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

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

September 2019

Progress and Plans

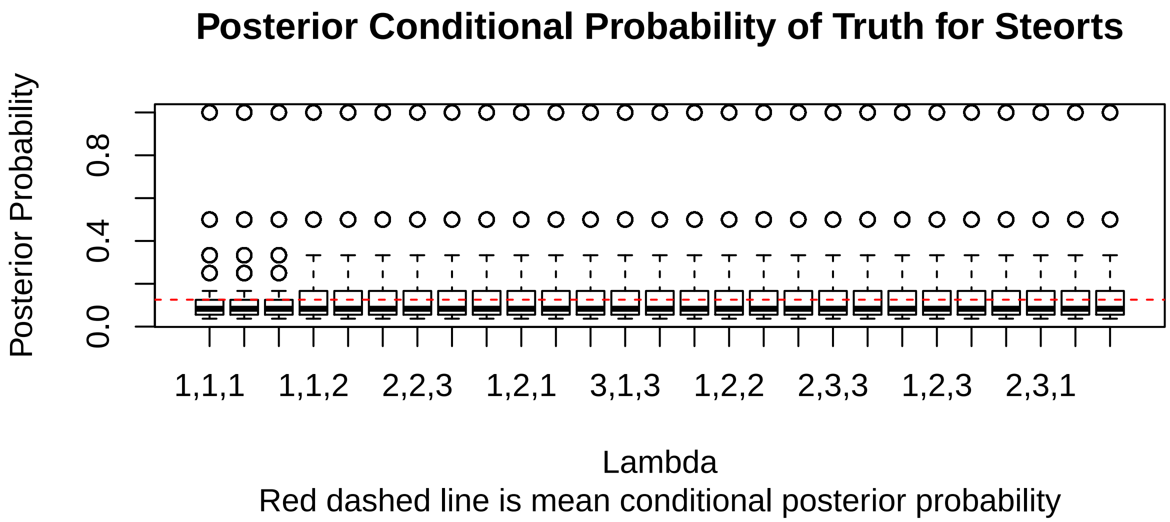
1. Description of progress made against each deliverable during the reporting period

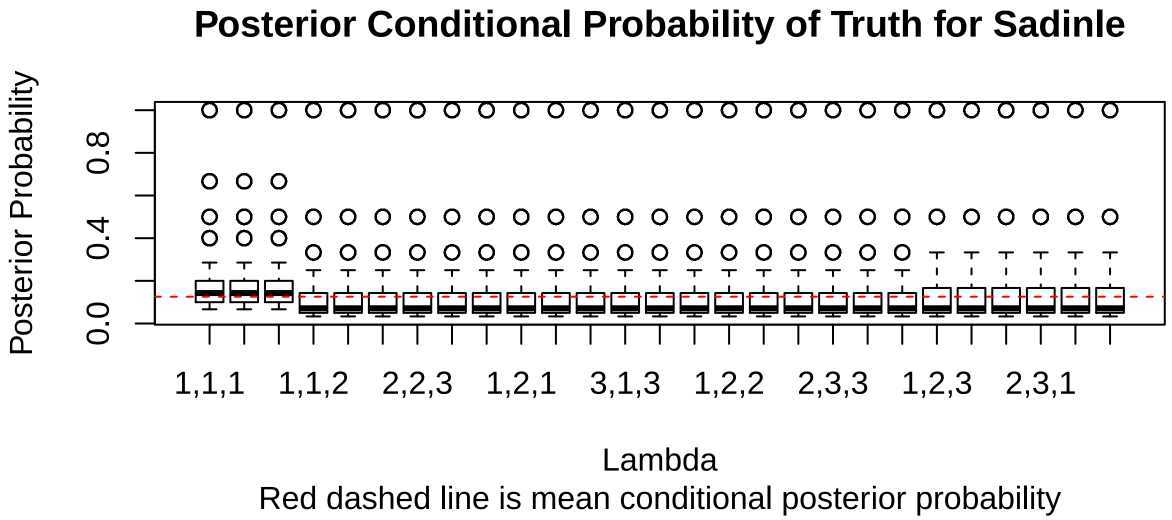
   Casey has created a framework to compare the resulting posterior probabilities of a Bayesian hierarchical deduplication model using the uniform prior of Steorts, Hall, and Fienberg (2016) and one using the partitioning prior of Sadinle (2014). We have spent September looking at numerical comparisons of small toy size models ($n = 3$) for completeness, as well as a larger, more realistic scenario (results to follow in the October report).

   Our goal with this exploration is to choose a suitable model for deduplication of the GDELT events data, while a secondary goal is to choose a suitable model that is a good candidate for extension to a streaming version.

2. Brief description of significant results

   Changing of the prior from uniform on the partitions (Sadinle) to uniform on matching of observations to entities (Steorts) resulted in minor changes in the full conditionals of matchings for a small sample size. For a given matching, we measured the posterior conditional probability of obtaining this true matching.





For matchings where the prior probabilities differed significantly between the two prior specifications, we found the posterior conditional probability to also have a relatively higher difference. The matchings where this occurred were matchings where all observations coresponded to the same underlying entity, in which cases the Sadinle prior put twice as much weight on these matchings as the Steorts prior. The following plots show the change in posterior probability of truth for Steorts and Sadinle priors, where each λ corresponds to a unique matching of observations to entities for $n = 3$.

3. Planned activities for the following reporting period

   We will choose a candidate model and do preliminary work to prepare the GDELT data for deduplication.

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 meet weekly for one hour.

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.