

Big Data Potential of the Global Database of Events, Language, and Tone (GDELT)

Tidewater Big Data Enthusiasts
Chuck Cartledge
Developer

April 23, 2020 at 12:56 Noon

Contents

| | |
|---------------------------------|----|
| List of Figures | i |
| List of Tables | ii |
| 1 Introduction | 1 |
| 2 Background | 1 |
| 3 Data sources | 1 |
| 4 Google Big Table / Big Query | 2 |
| 5 GDELT Event Data Explorations | 4 |
| 6 Conclusion | 5 |
| 7 References | 9 |
| A Misc. files | 11 |

List of Figures

| | | |
|---|--|---|
| 1 | Accessing GDELT via Google BigQuery. | 3 |
| 2 | GDELT event data for the world, 17 October 2017. | 6 |

| | | |
|---|--|---|
| 3 | GDELT event data for the world, 17 October 2017, event codes 50 - 57. . . . | 7 |
| 4 | GDELT event data for the north west quardasphere, 17 October 2017, event codes 54, 55, and 56. | 8 |

List of Tables

| | | |
|---|---|---|
| 1 | The six most frequently reported event codes. | 5 |
| 2 | Completeness of actor codes. | 5 |
| 3 | Most common source domains. | 9 |

1 Introduction

The Global Database of Events, Language, and Tone (GDELT) is the largest, most comprehensive, and highest resolution open database of human society ever created. Creating a platform that monitors the world’s news media from nearly every corner of every country in print, broadcast, and web formats, in over 100 languages, every moment of every day and that stretches back to January 1, 1979 through present day, with daily updates. The GDELT has created a database of a quarter billion georeferenced records covering the entire world over 30 years.

We’ll take a peek into GDELT and see what we can do with a standalone application, and also what can be done using Google’s Big Table technology.

2 Background

*“A Global Database of Society
The GDELT Project is an initiative to construct a catalog of human societal-scale behavior and beliefs across all countries of the world, connecting every person, organization, location, count, theme, news source, and event across the planet into a single massive network that captures what’s happening around the world, what its context is and who’s involved, and how the world is feeling about it, every single day.”*

GDELT Staff [4]

3 Data sources

We will be looking at the GDELT Event database, not the GDELT Global Knowledge Graph.

According to the GDELT website, the Event Database records over 300 categories of physical activities around the world, from riots and protests to peace appeals and diplomatic exchanges, georeferenced to the city or mountaintop, across the entire planet dating back to January 1, 1979 and updated every 15 minutes. For each of these events, nearly 60 attributes are automatically extracted, or derived from data about the event.

The GDELT event database comes in two different flavors (basically pre- and post- 19 February 2015. The event data fields pre February 2015 are a subset of the post February fields. Some of the potentially interesting aspects of the new data fields are¹:

1. Real-time Translation of 65 Languages.
2. Real-time Measurement of 2,300 Emotions and Themes.

¹<http://blog.gdeltproject.org/gdelt-2-0-our-global-world-in-realtime/>

3. High Resolution View of the Non-Western World.
4. Relevant Imagery, Videos, and Social Embeds.
5. Quotes, Names, and Amounts.

All of which have the potential to provide really, really interesting insights and data for a plethora of applications.

4 Google Big Table / Big Query

The GDELT database is available online at:

<https://bigquery.cloud.google.com/table/gdelt-bq:full.events>

Access to the GDELT data (see Figure 1 on the following page) is predicated on the “usual” terms to use Google tools[5]. Basically:

1. You must have a Google account.
2. The first 1TB of data processed per month is free.
3. After the first 1TB, you are charged for the number of bytes processed, or stored.
 - (a) Processed: \$5 per TB
 - (b) Storage: \$0.02 per GB per month

Full and current details are available online.²

4. Costs above the free tier are charged to your credit card on file with Google.

For many users, processing a TB of data seems like a large problem. But in the world of Big Data; things only start to get interesting above the 1TB boundary. As a specific example, the following BigQuery script locates and retrieves positional information from the GDELT Knowledge Graph (GKG) for 1 week. The script takes about 15 seconds to run, and processes 159GB of data. Running the script a few times to take care of typos, to identify the correct columns to work with, and ensuring that it is selecting the correct data will rapidly run over the 1TB threshold for the month.

²<https://cloud.google.com/bigquery/pricing>

The screenshot displays the Google BigQuery web interface. On the left sidebar, the 'gdelt-bq' project is expanded, showing a list of datasets including 'extra', 'full', 'crosswalk_geocountrycodetoh...', 'events', 'events_partitioned', 'gdeltv2', 'gdeltv2_ngrams', 'hathitrustbooks', 'internetarchivebooks', and 'sample_views'. Below this, 'Public Datasets' are listed. The main area shows the 'Table Details' for the 'events' table. The table schema is as follows:

| Field Name | Field Type | Field Mode | Description |
|---------------|------------|------------|---|
| GLOBALEVENTID | INTEGER | NULLABLE | Unique ID for each event |
| SQDATE | INTEGER | NULLABLE | Date the event took place in YYYYMMDD format |
| MonthYear | INTEGER | NULLABLE | Alternative formatting of the event date, in YYYYMM format |
| Year | INTEGER | NULLABLE | Alternative formatting of the event date, in YYYY format |
| FractionDate | FLOAT | NULLABLE | Alternative formatting of the event date, computed as YYYYFFFF, where FFFF is the percentage of the year completed by that day. This collapses the month and day into a fractional range from 0 to 0.9999, capturing the 365 days of the year. The fractional component (FFFF) is computed as (MONTH * 30 + DAY) / 365. This is an approximation and does not correctly take into account the differing numbers of days in each month or leap years, but offers a simple single-number sorting mechanism for applications that wish to estimate the rough temporal distance between dates |
| Actor1Code | STRING | NULLABLE | The complete raw CAMEO code for Actor1 (includes geographic, class, ethnic, religious, and type classes). May be blank if the system was unable to identify an Actor1 |
| Actor1Name | STRING | NULLABLE | The actual name of the Actor 1. In the case of a political leader or organization, this will be the leader's formal name (GEORGE W BUSH, UNITED NATIONS), for a geographic match it will be either the country or capital/major city name (UNITED STATES / PARIS), and for ethnic, religious, and type matches it will reflect the root match class (KURD, CATHOLIC, POLICE OFFICER, etc). May be blank if the system was unable to identify an Actor1 |

Figure 1: Accessing GDELT via Google BigQuery. <https://bigquery.cloud.google.com/table/gdelt-bq:full.events>

```

SELECT DATE, coord, cnt from (
SELECT DATE, coord, COUNT(*) as cnt
FROM (
select date, REGEXP_REPLACE(REGEXP_EXTRACT(
SPLIT(V2Locations,','))
,r'^[2-5]#.*?#.*?#.*?#.*?#(.*?#.*?)#')
, '^(.*)#(.*?)', '\\1;\\2') as coord
from [gdelt-bq:gdeltv2.gkg]
where DATE >= 20150322000000 and DATE <= 20150328999999 )
where coord is not null group by date, coord ORDER BY 3 DESC )
where cnt >= 3;

```

Because of the potential costs involved with exploring the GDELT data, a limited set of raw data files were downloaded to the local machine and processed using R.

5 GDELT Event Data Explorations

We decided to focus our efforts on the Event database. GDELT pragmatically analyses content and builds a single record that describes the event in the content[2]. Ideas and concepts that are important:

1. Actors 1 and 2. Actor 1 does something to Actor 2. Each actor is identified by a code, a name, a country code, a country name, a known group code and label, ethnic code and label, a religion code and label, a CAMEO[3] code and type.
2. Every event has action attributes. Whether or not this is a root event (answering the question about whether or not this event is the main reason for this report), an event code and description, classification of the event, a Goldstein scale[1] value for the event, number of mentions of this event across all sources, the number of sources, the number of articles reporting this event, and the average tone of the mentions.
3. Geographic information about actor 1 and actor 2. This meta data include: the resolution of the data, the actor's full name, country code, the associated FIPS10-4 code, the latitude and longitude of the actor.

Our explorations will focus on a single day 17 October 2017 from the GDELT database.³ All GDELT events were plotted for the single day (see Figure 2 on page 6). At the top of the image are all the event codes that are plotted, along with the total number of events. In this figure, there are 223,392 events. The blue circles are actor 1 locations, and the red circles are actor 2 locations. Actor 1 did something to actor 2. The event code is what actor

³The program used to query the GDELT and create all of the images used in this report is attached to this report.

Table 1: The six most frequently reported event codes.

| Code | Explanation | Reports |
|-------------|-----------------------------|----------------|
| 10 | Make public statement | 17,551 |
| 42 | Consult, make a visit | 16,373 |
| 43 | Consult, host a visit | 15,039 |
| 40 | Consult | 13,606 |
| 30 | Express intent to cooperate | 12,691 |
| 51 | Praise or endorse | 12,500 |

Table 2: Completeness of actor codes.

| | | Actor 1 | |
|---------------|-------------|----------------|------------|
| | | Good | Bad |
| Actor2 | Good | 143,294 | 20,272 |
| | Bad | 59,874 | 0 |

1 did. The size of the circle represents how many times actor 1 did something to actor 2. In the lower left is a summary of the positional data for actors 1 and 2. Good positional data means that both latitude and longitude data was available for the actor. Bad means that at least one piece of positional data was bad. In the figure, there are 19,861 events where the location of actor 1 was bad, and actor 2 location data was good. In many ways, there is too much data on the screen to be useful. Even reducing the event codes to the ones in the 50 range, doesn't help too much (see Figure 3 on page 7). Limiting the event codes to 54, 55, and 56 (see Figure 4 on page 8).

Other data pulled from this world view include:

- The most frequently reported event codes (see Table 1).
- How often actor codes are reported (see Table 2).
- Which are the most common source domains (see Table 3 on page 9).

6 Conclusion

The GDELT event database is a good place to start looking at how actors (countries, governments, NGOs, and other identifiable entities) interact with each other. At times the data is incomplete, either actor positional data, or identification data is missing. A link to the original source data included in each event record so when incomplete data is detected,

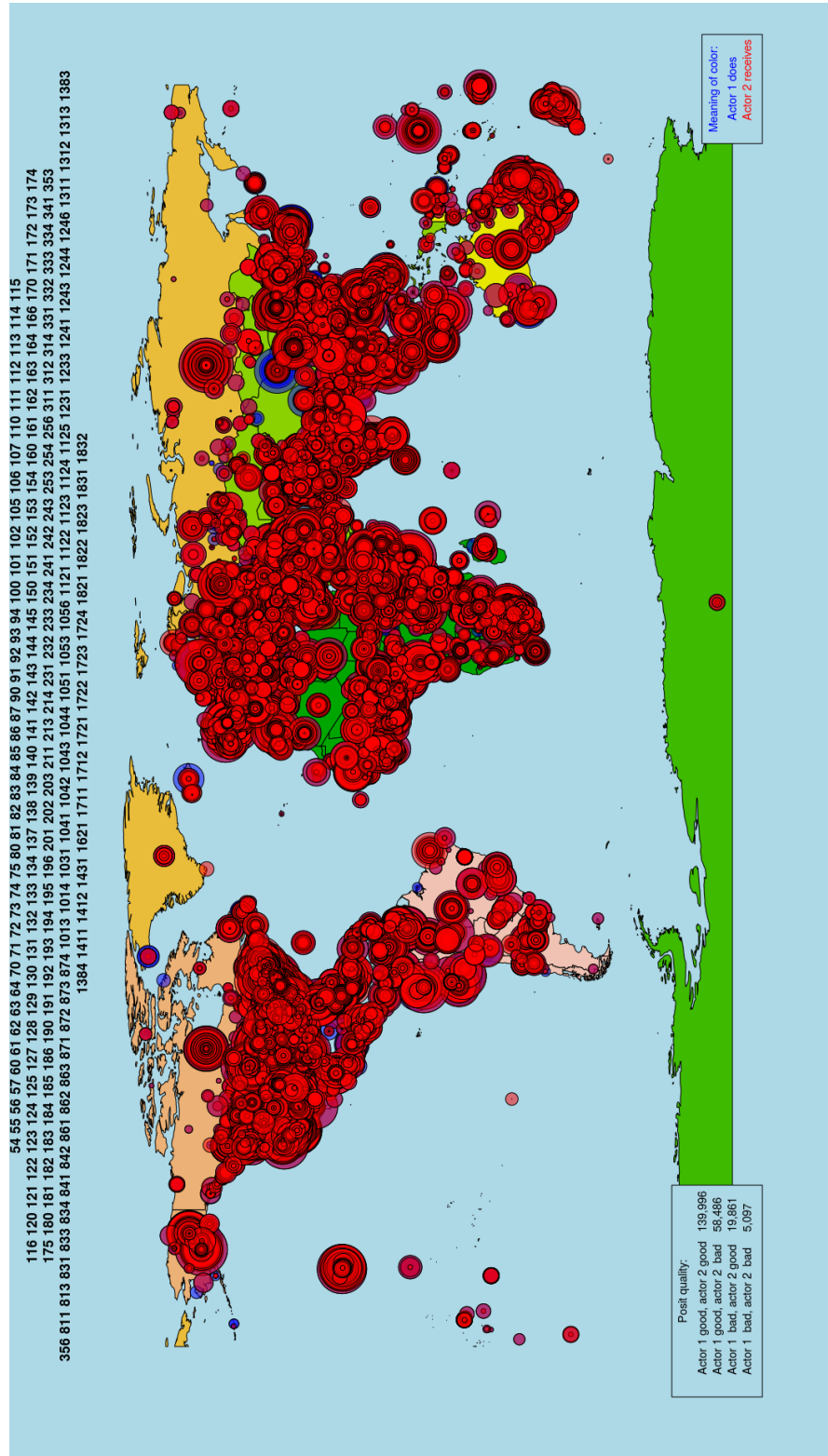


Figure 2: GDELT event data for the world, 17 October 2017.

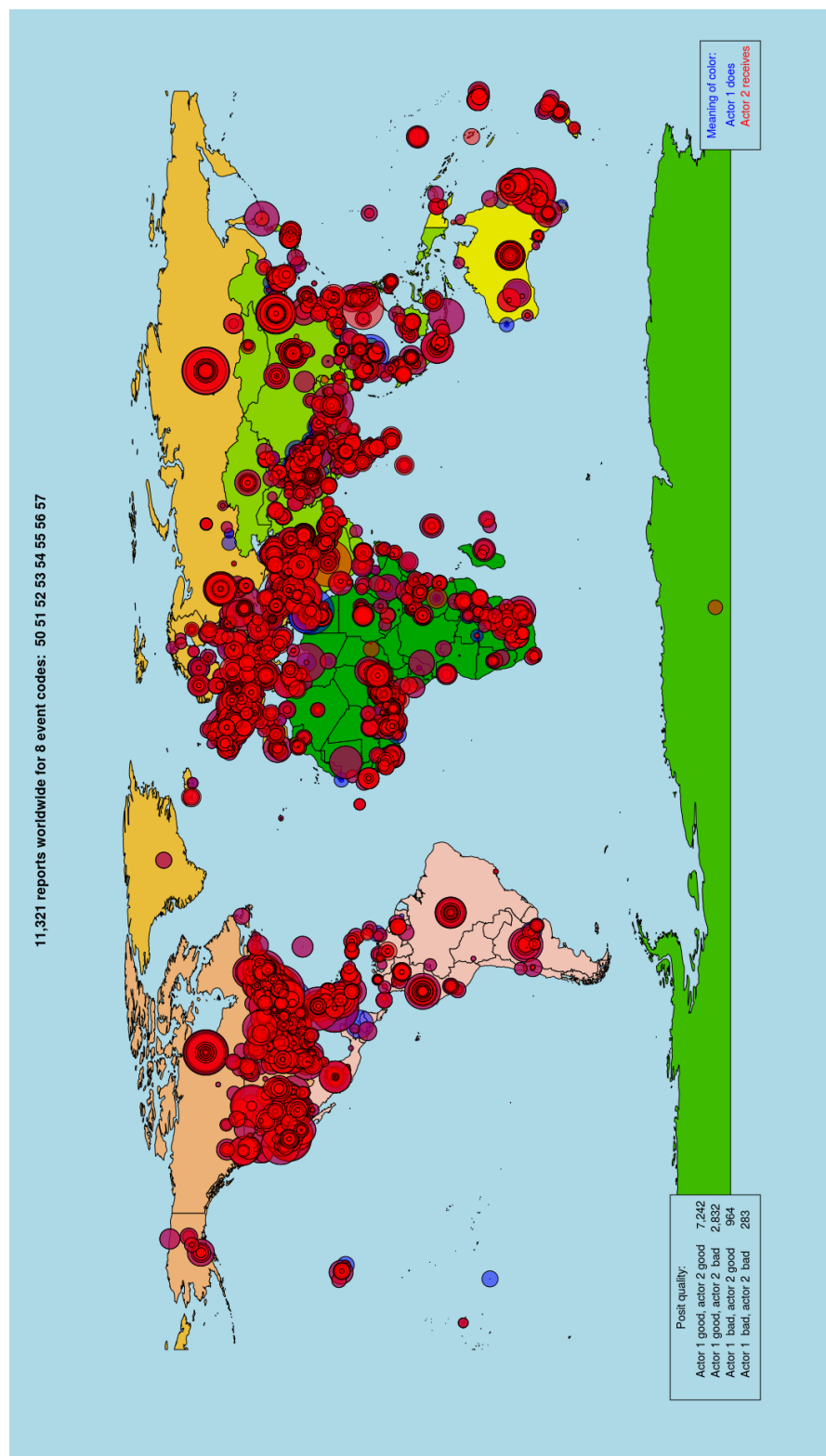


Figure 3: GDELT event data for the world, 17 October 2017, event codes 50 - 57.

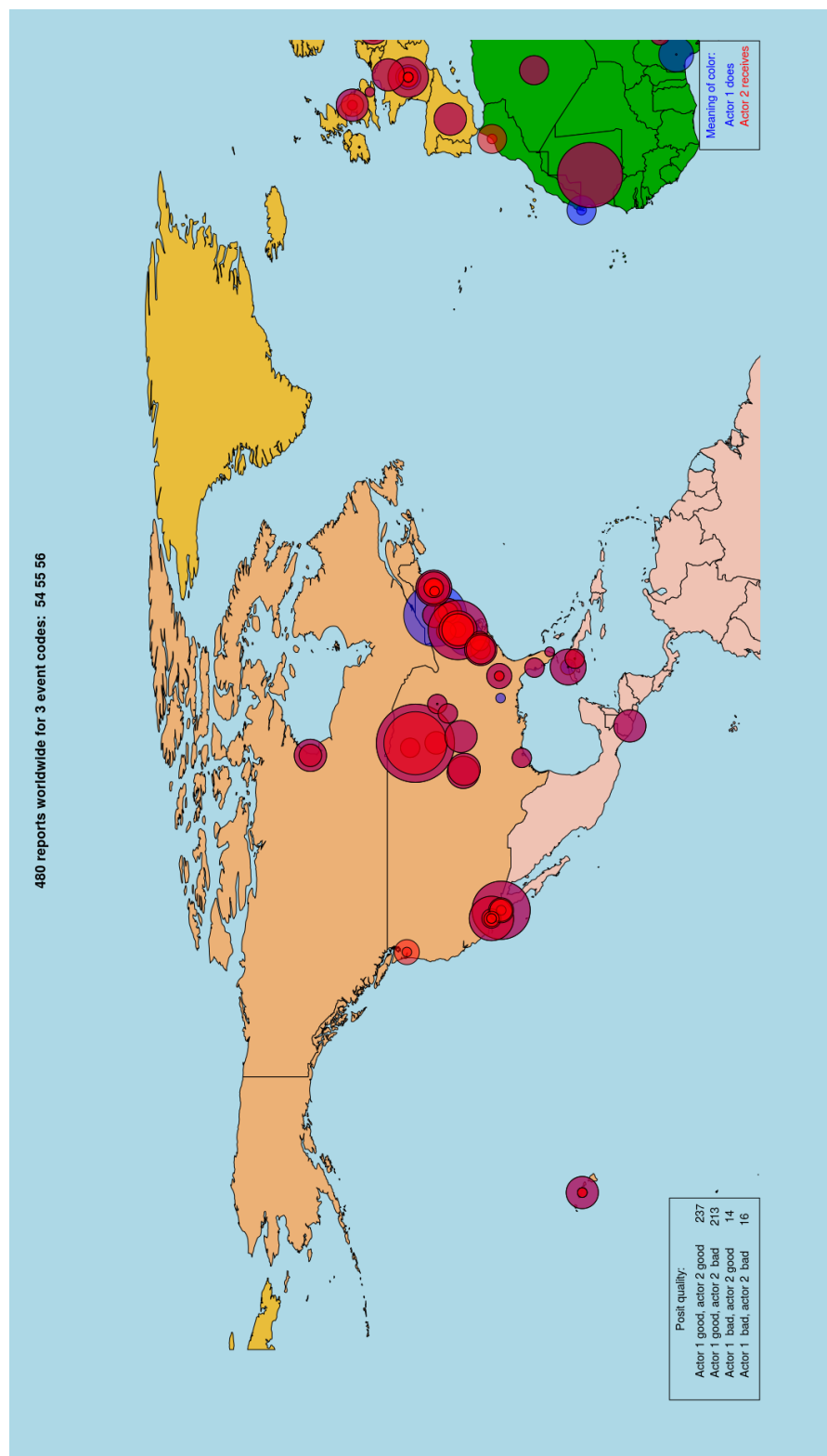


Figure 4: GDELT event data for the north west quardasphere, 17 October 2017, event codes 54, 55, and 56.

Table 3: Most common source domains.

| Source domain | Count |
|---------------------|-------|
| www.yahoo.com | 1,502 |
| allafrica.com | 1,328 |
| www.newsnow.in | 923 |
| www.dailymail.co.uk | 900 |
| www.4-traders.com | 891 |
| www.thehindu.com | 547 |

additional processing could be applied the source data in an effort to improve the quality of the analysis.

The number of events recorded in the single event file seemed low ($\approx 240K$). Based on the promotional information on the GDELT web site, I was expecting more. Approximately 14% of the events were reported by the top 6 domains, so if lots and lots of sources were be monitored and analyzed on a daily basis, it just seems like there should be more events, and more varied sources of data.

Now that I understand the some of the GDELT data, how it is collected, analyzed, and presented, I think there are many ways the data could be used to track the relationships between different levels of organizations over time. While any single day’s collection of event data may be “noisy” or less than perfect, in the long run the noise should cancel out.

Looking at GDELT data in the long run is where the Big Data aspects come into play. A single home machine is sufficient to handle one, or a few days worth of data, ($\approx 10MB$ of event data per day), to look at long term trends (months or years) would probably be out of scope. Looking at data over the long term, particularly trying to delve deeper into the original event data may raise other questions and concerns because “link rot” seems to be about 7-9% per year. So the older the source material is, the less likely it is to be around.


7 References

- [1] Joshua S Goldstein, *A conflict-cooperation scale for weis events data*, Journal of Conflict Resolution **36** (1992), no. 2, 369–385.
- [2] Kalev Leetaru and Philip A Schrodtt, *Gdelt: Global data on events, location, and tone, 1979–2012*, ISA Annual Convention, vol. 2, Citeseer, 2013.
- [3] Philip A Schrodtt, *Cameo: Conflict and mediation event observations event and actor codebook*, Pennsylvania State University (2012).
- [4] GDELT Staff, *The gdelt project*, <http://gdeltproject.org/>, 2016.

- [5] Jordan Tigani and Siddartha Naidu, *Google bigquery analytics*, John Wiley & Sons, 2014.

A Misc. files

The files used to create all these figures are attached to this report. They are:

1. gdeltReport.R – an R script to explore the GDELT data 
2. CAMEO.Manual.1.1b3.pdf – the Conflict and Mediation Event Observations Event and Actor Codebook 