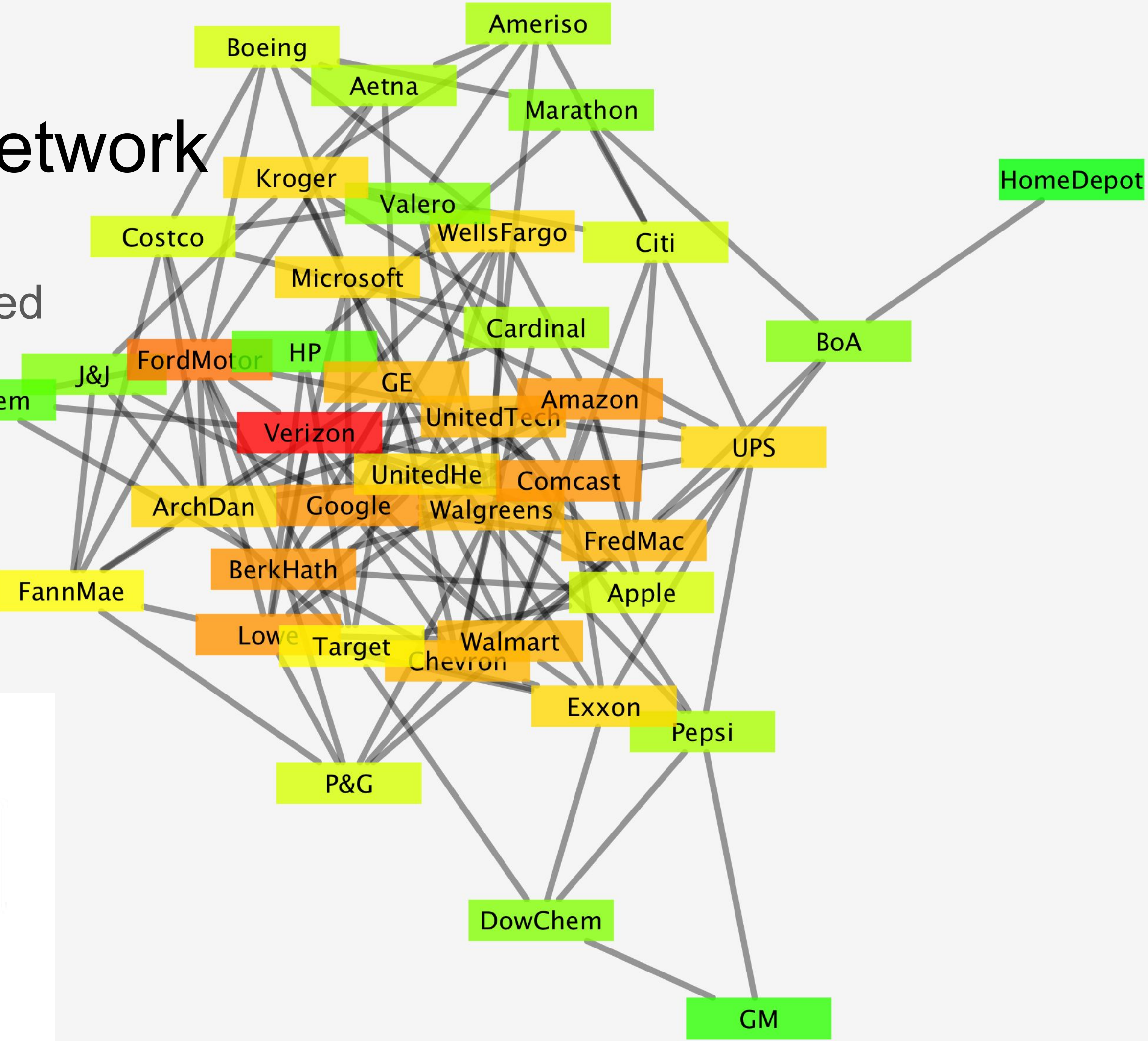


- Sadly, low correlation for tone. May require more filtering to narrow results.
- Also, very low correlation between tone & stock data.
- Let’s study the network anyway...

GDELT: Tone - Network

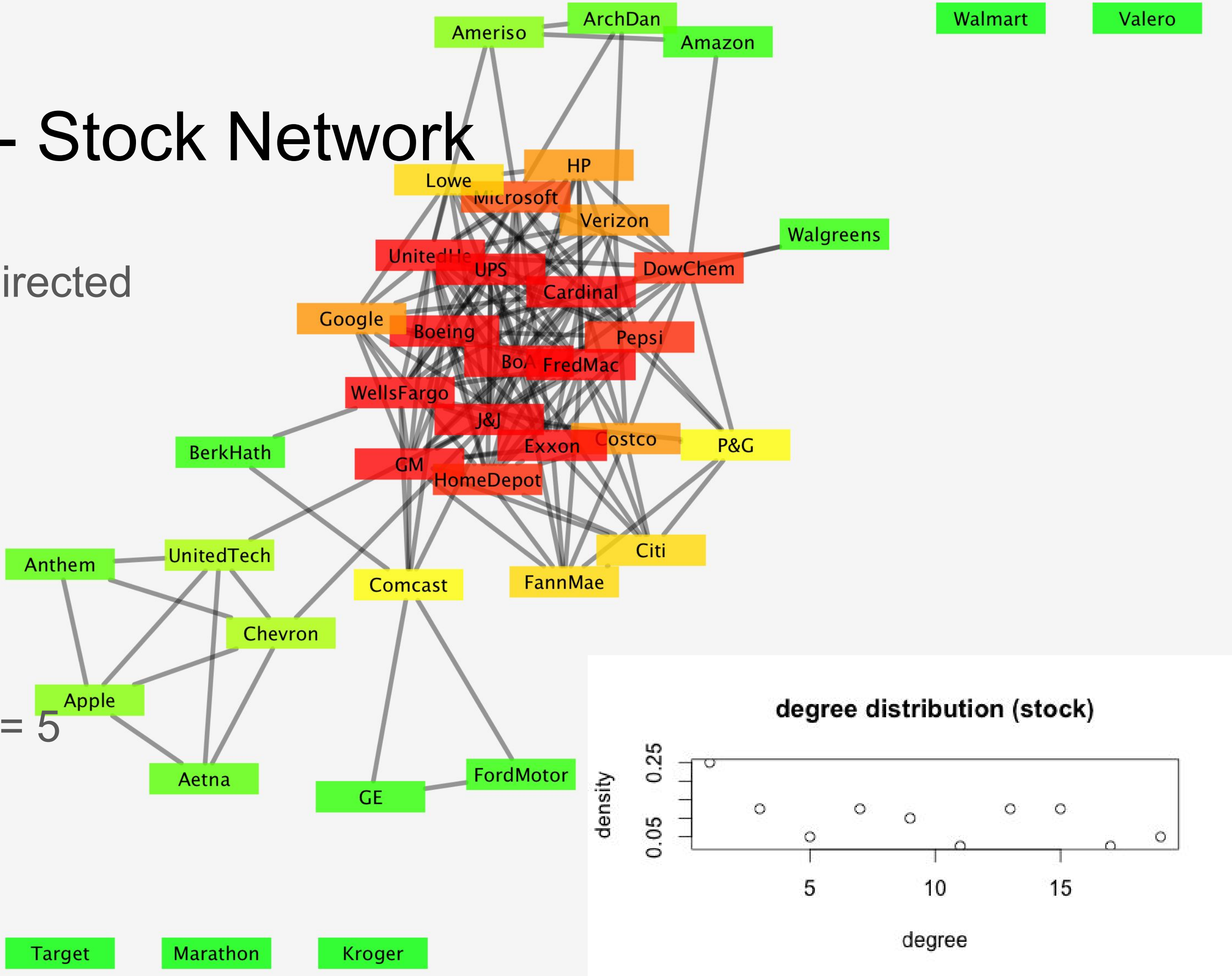
- Unweighted & undirected
  - Link if corr > 0.25
- 1 giant cluster
- <k> = 6.95
- Network diameter = 5





GDELT: Tone - Stock Network

- Unweighted & undirected
  - Link if corr > 0.5
- 1 giant cluster
- 5 isolated nodes
- <k> = 8.32
- Network diameter = 5



GDELT: Tone - Issues/Improvements

- Lack of correlation of tone data
  - Possible causes
    - Low-impact sources (exotic country/language)
    - Company mentioned is unrelated to the main story
    - Simple average of tone may not necessarily be a good measure
    - Low network size
    - Other factors that cause sudden changes in stock value - acquisitions/mergers, stock split, etc.
  - Improvements
    - Weight different sources or only consider articles from major sources
    - Weight for location of the company mention within the article (earlier = more relevant)
    - Incorporate tone “polarity”, a measure of how large positive & negative tones are
    - Incorporate more companies / baseline against stock indices

Event Database: China’s Material Cooperation with the World

http://analysis.gdeltproject.org/module-event-geonet.html

GDELT: EVENT Geographic Network Visualizer

- **GEOREFERENCED NETWORKS**
- **Node**: Two geographical <u>locations</u>, one initiator, one recipient
  - **TWO Location(node) weight**:
    - **Number Events**: total number of unique events, irrespective of how much news coverage each event received
      - **distribution** rather than importance
    - **Number Articles**: total number of news articles covering events found at that location.
      - **importance** rather than distribution
  - represent by **color** (light to dark)

GDELT: EVENT Geographic Network Visualizer

- **Link**: Events
  - color:
    - green: cooperation
    - red: conflict
  - Material / Verbal
  - Search Criteria:
    - event location
    - event code*: predefined type for events

TAMEO taxonomy: http://data.gdeltproject.org/documentation/CAMEO.Manual.1.1b3.pdf

GDELT: EVENT Geographic Network Visualizer

- **Cutoff Threshold**
  - in case the network runs too big or too small
  - default
    - Node:10
    - Edge:5
- **Date Range**
  - Limit the time period of analysis

Can see cities of the world are connected through events matching the search

Motivation

1. **China’s role in the world?**
2. **How does that evolves with time?**

*IMF agrees to include China's RMB in benchmark SDR currency basket

http://www.cnbc.com/2015/11/30/imf-agrees-to-include-chinas-rmb-in-benchmark-sdr-currency-basket.html

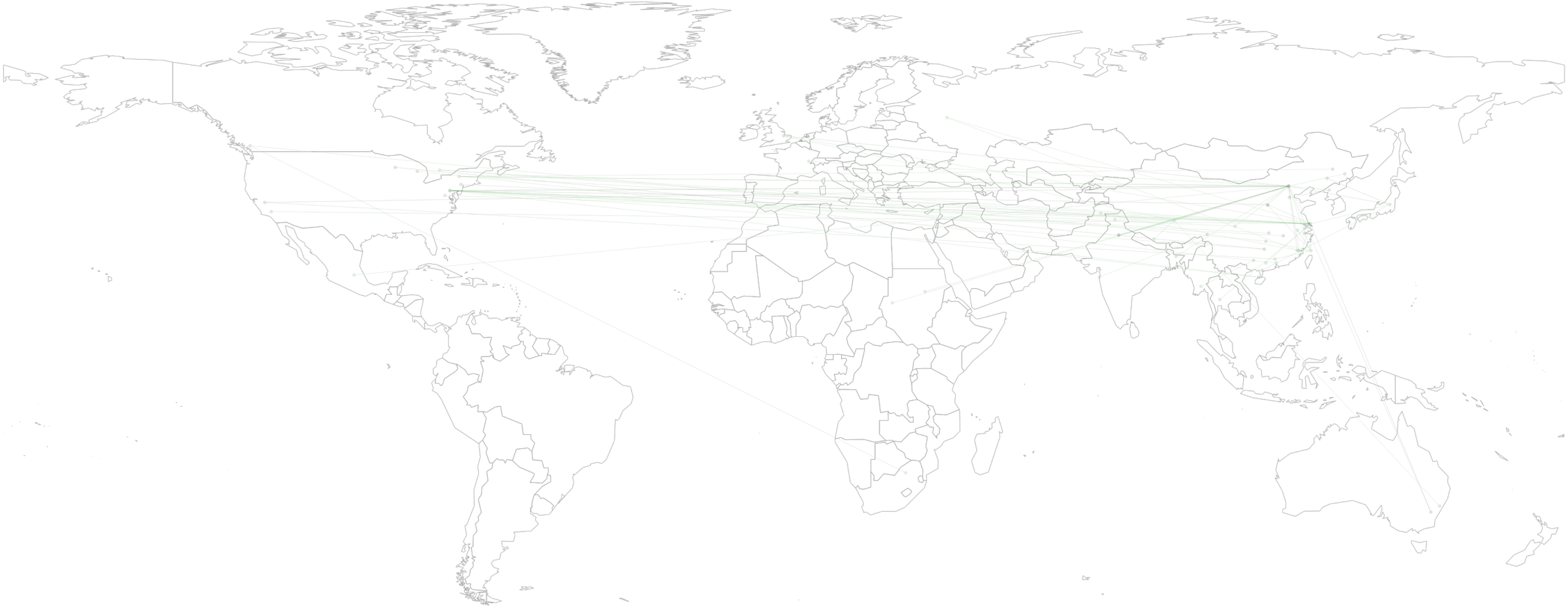
China’s Material Cooperation with the World

- Actor 1: China; Actor 2: Not specified
- Event (link) Class: Material Cooperation
  - could be economic, military, judicial, and etc.
- Period: 2 years
- From 2005 to 2015, 10 years
- Reference:
  - **commercial reports from Ministry of Commerce of the People’s Republic of China, Comprehensive Department**

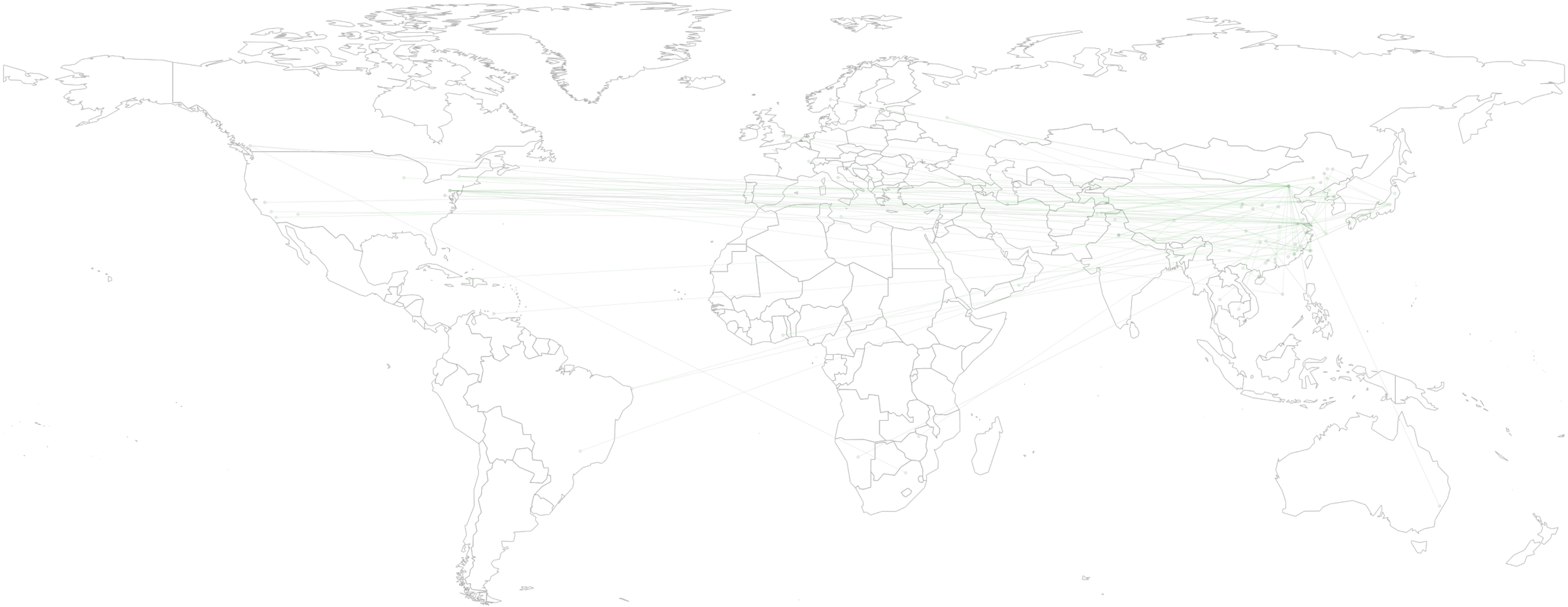
http://zhs.mofcom.gov.cn/article/cbw/

05-07

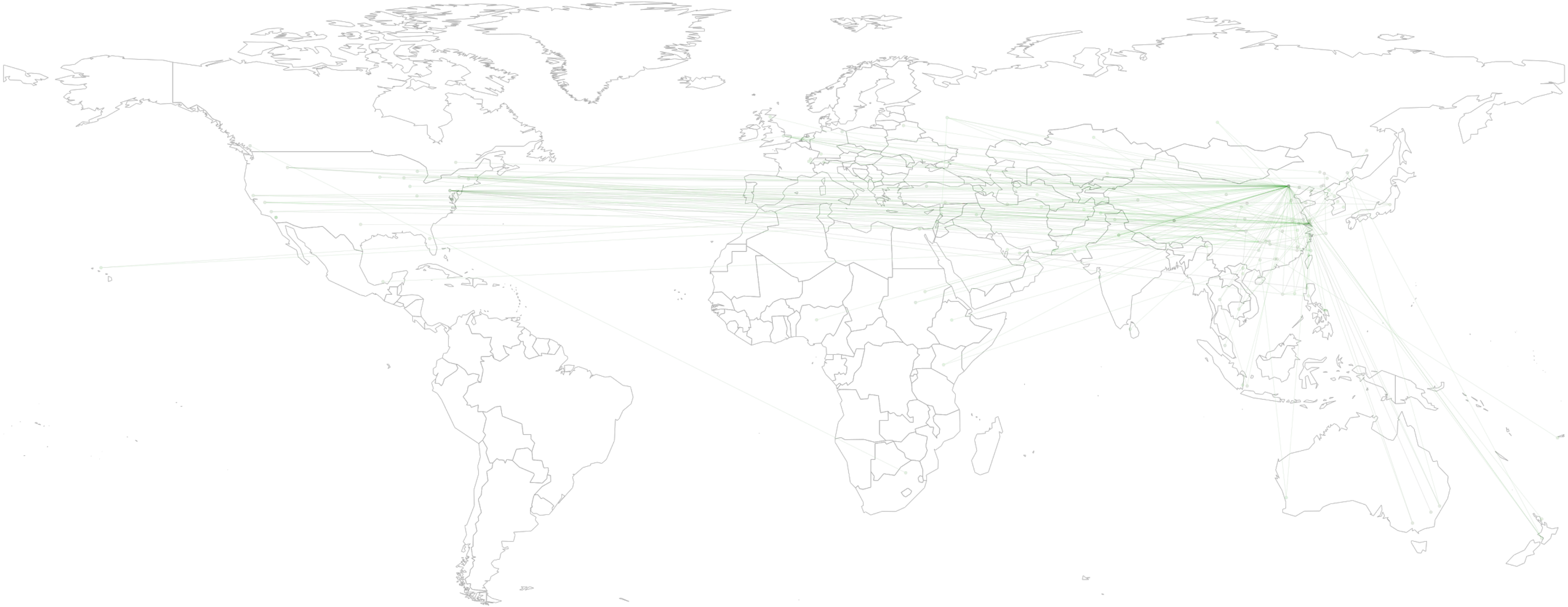
07-09



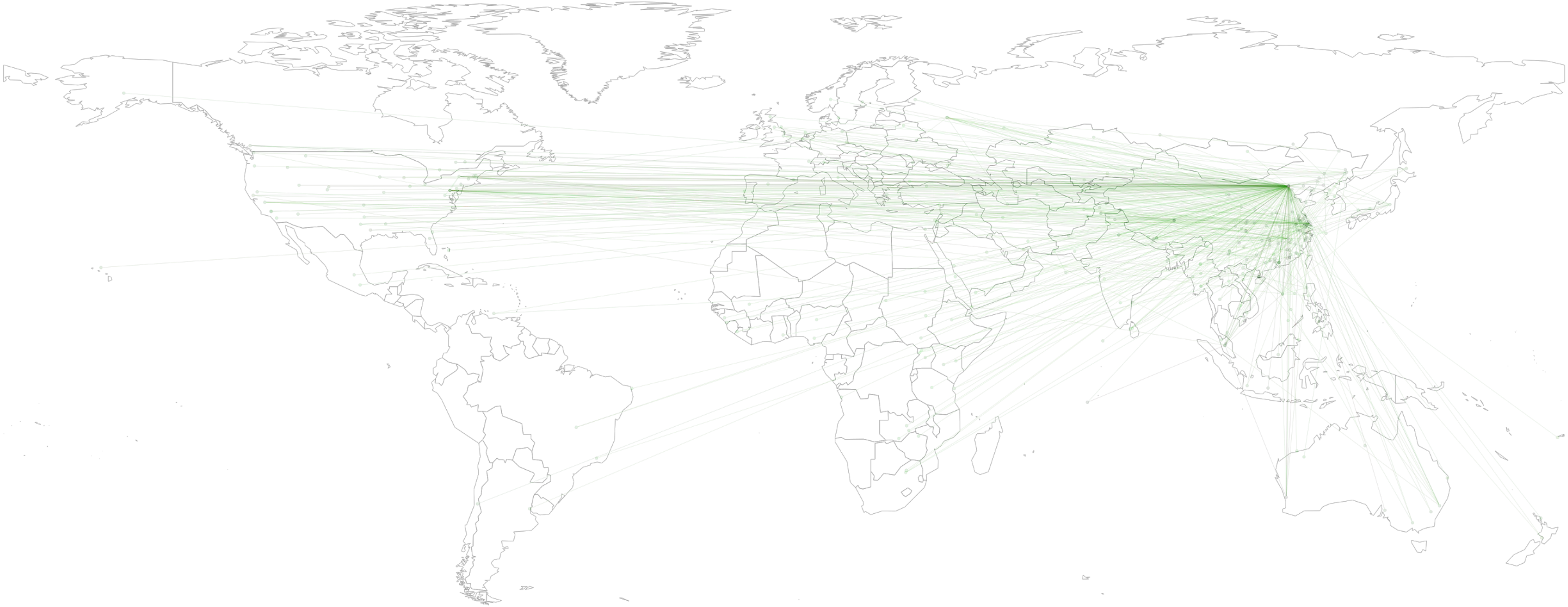
09-11



11-13



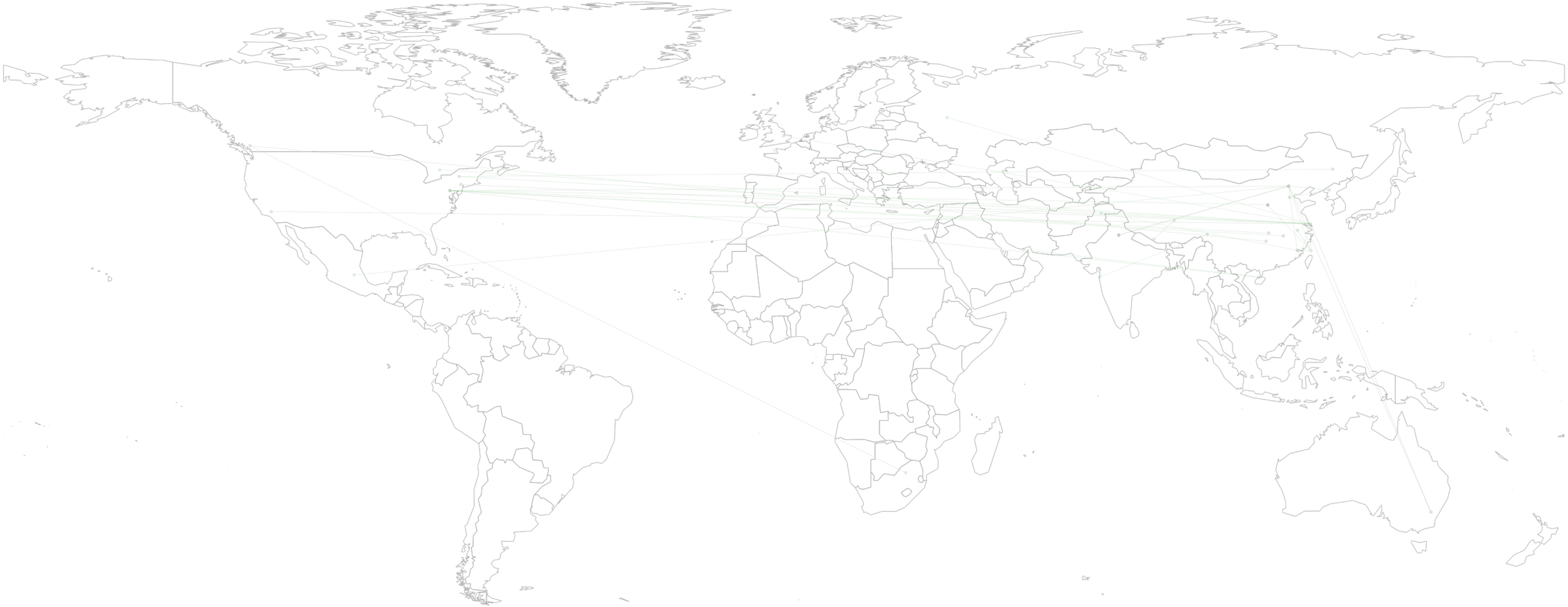
13-15



06-07

07-08

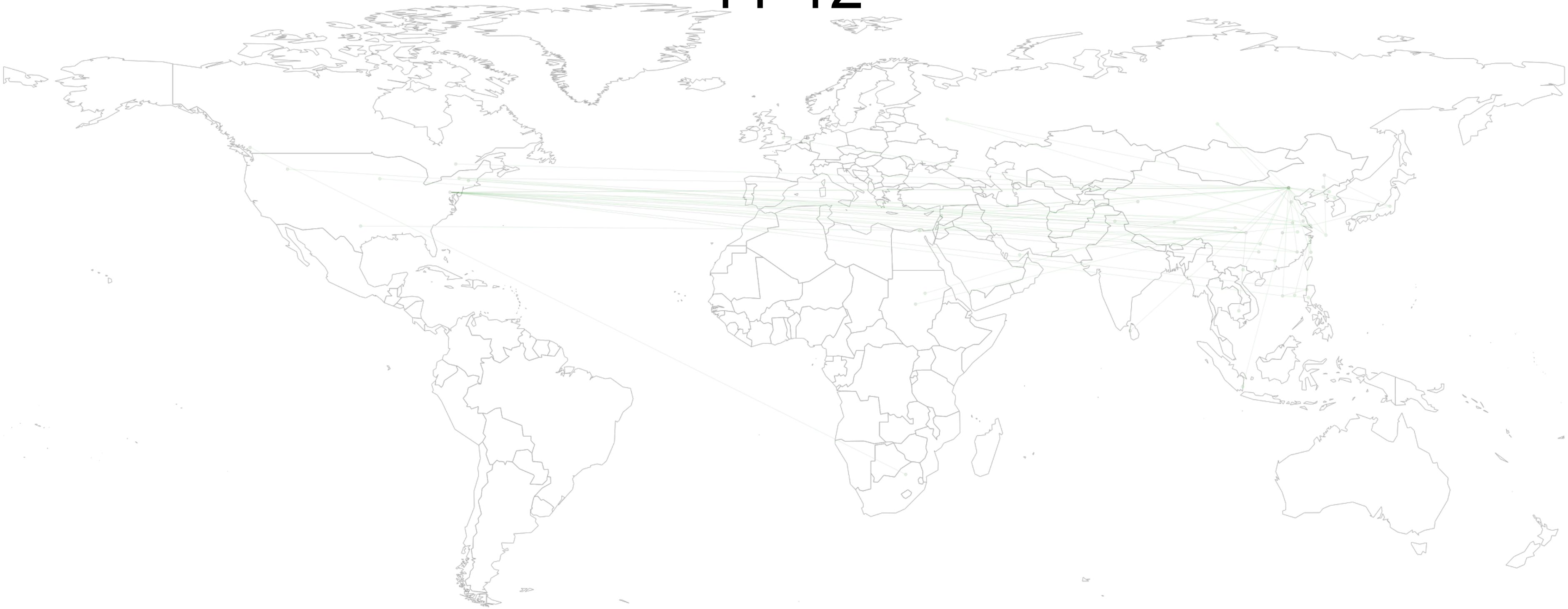
08-09



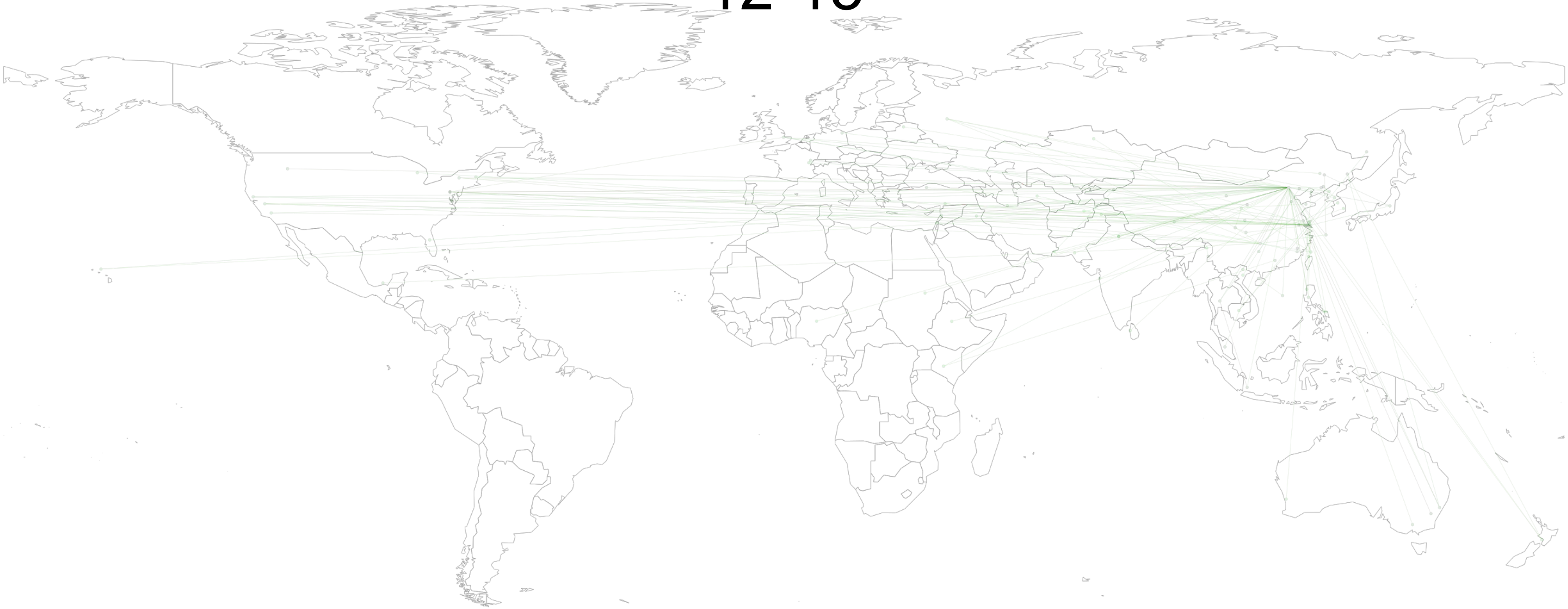
09-10

10-11

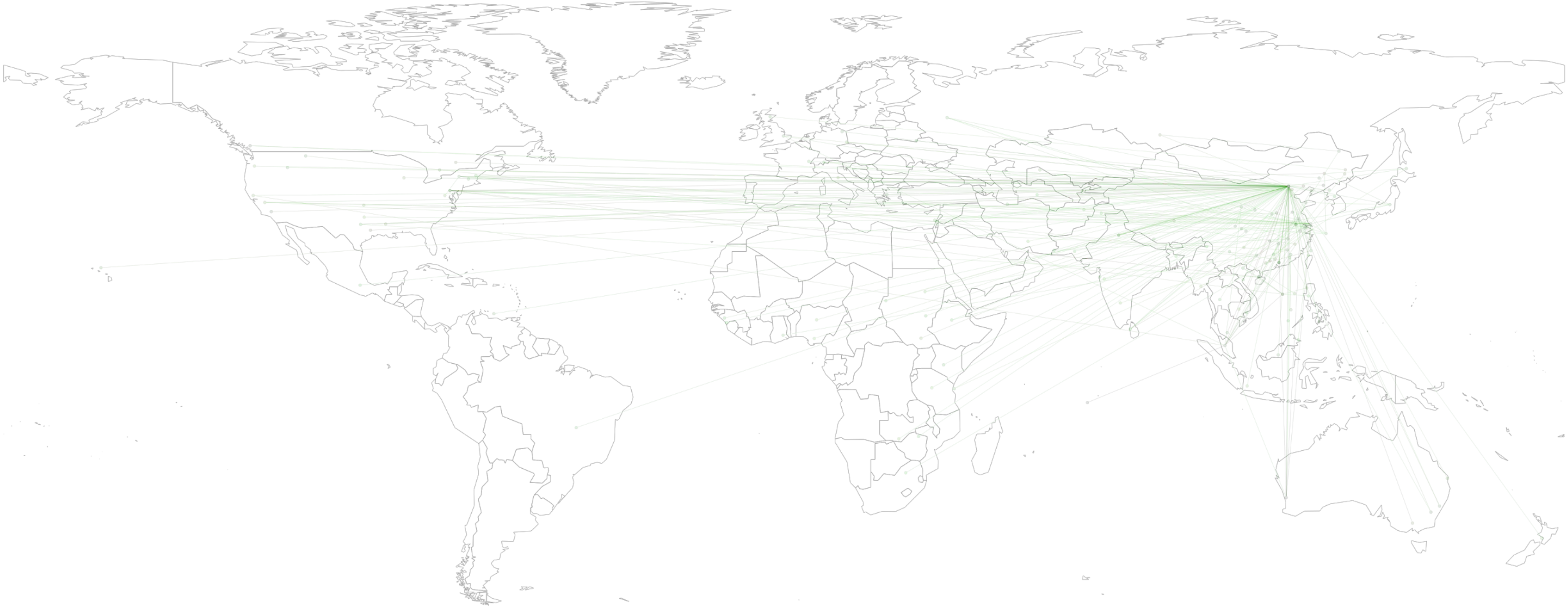
11-12



12-13



13-14



14-15