Record Linkage for the American Opportunity Study: Formal Framework and Research Agenda

Stephen E. Fienberg
Department of Statistics, Heinz College, and Machine Learning Department,
Carnegie Mellon University

Workshop on The Potential for Research Using Linked Census, Survey, and Administrative Data to Assess the Longer Term Effects of Policy

May 9, 2016

Exact vs. Statistical Matching

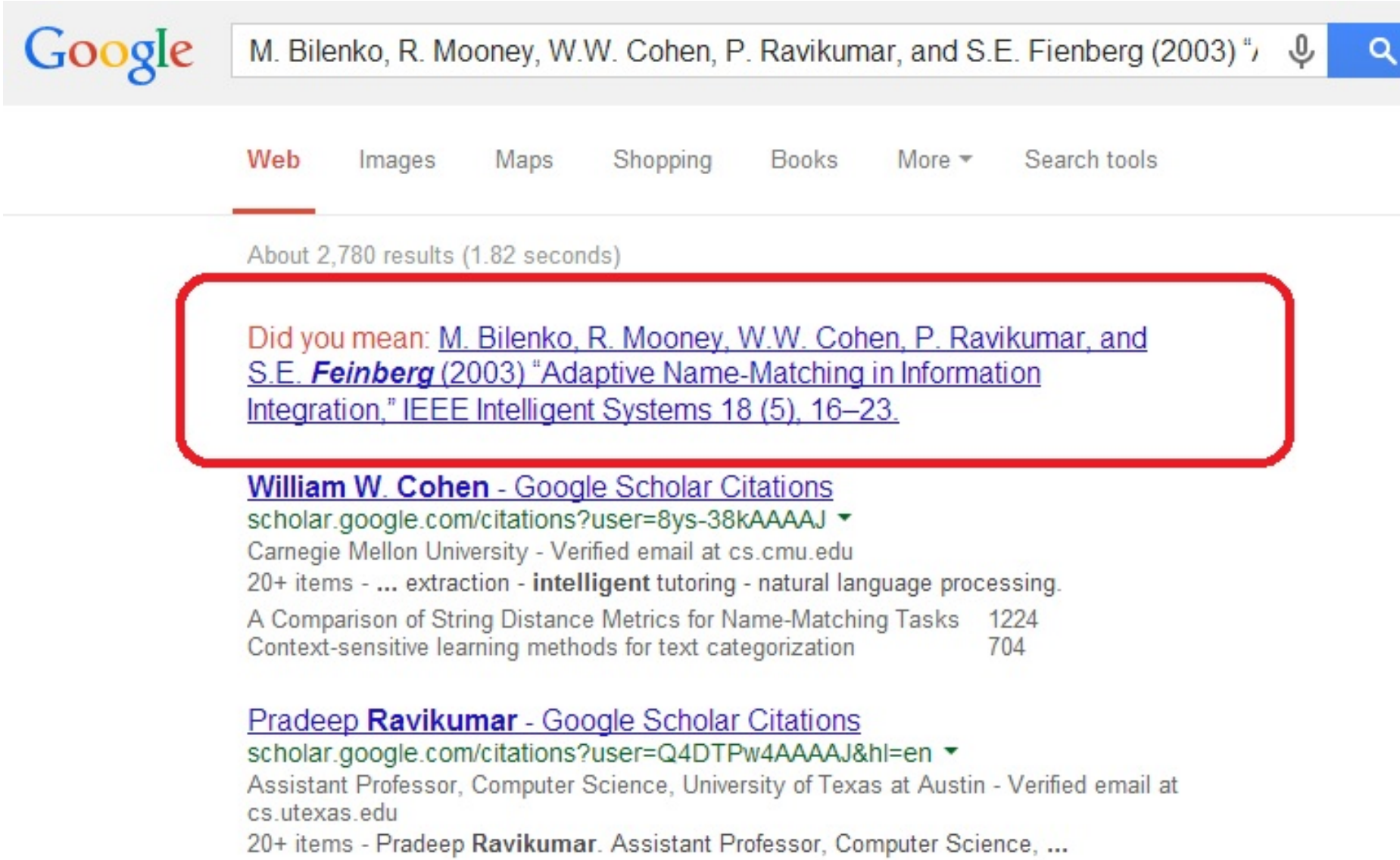
- Exact matching: Link (X, Z) with (Y, Z):
  - Updates to Social Security Administration Master Earnings File (MEF) and Numident file.
  - Electronic medical records.
- Statistical matching: Link (X, Z) with (Y, Z') where Z' is a noisy version of Z or vice versa:
  - Duplicate/misspelled names:






- Misspellings: Steve, Steven and Stephen; Fienberg, Feinberg, Fineberg, Fienburg, Feinburg, Steinberg, etc.
- Basically noisy matching data.

Google Does it Again



Where It Began: Foundational Work

- Ideas surfaced in multiple contexts with the rise of computational infrastructure in the 1950s:
  - Post-WWII welfare state and taxation system led to new administrative record systems
  - New computer technology
- Three key papers:
  1. H.B. Newcombe, J.M. Kennedy, S.J. Axford, and A.P. James (1959). "Automatic Linkage of Vital Records," *Science*, 130 (3381), 954–959.
  2. B.J. Tepping (1968). "A Model for Optimum Linkage of Records," *J. Amer. Statist. Assoc.*, 63 (324), 1321–1332.
  3. I.P. Fellegi and A.B. Sunter (1969). "A Theory for Record Linkage," *J. Amer. Statist. Assoc.*, 64 (328), 1183–1210.
- Public response: threat to individual privacy
  - R. Kraus (2013). "Statistical Déjà Vu: The National Data Center Proposal of 1965 and Its Descendants," *J. Privacy and Confidentiality*, Vol. 5, No. 1.

The Fellegi-Sunter Framework

- Represent every pair of records using vector of features that describe similarity between individual record fields.
- Place feature vectors for record pairs into three classes: matches (M), nonmatches (U), and possible matches.
  - Let $P(\gamma|M)$and $P(\gamma|U)$ are probabilities of observing that feature vector for a matched and nonmatched pair, respectively.
  - Perform record-pair classification by calculating the ratio $(P(\gamma|M))/(P(\gamma|U))$ for each candidate record pair, where γ is a feature vector for pair.
  - Establish two thresholds based on desired error levels—T_μ and T_λ—to optimally separate the ratio values for equivalent, possibly equivalent, and nonequivalent record pairs.
- Because most record pairs are clearly nonmatches, blocking databases so that only records in blocks are compared significantly improves efficiency.
- **1-1 linkage assumption often drives accuracy.**

The Fellegi-Sunter Framework: II

- Possible matches often go to clerical review in statistical agency context.
- **In AOS context we avoid clerical review.**
- Can do this parametrically using logistic regression or some other GLM, or non-parametrically.
- Supervised learning (with training data) vs. non-supervised learning
  - J.B. Copas and F.J. Hilton (1990). "Record Linkage: Statistical Models for Matching Computer Records," J. Roy. Statist. Soc. (A), 153, 287–320.
  - S. Ventura, R. Nugent, and E. Fuchs (2015). "Seeing the Non-Stars: (Some) Sources of Bias in Past Disambiguation Approaches and a New Public Tools Leveraging Labeled Records." *Research Policy,* Special Issue on Data. 44(9), 1672–1701
- Use string metrics and edit-distances for names and strings of numbers.
  - M. Bilenko, R. Mooney, W.W. Cohen, P. Ravikumar, and S.E. Fienberg (2003). Adaptive Name-Matching in Information Integration. *IEEE Intelligent Systems*, 18 (5), 16–23.

String-Comparator Illustration (Source: Winkler)

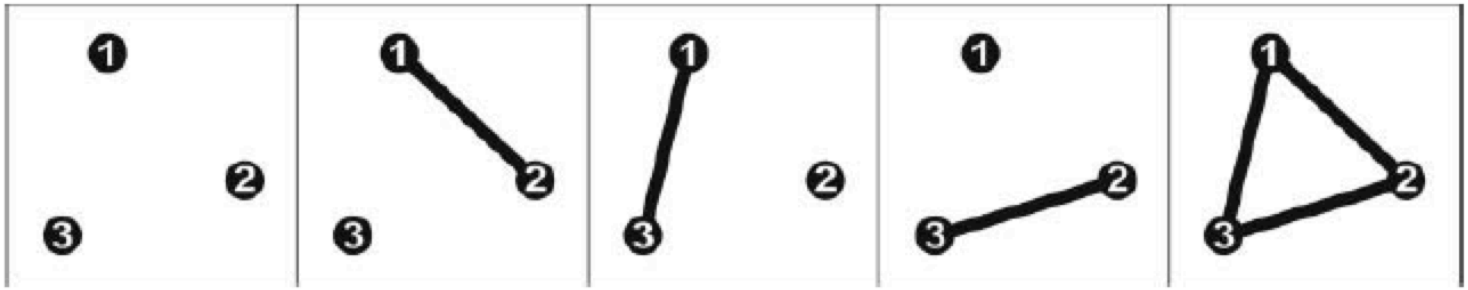
Two Strings		String Comparator Values			
		Jaro	Winkler	Bigram	Edit
SHACKLEFORD	SHACKELFORD	0.970	0.982	0.800	0.818
DUNNINGHAM	CUNNIGHAM	0.867	0.867	0.917	0.889
NICHLESON	NICHULSON	0.926	0.956	0.667	0.889
JONES	JOHNSON	0.867	0.893	0.167	0.667
MASSEY	MASSIE	0.889	0.933	0.600	0.667
ABROMS	ABRAMS	0.889	0.922	0.600	0.833
HARDIN	MARTINEZ	0.778	0.778	0.286	0.143
ITMAN	SMITH	0.467	0.467	0.200	0.000
JERALDINE	GERALDINE	0.926	0.926	0.875	0.889
MARHTA	MARTHA	0.944	0.961	0.400	0.667
MICHELLE	MICHAEL	0.833	0.900	0.500	0.625
JULIES	JULIUS	0.889	0.933	0.800	0.833
TANYA	TONYA	0.867	0.880	0.500	0.800
DWAYNE	DUANE	0.778	0.800	0.200	0.500
SEAN	SUSAN	0.667	0.667	0.200	0.400
JON	JOHN	0.778	0.822	0.333	0.750
JON	JAN	0.778	0.800	0.000	0.667

Fellegi-Sunter As a Missing Data Problem

- True match status unknown, and so impute.
- Latent variable representation.
- Bayesian and non-Bayesian formulations:
  - D.B. Rubin (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. & Econ. Statist.*, 4 (1), 87–94.
  - W.E. Winkler (1988). Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. *Proc. Section on Survey Research Methods, Amer. Statist. Assoc.*, 667–671.
  - T.N. Herzog, F.J. Scheuren, and W.E. Winkler (2007). *Data Quality and Record Linkage*. New York: Springer.
- Much subsequent work by Belin, Larsen, Zaslavsky and others.
- Recent work by Jerry Reiter's group at Duke.

Record Linkage With Multiple Lists

- Traditional approach used pairwise list record linkage.
- Problem: Non-transitive links: A (list 1) links to B (list 2) and B (list 2) links to (C (list 3), BUT A does not link to C!
- Lack of transitivity issue becomes greater as number of lists, K grows.
- Possible transitive linkage patterns with three lists:



- Generalized version of FS with logistic structure and EM gives principled transitive solution.
  - M. Sadinle and S.E. Fienberg (2013). A Generalized Fellegi-Sunter Framework for Multiple Record Linkage With Application to Homicide Record Systems. *J. Amer. Statist. Assoc.*, 108(502), 385–397.
  - Integration of 3 datafiles: 67 homicides recorded by Columbia Census Bureau, 62 by National Police, and 33 Forensics Institute.

Post-Enumeration Surveys and Undercount Estimation

- But what do we do about the errors associated with record linkage?
  - Gross Error in Census = Erroneous Enumerations + Omissions $\approx$ 10%.
  - Matching errors.
- Need to propagate uncertainty associated with these into the estimation process for linked files.
- In context of census-adjustment methodology, this was the thrust of the Bayesian multiple-imputation methods of
  - T.R. Belin and D.B.Rubin (1995). A Method for Calibrating False-Match Rates in Record Linkage. *J. Amer. Statist. Assoc.*, 90, 694–707.
  - M.D. Larsen and D.B. Rubin (2001). Iterative Automated Record Linkage Using Mixture Models. *J. Amer. Statist. Assoc.*, 96, 32–41.

American Opportunity Study

- Current methodology matches record from each source into SS Numident file, and then uses the ID for the latter to link files. PIKing.
  - Pairwise record linkage into Numident.
  - What to do with those in source files that don't match? For what % of the population are there unique SSNo.s?
  - Propagating errors from individual source files (10% gross error in the census; nonresponse errors) and matching uncertainty into the analyses of the merged files.
- What about alternative methods for record linkage?
- Title 13 and other Confidentiality provisions.
- Privacy, Privacy, Privacy!!!
  - The results from analyses implemented on linked files will be scrutinized and subjected to disclosure limitation review.

Record Linkage Approaches Based on Graphs

1. Non-parametric approach using bi-partite graph structure.
   - Propagates error into a confidence interval for statistical computation, e.g., regression, or MSE.
   - R. Hall and S.E. Fienberg (2012). "Valid statistical inference on automatically matched files," In *Privacy in Statistical Databases, PSD 2010* (J. Domingo-Ferrer and I. Tinnirello, eds.), Lecture Notes in Computer Science Vol. 7556, Berlin: Springer, 131–142.
2. Network block-models: Idea of a partition of a concatenated set of records.
   - M. Sadinle (2014). "Duplicate Record Detection Using a Bayesian Partitioning Model,"*Annals of Applied Statistics*, 8(4), 2404–2434.
   - M. Sadinle (2016). " Bayesian Estimation of Bipartite Matchings for Record Linkage," *J. American Statistical Assoc.*, forthcoming.
   - R. Steorts, R. Hall, and S.E. Fienberg (2016). "A Bayesian Approach to Graphical Record Linkage and De-duplication," *J. American Statistical Assoc.*, forthcoming.

Record Linkage and Statistical Estimation

Three critical components:

1. "Putting the lists together," or record linkage.
2. Statistical estimation and model selection on linked files.
   - Need to consider how to incorporate model uncertainty into some form of overall population estimates.
   - Since there is error in matching no matter how well done statistically, there could be both bias and added uncertainty.
3. Need to propagate uncertainty from the record linkage as an added component of uncertainty into statistical estimation for linked files.

We have methods for doing this for estimating the size of a population from multiple lists for some of the methods described above! We need to carry out this agenda for the kinds of analyses that all of you want to carry out.

AOS Research Agenda on Record Linkage

- Developing RL methods that scale:
  - No. of comparisons explodes with increasing K: $n_1 \times n_2 \times \ldots \times n_K$.
  - Innovative uses of blocking and approximate stratification:
    - J.S. Murray (2016). "Probabilistic Record Linkage and Deduplication After Indexing, Blocking, and Filtering," *J. Privacy and Confidentiality*, 7(1), 3–24.
- Implementing multiple record linkage approaches on AOS-related files.
- Propagating linkage error into uncertainty for subsequent modeling for all linkage methods.
- **Comparing methods designed for linking multiple files with CARA PIK method.**
- Learning to do record linkage on the fly:
  - For assigning individual online census forms to correct address and households.
  - Deal with duplicate separately or in real time.