Monthly Report on "An Extensible Model for Deduplication of the GDELT Events Database"

Andee Kaplan (andee.kaplan@colostate.edu; 832-526-7947)

August 2019

Progress and Plans

1. Description of progress made against each deliverable during the reporting period

   I started working with a graduate student (Casey Schafer) at CSU on Aug 16 with the goal of formulating, implementing, and applying an extensible deduplication model to the GDELT events data before Jan 1, 2020. Casey is a PhD student in Statistics at CSU with a strong background in computation and an interest in learning more about record linkage.

   We have spent the remainder of the month of August getting him up to speed on modern Bayesian methods for performing deduplication and discussed the pros and cons of the different model classes. We have discussed the following papers – Tancredi, Liseo, and others (2011), Sadinle (2014), and Steorts, Hall, and Fienberg (2016).

2. Brief description of significant results

   Our discussions have led us to focus on the methods of Sadinle (2014) and Steorts, Hall, and Fienberg (2016) because they have both been formulated for the deduplication (rather than the record linkage) task and they are both formulated for categorical data. We are currently deciding between the partition prior of Sadinle (2014) and the uniform linkage prior of Steorts, Hall, and Fienberg (2016). We have looked at numerical explorations of the differences in prior and will now turn to the expected posterior difference in the methods.

3. Planned activities for the following reporting period

   Once a method is chosen to move forward with, we will turn to computation and challenges faced specific to the GDELT data.

4. List of any LAS-funded trips during reporting period, with description of work presented

   None.

5. Description of any significant meetings/events held (e.g., a focused discovery activity, LAS Symposium) or conducted related to this task

   Casey and I have been meeting weekly and will continue to do so.

Issues

1. Progress on/results obtained related to previously identified problem areas

   None.

2. Proposed significant changes to your methodology, goals, milestones, or deliverables

   None.

3. New challenges affecting technical performance or schedule, with background

   None.

4. Has anything happened to impact your anticipated schedule?

   No.

References

Sadinle, Mauricio. 2014. "Detecting Duplicates in a Homicide Registry Using a Bayesian Partitioning Approach." *The Annals of Applied Statistics* 8 (4). Institute of Mathematical Statistics: 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). Taylor & Francis: 1660–72.

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