# Better Governing through Data Science- An Analysis of Inter-State Relations using Spark, with Information from GDELT

Chris Westerman, Tanner Johnson, Tom Shaw

## 1. Introduction

With globalization becoming a larger issue than ever before, government officials seek tools to accurately assess their relations with foreign actors. The 115th congress has so far spent over $10 million within the Foreign Affairs Committee (Opensecrets.org) which has the role of informing our congress on issues relating to international trade, foreign relations, and cultural awareness. As data production exceeds human comprehension, special tools are needed to rationalize the torrent of geopolitical news. Since it's nearly impossible to even the most decorated political advisor to parse through the hundreds of daily news articles, we provide a data visualization tool to coalesce every news article published on a daily basis. With properly informed politicians, the whole of a nation can prosper. Our research aims to provide a broad sweeping visualization of a specific actor's relationship to each other nation on a daily basis which helps to provide a birds eye view of global events.

By leveraging information produced by the GDELT event database, our program sifts through over 200,000 articles searching for ones that relate an actor to every country in the UN's database. The GDELT data set provides a useful field referring to the tone of the article which can be used to get a generalized view of the article without having to even read the article; using these metrics, we can provide a first look a the daily interactions between countries. From the data pulled out from various days of GDELT's information, we provide an informative chart that can be read by political officials. These charts can point out glaring holes with our foreign relations that can inform the future interactions with negatively perceived countries. Positively viewed countries can be seen as representing successful ventures between different actors. Since the graphs represent a window spanning several days or months, patterns of excessively negative interactions can be picked out visually and influence foreign policy decisions.

In addition to our snapshot graphs, we provide a day by day choropleth, essentially a geographically bound heat map, that visually expresses tone using a red to blue scale overlayed on a map of the world's various countries. This provides a fun and informative representation of the data collected that is useful for displaying ad hoc analysis of the global political climate.

It's reasonable to assume the Foreign Affairs Committee and its counterparts across the globe don't have access to the computing power or knowledge of data processing to be able to do these kinds of distributed analytics; the political advisers who do may not provide the visualizations to make the data comprehensible. As most people understand the implications of graphical data, we can be certain that the information provided by the output of our project will be easily digested by our elected officials.

2. **Problem characterization**

With the advent of social media, the overwhelming production of data reduces the overall quality we're presented with. Humans now are given a horrendously massive amount of choice when it comes down to every facet of life. On average, an individual can only "efficiently compare five different options" and when presented with more they shut down or make a hasty decision (Psychology Today). Five options represents an infinitesimal fraction of the 500+ global news distributors in addition to the limitless number of personal blogs and various talk shows that provide similar information. Most people, and by extension politicians, will select a limited subset of news outlets and disregard opposing information. Our politicians in the United States are notorious for locking themselves into bias based on their selection of news distributors.

The second problem with the massive amount of information presented in today's political environment is that human processing of international news costs an excessive amount of money and time. GDELT's main page indicates that just a year of news pulled into the GDELT database represents 2.5TB+ of condensed data (gdeltproject.org) which doesn't include the actual contents of the article. Governments won't be able to process every incoming news article at the rate at which they are coming in. As our data production grows exponentially it's unreasonable for any organization to process these records with humans alone. The need for an automation of this process will only grow. While human officials will still be necessary in the governing process, some sort of aid is needed.

According to the 2017 Congressional Research Survey annual report, congressional staff have viewed the CRS website 1.6 million times (CRS 2017). The United States Congressional Research Survey aims to inform the senatorial and congressional staff about pressing issues within and outside of the United States. Other countries have similar institutions like the CRS, an example being the British Library Advisory Council, which seek out global information to inform their legislatures. The sheer number of request illustrates the massive importance of providing a more refined method of scouring records during the information age.

In addition to helping government officials, students, researchers, and citizens currently have no way of receiving an immediate unbiased snapshot of the global political climate. By being able to immediately see a country's global influence, individuals can make more granular investigations about global events without having to sift through insignificant interactions. In order to construct a more politically educated and non-biased populace, data needs to be represented in an easy to read and gather way.

3. **Dominant approaches to the problem**

For many countries in the past the main way that countries determined relations was to exchange diplomats and establish embassies. In this way a country would have a specialist that would have a significant understanding of how relations are fairing. On special occasions there would be Top-Level Visits which constitute the peak importance

of state relations (Orav). However this methodology fails to adequately accomplish quantifying relations since the perspective of a country is heavily dependent on the perceptions of a small group of people in each country. As a result of this many personal biases can develop and 'taint' how different states interact, including the high cost of maintaining a diplomatic mission in every country both in man power and monetarily. In addition this method is only qualitative in nature by nature of the fact that the measure of relations is judged by a person and not a standard metric. However the nature of this understanding is more in depth and nuanced than can simply be expressed through raw data.

In order to address the need for a quantitative metric other studies have approached the problem of quantifying soft power dynamics by measuring their effects on financial interchange such as investment and financial aid (Anguelov). This study sought to investigate how the United States affected Middle Eastern countries by measuring how the fiscal connections changed. However this study fails to account for other resulting variables such as political statements and immigration reports which would also describe foreign relations. This study does succeed at generating metrics that are quantitative in nature and are removed from personal biases. What this study lacks in nuance and multifaceted information, it gains in both accuracy and precision to the metrics it attempts to describe.

## 4. Methodology

When designing our system, we wanted to make it compatible with both the GDELT 1.0 dataset and the GDELT 2.0 dataset. Because the GDELT 2.0 updates every 15 minutes, we needed to create a framework that can rapidly process incoming data and produce data. Apache Spark lends itself well to such processing; unlike Hadoop, it allows for flexible transformations of data. By using the Python version of Spark, Pyspark, this flexibility and expressiveness was even easier to exploit to rapidly develop a system that reflects the current international relationships of any particular country a user is concerned about.

We set the requirement that the software should be capable of being hooked up to a subscription platform such as Apache Kafka with relative ease. We worked with static data for creating a proof of concept, but we nonetheless had to consider the processing overhead when building the system. As such, the computations we perform are fairly simplistic. Our proof of concept is mostly focused on visualizations, because raw data is less appealing to a politician that needs to make quick decisions.

Currently, the system is designed to work with a directory containing GDELT 1.0 daily snapshot data files, each of which is about 50 to 100 megabytes in size. Though each file individually is rather small, the goal was to design a system that can work with years of data if necessary. Furthermore, these snapshot files are actually just a fraction of the actual data going into GDELT 2.0. In a streaming environment, we would also need to store previous data up to a certain user-specified threshold. The software is designed to read, process, and write its data and visualizations as quickly as possible to avoid running out of memory or falling behind on incoming data.

To accomplish such a streamlined system, we first filter out as much irrelevant information as possible. Our software is focused on providing a snapshot of international relations for a single actor in the GDELT dataset, such as the United States or Russia. A global look would be exponentially more expensive to produce, and it would ultimately be less meaningful to a politician or analyst using our tools. A company, organization, or government interested in the current international relationships of multiple countries can easily run multiple instances of the software across their cluster. This is an inherent benefit of providing a lightweight system. Given a single country to process, we are able to instantly filter on the “ACTOR 1” field of a GDELT file, which describes which country initiated an action that was reported about and logged into the dataset. We filter out any row of data that does not feature the user-specified primary actor.

The GDELT dataset provides information about not just international actions but intranational ones as well. For example, in cases of political tension, civil war, or protests, the article reported in the GDELT will have the same primary and secondary actor. Our software is focused on providing a glimpse at international relations. Therefore, we are also able to ignore information where the specified primary actor is also the secondary actor in an event. Should we wish to look at intranational circumstances, which may be useful for senators or ambassadors, we could remove this filter and examine the additional actor fields that describe characteristics such as religious or organizational affiliation to generate additional insights and visualizations. Spark’s RDD operations make filtering of such a huge dataset relatively simple, so a more advanced range of possible user input should be possible in the future.

Our analysis of the filtered data focuses on the average tone field that the GDELT provides. This field analyzes the tone of language based on a number of keywords in each article reporting a certain event to calculate a mean representing the severity of an event. GDELT assumes that articles with more severe language mean a particular event is more important to a country than ones with articles that use milder descriptors. The media plays a large role in the perceptions both citizens and governments have of foreign entities (Coban), so we believe this field fairly accurately reflects the current relationship between countries. To predict the relationship between the primary actor and each secondary actor, we sum and average this field daily. It was important to distill all of the GDELT data to this concentrated field, because creating visualizations required that all of the necessary data be collected at the Spark driver. Collects are dangerous operations if too much data is brought back, which why Spark’s filtering and data transformations were used so liberally.

The core of our project is in visualizations. Without useful visualizations, information is not particularly relevant or powerful in decision making. We leveraged the Python library matplotlib to automatically generate our visualizations. We currently generate a by-country and international view for the input files the user specifies. Each secondary actor that occurs in the timeframe of files given by the user produces a line graph showing the average article tone by day using the data generated in the previous step. As we write the data produced for each set of input to HDFS, we can still easily produce these graphs in a streaming environment.

For the global visualization, we use a choropleth, which is essentially a heatmap that is shaded based on geographic boundaries. To do so, we first create an average of tone per country for all days in the user provided timeframe. This information is then joined with a vector image of the world, provided publicly by Natural Earth, based on country code. Both the GDELT and country vector use ISO 3166 alpha-3 codes to specify country, allowing the join process to be completely automated. From there, each country is plotted onto the vector with color mapping, with more negative tones showing up as a dark red, neutral tones showing up as white, and positive tones showing up as dark blue. This provides a quick overview of international relations. Some example graphs are available at the end of this report.

## 5. Experimental Benchmarks

In determining the efficacy of our solution at quantifying relations with foreign Actors it is impossible to draw immediate comparisons as there are no other solutions that accomplish what we set out to do. As such we cannot draw quantitative to quantitative comparisons to judge the effectiveness. However we can analyze the trends that can be seen in global news. For example our data shows that Russian relations are poor with the United States, which is corroborated by various news sources such as The Washington Post which describes US Russian relations to be “chilly” and “worse than any post Cold War period” (Washington Post). As can be clearly seen relations are not friendly and therefore we can conclude that relations would not be positive, however there are not military acts so relations would not be too far into the negatives.

We can also appeal directly to the source of United States’ foreign relations, the State Department. In their description of bilateral relations with Russia the State Department states “ the United States downgraded the bilateral political and military relationship and suspended the Bilateral Presidential Commission” which is representative of negative relations between the two countries. Later in the analysis the State department discusses the United States openness to building better relations given certain actions from Russia. This gives us more information, that relations should be relatively neutral but on the negative side. Our choropleth shows that relations are around -3.5 meaning that the two countries are not pleased with each other but are not openly hostile. This result matches what other sources are saying qualitatively, of course this is not an exact science but the overall trend is present.

As another example relations in the Middle East also appear to be unfavorable which is to be expected considering the recent history of violence in the area. Considering the presence of military actions by the United States in the Middle East we might expect far more negative results than what we produced; however, the trend is still present. One possible explanation of this phenomena is that after prolonged tensions, the severity of language in news articles may lessen.

Our software was intentionally designed to be somewhat simplistic in the computations it performs so that it would be capable of processing massive amounts of data in a subscription-style environment. Fast processing is a must in such an environment. Our experimental setup was a Spark cluster running on top of HDFS,

consisting of 15 machines in the CSB 120 computer lab. Each machine contains an Intel Xeon E5-2650 v2 eight core processor, clocked at 2.60 gigahertz and 32 gigabytes of RAM. The machines in our setup are usually in use by multiple people simultaneously, so our benchmarks may be inaccurate due to unavoidable noise during benchmarking.

When analyzing the United States, one of the most common primary actors in the GDELT data, across a month of GDELT 1.0 data (approximately two gigabytes worth), execution, including generation of visualizations, takes an average of about 75 seconds. Similar results were obtained for other primary actors on the same data, with Russia also taking about 75 seconds, and Afghanistan taking an average about 70 seconds. Most of the execution time seems to be spent reading the data from HDFS. These numbers are within the performance benchmarks we were hoping to achieve.

## 6. Insights Gleaned

The GDELT dataset provided large sums of data that we did not address using our tool. However we were able to discern some interesting points through our analysis due to the amount of data produced by GDELT and the modularity of how we process this data. For instance in the included chart, relations with the Vatican City (VAT) tend towards negative tones. Which in normal terms means that the Vatican tends to be displeased with most other Actors. However as the chart also shows, this tone can change by 7 point in a single day which goes to show how volatile relations are, though this decline in opinion was likely caused by a global event which the Vatican was clearly not in agreement with. However this proves the efficacy of our process, since this drop in opinion is likely due to some event, we were able to observe this through our outputs.

The choropleth of the United States relations with foreign Actors on the other hand shows a lot more information. For example the general tone of interactions tends to be negative meaning that the United States is not currently happy with most other countries. However there are no severe relations as most relations are in the range of -5 to 5 on a scale of -100 to 100 meaning that the opinions expressed are minor inclinations but are statistically significant.

Looking at continents instead of globally it can be seen that the United States has relatively neutral relations across Europe and South America. This being the result of averaging somewhat positive relations and somewhat negative ones. However most of Asia, the Middle East, and Northern Africa are not favorably viewed by the United States, as can be seen by the dark red present in many of these regions. That being said it's interesting to see how neutral the United State's relations are with foreign countries. As previously discussed most of the relations can only be classified as mildly displeased or mildly friendly. Considering the past between the US and England for example shows a long history of being allies and yet relations are effectively neutral if not mildly displeased.

## 7. How the problem space will look like in the future

Currently, the scope of our analytics is limited to data provided by the operator. In order to supply a consistent flow of data, we most likely pull data on a daily basis from

GDELT and place it within a streaming environment. By removing the human element from operating our program, it becomes easy for any agency to start collecting information without having to understand the specifics of the program. This goal could be attained by using Apache Kafka to pull data and run streaming jobs on newly fetched data. Using the streaming environment would also enable our application to fully take advantage of the GDELT 2.0 data sets. In addition to all the GDELT 1.0 information, version 2.0 adds a few new columns and is updated in 15 min intervals. We can then represent data in real time over a period of hours rather than days.

As with most data visualization tools, getting the information in a reliable and user friendly way is paramount. Eventually, we can build a web front-end to house all of the analytic data and provide an easy way to interact with the program without having to access the command line. From here, we could present a user with a dashboard of information including a personalized time window snapshot or various choropleths in tandem. Such a view would go hand in hand with a streaming environment.

The program produced for this project is mostly a proof of concept. Its calculations are fairly simplistic; future attempts at automated global relationship calculations will likely go in the direction of machine learning and deeper mathematical analysis to provide more accurate numbers. There are some concerns about bias in the GDELT (Weller and McCubbins); additionally, reports of the same events end up in the GDELT multiple times, which can lead to oversights in analysis (Source). Future attempts at the analysis we performed should find ways to assign and predict bias scores and combine duplicate reporting to make information more accurate. The use of data and analytics in politics will likely only grow. Computer science is permeating every sector, and politics are no exception. Datasets like GDELT are evolving and will only become more useful.

## 8. Conclusions

Having set out to create a tool that can effectively quantify foreign relations, by and large we succeeded, with reasonably accurate representations and quick processing times. However our solution is incomplete as we are only making use of a single source of data (albeit a meta source). As such we can only say with confidence that we are addressing some of the aspects that define how one actor may relate to another. However creating a thorough and nuanced analysis of relations is not what our goal was, we wanted to quantify the general tone and provide a quick reference to those who need to act upon such data. To this end we were successful and with the inclusion of additional technologies our solution would be able to produce more statistically significant results, more frequently.

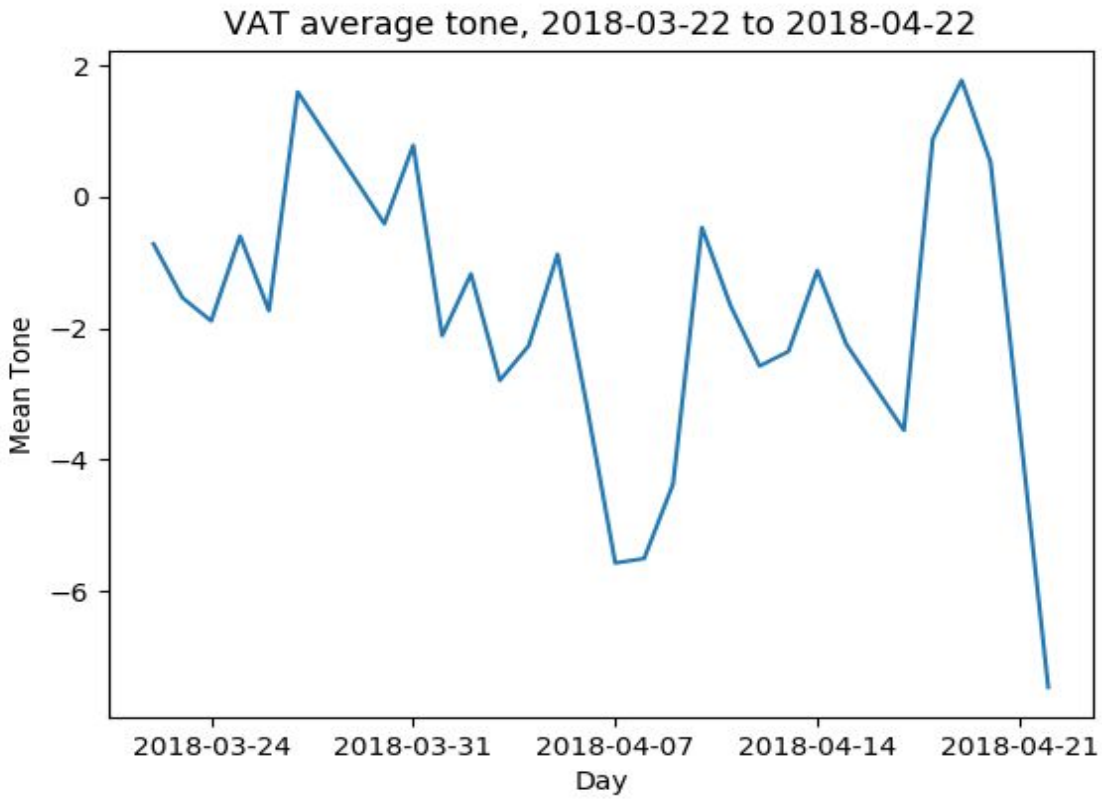
Any discussion of news analysis should also mention that current news outlets may be subject to false leads and inconsistent reporting (ie fake news), while we expect the GDELT dataset to be factually accurate, general tones may be affected by the climate which can at times be impacted by such news.

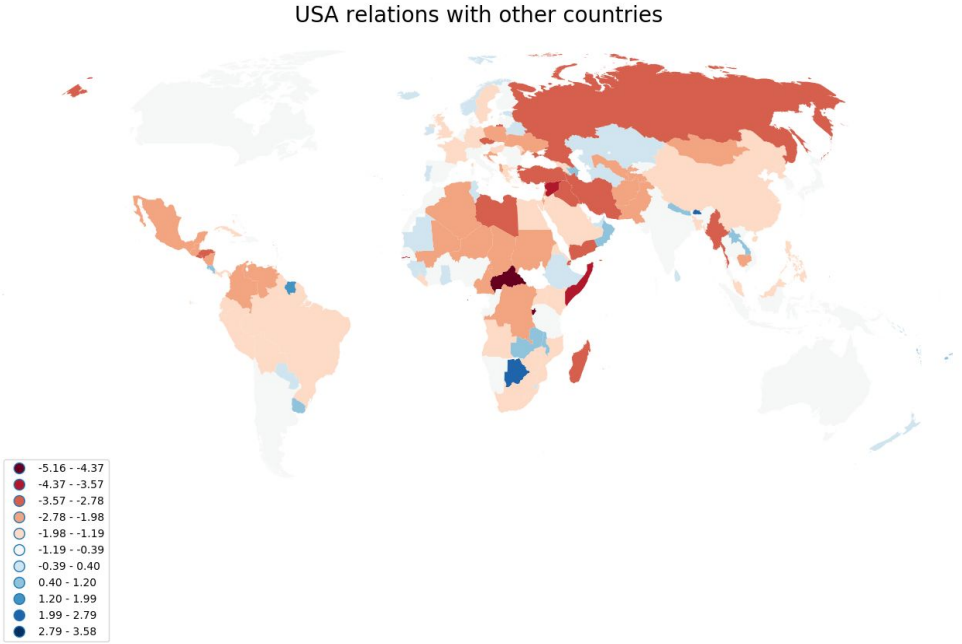
Of course our results are only as valid as the data with which we performed our analysis. We believe that regardless of the validity of the data, that our analysis is sound and can be applied to other similar datasets. However due to the volume of data being

processed we conclude that the data itself possesses enough data points to average out for any mistakes in the formula used by GDELT to quantify the tone of news articles, and as such we can conclude that our analysis can be used with reasonable certainty.

Further work should concentrate on adding additional sources of data such as immigration data, import and exports, and military operations, which will paint a more holistic representation. Our work here is a beginning to a solution to the problem we set out to accomplish. We do not believe our solution to be the ultimate form that a quantifiable metric should be, but it does begin to address the problem. As with all data related projects, more data and of better quality will always produce better results.

**Appendix: Sample graphs**



**Bibliography**


Hunter, J. D. Matplotlib: A 2D graphics environment. Computing In Science & Engineering. 9. 3. 90-95. IEEE COMPUTER SOC. 10.1109/MCSE.2007.55. 2007.

Natural Earth. 1:10m Cultural Vectors. naturalearthdata.com. http://www.naturalearthdata.com/downloads/10m-cultural-vectors/. April 24, 2018.

ISO. Glossary for ISO 3166 - Codes for countries and their subdivisions. iso.org. https://www.iso.org/glossary-for-iso-3166.html. April 22, 2018.

Nicholas Weller and Kenneth McCubbins. Raining on the Prade: Some Cautions Regarding the Global Database of Events, Language, and Tone Dataset. politicalviolenceataglance.org. http://politicalviolenceataglance.org/2014/02/20/raining-on-the-parade-some-cautions-regarding-the-global-database-of-events-language-and-tone-dataset/. February 20, 2014. April 25, 2018.

Source. GDELT and the Problem of Decontextualized Data. source.opennews.org. https://source.opennews.org/articles/gdelt-decontextualized-data/. May 14, 2014. April 25, 2018.

Library of Congress. CRS Annual Report. loc.gov. http://www.loc.gov/crsinfo/about/crs17_annrpt.pdf. April 21, 2018.

OpenSecrets. House of Foreign Affairs 115th Congressional Report. opensecrets.org. https://www.opensecrets.org/cong-cmtes/overview?cmte=HINT&cmtename=House+International+Relations+Committee&cong=115. April 20th, 2018.

Liraz Margalit. The Psychology of Choice. Psychology Today. https://www.psychologytoday.com/us/blog/behind-online-behavior/201410/the-psychology-choice. October 03, 2014. April 23, 2018.

GEDLT. gdeltproject.org. https://www.gdeltproject.org/#downloading. April 20th, 2018.

Aivo Orav. Measuring Foreign Relations. http://vm.ee/sites/default/files/content-editors/web-static/175/Aivo_Orav.pdf. April 24th, 2018.

Nikolay Anguelov, Tiffany Kaschel. Tpward quantifying soft power: the impact of the proliferation of information technology on governance in the Middle East. Palgrave Communications. https://www.nature.com/articles/palcomms201716. March 14th, 2017. April 23rd, 2018.

Filiz Coban. The Role of the Media in International Relations: From the CNN Effect to the Al –Jazeere Effect. Journal of International Relations and Foreign Policy. http://jirfp.com/journals/jirfp/Vol_4_No_2_December_2016/3.pdf. December 2016. April 23rd, 2018.

Karen DeYoung. Putin speech adds to freeze in U.S.-Russia relations. Washington Post. https://www.washingtonpost.com/world/national-security/putin-speech-adds-to-freeze-in-us-russia-relations/2018/03/01/ffab9174-1d8d-11e8-ae5a-16e60e4605f3_story.html?noredirect=on&utm_term=.8ed3b37c4151. March 1st, 2018. April 25th, 2018.

State Department. US Relations With Russia. state.gov. https://www.state.gov/r/pa/ei/bgn/3183.htm. April 23, 2018. April 24th, 2018.