Department of Statistical Science
Duke University
Durham, NC 15213-3890
Phone: 919-684-4210
Fax: 919-684-8594

Rebecca C. Steorts
Assistant Professor
919-684-4210
beka@duke.stat.edu

November 29, 2021

Dear Editor Guindani:

I am a PhD student at Duke University in the Department of Statistical Science. I am writing to submit my paper with my coauthors Jerome Reiter and Rebecca C. Steorts entitled "Efficient and Scalable Bayesian Bipartite Matching with Fast Beta Linkage (fabl)" to *Bayesian Analysis*.

Very often information about social entities is scattered across multiple databases. Combining that information into one database can result in enormous benefits for analysis, resulting in richer and more reliable conclusions. The application areas touched by record linkage include voter registration, official statistics, human rights violations, patent inventions, health applications, among many others. Disciplines that utilize both applications and methods include statistics, biostatistics, computer science, machine learning, political science, and others. The goal of our review article is to review nearly almost all of entity resolution in such a way that it can be read and understood by as many communities as possible given that the current state of the literature is mathematical and overly technical, making the entry point difficult for those in the computational social sciences.

Recently, researchers have developed Bayesian versions of the Fellegi Sunter model for record linkage. These have the crucial advantage of quantifying uncertainty from imperfect linkages. However, current implementations of Bayesian Fellegi Sunter models are computationally intensive, making them challenging to use on larger-scale record linkage tasks. We propose and investigate a variation on Bayesian Fellegi Sunter models that we call fast beta linkage, or fabl. Specifically, in fabl we propose independent prior distributions over the matching space, allowing us to use hashing techniques that reduce computational overhead. This also allows us to complete pairwise record comparisons over large datasets through parallel computing and reduce memory costs through a new technique called storage efficient indexing. Through simulations and two case studies, we show that fabl has markedly increased speed with minimal loss of accuracy.

To help the editorial office, we have listed a number of potential reviewers that do not have any conflicts with the authors:

Mauricio Sadinle, msadinle@uw.edu, University of Washington; Serge Aleshin-Guendel, aleshing@uw.edu, University of Washington; Yves Thibaudeau, Yves.Thibaudeau@census.gov, United States Census Bureau; Megan Price, meganp@hrdag.org, Human Rights Data Analysis Group; Ted Enamorado, ted@wustl.edu, University of Washington St. Louis; Nick Monath, nmonath@cs.umass.edu;

Dinusha Vatsalan, Dinusha.Vatsalan@data61.csiro.au, Ben Rubinstein; benjamin.rubinstein@unimelb.edu.au; Neil Marchant; nmarchant@unimelb.edu.au; Andy Toole; Andrew.Toole@uspto.gov; jph34@stat.duke.edu; Giacomo Zanella, zanella.gcm@gmail.com, Brunero Liseo, brunero.liseo@uniroma1.it, Andrea Tancredi, andrea.tancredi@uniroma1.it; Peter Christen, peter.christen@anu.edu.au, Partha Lahiri, plahiri@umd.edu, Rebecca Nugent, rnugent@stat.cmu.edu, Kayla Frisoli, kaylafrisoli@gmail.com.

We appreciate consideration of our paper, and please do not hesitate to reach out if you have any questions.

Sincerely,

Rebecca C. Steorts