GeoGraph - An international level analysis of alliance and conflicts in GDELT

Antonio Lopez, Artificial Intelligence, 0001083382
Matteo Vannucchi, Artificial Intelligence, 0001084557

1 Introduction

Geopolitics encompasses the dynamic interplay of geography, history, and power between nations. It delves into the intricacies of international relations, analyzing the influences and power dynamics that shape global interactions. As the world becomes increasingly interconnected, the boundaries of geographical distance blur, rendering geopolitical analysis indispensable for comprehending the complex web of modern international affairs. Over time, nations forge alliances and develop enmities, tracing the historical evolution of international relations. Some partnerships persist, while others transform. This constantly evolving geopolitical landscape is actively influenced by major players like the United States, Russia, and China, each striving to shape the global balance in their favor.

Newspaper source turns out to be effective in the description of international events, their continuous streams and capillarity provide an awesome tool to explore and analyze the evolving of conflicts and social relations not only at an international level but also at national and regional. The Global Database of Events, Language, and Tone (GDELT) is a comprehensive dataset built from news articles around the world, in particular, is event-based: each event is described by a source actor, a target actor, and the kind of event that happened. These events range from diplomatic exchanges, and economic aid to threats, intelligence operations, and even military interventions. This allows for diverse analyses, ranging from pinpointing and analyzing a specific conflict [3], to the prediction of future peace [4] and violence [6].

In this project, we aim to construct and analyze a network graph derived from GDELT to investigate the intricacies of international relations. Our focus will be on identifying shifts in diplomatic ties, pinpointing significant conflicts, and delineating the geopolitical blocs that have emerged over time. Through the application of various metrics and algorithms, we intend to uncover patterns and trends that reveal the dynamic nature of global political alliances and confrontations.

A dynamic web visualization can be seen here that will be kept online for at least one month, while here is a link to the GitHub repository.

2 Problem and Motivation

Analyzing and understanding the main relation between nations and geopolitical blocks is crucial for governments, international organizations, businesses, and also for the common citizen. Navigating the complexities of geopolitical analysis poses a significant challenge, not only for the layperson but also for experts in the field. The intricate strategies employed by nations and

international actors add layers of complexity that require subtle understanding and interpretation. Moreover, the sheer volume of data, encompassing both official and unofficial sources, exceeds the capacity for traditional human analysis.
Previously cited works [3][4][6] have sought to incorporate data into the analytical process, focusing on specific case studies to shed light on complex geopolitical dynamics. Similarly, the GDELT project offers a tool for analysis [1], primarily designed for data visualization and retrieval. However, its application is largely confined to presenting data, rather than focusing on deeper investigative analysis or facilitating the extraction of interesting insights. Differently, we propose a data-driven approach that aggregates data to model a graph capable of grasping world-level interactions. In particular, we focus on analyzing friendships and enmities, investigates the dynamics of relationships between two nations over time, identifies the most significant conflicts, pinpoints the key global actors, and delivers meaningful visualizations to succinctly convey our findings. This method does not want to replace human analysis, instead, it provides insight to experts in the field giving them an effective way to take advantage of thousands of data that otherwise would be difficult to interpret.

3 Datasets

GDELT, short for Global Data on Events, Location, and Tone, is "the largest, most comprehensive, and highest resolution open database of human society ever created". This dataset is continually updated every 15 minutes by monitoring the world's broadcast, print, and web news from over 100 languages. This method ensures the dataset captures a wide array of media sources, mitigating biases towards specific types of news outlets and giving a global perspective. Given the immense volume of data involved, direct utilization of the raw GDELT dataset presents significant challenges. For context, the dataset's volume from the previous year alone exceeds 2.5 TB which is far beyond our resource availability. However, GDELT offers a solution through a compressed and aggregated version of its dataset, spanning from January 1, 1979, to February 17, 2014. This streamlined version simplifies the data by aggregating similar events; for instance, all protests occurring in Russia on any given day are merged into a single record. This makes the dataset much more manageable without sacrificing too much information.

3.1 Dataset format

The compressed dataset has the following features for each record:

- **Date**: date of the event in the YYYY-MM-GG format.
- **Source**: the complete raw CAMEO code for the source actor.
- **Target**: the complete raw CAMEO code for the target actor.
- **CAMEOCode**: CAMEO action code describing the action that the source performed upon the target.
- **NumEvents**: number of events, as mentioned above more entries of the same type are collapsed in a single row.

[1] GDELT analysis tool

- **NumArts**: this is the total number of source documents containing one or more mentions of this event.
- **Goldstein**: each CAMEO event code is assigned a numeric score from -10 to +10, capturing the theoretical potential impact that type of event will have on the stability of a country.

Date	Source	Target	CAMEOCode	NumEvents	NumArts	Goldstein
1994-01-01	AFG	AFGELI	61	1	2	6.4
1994-01-01	AFG	GOV	173	2	8	-5.0
1994-01-01	AFGGOV	GOV	30	2	19	4.0
1994-01-01	AFG	LBR	193	1	4	-10.0
1994-01-01	AFG	MIL	151	1	9	-7.2

Table 1: Collapsed data format used

3.2 Preprocessing

The data are processed following these steps:

- **Country code extraction**: since the aim is to model nation-wise links we need to reduce the actors' CAMEO codes to their root to retrieve the nation code. We utilize the first three characters of each code, aligning with the ISO-3166 alpha 3 standard for country identification. While this approach sacrifices some data granularity, it significantly reduces data complexity and facilitates efficient analysis. For example, a code like "ITANGO" that represents a non-governmental organization working in Italy would be associated with Italy.
- **Undesired code removal**: several actor's codes are linked to organizations that cannot be traced back to any specific nation. An example is international organizations like Amnesty International (identified by NGOHRIAMN).
- **Data weights**: to capture the varying importance of events in our model, we assign weights to each data point proportional to $NumEvents \cdot NumArts$. This weighting compensates for the limitations of the Goldstein scale, which lacks consideration for event frequency and media coverage. Ideally, news generating significant media attention, reflecting broader impact, should be deemed more relevant in our analysis.
- **Aggregation**: each record is then grouped by source and target. For each group the total sum, mean, standard deviation, and event count are computed for the Goldstein value. This allows us to define the relationship between a source country and a target country using only one record.

At the end of preprocessing the data have the format shown in Table 2. The last step is to select the most relevant pairs from the data sorting our pairs by *count* and keeping the top 20%.

Source	Target	Goldstein sum	Goldstein mean	Goldstein std	Goldstein count
AFG	AFG	-3282.5	-1.5	6.1	2197
AFG	AUT	-32.0	-4.0	0.0	8
AFG	BGR	461.9	5.4	2.6	85
AFG	BLR	20.3	1.5	0.5	14
AFG	CAN	22.0	1.0	0.0	22

Table 2: Aggregated data

3.3 Network creation

We construct an undirected weighted graph from a dataset where countries are represented as nodes. The edges between these nodes denote the relationships between countries, disregarding the directionality to simplify the analysis. The weight of each edge is determined by taking into account two main factors, reflecting both national and global significance of the interactions. We first need to define three important values:

- **Relation sum**: this measures the aggregate impact of the interactions between two specific countries, source s and target t. It is calculated as the sum of the *Goldstein sum* column for the interactions in both directions ($G_{s,t}$ and $G_{t,s}$). Mathematically, it's given by:

$$r_{\{s,t\}} = G_{s,t} + G_{t,s}$$

- **Total sum**: this represents the overall significance of two countries in the dataset, considering all of their relations. It's calculated as

$$t_{\{s,t\}} = \sum_{(s,x)} G_{s,x} + \sum_{(x,s)} G_{x,s} + \sum_{(t,x)} G_{t,x} + \sum_{(x,t)} G_{x,t} - G_{s,t} - G_{t,s}$$

where we subtract $G_{s,t}$ and $G_{t,s}$ since they would be counted twice.

- **Global sum**: this represents the overall significance of all relations in the dataset. It's calculated by summing the *Goldstein sum* for each source-target pair and serves as a baseline to evaluate the global relevance of individual relationships:

$$g = \sum_{(x,y)} G_{x,y}$$

We can then calculate the two main components:

- **National relevance**: this metric assesses the importance of a specific relationship relative to the total international engagements of the involved countries. It is a measure of how relevant is a relation for the specific countries at play. Calculated as:

$$national_{\{s,t\}} = \frac{r_{\{s,t\}}}{t_{\{s,t\}}}$$

- **Global relevance**: this values the significance of a relationship in the context of worldwide interactions. It views the impact of a specific relation compared to the total global interactions. Defined as:

$$global_{\{s,t\}} = \frac{r_{\{s,t\}}}{g}$$

The final weight is then calculated as:

$$w_{\{s,t\}} = \beta \cdot national_{\{s,t\}} + (1-\beta) \cdot global_{\{s,t\}}$$

where β determines how much importance to give to national relevance compared to global relevance, for the following experiments we set $\beta = 0.2$. In this way, we have that:

$$-1 \leq w_{\{s,t\}} \leq 1$$

In addition, in our analysis, we utilize specific tunable parameters to refine the results and have a more precise examination. These tunable parameters are:

- **Quantile value**: this parameter determines how many of the original relationships to keep. By setting a quantile value we effectively filter the dataset to retain a specific percentage based on the 'Goldstein_count' column. For our experiments, the quantile value is set to 20%. This means we only keep the top 20% of the original relationships discarding the less significant ones, reducing noise, and enabling clearer insights and visualization.

- **Map Type**: the nature of the relationships considered can be further refined by selecting among different types of maps, offering different views on the data. There are three different options:

  - **Aggregated Map**: this type includes the entire dataset, including all events regardless of their nature. It provides a more realistic view of the world. An example can be seen in Figure 1.
  - **Only Positive Map**: this selection filters the dataset to include only positive events. This allows for an analysis centered on cooperation and alliance between nations.
  - **Only Negative Map**: in contrast, this map type restricts the dataset to only negative events. This allows a better focus on conflicts, disputes, and negative relations.

4 Validity and Reliability

In the context of analyzing and identifying the most relevant interactions we tested if our method was able to recognize important events across several years. Specifically, our model was able to highlight the most important conflicts like the war in Iraq, the Falkland conflict, the tensions in the Middle East, and also other minor conflicts. The data contained in GDELT indeed captures the global trends as we can visualize ongoing conflicts and alliances. Furthermore, the validity of the data is ensured by the wide use of these resources. On the GDELT project page, a wide list of projects and research based on these data are available. The entire process was carried out to ensure its reproducibility by following some fundamental indications:

- The dataset is open and publicly available and has not undergone any further changes compared to those carried out during the preprocessing phase described in the previous section.

- The code developed was entirely written in Python using open-source libraries and it is available on GitHub.

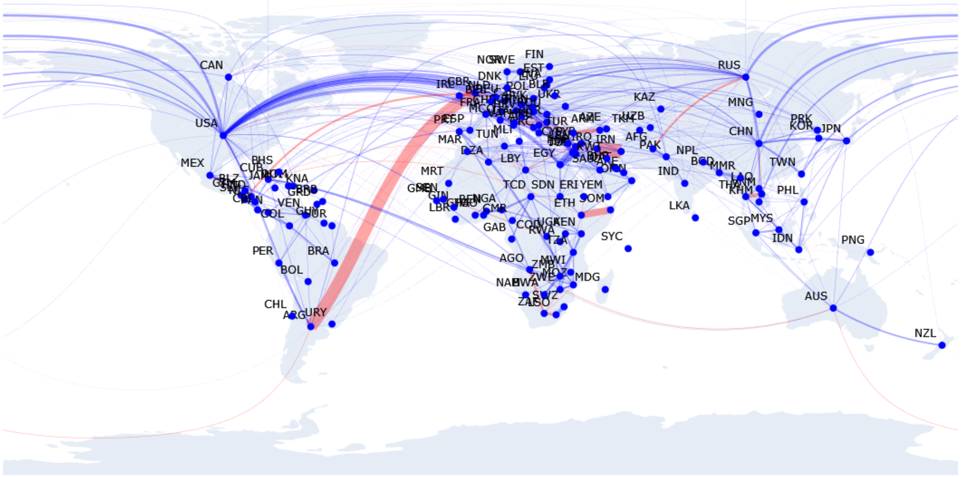

Figure 1: Aggregated map of the 1982. The Falkland War and the conflicts between Somalia and Ethiopia are immediately evident. Also, it's possible to see how the USA and Russia are the major nations with their influence spreading across the globe.

5 Measures and Results

In this chapter, several measures are applied to assess the quality of our model. We investigated several aspects of the built network to extract knowledge and valuable insight. In addition to the presented measure, we experimented with additional metrics such as clique, k-components, clustering coefficient, core-periphery, and dominating set. However, we decided to exclude them from our study since their results were not interesting. We also tried the small-world alpha and omega coefficients, but we could not get a result in a reasonable amount of time. Furthermore, metrics like assortativity and reciprocity did not apply to our graph model.

5.1 Centrality

Centrality is an important concept in graph analysis for identifying important nodes. Since each node could be important, depending on the angle and from what we want to emphasize, different centrality measures are applied.

- **Degree centrality**: degree centrality metric defines the importance of a node in a graph as being measured based on its degree, the higher the degree of a node, the more important it is in a graph. For our analysis, we would like to know which countries have a high number of connections. A high degree of centrality would imply that a state maintains relations with many others and that aims to be central in international relations.

- **Eigenvector centrality**: this metric measures the importance of a node in a graph as a function of the importance of its neighbors.

- **Closeness centrality**: differently from the previous centrality measures, based on nodes' degree, closeness centrality uses the shortest paths in networks, measuring the mean distance from a node to other nodes. Further, the closeness centrality metric defines the importance of a node in a graph as being measured by how close it is to all other nodes in the graph. To apply this measure we used as distance:

$$d_{\{s,t\}} = \frac{1}{w_{\{s,t\}} + 1 + \varepsilon}$$

The reason is that the more a weight is positive, the closer the two nations are; similarly, the more a weight is negative, the more distant the nations are. The ε is just for numerical stability.

- **Betweenness centrality**: it is also based on shortest paths, and measures the extent to which a node lies on paths between other nodes. It measures a node's importance in facilitating connections between other nodes, considering the most efficient paths available. For our study, it is important because nations with high centrality have good relations with many others and they could be considered as an intermediary in a geopolitical context. As distance function, we used the same as in the closeness.
- **Pagerank**: PageRank algorithm measures the importance of each node within the graph, based on the number of incoming relationships and the importance of the corresponding source nodes. The underlying assumption, roughly speaking, is that a nation's importance is determined by the importance of the nations it is connected to

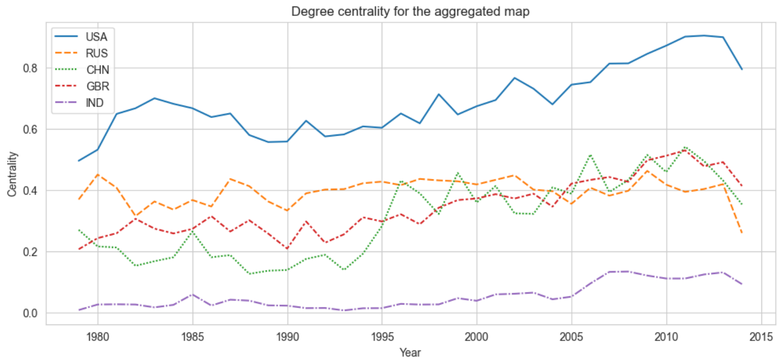

Figure 2: Plot of the centralizes over time for some selected countries

Nation	Degree	Closeness	Betweenness	Eigenvector	PageRank
USA	0.667	0.718	0.654	0.170	0.063
RUS	0.315	0.526	0.069	0.050	0.026
CHN	0.153	0.481	0.016	0.040	0.022
GBR	0.306	0.534	0.043	0.108	0.036
IND	0.027	0.427	0.000	0.004	0.004

Table 3: Centrality Measures and PageRank for Different Nations in 1982

5.2 Community

Communities are a property of many networks in which a particular network may have multiple communities such that nodes inside a community are densely connected. Applying community measures we would like to explore how different nations are grouped, if we took into consideration only positive relations we should be able to understand geopolitical blocs. On the other hand, taking into consideration negative links we could understand the nations present in various conflicts. While using the aggregated map, we should have a mixed view. To conduct this analysis and have robust results we used the following algorithms:

- Louvain Community Detection [2]: the algorithm works in 2 steps. In the first step, it assigns every node to be in its community and then for each node, it tries to find the maximum positive modularity gain by moving each node to all of its neighbor communities. If no positive gain is achieved the node remains in its original community.

- Clauset-Newman-Moore: greedy modularity maximization to find the community partition with the largest modularity [1].

- Label propagation: the label propagation algorithm is described in [5].

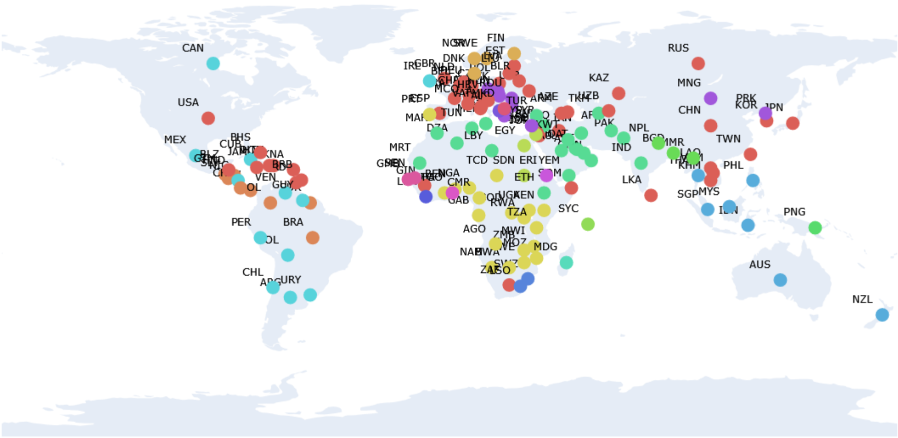

Figure 3: Communities for 1982 on the aggregated map computed by the Louvain algorithm.

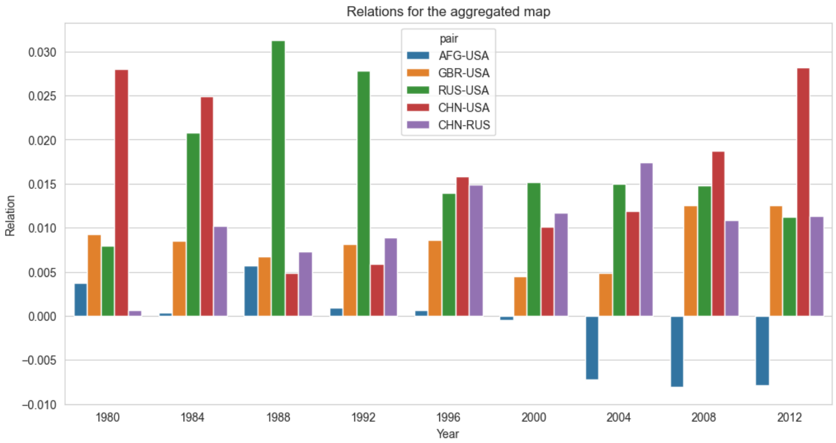

Figure 4: Relations between some selected nations every four years

6 Conclusion

In this section, we will present analyses concerning the results obtained by our model. The flexibility of the model allows for tweaking different parameters, including quantile selection, the metrics considered, the specific year of analysis, and the β value. Due to the limited space and resources, we present the most interesting results using the configuration: quantile 20%, β as 0.1, and analyzing the year 1982.

From the map 1 the major relations between nations are easily recognizable. In red we have negative relations and in blue positive ones, the line thickness indicates the importance of the link. Major conflicts immediately pop up with ticker lines while less important relations have thin edges. It turns out that also for other years, the map is a valuable tool to identify the major ongoing events. Figure 4 shows how the relations between some central countries have changed over the years. Notably, the relationship between the USA and Afghanistan captured immediate attention, primarily due to the start of war in 2001. Additionally, the dynamics between Russia and the USA, although remaining positive overall, have experienced a gradual decline over the years while the China-USA, after a downturn, returned to the same level.

Table 3 presents centrality measures for selected countries, revealing expected trends, notably the USA's dominance across all metrics, especially in Betweenness centrality, where it significantly outpaces other nations. This dominance underscores the USA's extensive global connections, positioning it as a central intermediary in numerous international interactions. Interestingly, while China exhibits a lower Degree of centrality, indicating fewer direct connections, it scores high on Closeness centrality. This suggests that despite fewer connections, China maintains a relatively strong link with other nations. An interesting analysis is how the centrality of a nation changes over the years. This type of investigation shows how some nations could have lost or gained importance over the years. Figure 2 reports the evolution of the degree of centrality over the years. From the plot, we can affirm that the USA and UK remained quite stable with a small increase in the last years while China had a significant increase around 1996 and in the last years it has reached the same level as the UK and Russia.

Figure 3 shows different communities obtained by the Louvain algorithm [2]. The results obtained by this, but also from other communities' algorithms, are not easy to interpret. First of all the resolution parameter can drastically change the communities, where higher values lead to more granular communities. Even though we do not feel confident to draw some conclusions, from these communities it's interesting to see the formation of groups. Different communities can be observed in Western countries, Arab nations, Eastern nations, South American countries, and South African regions. However, it is not clear whether this is the result of geographical, cultural, or political vicinity.

7 Critique

In this work, we provide a comprehensive analysis of international relations through a methodological framework designed for the ingestion, aggregation, and insightful analysis of event data. Some interesting future developments could be:

- **Directed graph**: creating a directed graph could provide a more specific view of the dynamic of international relations.

- **Event filtering**: an analysis that filters events based on specific CAMEO code categories could yield more targeted insights.

- **Real-time data**: due to our resource limitation we have not explored the possibility of using real-time data to perform a real-time analysis. Furthermore, it is also interesting to try to include all the data available on GDELT.
- **International organization**: expanding the analysis to include entities such as the UN, EU, BRICS, and others could provide a better understanding of global diplomacy.

For the objectives we have set ourselves, we are satisfied with the results achieved, however, we think that the aggregation at the national level cut out a lot of nuances and interesting relations between the different organizations that compose a nation.

References

[1] Aaron Clauset, Mark EJ Newman, and Cristopher Moore. Finding community structure in very large networks. *Physical review E*, 70(6):066111, 2004.

[2] Nicolas Dugué and Anthony Perez. Directed Louvain : maximizing modularity in directed networks. Research report, Université d'Orléans, November 2015. Please cite the following published version: https://doi.org/10.1016/j.physa.2022.127798 rather than this one.

[3] Swetha Keertipati, Bastin Tony Roy Savarimuthu, Maryam Purvis, and Martin Purvis. Multi-level analysis of peace and conflict data in gdelt. In *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*, pages 33–40, 2014.

[4] Kalev Leetaru and Philip A. Schrodt. Gdelt: Global data on events, location, and tone. *ISA Annual Convention*, 2013.

[5] Usha Nandini Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical review E*, 76(3):036106, 2007.

[6] Vasiliki Voukelatou, Ioanna Miliou, Fosca Giannotti, and Luca Pappalardo. Understanding peace through the world news. *EPJ Data Science*, 11(1):2, 2022.