Empirical Design of System of Systems Using Open Data

A submission to the SBP-BRiMS 2016 Grand Data Challenge

This submission proposes a methodology for correlating a set of known geopolitical events and corporate social ties to predict the outcome of future corporate agreements in the European oil and natural domain. Specifically, we use an assessment of historical case studies through an interactive visualization system to draw conclusions about "gray zone" warfare tactics present in the European natural gas ecosystem and the future of gas supply agreements between the European states and potential suppliers. Additionally, we explore the role of the human analyst in successfully employing data driven approaches.

The methodology developed here allows two separately structured data sets to be meshed into a common structure for analysis. Two datasets are leveraged in this project for correlating geopolitical events and corporate relationships: the GDELT global event repository and the Thompson Reuters SDC Platinum set of corporate relationships and ties. In order to structure a search in a targeted approach, the methodology requires careful consideration of how to define the case studies to answer the questions of interest. First, a specific case study is identified by a human analyst who then writes a narrative, or structured story, on the scenario. This narrative is the means by which the analyst will guide the data search process. From the written narrative, the key words and phrases for data search will be computationally isolated. These words and phrases are then the parameters around which the relevant data from GDELT will be pulled to mesh with the SDC Platinum data. Since each dataset is mapped through time, the combined data provides records of known formation and duration of company ties and a record of geopolitical events that occurred around the time for the case study. The SCD Platinum data sets provides a foundational set of nodes and edges formed by the emergence of firm-to-firm relationships over time. The GDETL dataset provides a quantitative assessment of time-based events that is meshed with the SCD dataset based on the targeted key word search. Merging these datasets provides a means to test the correlation of corporate relationships with geopolitical events; that is, how corporate events and relationships drive and impact specific global events.

Using an interactive visualization system, we explore the data in four separate views: a Path view, Category view, Geography view, and a ScatterNet view. The Path View is a topology-focused connectivity view that uses a cluster-emphasizing approach to assess nodes and their groups of partners. The Category view positions nodes relative to their categories, as depicted through colored bands on an arc, providing a depiction of the time-base evolution of node-category memberships. The Geography view places nodes on an interactive world map corresponding to their latitudes and longitudes while maintaining the edges between the nodes. Finally, the ScatterNet view combines a scatter plot with network functionalities, allowing the data to be assessed against an x-axis and y-axis, based on variables derived from the network.

Each of these views leverages the time-dependent nature of the data. The team is able to visually assess the formation and evolution of networks as relationships enter the ecosystem data space and events occur with the progression of time. The playback feature of the visualization system allows the team to navigate through time and animation the evolution of the networks of relationship ties and events. Most notably, the Geography view is a novel way to study geospatially and temporally distributed entities – both global events as well as global actors. In this way, the team can explore the correlation of the SDC Platinum social corporate ties with the event data from GDELT, mapping peaks of events to the development of corporate networks.

This methodology is tested through a set of regional case studies that describe historical events in the European oil and natural gas market. These case studies involve a finite timeline of events and relevant actors where the results of the data analysis can be compared to known outcomes. The case studies are divided into four sub-regions. Within each case study, a historical scenario, such as the Russia and Ukraine gas conflict in 2006, is detailed in a narrative. Each narrative must be carefully constructed to provide the key phrases for meshing the GDETL and SDC Platinum data sets. The four regions are as follows:

- [Gazprom Corporate Ecosystem]: All of EU28, U.S., and Canada, Gazprom
- [Northern Europe] Germany, Norway, Poland, Finland, Lithuania, Latvia, Estonia
- [Central Europe]: Italy, Hungary, Austria, Croatia, Romania, Bulgaria, Greece
- [Southern Sources]: Turkey, Azerbaijan, Iran, Turkmenistan

Following demonstration through the historical case studies, the methodology will then be applied to understanding the prospect of corporate ties while monitoring global events in near real time. Specifically, the investigation centers on understanding the future of corporate ties associated with the State Oil Company of Azerbaijan Republic (SOCAR) and the European Union. Volumetrically, SOCAR does not compare to its Russian counterparts, Gasprom and Rosneft. However, with respect to corporate ties and relationships, SOCAR has very strong social ties, allowing it to complete deals in key areas in competition with Russian gas companies. With the methodology validated through the case studies, and lessons on meshing data, this approach is used to explore the strength of ties in SOCARs role in the region. We aim to understand better the priority of the quantity of gas flow versus the social capital when forging gas contracts and infrastructure development in the EU. The use of GDELT allows the team to measure peaks in geopolitical events associated with social unrest – specifically, those timed with SOCAR agreement signing, possibly “gray zone” warfare measures instigated by Russians operatives in the region as a means to undermine any agreement.

The success of the project contributes to both the data science and international security fields by providing a novel means for assessing large amounts of time dependent data. The team conducting the research is made up of researchers across multiple disciplines from the Georgia Institute of Technology, in Atlanta, Georgia. Researchers from the applied research arm of Georgia Tech, the Georgia Tech Research Institute (GTRI), have partnered with researchers from the College of Computing on the academic arm of the Institute. Ms. Lindsey Sheppard is leading the project as the Principal Investigator; she is a research engineer with the Electronic Systems Laboratory (ELSYS) at GTRI. Mr. Tom McDermott and Ms. Molly Nadolski, both with ELSYS, serve as the advisor and expert on the narratives and regional case studies. Mr. James Arruda and Ms. Melissa Rost, of ELSYS and the Information and Communication Laboratory (ICL), respectively, are contributing to the project as data science analysts. Finally, Dr. Rahul Basole and his students provide contribution in the form of a data visualization and analysis interactive system. Dr. Basole is an Associate Professor with the School of Interactive Computing and Associate Director of the Tennenbaum Institute at Georgia Tech. The email addresses for the team are listed alphabetically as follows: james.arruda@gtri.gatech.edu; basole@gatech.edu; tom.mcdermott@gtri.gatech.edu; molly.nadolski@gtri.gatech.edu; Melissa.rost@gtri.gatech.edu; lindsey.sheppard@gtri.gatech.edu.