An Extensible Model for Deduplication of the GDELT Events Database

*Principal Investigator: Andee Kaplan**

Department of Statistics, Colorado State University

The GDELT (Global Data on Events, Location and Tone) 1.0 Event Database ("The GDELT Project" 2019) contains recorded CAMEO-coded global political events collected by the global news media. Records are collected monthly from January 1979 through March 2013 and daily after this date. Each news story is automatically processed (extracted) prior to inclusion in the database, and basic information is pulled out of each article, including actors, dates, locations, and event types (among others). The database is openly accessed via downloadable tab-delimited CSV files as well as a real-time API.

Currently, the GDELT 1.0 Event Database is used by a LAS Computational Social Sciences and Triage (CSS&T) project team led by Soumen Lahiri for the project year 2019 in an effort to predict the status (stable, fragile, failing, or failed) of countries based on the status of neighboring countries and the (temporal) number of higher-severity events (as coded by CAMEO). One goal of this project is to build a spatio-temporal model that accurately captures the relationship between fragile and failing states and the occurrence of high-severity events that lead to the fragile and failing classification. The creation of such a model relies on the existence of clean GDELT data.

A main challenge to utilizing the GDELT Events Database as a data source for modeling is that events in the database are not unique, and there is no unique key to identify duplicate events. As such, when aggregating the number of events that occurred in a particular country by month (as is done to create an explanatory variable in the model of the Lahiri group), these values will be over-inflated, affecting the model fit to these data. Thus, to ensure an accurate representation of the relationship between political event occurrence and fragile and failing states, a clean (deduplicated) version of the GDELT event data would be very useful. The goal of this project is to provide a method for deduplicating the existing GDELT 1.0 Event Database (an undertaking of a scale rarely seen in the current literature), as well as lay the groundwork for a model that can be updated in real-time as more records are added to the database.

Proposed Methods

I propose using Bayesian record linkage methods (see Tancredi and Liseo 2011; Steorts, Hall, and Fienberg 2016; and Sadinle 2014 for examples of current widely used models) in an effort to deduplicate the GDELT event database. Broadly, there are two main challenges to deduplication that must be addressed for this data. First, the categorical nature of the data will make deduplicating the GDELT data a challenge, due to the absence of string fields that can make identification of duplicates somewhat more discriminating. Secondly, the overwhelming size (over a quarter-billion records) of the data must also be taken into account.

Deduplication

Deduplication (as well as related concepts record linkage and entity resolution) aims to remove duplicate records from data sets in the absence of a unique identifying attribute. This process can be performed in a deterministic fashion using data set specific rules (Christen 2012) or by utilizing probability ideas like likelihood ratio tests (cf. Fellegi and Sunter 1969) or latent variable clustering via hit-and-miss models (Copas and Hilton 1990; Tancredi and Liseo 2011; Steorts, Hall, and Fienberg 2016; Sadinle 2014).

I intend to employ a Bayesian latent variable clustering method for deduplication that uses a flexible class of priors to allow for a fat tailed distribution of duplication, first introduced in Miller et al. (2015) and

*andee.kaplan@gmail.com; http://andeekaplan.com; (832) 526-7947

Betancourt et al. (2016). The two main benefits to this model are that, first, it is flexible to the distribution of duplication and second, it is in a Bayesian paradigm. The inherent benefits of utilizing the Bayesian paradigm include an ability to be used in an unsupervised setting (necessary for the GDELT data, which has no ground truth), implicit uncertainty quantification via the posterior distribution for linkage, and the ability to be used in a streaming setting by using the posterior as a prior as more data become available (see Section on Extensibility for reasoning on the importance of this feature). The model framework of Betancourt et al. (2016) with nonparametric priors is formulated for use with categorical (not string) data and has shown success in smaller applications; thus it is a good candidate for use with the GDELT 1.0 Events Database.

Blocking

The size of the data is another challenge that must be considered in the proposed approach. With over a quarter-billion records available, this is a large computational problem, especially when considering that record linkage is in general of order 2^n, where n is the number of records in the data set. One approach to ease this computational burden is through the use of blocking (Christen 2012; Steorts et al. 2014), whereby the deduplication problem is broken into many smaller deduplication problems (or blocks) based on a variable (or combination of variables). I intend to employ deterministic blocking because it is fast to implement and easy to update as more data are available.

Extensibility

In discussions with the CSS&T group at LAS, there has been interest in moving from batch files of the GDELT (collection of large sets of records at once) to the use of a real-time updating streaming version via the existing GDELT API. This move would require the accompaniment of real-time streaming deduplication to the GDELT 1.0 Events Database. The extension of a model from single batch-processing to real-time updating would be a large step forward in the record linkage and deduplication literature and would allow for the application of deduplication on larger, more complex, and more dynamic data sets. As such, the proposed methodology will need to be created in such a way to allow for this expansion of the model into a broader use case. By utilizing Bayesian methods for deduplication and simple blocking methods, I am proposing laying the groundwork of a model for deduplication of the GDELT data that can be extended in the future.

Proposed Deliverables and Future Work

The proposed deliverables for this project include:

1. an extensible model for deduplicating the GDELT 1.0 Events Database,
2. a clean (deduplicated) GDELT 1.0 Events Data Set for use in any LAS projects that require it,
3. extensive documentation on the cleaning process and the model used, and
4. reproducible open source code that can be extended easily for use with more recent GDELT data.

With thoughts to potential future work, the model and code deliverables in this project will be developed with extensibility in mind. There are two main considerations, computational and methodological. Computationally, I will maintain an agnostic source of data to the code fitting the model, allowing for a future shift to the existing GDELT API as a data source. The methodological considerations shape the type of deduplication that is performed on the GDELT data in the current project cycle with a streaming update remaining open as a future possibility. Two of these considerations include easy updating of the deduplication model based on new data (Bayesian framework) and easy application of blocking scheme on new data (simple blocking rules).

I intend to propose this extended streaming framework in the 2020 Call for Proposals with Brenda Betancourt (Department of Statistics, University of Florida).

Budget

I am requesting funding for one graduate research assistant (20 hours/week) for the Fall 2019 semester.

References

Betancourt, Brenda, Giacomo Zanella, Jeffrey W Miller, Hanna Wallach, Abbas Zaidi, and Rebecca C Steorts. 2016. "Flexible Models for Microclustering with Application to Entity Resolution." In *Advances in Neural Information Processing Systems*, 1417–25.

Christen, Peter. 2012. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Science & Business Media.

Copas, JB, and FJ Hilton. 1990. "Record Linkage: Statistical Models for Matching Computer Records." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 153 (3): 287–312.

Fellegi, Ivan P, and Alan B Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association* 64 (328): 1183–1210.

Miller, Jeffrey, Brenda Betancourt, Abbas Zaidi, Hanna Wallach, and Rebecca C Steorts. 2015. "Microclustering: When the Cluster Sizes Grow Sublinearly with the Size of the Data Set." *arXiv Preprint arXiv:1512.00792*.

Sadinle, Mauricio. 2014. "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach." *The Annals of Applied Statistics* 8 (4): 2404–34.

Steorts, Rebecca C, Rob Hall, and Stephen E Fienberg. 2016. "A Bayesian Approach to Graphical Record Linkage and Deduplication." *Journal of the American Statistical Association* 111 (516): 1660–72.

Steorts, Rebecca C, Samuel L Ventura, Mauricio Sadinle, and Stephen E Fienberg. 2014. "A Comparison of Blocking Methods for Record Linkage." In *International Conference on Privacy in Statistical Databases*, 253–68. Springer.

Tancredi, Andrea, and Brunero Liseo. 2011. "A Hierarchical Bayesian Approach to Record Linkage and Population Size Problems." *The Annals of Applied Statistics* 5 (2B): 1553–85.

"The GDELT Project." 2019. https://www.gdeltproject.org.