FINAL REPORT

Person Validation and Entity Resolution Conference Report

PRESENTED TO:
U.S. Census Bureau
4600 Silver Hill Road
Suitland, MD 20746

PRESENTED BY:
NORC at the
University of Chicago
55 East Monroe Street
20th Floor
Chicago, IL 60603
(312) 759-4000
(312) 759-4004

JULY 28, 2011

NORC at the UNIVERSITY of CHICAGO

Table of Contents

Introduction

The Person Validation and Entity Resolution Conference was an information exchange among government, academic, and private sector practitioners of record linkage. The goal of the conference was to discuss applications of record linkage techniques with researchers and program managers from U.S. federal statistical agencies and other interested parties. Recent technological innovations have greatly advanced the process of linking survey and census data with administrative records data.

This report is a summary of the conference sessions and is broken into four sections and three appendices. **Sections 1 – 3** provide presentation synopses from the first three conference sessions. These sections also include summaries of discussant comments, and answer to questions that were posed by conference attendees. **Section 4** is a summary of the conference's final panel session that featured the comments from the PVS Assessment project's Technical Advisory Panel. **Appendix A** is a list of the conference attendees' organizations. **Appendix B** is a compilation of the presentation slides from each session, and **Appendix C** is a compilation of the papers upon which the presentations were based. Conference materials are available electronically at NORC's PVS Assessment website. [1]

[1] The full URL is http:/www.norc.org/Research/Projects/Pages/census-personal-validation-system-assessment-pvs.aspx.

Session 1: International Papers

The international papers session started with Antoine Chevrette (Statistics Canada). He discussed Statistics Canada's probabilistic record linkage system called G-Link, and how the software evolved from the Generalized Iterative Record Linkage System-GIRLS. Mr. Chevrette described how users can easily implement various record linkage steps with the G-Link system.

Lisa Jackson-Pulver (The University of New South Wales) and Richard Madden (The University of Sydney) followed with a presentation on the estimation of the life expectancy of Australia's indigenous peoples, based on a linkage of indigenous deaths registration and census records. They found the Australian Bureau of Statistics census estimates of indigenous deaths were consistently lower compared to the record linkage based estimates.

The last session presentation was given by Andrew Bate (Pfizer, formerly with the World Health Organization (WHO)). Dr. Bate discussed the usefulness of a hit-miss model for duplicate detection in the WHO drug safety database. Dr. Bate described the performance of their model in the context of Norwegian reports on adverse drug reactions.

Fritz Scheuren (U.S. Census Bureau) was the discussant for the session. Dr. Scheuren began by expressing his belief that the field of record linkage is quite well developed, although not finished. While comparing the three papers, he thought each had very diverse goals, although there were some commonalities. The problem of under-identification of population subgroups in many administrative data sets, as discussed in the Jackson-Pulver and Madden presentation, is common to other countries (U.S., Canada, etc.) due to cultural differences between the aboriginal people and the conquering countries, and this leads to linkage errors—particularly false negatives. Record linkage researchers need to address this challenge in addition to privacy and confidentiality concerns. Cultural issues present more of a problem to record linkage than previous hurdles such as a lack of technology. Dr. Scheuren liked that all the presentations included interesting graphics, and he encouraged the use of visualization tools—separation of linkage weights using histogram, etc.—to aid in making linkage process decisions, as visualization is connected with intuition. He also suggested some initiatives to resolve the multiple matches obtained from a record linkage procedure. Un-duplicating the two files, before linkage, might not always be recommended, as it could reduce the number of matches. He also suggested the results from G-Link be compared with that of the Big Match software developed by Bill Winkler at the US Census Bureau. He concluded his discussion by predicting that national statistical offices around the world will continue linking large administrative databases in spite of statistical and computational challenges.

Session 1 Q&A Summary

A short question and answer session was lead by Dr. Scheuren. The first question was asked by Tom Belin (UCLA), and it pertained to the Jackson-Pulver and Madden presentation. Dr. Belin wanted to better understand the relationship between false positives, false negatives and the cultural issue of identifying oneself as indigenous). He wondered if only the number of false negatives is increasing during the linkage procedure. Jackson-Pulver and Madden responded by noting that the linkage procedure certainly produces a small number of false positives—records identified as indigenous for people that do not have an indigenous ancestry—but it is the false negatives that lead to significant underestimation of death counts among indigenous people in the Australian Bureau of Statistics estimation methodology. The more common scenario is indigenous people do not want to identify themselves as indigenous during census, but they are identified as indigenous by their close relatives at their death, leading to false negatives.

Dr. Scheuren added that in the U.S., during 2010 census, there was a substantial increase in the count of Native American people. He felt that was a grand success of 2010 census.

Surajit Chaudhuri (Microsoft) then asked about blocking and linkage rules as it pertained to the Statistics Canada record linkage system. Dr. Chaudhuri wanted to know if G-link has any particular sequencing of blocking steps and rule creation steps, or whether they are done simultaneously. Mr. Chevrette responded by stating that blocking is done to get the initial pairs that would go through the linkage procedure. Then rules are created based on certain variables other than the blocking variable. They are not done together, its two completely separate processes.

Finally, Rod Little (U.S. Census Bureau) asked whether or not the record linkage literature contains any Bayesian approaches in order to propagate the uncertainty from linkages. Dr. Little stated that in principle, using a Bayesian model, one can create multiply imputed linkages and do a multiple imputation analysis that propagates the linkage uncertainty. As far as he can tell, the current linkage procedures with the way a best match is created, lacks a principled approach, as it does not account for the uncertainty properly. Responses to the question topic were mixed.

Mr. Chevrette and Statistics Canada colleague Mr. Wenzowski stated that they did not believe a Bayesian approach was incorporated in G-Link, but that they would need to confirm that with the project statistician at Statistics Canada. An attendee (unidentified at this time) stated that the U.S. Census Bureau does consider a Bayesian probabilistic match for the Longitudinal Employer-Household Dynamics (LEHD) program. However, it is not feasible in practice to account for linkage uncertainties through multiple imputation when the matching is done based on multiple variables. The best possible imputed dataset is used to determine the matching status. Dr. Bate concluded the discussion by stating that there is

Bayesian thinking behind the WHO drug-monitoring project. To calculate the weights corresponding to the report pairs, a posterior distribution is used, after assuming a non-informative prior for the hyper-parameters. For the same project, other researchers considered an empirical Bayesian approach to model the probability of matches.

Session 2: Fuzzy Matches and Blocking

The Fuzzy Matches and Blocking session started with a presentation by Theresa Lamagni (Health Protection Agency, U.K.). She discussed the usefulness and challenges of linking the health datasets of different organizations. Ethical concerns related to the disclosure of personal information—data can only be shared for public benefit—and technical barriers such as large datasets and the lack common unique identifier are some of the challenges researchers face. In the context of estimating the mortality rates associated with meticillin-resistant Staphylococcus aureus (MRSA) bacteraemia infection, Dr. Lamagni discussed the performance of a probabilistic linkage procedure while linking MRSA records to death registrations.

In the second presentation, Surajit Chaudhuri (Microsoft) focused on the record matching part of a data cleaning technology heavily used by computer scientists. In his experience, two challenges faced while matching records are: 1) the large size of the reference file (more than 10 million records), and 2) a lack of customizability. To deal with the challenge of large reference table, he discussed the Jaccard similarity function that reduces the number of records from the reference file significantly. He also discussed some transformation rules to handle the large number of possible variations among entries due to the lack of customizability.

Next, Prasenjit Mitra (The Pennsylvania State University) spoke about integrating and cleaning information obtained from websites via schema matching and object mapping. Most of the current data cleaning solutions assume the data values referencing the same object bear some textual similarity. However, this assumption is often violated in practice. “Two-door front wheel drive” can be represented as “2DR-FWD” or “R2FD”, or even as “CAR TYPE 3” in different data sources. To address this problem, Dr. Mitra discussed a novel two-step automated technique that exploits statistical dependency structures among objects which is invariant to the tokens representing the objects.

The last session presentation was given by Craig Knoblock (University of Southern California) on the selection of blocking variables based on objective criteria. Dr. Knoblock discussed two goals behind the selection of blocking attributes, and suggested two criteria to measure them. Using a high Reduction Ratio measures enables one to choose blocking variables so that the number of candidate pairs are reduced. Using a high Pairs Completeness measure enables the choice of variables that maximizes true matches. Dr. Knoblock and his colleagues have developed a blocking variable selection algorithm to simultaneously control the measures, and thereby find blocking variables that meet the objective goals.

This session did not have a formal discussant, but Mary Batcher (Ernst and Young) provided some comments. She mentioned all the papers presented in the session were prime examples of efficient applications that highlight policy and privacy issues, and also present a glimpse of the business world that are using different record linkage methods. Moreover, she thought invitations from the computer scientists to work together with the statisticians would help the field develop further. Some additional discussion among attendees took place but was not adequately recorded for summary in this report.

Session 3: Dynamic Processing

The Dynamic Processing session started with a presentation by Lisa Getoor (University of Maryland). She presented a unique perspective on entity resolution wherein relational similarity between records is used to enhance the attribute similarity. Her 'greedy clustering algorithm' for entity resolution was based on an objective function that includes both similarity between attributes and similarity based on relational edges.

Next, Marianne Winglee (Westat) discussed a record linkage case study. She presented a simulation approach, called SimRate, to determine thresholds needed to classify record pairs as either matched or unmatched and to estimate linkage errors—false match and false nonmatch rates—corresponding to the particular thresholds. SimRate uses the observed distribution of data in matched and unmatched pairs to generate a large simulated set of record pairs, assigns a match weight to each pair based on specified match rules, and uses the weight curves of the simulated pairs for error estimation.

The last session presentation was given by Luna Dong (AT&T). She discussed discerning which information on a topic obtained from the internet is reliable. In many cases, the same information is available from multiple sources and there is some dependence between data sources. She presented a Bayesian approach to incorporating the dependence between data sources for discovering the truth.

Bill Winkler (U.S. Census Bureau) was the discussant for the session. He started by talking about his three main issues related to a basic record linkage process: i) the cleanup of individual files, which may be administrative lists, for use in the linkage process; ii) the use of non-unique quasi-identifiers, e.g. name, address, and date of birth, to improve record linkage; and iii) performing a linkage error adjustment analyses based on linked records. There are number of tools and references available for addressing these issues.

With these three issues as a backdrop, Dr. Winkler went on to comment on the presentations. Dr. Getoor's methods use a Markov Chain Monte Carlo (MCMC) methodology that is related to the method of Larsen and Rubin, which can be found in the Journal of the American Statistical Association (JASA, 2001). Sometimes MCMC methods are not computationally tractable. Also, Dr. Winkler notes that Dr. Geetor's work on resolving references using multiple authors is analogous to Federal agency record linkage programs, where there are multiple addresses (or telephone numbers) attached to individuals. The graphical model type techniques would help improve certain issues, however the method would have to be scaled for large files.

Regarding the case study paper by Dr. Winglee, Dr. Winkler mentioned that the method should be applied to several datasets. Dr. Winglee had parameter estimates that were plugged in the model, assuming conditional independence, but the conditional independence assumption is questionable in many situations. She calculated the conditional probability that a record pair agrees on a matching variable given that the pair is a nonmatch (U-probabilities), for pairs based on a sample of records from two files. Dr. Winkler mentioned the sampling might not always work well. On the other hand, if the model is not correctly specified, then calculation of an M-probability—the conditional probability that a record pair agrees on a matching variable given that the pair is a true match—based on training data might not always lead to a correct false nonmatch rate. While estimating error rates it is advisable to look at couple of alternative methods and implement the methods on several datasets to evaluate the performance.

With respect to Dr. Dong's methodology, Dr. Winkler's initial thoughts were that the methodology might not be applicable to Census Bureau project, as they deal with several independent administrative datasets, which are not copied from other sources. However, he realized that sometimes a reference file is created after merging various administrative datasets as well as some commercial datasets. The commercial files might be copied from some other sources and make the same errors as the source file. After merging, Dr. Dong's dependence detection algorithm can be applied to the reference file to discover the truth.

Dr. Winkler concluded his discussion by mentioning the Census Bureau deals with extremely large files such as matching 300 million records to a reference file having 8 billion records. Hence they need an algorithm that works reasonably fast. The Census Bureau's BigMatch software serves that purpose, and this software is 40-50 times faster than the parallel software available in the computer science literature.

Session 3 Q&A Summary

A short question and answer session was lead by Dr. Winkler. The first question, directed to Lisa Getoor, was asked by Yves Thibaudeau: (U.S. Census Bureau). He wanted to know how many clusters are typically considered in the latent Dirichlet allocation entity resolution problem, and the decision rule for labeling the cluster. Dr. Getoor noted that in her bibliographic dataset example, there were roughly 200-400 clusters for affiliation groups. Although clusters are characterized by having entities that appeared together in references, they are not labeled.

Surajit Chaudhury (Microsoft) asked Bill Winkler to clarify the comment in his discussion regarding the performance of the BigMatch software. Dr. Winkler confirmed that BigMatch that is used for matching records at Census Bureau is 40-50 times faster than some of the parallel software available in computer science literature (one being P-Swoosh). He and some his students performed that comparison.

Rod Little (U.S. Census Bureau) interjected comments that draw an analogy between the record linkage problem and the nonresponse error problem. If we have a training sample for which we know the match and nonmatch status, we can model the probability of a match based on the covariates available from two data sources. Using the parameter estimates, we can then draw matches (or nonmatches) for the record pairs for which we don't know the match status multiple times and use that multiply-imputed dataset for inference purposes. This multiple imputation paradigm would avoid the need to determine the cut-off points, which are pivotal in making decisions about match and nonmatch status.

Marianne Winglee concluded the Q&A session discussion by addressing some of the issues Dr. Winkler raised in his discussion. Regarding the conditional independence assumption, she mentioned the assumption worked reasonably well in their example. She and her co-authors also explored an alternative multinomial modeling approach to incorporate the dependencies among the matching variables. On Dr. Winkler's comment about being careful while estimating match rates, Dr. Winglee agreed with the need to have a reliable training dataset to obtain precise error rate estimates. However, in the paper she and her co-authors refined this uncertainty by considering a simulation approach where they generated match and nonmatch pairs repeatedly to evaluate the performance of the error rate estimates.

Session 4: Assessing and Improving the Census Bureau PVS Methodology - Conclusions, Recommendations & Next Steps from Today

The final session of the conference was a panel discussion among four experts on various record linkage methods and their applications leading to improved public policies in different fields. The technical advisory group panel consisted of Robert Goerge, Ivan Fellegi, Thomas Belin, and Michael Davern. The focus of the session was to discuss and assess various applications of record linkage techniques. In particular, the panel would focus on the Census Bureau's PVS methodology and provide recommendations for further improvements.

Dr. Robert M. Goerge is a Senior Research Fellow with Chapin Hall at the University of Chicago. He has over 25 years of research experience with a focus on improving the available information on children and families, particularly those who require specialized services related to maltreatment, disability, poverty, or violence. Goerge developed an integrated database on Child and Family Programs in Illinois by linking various administrative data to provide a comprehensive picture of child and family use of publicly provided or financed service programs. His work has provided high-quality information to policymakers to improve the programs serving children and their families.

Dr. Ivan P. Fellegi was Chief Statistician of Canada for Statistics Canada from 1985 - 2008, and, since his retirement, has the title Chief Statistician Emeritus. His dedication to public service and his extensive contribution in serving Canadians over the years has been exemplary. He also spent time in the service of the U.S. Government as a member of President Carter's Commission on the Reorganization of the U.S. Statistical System. He is considered as one of the founding fathers of record linkage. Many record linkage techniques currently used in practice are based on the probabilistic model developed by Fellegi and Sunter in their seminal paper in 1969.

Dr. Thomas R. Belin is Professor with joint appointments in the Department of Biostatistics, UCLA School of Public Health, and the Department of Psychiatry and Biobehavioral Sciences, David Geffen School of Medicine at UCLA. In the past, he has worked at the U.S, Census Bureau in the Statistical Research Division and the Statistical Support Division. His research interests have focused on incomplete data due to nonresponse, incomplete data in longitudinal studies, and the study of models that incorporate unobserved latent variables—such as hierarchical and mixture models—in order to motivate a solution to an estimation problem,. He has also pioneered the use of mixture models for performing record linkage of multivariate records.

Michael Davern is Senior Vice President and Director of the Public Health Research Department for NORC at the University of Chicago. He has built strong working relationships with producers of population-based health data at the national level, including the U.S. Census Bureau, the Agency for Healthcare Research and Quality, and the National Center for Health Statistics. A major focus of his work has been applying state-level data to health policy issues and assisting states monitor trends in health insurance coverage rates.

Census Bureau's PVS and NORC Assessment Report

The Person Identification Validation System (PVS) is the Census Bureau's production capability to verify and search for Social Security Numbers (SSNs) or Protected Identification Keys (PIKs) for person records in demographic surveys, censuses, or administrative records. PIK's are internal Census identifiers that correspond one-to-one with the set of nine-digit numbers from 000000000 to 999999999. Thus, a Social Security Number (SSN), which is a nine-digit number, corresponds one-to-one with a PIK and represents a unique individual. The PIK is assigned independently and randomly to protect the privacy of the individual person. Used as unique person identifiers, PIKs facilitate record linkage across files while enhancing data confidentiality and privacy. The quality of the PVS research files depends on the technical ability to assign the correct person identifier across linked files.

The PVS verifies SSNs and assigns PIKs by comparing person characteristics from an incoming file to the characteristics of records in the PVS reference files. NORC has conducted an assessment review of the Census Bureau's record linkage methods associated with the PVS, as well as an environmental scan of record linkage methods used by other government agencies—both within and outside of the U.S.—and private enterprises. NORC's review of the system focused on issues related to the efficiency of the matching algorithm, the quality of the input file, and the coverage of the reference files. The complete report of NORC's evaluation of the PVS can be found at the NORC PVS Assessment[2] website.

Technical Advisory Group Panel Comments

Dr. Robert M. Goerge

Dr. Goerge started the session by stating the papers presented at the conference demonstrated that challenges related to record linkage were less technical in nature–both in terms of theory and computing power—than challenges related to moving data around and getting access to data. The most difficult issue is how to allow access to extremely rich administrative data while maintaining the privacy of individuals. In Goerge's opinion, researchers are doing fantastic on record linkage theory and implementation, but

[2] The full URL is http://www.norc.org/Research/Projects/Pages/census-personal-validation-system-assessment-pvs.aspx.

need to address privacy and security better, a topic that was, at times, discussed in the conference sessions.

The use of administrative records from state agencies is rapidly expanding in the human services fields to evaluate and analyze public policies. Therefore, Goerge expressed that it important for the Census Bureau to show leadership in building high quality data that combines information from several sources. His hope was that the Census Bureau would respond well to states' requests in merging datasets and assigning PIKs.

Overall, Goerge believed that with advent of the PVS, the Census Bureau is doing a great job in helping researchers link administrative data sources. His experience in working with Census, in linking data sources, has been good. He and his colleagues at Chapin Hall at the University of Chicago have linked administrative data from state agencies in the State of Illinois. For example, one study included all families who have been Illinois Department of Children and Family Services cases, [3] and all food stamp cases and Temporary Assistance for Needy Families (TANF) cases. For these families, the researchers linked data from five important domains: foster care, mental illness in adults and emotional disorders in children, substance abuse, adult incarceration, and juvenile incarceration. Then, they linked the combined population level family data with additional programs such as the Food Stamp Program, TANF, Medicaid, and the Women Infants and Children program to better understand the children from families in human service system. A number of published reports have been produced from studies that used this data.

Goerge mentioned there are some issues associated with using only administrative data for policymaking. Nonparticipation is a widespread problem across many federal, state, and local social welfare programs. To understand both the prevalence and concentration of nonparticipation in these programs, Goerge and his colleagues linked administrative data with the general population of families, in Illinois, Texas, Maryland, and Minnesota using the American Community Survey (ACS) data.

An important issue in the academic world has been data quality—how the record linkage is done, what the issues are, evaluation of the methods, getting good statistics demonstrating the quality of the record linkage etc. NORC's report on the assessment of PVS is an important step in getting good statistics to share with the academic community researchers. Accuracy is paramount for many academic researchers, whereas relevance is paramount for the policy researchers. Goerge's work at Chapin Hall has been trying to combine these two; doing high quality, relevant research.

[3] A family where any member has had a substantiated case of abuse or neglect.

Dr. Ivan P. Fellegi

Dr. Fellegi's comments can be classified into three categories: record linkage theory, a data collection suggestion, and a philosophical point.

Record Linkage Theory

Fellegi reminded everyone that the idea of Fellegi-Sunter (1969) is that there is a space consisting of all comparisons of records in the two files. One calculates for each comparison two conditional probabilities: that of the given comparison occurring given that the two records on which they are based are truly matched, and that they are truly unmatched. The ratio of these two conditional probabilities (or more often its logarithm) is the test statistic. There are two thresholds, and then the main theorem states that the optimum record linkage is one where all records are linked that have a higher odds ratio than the bigger threshold, while all records are unlinked that have an odds ratio lower than the smaller threshold. These thresholds are determined depending on the tolerance we have for the two types of error: linking unmatched records, and failing to link matched ones.

Based on this basic version of the theory, Fellegi was looking for some explicit statement of how thresholds were established for PVS based on the tolerance for the two kinds of error rates. He was also looking for an explicit statement about what happens to the records that fall between the two thresholds. Or are the two thresholds set to the same level, in which case a record linkage system is implicitly deciding that all records that are not positively linked are unlinked. Fellegi saw no consideration of these thresholds within the PVS documentation that he reviewed—the NORC PVS Assessment Report—particularly in relation to the two kinds of error and "fitness to use" relative to the subsequent analyses.

Fellegi noted that if there is, in fact, only one threshold, then he feels that there should be some explicit consideration of the implication of this decision for the two types of error tolerances. Setting a single threshold means that one can only control the probability of false linkages or the probability of failure to link matched records—but not both. From the NORC investigation it appears that the probability of failing to match is small; but then is it possible that the common threshold had been set too low and therefore the probability of false linkages is higher than it needs to be. It is hard to tell from the abstract, but Fellegi feels that the NORC report might provide a promising approach to the setting of thresholds. More generally, Fellegi argued that the middle range represents the cases where we could not decide on the basis of available information, and at the specified error levels, whether the two records that are compared are matched. In this sense, the middle range represents the degree of uncertainty, or possible ignorance; and it is always dangerous to ignore ignorance.

Another observation that is made in Fellegi-Sunter is that when two records agree on a variable, the positive weight contributed by that variable depends on its discriminating power; and when the records

disagree on the variable, the corresponding negative weight depends primarily on the error with which that characteristic is reported. Depending on one's tolerance of the two types of error, one may either look for high discrimination—which, on the other hand might increase the error rate of reporting—or conversely. Fellegi would like an assessment of how variables were structured and "standardized" to achieve a good balance between discrimination and reporting errors.

Because of the vast number of comparisons that record linkage theory calls for, there is a clear need for a blocking strategy. Blocking eliminates large numbers of comparisons, but this implicitly assumes that the records not compared are not matched. It flows from the theory that blocking should be done using variables that have at least moderate discrimination, but low reporting error. Blocking is clearly important and there is potential offered by the "Learning Blocking Schemes" methods presented by Craig Knoblock from the University of Southern California's .Information Sciences Institute.

A Data Collection Suggestion

Based on his experience at Statistics Canada, Fellegi suggested a way to collect important financial information from ACS and other federal survey participants. To reduce reporting burden, in place of a series of questions on income, Statistics Canada provides the respondents with the option to permit Statistics Canada to access their income tax records. In Fellegi's experience, an overwhelming proportion—some 90 percent—choose to exercise this option. Of course, one could ask for the SSN at the same time.

Fellegi suggested that the Census Bureau might want to try doing the same, thereby acquiring SSNs for some of its most important survey files, including ACS, while at the same time reducing reporting burden and data collection costs. This would also allow the Census Bureau to evaluate the PVS methodology more efficiently in the presence of a truth deck consisting individuals of reporting their SSN.

A Philosophical Point

Fellegi made an important cautionary comment, from an ethical perspective, that was prompted by the paper "Application of Probabilistic Linkage Methods to Join Infectious Disease Surveillance Records to Death Registrations," presented by Theresa Lamagni from the U.K. Health Protection Agency Centre for Infections, and its statements that record linkage is a "powerful and cost-effective tool—hence it is feared."

The UK paper added that it needed inter-agency cooperation and legislative support – all of which is true. But, Fellegi thinks, given the power and utility of the technique, as well as public concerns about privacy, more defensive efforts are warranted. Any record linkage involving personal records is intrinsically privacy intrusive – since information about people is used without their authorization and control.

Therefore, every record linkage project must be individually justified on the basis that the benefits truly outweigh the privacy invasion. Furthermore, there must be clear evidence that the purpose is statistical; that no individual or collective harm can result to the people whose records are involved; sensitive linkages need to be cleared through some independent committee (like the Privacy Commissioner of Canada); and every record linkage, and its justification, should be displayed on the corresponding organization's web site. The objective is, of course, transparency, but also the demonstrable basis of clear principles and the auditable presence of sensible processes.

Dr. Thomas R. Belin

Dr. Belin's comments focused on the calibration of false-match rates in record linkage (Belin and Rubin 1995). As mentioned in Fellegi's comments, the setting of thresholds is related to error rates. If the model is known, then one can estimate the conditional probabilities of agreement given match and nonmatch rates. Then, from the ordered pattern of the cumulative probabilities, one can set the thresholds depending on the specific tolerance level of false match rate and false nonmatch rate. This was the basic principle behind the Fellegi-Sunter idea of determining the thresholds.

In his experience working with the Census Bureau's post enumeration data, Belin observed that the Fellegi-Sunter procedure for estimating false-match rates does not work well, because the data being matched did not conform well to a model of mutual independence in the matching variables. To improve the estimation of false match rates, Belin suggested the idea of calibration that can be implemented using training data. In this approach, the false-match rate is calculated assuming a mixture pattern of composite scores (weights) for each possible cut-off level (thresholds). In other words, the distribution of weights for true matches and false matches are—after suitable transformations—two different normal distributions. Several repetitions of the method lead to a single estimate of false-match rate along with its measure of uncertainty.

The theme of this discussion was related to the principle that Dr. Roderick Little eluding to during a discussion earlier in the day: capturing the uncertainty of linkage through multiple imputation. This might be possible when we have an explicit estimate of false match rates. To avoid a time consuming clerical review process, particularly in the context of large scale linkage problem that Census Bureau handles, the multiple imputation approach is often approximated by setting a single threshold.

Belin emphasized the use of training data as a powerful tool that has also been exploited in computer sciences fields such as machine learning. To illustrate the utility of training data, he referred to one specific application he was involved in—the Survey of Street Prostitutes in Los Angles. Approximately 1,000 prostitutes participated in the survey. To increase the participation rates, prostitutes were not asked their names, but were asked a series of 20 questions that could be used to distinguish individuals. The

survey offered an incentive to the participants. The methodology resulted in a certain percentage of duplicates and triplicates—raising question about the validity of the study. Belin suggested that known duplicates and triplicates could be used as a training sample for comparing individuals based on the 20 identifying questions asked in the survey. The distribution of weights for the false matches and false nonmatches would follow a mixture pattern and that would lead to the estimation of the false match rates.

Belin discussed the privacy issue as well. The more attributes you have in a dataset, the better you perform with the linking procedure. It raises the question about the variables kept in public use files, and the care needed to maintain privacy.

Dr. Michael Davern

Dr. Davern's comments focused on the usage of data that have been linked using record linkage, and the importance of studying the uncertainty associated with different attributes. The Census Bureau's PVS methodology—and record linkage in general—helped him to understand different dimensions of the uncertainty that exist in the real world. In this subtle and complex world there can more than one right answer to a research question.

To elaborate, Davern discussed one of his collaborative research projects with the Census Bureau and other organizations where he studied the accuracy of survey estimates. Medicaid data from the Centers for Medicare & Medicaid Services (CMS) was linked to the Current Population Survey's Annual Social and Economic Supplement (CPS ASEC), which was, at the time, the most prominently used survey for policy research that measures health insurance coverage. The PVS was used to assign PIKs to individual records in both the CMS and CPS ASEC data sets. The PIKs were then used to perform the linkage between records in each data set. After the linkage, the initial concern was the evaluation of the linking procedure, i.e., how the PVS system performed in terms of verification and validation rates. In that regard, the PVS performed well.

However, discordance between the survey estimates of public health insurance program enrollment and the estimates compiled from administrative data was observed. This difference became evident only after combining two completely different pieces of information; the type of insurance coverage was different between the survey response and the Medicaid administrative records for approximately 40 percent of the linked to the Medicaid administrative data. The research team's initial response was to consider the differences in reported survey Medicaid enrollment as survey measurement error and to use a model to replace the "wrong" answers in the survey with those from an imputation model informed by the linked survey and Medicaid Administrative data. The substitution of the apparent wrong answers with the imputed values was thought to be very useful for some policy purposes.

Although he initially did not recognize the issue, the research team discovered when he looked at the insurance coverage rates at the state level there was an additional concern. The Current Population Survey is conducted in a similar manner across the states; the interviewers get similar training, the same questionnaire is presented to the respondents, and same protocol is used for proxy reporting in all states. Despite of all these commonalities, there was a large amount of state variations in reporting Medicaid coverage by state and Medicaid is a stat run health insurance program. The research team noticed that the variations across states remained even after controlling other components of the survey administration (e.g., whether a proxy interview was conducted). The team then began to think that the state variations in reporting being enrolled could be used to evaluate the state run Medicaid program. If the health benefits assumed to derive from health insurance coverage (e.g. seeking preventative care) only go to those who know, they are enrolled or have people in the household who know they are enrolled then having state by state variation in accuracy can actually have health impacts. One mechanism that could be causing state differentiation is the level of state communication regarding health insurance coverage by the Medicaid agency. For policy purposes, therefore, two pieces of information obtained from his research are valuable. First, the discordance between survey estimates and the administrative records could be used to help minimize the systematic reporting error and lead to suggestions on possible improvements to CPS and other surveys. Second, the state variations could be useful for evaluating the Medicaid themselves program in various states.

In summary, he thought looking at the nuances of reality a good idea. His suggestion was to look at various truths that follow from research instead of searching for a single “truth”. To illustrate his point further, he discussed the “construction of institutional affiliation database” example described by Luna Dong from AT & T. In that context, while trying to discover the true affiliation of a person, instead of assuming that there is one right answer, he recommended using different pieces of information in innovative ways. It might be that a person has multiple affiliations and looking for only one right answer might obscure additional information that could prove useful and even, at least, partially correct.

Post Conference Comments

As communicated to NORC, the Census Bureau was pleased with the conference and will consider future research to see if the record linkage and entity resolution theories and techniques presented at the conference could be helpful to the Census Bureau’s record linkage programs, in particular, the PVS. The comments made by Technical Advisory Group may help guide this future research. Of this group’s comments, Dr. Fellegi’s had more to do with how the current PVS implements record linkage, and therefore the Census Bureau feels that some clarification is needed concerning aspects of the PVS that Dr. Fellegi discussed.

Deborah Wagner and Fritz Scheuren have put together the following list of issues related to Dr. Fellegi's comments.

1. One of the main uses of the PVS is to bring disparate data systems together. As Dr. Fellegi discussed, the possibility of a single "tight" upper bound to define a "True Link" may have utility, though it leaves unaddressed the handling of "False Nonlinks." The question of how to address Indeterminate or "Possible Links" requires additional research. At present, the Census Bureau deals with Indeterminate/Possible Links on an application-by-application basis.

2. In its assessment report, NORC did not cover all the details related to choosing the two linkage cutoff thresholds—defining the upper limit for True Links and the lower limit for Non-Links. The production version of PVS currently has the same number for both thresholds. The PVS is a one-to-many match process to assign a unique person ID to records, which serves as a person linkage key used to link to other datasets. This unique person ID should be a high quality linkage key to enable person-level research. The size of our files is too large to clerically review the Indeterminate set during production. However, during each PVS application, we review listings as we set our parameters. It is true that all records above the cutoff are linked (receive a non-blank PIK), while all records below the cutoff are not linked (receive a blank PIK). The current PVS is supported by a set of SSN-based reference files, so there are legitimate records that should not receive a PIK if that person does not have an SSN.

3. The Census Bureau PVS aims to achieve a low false match rate, and will accept a higher false non-match rate. The current cutoffs are conservatively set in the hopes of obtaining a low false match rate. The NORC report mentioned not finding many new links within the current set of non-match cases (based on the current set of reference files). The report showed problems with missing data, and geographic problems which could lead to name edit/parsing issues, making it hard to determine the appropriate link. The Census Bureau will attempt to mitigate these issues by obtaining other reference files and developing additional linkage techniques to reduce the false non-match rate while not increasing the false match rate.

4. The Census Bureau has assumed that errors in the True Link rates are small and relatively consistent across data sources processed through PVS. We speculate that False Nonlink rates, (unlike False Link rates), could vary greatly across data sources. The Census Bureau intends to

analyze False Nonlink rates using post enumeration survey data. Beyond informing possible PVS improvements, this work could contribute to the literature on multiple systems estimation.[4]

5. With respect to Dr. Fellegi's suggestion that the Census Bureau present respondents with the option to permit the government to obtain information directly from administrative records such as Social Security and tax records, there is concern that this could potentially lower response rates. The Census Bureau made the decision to stop asking for respondent SSNs in surveys in 2006. The impact of this decision was a reduced number of refusal/opt-out cases.

[4] If the number of PVS False Links is *de minimis*, then under standard independence assumptions (e.g., Bishop, Fienberg and Holland 1975) a multiple systems estimate can be obtained of the total (capturable) US population. Traditionally the US Census has done this using just two systems (e.g., Sekar and Deming 1949). This is the so-called "dual systems" approach, where one of the systems being employed is the Decennial Census itself. It has long been clear that the independence assumption underlying this dual systems approach is too strong to be tenable and must be supplemented. The PVS is a tool that offers a way, employing three or more multiple systems for weaker independence assumptions to be made and consequently a stronger approach offered.

Appendix A: Attending Organizations

Agency for Healthcare Research and Quality

AT&T

Bureau of Justice Statistics

Bureau of Labor Statistics

Centers for Medicare & Medicaid Services

Chapin Hall

Congressional Budget Office

Discovery Logic

Economic Research Service

Energy Information Administration

Ernst and Young

George Washington University

Government Accountability Office

Gunnison Consulting Group

Housing and Urban Development

Mathematica Policy Research

Microsoft

National Center for Education Statistics

National Center for Health Statistics

National Institute on Aging

National Institutes of Health

National Science Foundation

NORC at the University of Chicago

Office of Management and Budget

Pfizer

Social Security Administration

Statistics Canada

Survey Research Center at U Michigan

The Pennsylvania State University

U.K. Health Protection Agency

University of Baltimore

University of California at Los Angeles

University of Maryland

University of New South Wales

University of Southern California

University of Sydney

United States Census Bureau

Westat