The Annals of Applied Statistics
2016, Vol. 10, No. 4, 1907–1929
DOI: 10.1214/16-AOAS925

CATEGORICAL DATA FUSION USING AUXILIARY INFORMATION[1]

BY BAILEY K. FOSDICK*, MARIA DEYOREO† AND JEROME P. REITER†

*Colorado State University** *and Duke University*†

In data fusion, analysts seek to combine information from two databases comprised of disjoint sets of individuals, in which some variables appear in both databases and other variables appear in only one database. Most data fusion techniques rely on variants of conditional independence assumptions. When inappropriate, these assumptions can result in unreliable inferences. We propose a data fusion technique that allows analysts to easily incorporate auxiliary information on the dependence structure of variables not observed jointly; we refer to this auxiliary information as glue. With this technique, we fuse two marketing surveys from the book publisher HarperCollins using glue from the online, rapid-response polling company CivicScience. The fused data enable estimation of associations between people's preferences for authors and for learning about new books. The analysis also serves as a case study on the potential for using online surveys to aid data fusion.

1. Introduction. In many applications in marketing, analysts seek to combine information from two or more databases containing information on disjoint sets of individuals and distinct sets of variables [Gilula, McCulloch and Rossi (2006), Kamakura and Wedel (1997), Kamakura et al. (2003), van Hattum and Hoijtink (2008), van der Putten, Kok and Gupta (2002)]. For example, a company has one database on customers' purchasing habits and another database on individuals' media viewing habits, and seeks to find associations between viewing and purchasing habits [Gilula, McCulloch and Rossi (2006)]. This procedure, known as data fusion [Rässler (2002), pages 60–63], arises in other contexts, including microsimulation modeling in economics [Moriarity and Scheuren (2003)] and government statistics [D'Orazio, Di Zio and Scanu (2002)]. For applications in other areas, see Kadane [(2001), reprinted from a 1978 manuscript], Rodgers (1994), Moriarty and Scheuren (2001) and D'Orazio, Di Zio and Scanu (2006).

Typical applications of data fusion rely on strong and unverifiable assumptions about the relationships among the variables. To see this, consider fusion of two databases, D_1 and D_2, with disjoint sets of individuals. Let A denote the set of variables common to both databases, such as demographics; let B_1 denote the

Received June 2015; revised December 2015.

[1]Supported in part by grants from the National Science Foundation (NSF-SES-11-31897 and DMS-1127914). This research was approved by the institutional review boards at Duke University (IRB protocol number D0489) and Colorado State University (IRB protocol number 081-17H).

Key words and phrases. Imputation, integration, latent class, matching.

set of variables unique to D_1; and let B_2 denote the set of variables unique to D_2. Since $\{A, B_1, B_2\}$ are never observed simultaneously, the joint distribution of $\{A, B_1, B_2\}$ is not identifiable based on (D_1, D_2) alone. Neither is the distribution of $\{B_1, B_2\}$, either marginally or conditionally on A. Put another way, many possible specifications of the joint distributions of $\{A, B_1, B_2\}$ may be consistent with the marginal distributions of $\{A, B_1\}$ in D_1 and $\{A, B_2\}$ in D_2. The data provide no information on which specifications to favor.

For data fusion to proceed, analysts must make some assumption about the joint distribution of $\{A, B_1, B_2\}$. The most common assumption is that the variables in B_1 are conditionally independent of those in B_2, given the variables in A [D'Orazio, Di Zio and Scanu (2006), Gilula, McCulloch and Rossi (2006), Kiesl and Rässler (2006)]. For example, an analyst might assume that every person with the same age, gender, occupation, race, county of residence, etc., has the same probability of purchasing the product, regardless of their media viewing habits. While this assumption could be reasonable in some contexts with rich A variables, it also could be grossly incorrect. For example, in some demographic groups, people who watch advertising infrequently may be less likely to purchase the product. When this is the case, assuming conditional independence can result in inferences about $\{A, B_1, B_2\}$ that do not accurately reflect the underlying relationships in the population.

To reduce reliance on conditional independence assumptions, analysts require some form of auxiliary information. For example, analysts can use knowledge about the joint distribution of $\{B_1, B_2\}$ from other sources to bound the joint distribution of $\{A, B_1, B_2\}$ [D'Orazio, Di Zio and Scanu (2006)]. Another possibility is to collect additional data that provides information on unknown features of the joint distribution of $\{A, B_1, B_2\}$. Historically, such surveys have been untimely and prohibitively expensive. However, in recent years technological advances have opened the door to fielding rapid response, low-cost surveys [Gilula and McCulloch (2013)]. Questions then arise as to how analysts can leverage the information in such surveys for more accurate data fusion.

In this article, we propose a novel data fusion approach that allows analysts to incorporate auxiliary information on arbitrary subsets of $\{A, B_1, B_2\}$ with at least one variable in B_1 and one in B_2 jointly observed. While auxiliary information in the form of complete data has been considered previously, the modeling framework presented here accommodates a much wider variety of forms. We refer to this auxiliary information as *glue*, since it serves to strengthen the connection between B_1 and B_2. We present the approach for the common setting of all categorical variables, although similar strategies could be used for numerical variables. The basic idea is to collect or construct a dataset that represents the auxiliary information, append this dataset to the concatenated file (D_1, D_2), and fit an imputation model to predict missing B_1 in D_2 and missing B_2 in D_1. As the engine for imputation, we use a Bayesian latent class model [Dunson and Xing (2009), Si and Reiter (2013)]. Using simulation studies, we illustrate how to accommodate glue of various sizes

and on various variable subsets, and demonstrate the potential for glue to improve accuracy relative to fusion procedures that assume conditional independence. We discuss problems that can arise when using glue from a nonrepresentative sample, and propose an approach to incorporating nonrepresentative glue in data fusion. We illustrate the methodology using a data fusion experiment in which we obtain glue from the internet polling company CivicScience, and use the glue to fuse surveys fielded by the book publisher HarperCollins Publishers on author preferences and author discovery tendencies.

The remainder of the article is organized as follows. In Section 2, we introduce the HarperCollins data fusion context and review typical approaches to data fusion in the literature. In Section 3, we describe how to adapt Bayesian latent class models for data fusion to accommodate glue. The approach allows for both the creation of completed data files, that is, as in multiple imputation [Rubin (1986, 1987), Reiter (2012)], as well as parameter inference. We focus on creating completed datasets, which can be subsequently analyzed using the techniques of Rubin (1987). We also summarize results of simulation studies that demonstrate the benefits of leveraging glue in data fusion. In Section 4, we present results of the HarperCollins Publishers and CivicScience data fusion. In Section 5, we conclude with a discussion of open questions and future research directions.

2. Background.

2.1. *HarperCollins data and CivicScience glue.* HarperCollins Publishers routinely administers surveys to the public to learn about their behaviors and opinions, relying on this information to guide business decisions. The surveys typically include questions about basic demographics (e.g., age, income, gender) and reading habits, as well as questions on focused topics such as technology usage or author preferences. Usually around 10% of questions in the surveys address basic demographics and reading habits, and the remaining 90% are specific to the survey. We seek to fuse data from two HarperCollins surveys, one including questions on the authors people read and the other including questions on where people discover new authors (e.g., Facebook and Best Sellers lists). The first survey comprises 4001 respondents and 734 variables; we use only a subset of questions related to discovery and demographics. The second survey comprises 5015 respondents and 1433 variables; we use only a subset of questions relating to author readership and demographics. The surveys were administered by an independent company to a random sample of people residing in the United States, with prespecified numbers of individuals in specific categories based on age, gender, ethnicity and geographic regions.

HarperCollins is interested in understanding the demographics of readers of particular authors and how to reach them. For example, if HarperCollins publishes a new book by the author Lisa Kleypas, will they reach more of her readers by advertising the new book in bookstores or on Facebook? Furthermore, who should be the

target audience (age, gender, etc.) of the advertisements? Leveraging the connections between author readership, book discovery and demographics across surveys can help publishers such as HarperCollins pursue profitable marketing strategies.

To obtain glue for the data fusion, we collaborated with internet polling company CivicScience. Internet polling companies are potentially ideal glue collectors, as they are able to quickly survey thousands of people at a low cost. As a case in point, CivicScience collects hundreds of thousands of responses per day and has information stored on millions of respondents. CivicScience is routinely paid by other companies to canvass the public on marketing and business decisions.

CivicScience obtains information by posting short surveys, typically three or four questions, on the sidebar of popular websites. Participation is purely voluntary (raising the potential for selection bias, which we return to later). CivicScience entices participation by beginning each survey with an engagement question that people are often willing and eager to share their opinion on (e.g., "Who will win the Superbowl?"). The next question(s) is a value question asked on behalf of a paying client. The final question inquires about respondent demographics. After completing the short survey, participants are offered the option to answer additional questions. CivicScience uses participants' computer IP addresses to link responses from the same individuals (more accurately, from the same computer).

For our application, CivicScience ran numerous three-question surveys on author readership and discovery. The first question was an engagement question to solicit participation; the second question was about either author readership or discovery; and the third question was about either the respondent's age or gender. Many participants completed more than one survey, allowing CivicScience to link responses on author readership, discovery, age and gender. We use these linked data in the fusion of the HarperCollins surveys.

2.2. *Common data fusion methods*. The most widely used data fusion technique in practice is statistical matching [van der Putten, Kok and Gupta (2002), Wicken and Elms (2009)]. The analyst divides the observations in (D_1, D_2) into groups based on the similarity of values in the A variables. Within each group, the analyst imputes missing B_1 values for records in D_2 by sampling from the empirical distribution of B_2 in that group. The analyst imputes missing B_2 values for records in D_1 in a similar manner. Often one cannot find groups of records in D_1 and D_2 with exactly the same values on all of A, particularly when the contingency table implied by the variables in A has a large number of cells. In such cases, analysts form groups based on some subset of A variables. Alternatively, analysts specify some distance function that quantifies how "close" the A values are for a given pair of observations from D_1 and D_2, and form groups based on the close matches. Regardless of how the analyst forms groups, these approaches all make the unverifiable assumption that B_1 is independent of B_2 within the analyst-specified groups.

A second approach to data fusion is to estimate regression models for the distributions of $(B_1|A)$ from D_1 and $(B_2|A)$ from D_2, and set $f(B_1, B_2|A) = f(B_1|A)f(B_2|A)$, that is, assume conditional independence between B_1 and B_2 [Gilula, McCulloch and Rossi (2006), Rodgers (1994)]. One then imputes missing values of B_1 using the estimated model for $(B_1|A)$, and imputes missing values of B_2 using the estimated model for $(B_2|A)$. Gilula, McCulloch and Rossi (2006) describe how to adapt this regression-based approach to incorporate auxiliary information about the dependence between a single binary B_1 and a single binary B_2.

A third approach is to estimate models for the entire joint distribution of $\{A, B_1, B_2\}$. For example, one could use a multinomial distribution with probabilities constrained by a log-linear model that excludes terms involving interactions between B_1 and B_2. This also assumes conditional independence between B_1 and B_2. D'Orazio, Di Zio and Scanu (2006) describe how this conditional independence assumption can be relaxed in log-linear models by incorporating auxiliary information on marginal probabilities for (B_1, B_2). Alternatively, one could estimate the joint distribution of $\{A, B_1, B_2\}$ with a latent class model [Goodman (1974)], as suggested by Kamakura and Wedel (1997) and as we do here.

Unlike log-linear models, latent class models can capture complex associations among the variables automatically, avoiding the difficult task of deciding which interactions to include from the enormous space of possible models [Si and Reiter (2013), Vermunt et al. (2008)]. Latent class models also easily handle missing values in D_1 and D_2 due to item nonresponse within the surveys, assuming nonresponse is missing at random [Rubin (1976)]. While the latent class model has been used in the context of data fusion, it has never been developed for incorporating auxiliary information in data fusion. Furthermore, while others have proposed methodology for using auxiliary information in the form of complete observations on $\{A, B_1, B_2\}$, we introduce more general methodology able to accommodate any additional, not necessarily complete, observations that contain variables not previously observed simultaneously.

3. Methodology.

3.1. *Bayesian latent class models for categorical data fusion.* Suppose that we seek to fuse database D_1 comprising n_1 individuals with database D_2 comprising n_2 individuals. Let $Y_{ij} \in \{1, \ldots, d_j\}$ be the value of variable j for individual i, where $j = 1, \ldots, p$ and $i = 1, \ldots, n_1 + n_2$. Let $Y_i = (Y_{i1}, \ldots, Y_{ip})$ for all i. The p variables form a contingency table with $\prod_{j=1}^{p} d_j$ cells. For variables $j \in A$, we observe Y_{ij} for all $n = n_1 + n_2$ individuals; for variables $j \in B_1$, we observe Y_{ij} for only the n_1 individuals in D_1; and, for variables $j \in B_2$, we observe Y_{ij} for only

the n_2 individuals in D_2. We note that, in practice, item nonresponse will result in unintentionally missing values within D_1 and D_2 as well.

In latent class models for categorical data, we assume that each individual is a member of one of N unobserved classes. Let $Z_i \in \{1, \ldots, N\}$ denote individual i's class membership, and let $\pi_l = \mathrm{P}(Z_i = l)$ be the probability that individual i is in class l. We assume that $\pi = (\pi_1, \ldots, \pi_N)$ is the same for all individuals. Within each class, we assume the variables follow independent categorical distributions with variable-specific probabilities $\phi_l^{(j)} = (\phi_{l1}^{(j)}, \ldots, \phi_{ld_j}^{(j)})$, where $\phi_{ly}^{(j)} = \mathrm{P}(Y_{ij} = y | Z_i = l)$. As a flexible and computationally convenient prior distribution on π and $\{\phi_l^{(j)}\}$, we use the truncated version of the Dirichlet Process (DP) prior [Sethuraman (1994)]. The complete model, referred to as the DP mixture of products of multinomials (DPMPM), can be expressed as

$$Y_{i1}, \ldots, Y_{ip} | Z_i, \phi \overset{\text{ind.}}{\sim} \prod_{j=1}^{p} \text{categorical}\big(Y_{ij}; \phi_{z_i 1}^{(j)}, \ldots, \phi_{z_i d_j}^{(j)}\big), \tag{3.1}$$

$$\begin{aligned} Z_i | \pi &\overset{\text{ind.}}{\sim} \text{categorical}(\pi_1, \ldots, \pi_N), \qquad i = 1, \ldots, n, \\ \pi_l &= V_l \prod_{r=1}^{l-1} (1 - V_r), \qquad \pi_N = 1 - \sum_{l=1}^{N-1} \pi_l, \\ V_l | \alpha &\overset{\text{i.i.d.}}{\sim} \text{beta}(1, \alpha), \qquad V_N = 1, \qquad l = 1, \ldots, N-1, \\ \phi_l^{(j)} &\overset{\text{ind.}}{\sim} \text{Dir}\big(a_1^{(j)}, \ldots, a_{d_j}^{(j)}\big), \qquad l = 1, \ldots, N, \, j = 1, \ldots, p, \end{aligned} \tag{3.2}$$

$$\alpha \sim \text{gamma}(a_\alpha, b_\alpha). \tag{3.3}$$

The parameter α plays a central role in determining the number of effective components in the mixture, with smaller values favoring fewer components. A hyperprior on α allows the data to inform the number of components. In our applications, we fix a_α and b_α equal to 0.5 in the prior distribution in (3.3), which represents a relatively noninformative prior. We set $a_1^{(j)} = \cdots = a_{d_j}^{(j)} = 1$ for all j.

We estimate the DPMPM model using Markov chain Monte Carlo (MCMC) posterior simulation techniques [Ishwaran and James (2001), Ishwaran and Zarepour (2000)]. The missing Y_{ij}, unforeseen from item nonresponse and expected due to the structure of data fusion, are imputed as part of the MCMC. Given a draw of model parameters $(\alpha, \{\phi^{(j)}\}, Z, V, \pi)$, we sample a value for each missing Y_{ij} from the relevant independent categorical distribution in class Z_i. Further details on the sampling algorithm are provided in the Appendix.

The probability model defined in (3.1) and (3.2) is the same as that used by Kamakura and Wedel (1997). However, rather than use a fully Bayesian estimation approach, they maximize the likelihood function obtained from equations (3.1) and (3.2). Additionally, Kamakura and Wedel (1997) use heuristics to determine

some optimal number of classes, whereas with the DPMPM one simply can fix the truncation level N to a large value [Ishwaran and James (2001)]. To ensure that N is large enough, the analyst confirms that the number of occupied classes n^* is always significantly less than N across MCMC samples. If the posterior distribution for n^* places significant mass near N, then N should be increased. In the analyses in this article, $N = 30$ is always sufficiently large.

Even though variables are independent within the latent classes, variables still can be marginally dependent across the set of classes. For example, for any pair of variables j and j', we have

$$P(Y_{ij} = y, Y_{ij'} = y' | \pi, \{\phi^{(j)}\}) = \sum_{l=1}^{N} \pi_l \phi_{l,y}^{(j)} \phi_{l,y'}^{(j')}. \tag{3.4}$$

In general, the expression in (3.4) is not identical to the product of the two marginal probabilities, $(\sum_{l=1}^{N} \pi_l \phi_{l,y}^{(j)})(\sum_{l=1}^{N} \pi_l \phi_{l,y'}^{(j')})$, implying Y_{ij} and $Y_{ij'}$ are independent conditional on Z_i and $\{\phi^{(j)}\}$, but dependent upon marginalization over Z_i. Expression (3.4) can be used for model-based inferences about probabilities.

As suggested by Gilula, McCulloch and Rossi (2006) when discussing the model used by Kamakura and Wedel (1997), estimates of the joint distribution of $\{A, B_1, B_2\}$ from latent class models may not be concordant with conditional independence. In our simulations, we found that the DPMPM favors somewhat stronger correlation between B_1 and B_2 than is implied under conditional independence. This results from the clustering engendered by the DP prior specification since the data contain no information about $\{B_1, B_2\}$ jointly. This finding underscores the potential benefits of using glue when using latent class models for data fusion.

3.2. *Incorporating glue in data fusion.* Schifeling and Reiter (2016) developed a strategy for incorporating prior information about marginal probabilities into the DPMPM. They suggest constructing a hypothetical dataset that represents prior beliefs, appending it to the collected data, and estimating the latent class model with the concatenated real and hypothetical data. As an example, if one knows only that the true proportion of women in a population is exactly 50%, one can append a large hypothetical dataset with equal numbers of men and women with all other variables missing. Schifeling and Reiter (2016) show that this approach fixes the posterior probability of being female at 50% without distorting the conditional distributions of other variables on gender.

We adapt this strategy to incorporate glue in data fusion. We assume that the analyst has glue data, D_s, in which some subset of the $\{B_1, B_2\}$ variables, possibly with A, is measured. For individuals $i = 1, \ldots, n_s$ in D_s, let Y_i be the $p \times 1$ vector of measurements for the ith individual. In most data fusion scenarios, each Y_i will be incomplete by design, in that only some variables are available in D_s. We assume that Y_i for individuals in D_s follows the model in (3.1)–(3.3). Thus,

we concatenate (D_1, D_2, D_s) in one file, and estimate the DPMPM model using MCMC. The information on $\{A, B_1, B_2\}$ available in D_s influences the parameter estimates, resulting in imputations of missing B_1 variables in D_2 and B_2 variables in D_1 that reflect the dependence relationships in the glue. For computational convenience, when fitting the MCMC we impute missing values in D_1 and D_2, but not those in D_s.

The ideal glue includes data on all variables in (A, B_1, B_2) and is a sample from the distribution of (A, B_1, B_2) in the population of interest. In practice, glue may be available only on subsets of variables, such as (B_1, B_2). In addition, D_s may not be representative of the population. For example, in the HarperCollins and CivicScience data fusion, only the conditional distributions $P(B_1|A, B_2)$ can be plausibly considered representative.

To investigate the potential benefits of glue in these scenarios, we use three sets of simulation studies. First, we add glue on different subsets of variables to explore the intuition that richer glue (i.e., glue that contains more variables simultaneously observed) results in larger improvements in inference. Second, we analyze the sensitivity of inference to the addition of varying amounts of data subjects in the glue. Third, we study the validity of inferences when using glue that is not representative of the population distribution of (A, B_1, B_2). We also present a method for appropriately incorporating such information. We note that each of these issues arises when using the CivicScience data as glue.

3.3. *Simulation studies with representative glue.* We simulate fusion settings using a third HarperCollins survey containing 4000 respondents and 1056 variables. As the A variables, we select demographics variables gender, age, work status and income. As the B_1 and B_2 variables, we select eBook reader ownership and number of hours spent reading per week, respectively. Table 1 describes the variables in detail. To generate simulated data, we create D_1 by randomly selecting half of the 3567 complete cases and removing reading hours, and create D_2 as the remaining half of the complete case data with eBook reader ownership removed. This process is repeated 10 times in order to obtain 10 different sets of (D_1, D_2). We are interested in fusing D_1 and D_2 to estimate the relationship between eBook reader ownership and reading hours per week, conditional on specific demographics variables. Because we have the complete observations of $\{A, B_1, B_2\}$ in the original data, we can compare results from data fusion to the ground truth.

To quantify the potential for glue in this example, we investigated the Fréchet bounds [D'Orazio, Di Zio and Scanu (2006)] on $P(B_1 = j, B_2 = k)$ for $j = 1, 2$ and $k = 1, 2, 3$, as implied by the marginal distributions $P(A, B_1)$ and $P(A, B_2)$. If these bounds are tight, signifying the probabilities are highly constrained by the observed marginal probabilities $P(A, B_1)$ and $P(A, B_2)$, then little is to be gained from incorporating glue. Conversely, if the bounds on the cell probabilities of $P(B_1, B_2)$ are wide, then glue has the potential to greatly improve inferences

TABLE 1
Variables contained in the HarperCollins survey used for simulations. Level labels correspond to the ordering of categories listed

Variable	Group	No. levels	Levels
Gender	A	2	male, female
Age	A	6	18–24, 25–34, 35–44, 45–54, 55–64, 65+
Work status	A	6	emp FT, emp PT, homemaker retired, self-emp, other
Income ($1000)	A	6	<25, 25–45, 45–75, 75–99, 100+, won't say
Ebook	B_1	2	yes, no
Hours	B_2	3	< 1, 1–4, 5+

based on $P(B_1, B_2)$. Note that the marginal distributions $P(B_1)$ and $P(B_2)$ themselves constrain $P(B_1, B_2)$. The Fréchet bound widths on the six cell probabilities ranged from 0.163 to 0.169. This implies that even with observing $\{A, B_1\}$ and $\{A, B_2\}$ there remains a lot of uncertainty about $\{B_1, B_2\}$ and potentially much to be gained from collecting glue.

3.3.1. *Glue richness.* We consider four types of glue for D_s. In increasing order of richness, these include only the marginal distribution $\{B_1, B_2\}$, the joint distribution of $\{A_g, B_1, B_2\}$ where A_g represents gender, the joint distribution of $\{A_a, B_1, B_2\}$ where A_a represents age, and the joint distribution of $\{A_g, A_a, B_1, B_2\}$. In each case, we create glue by duplicating the appropriate variables for all respondents in the original survey; thus, $n_s = 3567$. We run the MCMC chains long enough to obtain 120,000 posterior samples of all parameters. From these runs, we sample $m = 50$ completed datasets, (D_1^*, D_2^*), which we use in multiple imputation inferences.

To evaluate the impact of glue richness, we compare Hellinger distances, which are commonly used to quantify the similarity between two probability distributions [Gibbs and Su (2002), Pollard (2002)]. Hellinger distances based on $\{A, B_1, B_2\}$ reflect the accuracy of the entire estimated joint distribution $P(A, B_1, B_2)$, which arguably is the most important level of validity a fusion process can achieve [Rässler (2004)]. For two discrete distributions P and Q taking on k values with probabilities $(p_1, \ldots, p_k)$ and $(q_1, \ldots, q_k)$, the Hellinger distance is given by $2^{-1/2}\sqrt{\sum_{i=1}^{k}(\sqrt{p_i} - \sqrt{q_i})^2}$. This quantity is between zero and one, where smaller values imply more similarity between the distributions. Because the richest type of glue contains observations on $\{A_g, A_a, B_1, B_2\}$, for each simulation we compute Hellinger distances between the empirical distribution of (A_g, A_a, B_1, B_2) based on the original complete survey and that based on imputed data files. Calculations of distances based on the joint distribution (A, B_1, B_2) including all demographic variables, rather than just (A_g, A_a, B_1, B_2), yield similar patterns.

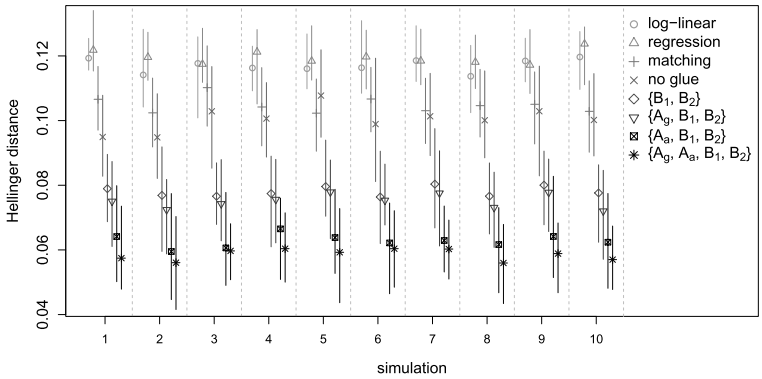

FIG. 1. *Hellinger distance between the empirical distribution of* (A_g, A_a, B_1, B_2) *based on the original complete survey and that based on completed data files using supplemental glue* D_s *with five varying levels of richness. For each of the* 10 *replications, the point represents the mean Hellinger distance across the* 50 *imputed datasets, and the endpoints of the vertical bars represent the minimum and maximum distances.*

In addition to the latent class model with and without glue, for comparison we implement the three common data fusion techniques outlined in Section 2.2: matching, regression, and log-linear models. For each approach, we create 50 completed datasets using multiple imputation. For matching, we match each record in D_1 to a record in D_2, and vice versa, with the exact same A variables, choosing one record at random when the equivalence set has multiple records. In the regression approach, the imputation models include a logistic regression for $P(B_1 = 1|A)$ estimated with the data in D_1, and a multinomial logit model for $P(B_2 = j|A)$, where $j = (1, 2, 3)$, estimated with the data in D_2. Each regression model contains only main effects terms for A because none of the interaction terms were statistically significant. For the log-linear model, we use the model that includes all main effects and interactions except those involving (B_1, B_2) simultaneously, which are inestimable without glue.

Figure 1 displays the minimum, maximum and mean Hellinger distances between the empirical distribution of (A_g, A_a, B_1, B_2) and the distributions estimated from 50 imputed datasets. The results indicate that using any type of glue yields significant gains in accuracy, with increasing gains with richer glue. These results also suggest that gender offers smaller gains than age, a consequence of the fact that the distribution of $\{B_1, B_2\}$ is more similar across gender than age. This finding is evident in all of the evaluations that follow. For matching, the empirical joint probability distribution is comparable to that produced from the latent class model with no glue. The regression and log-linear modeling approaches are the least accurate in terms of Hellinger distance.

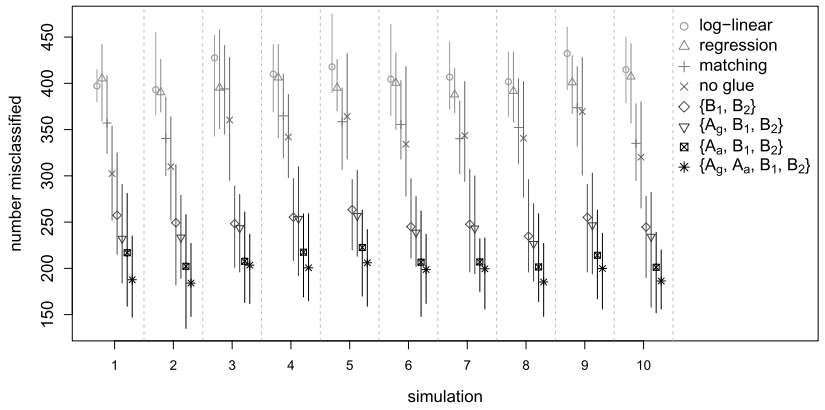

FIG. 2. *Number of individuals in the incorrect cell of the contingency table under five different glue scenarios and three existing fusion methods with no glue. For each of the* 10 *replications*, *the point represents the mean number of misclassified individuals across the* 50 *imputed datasets*, *and the endpoints of the vertical bars represent the minimum and maximum number of misclassifications*.

We also compare the sum of the absolute differences between the counts in the true contingency table for $\{A_g, A_a, B_1, B_2\}$ based on the original complete data file and those based on imputed complete data files. These counts, when divided by two, indicate how many individuals the model places in incorrect cells of the empirical contingency table. We approximate the expected number of "misclassified" individuals in an imputed dataset with the empirical average over 50 imputed data files. Mathematically, the approximation for the expected number of misclassified individuals can be expressed as

$$\mathrm{E}\left(0.5\sum_{j=1}^{\prod_{k=1}^{p} d_k} |n_j - \hat{n}_j|\right) \approx \frac{1}{50}\sum_{m=1}^{50}\left(0.5\sum_{j=1}^{\prod_{k=1}^{p} d_k} |n_j - \hat{n}_j^{(m)}|\right),$$

where $\hat{n}_j^{(m)}$ is the number of individuals in cell j in the mth imputed dataset and n_j is the true number of individuals in the original complete dataset. Figure 2 shows similar patterns as in Figure 1: using glue improves over existing approaches that assume conditional independence, with increasing gains as the glue becomes richer. We note that adding gender information to glue already containing age does not lead to much improvement in imputation accuracy.

As a more focused evaluation, we use the completed datasets corresponding to a single generated (D_1, D_2) to estimate a logistic regression of eBook reader ownership on reading hours and the demographics variables. The model includes terms for all main effects for all predictors, pairwise interactions between reading hours and gender and reading hours and age, and the three-way interaction among reading hours, gender and age. Letting A_i represent income and A_w represent work

status, the link function can be expressed as

$$\begin{aligned}\operatorname{logit}\big(P(B_1=1)\big) &= \beta_0 + \beta^g 1(A_g=2) + \sum_{k=2}^{6}\beta_k^a 1(A_a=k) + \sum_{k=2}^{6}\beta_k^w 1(A_w=k)\\ &\quad + \sum_{k=2}^{6}\beta_k^i 1(A_i=k)\\ &\quad + \sum_{k=2}^{3}\beta_k^h 1(B_2=k) + \beta^{gh}1(A_g=2, B_2=3)\\ &\quad + \beta^{ah}1(A_a=6, B_2=3) + \beta^{gah}1(A_g=2, A_a=6, B_2=3).\end{aligned}$$

We estimate the coefficients from the 50 completed datasets for one of the simulations using the standard multiple imputation combining rules [Rubin (1987)]. Because the imputations are extremely similar across the 10 replications, we show these results only for one simulation of (D_1, D_2). As displayed in Figure 3, 18 of the 22 regression coefficients based on the original data are contained in the 95% MI confidence intervals under the data fusion model applied with no glue. All intervals contain the original data coefficients when glue includes $\{A_a, B_1, B_2\}$ as well as $\{A_g, A_a, B_1, B_2\}$. Adding glue with only $\{B_1, B_2\}$ improves the estimates of the main effects associated with B_2 (reading hours). Adding glue with at least $\{A_a, B_1, B_2\}$ results in further improvements, in particular resulting in more reliable estimates of the interaction term associated with $A_a \times B_2$ (age $\times$ hours). Clearly, even targeted inferences can be improved by collecting glue, with generally increasing gains with richer glue.

3.3.2. *Glue size.* In Section 3.3.1, the glue sample size was equal to the total survey sample size, that is, $n_s = n = 3567$. Generally, this will not be the case. To evaluate the role of glue sample size, we repeated the simulations using $\{A_g, A_a, B_1, B_2\}$ as glue with different sample sizes for D_s in each of the 10 simulations. As shown in Figure 4 and Table 2, as expected, more high quality glue observations result in more accurate estimates with less uncertainty. Data fusion with $n_s = 1784$ glue cases yields inferences that are close to the ground truth and to the inferences produced with more glue cases, suggesting that even modest amounts of glue can improve inferences.

3.3.3. *Nonrepresentative glue.* While glue obtained from nonprobability samples like CivicScience polls is convenient and inexpensive, it might not be representative of the joint distribution of $\{A, B_1, B_2\}$ in the target population for (D_1, D_2). For example, D_s may disproportionately represent some demographic groups compared to their shares in (D_1, D_2). When the concatenated data (D_1, D_2, D_s) is not an (incomplete) draw from $P(A, B_1, B_2)$, the posterior distributions of the DPMPM model parameters will not produce accurate estimates

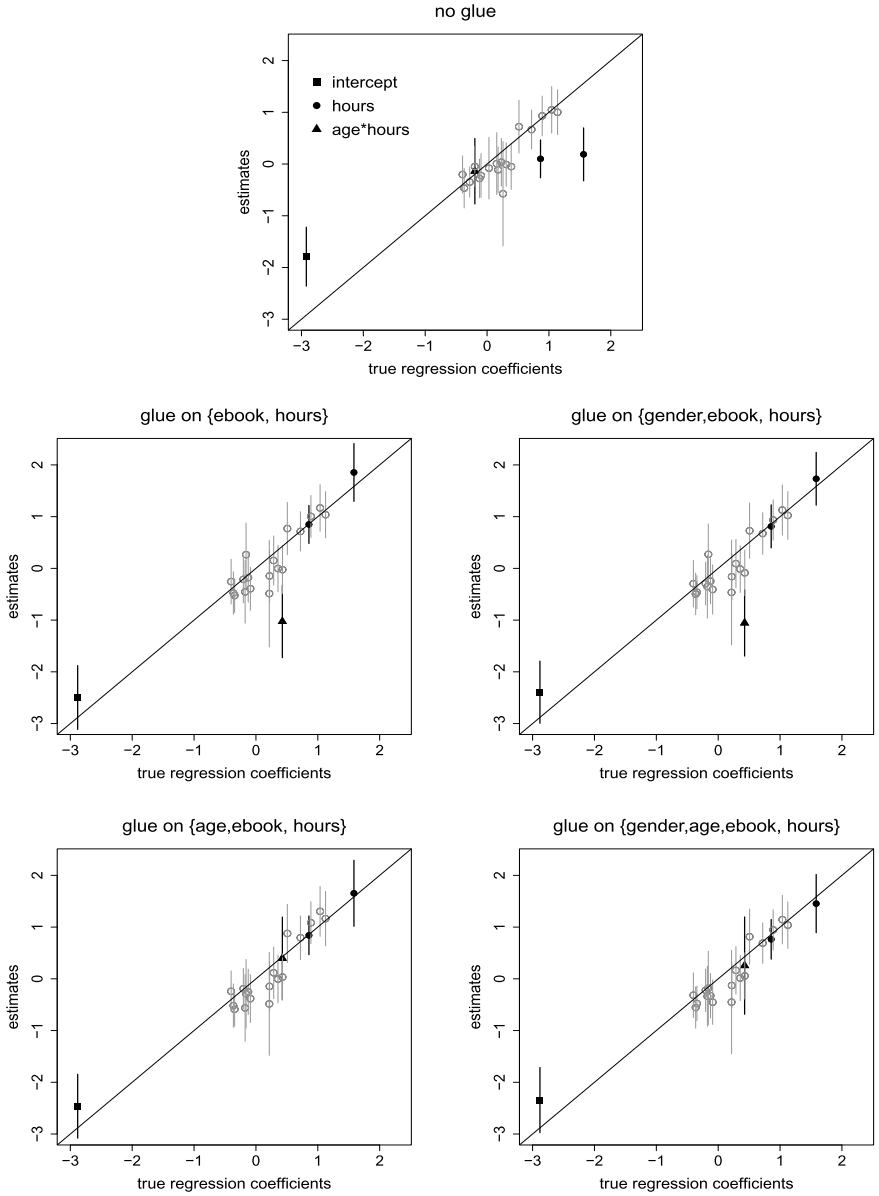

FIG. 3. *Point estimates and* 95% *confidence intervals for estimated versus true regression coefficients under five different glue scenarios for one of the simulations. The first plot refers to the no glue scenario, and highlights terms which are affected by adding glue. These same* 4 *terms are highlighted in the remaining plots as more glue is added.*

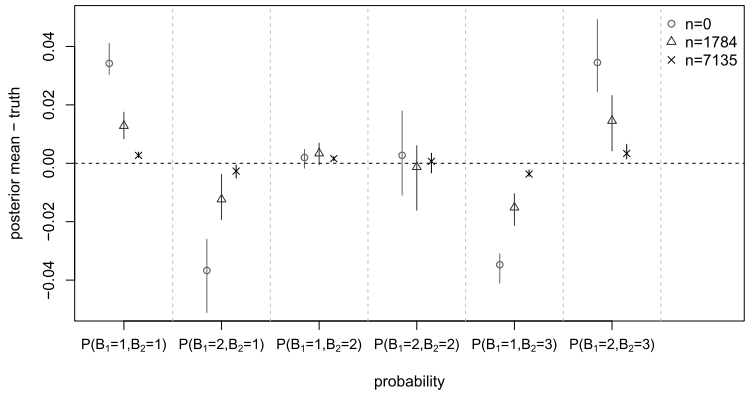

FIG. 4. *Difference between the posterior mean estimates for the marginal bivariate distribution of* $P(B_1, B_2)$ *under three different glue sample sizes and the empirical distribution based on the original complete survey, labeled truth. The points represent the average difference between the posterior mean and the truth over the* 10 *data fusion settings, and the vertical bars represent the minimum and maximum difference over simulations.*

of $P(A, B_1, B_2)$. Therefore, the resulting imputations will be draws from a biased estimate of $P(A, B_1, B_2)$, which can diminish or even negate the benefits of using glue.

To illustrate this phenomenon, consider a scenario where the glue includes $\{B_1, B_2, A_g, A_a\}$ and only responses from people aged 55+. We construct this scenario by discarding all survey responses for which age is less than 55 from the CivicScience data, and randomly sample from the remaining records to create a D_s of the same size as the original survey. When the DPMPM model is fit to the concatenated data, the average number of misclassifications is 245.2 over 50 imputations. Comparing this to the average number of misclassifications of 195.2 obtained with representative glue, we see that blindly including nonrepresentative glue can degrade inferences substantially.

TABLE 2

Average posterior mean and width of 95% *credible intervals over the ten simulations for the marginal bivariate distribution of* $P(B_1, B_2)$ *under three different glue sample sizes*

	Truth	$n_s = 0$	$n_s = 1784$	$n_s = 7135$
$P(B_1 = 1, B_2 = 1)$	0.037	0.071 (0.020)	0.050 (0.017)	0.040 (0.009)
$P(B_1 = 2, B_2 = 1)$	0.363	0.326 (0.041)	0.351 (0.032)	0.360 (0.020)
$P(B_1 = 1, B_2 = 2)$	0.064	0.066 (0.017)	0.067 (0.017)	0.066 (0.011)
$P(B_1 = 2, B_2 = 2)$	0.252	0.255 (0.037)	0.251 (0.029)	0.253 (0.018)
$P(B_1 = 1, B_2 = 3)$	0.096	0.061 (0.018)	0.081 (0.020)	0.092 (0.012)
$P(B_1 = 2, B_2 = 3)$	0.186	0.220 (0.037)	0.201 (0.029)	0.189 (0.017)

When D_s is not representative of the population, one still can construct useful glue provided that either $P(B_1|B_2, A)$ or $P(B_2|B_1, A)$ in D_s is a draw from the corresponding conditional distribution in the population. The analysis proceeds as follows:

1. Fit the DPMPM model to D_s alone to estimate $P(A, B_1, B_2)$, from which one can obtain $P(B_1|A, B_2)$ and $P(B_2|A, B_1)$.
2. Construct glue D_s^* by duplicating or sampling records $\{A, B_1\}$ with replacement from D_1, or duplicating or sampling records $\{A, B_2\}$ with replacement from D_2, and imputing the missing values of B_2 from $\{B_2|A, B_1\}$ and the missing values of B_1 from $\{B_1|A, B_2\}$ based on the conditional distributions from step (1).

In this way, the constructed glue appropriately reflects the marginal distribution of A and the information in the conditional distributions. With glue representing the appropriate joint distribution, we are in the scenarios described in Sections 3.3.1 and 3.3.2.

To assess the validity of the assumptions that $P(B_1|A, B_2)$ and $P(B_2|A, B_1)$ from D_s are representative of the population of interest, analysts can compare the empirical distributions of the sampled B_1 and B_2 variables in step (2) to those from D_1 and D_2. When these empirical distributions differ greatly, the assumptions of conditional representativeness of the glue may be inappropriate, and the glue is not useful for data fusion. When only one conditional distribution, either $P(B_1|A, B_2)$ or $P(B_2|A, B_1)$, seems reasonable, the glue can be constructed using that conditional distribution only. Analysts can choose the number of records in the constructed D_s^* to reflect their level of certainty about the conditional distributions.

We now illustrate that this diagnostic procedure can detect whether or not glue is representative on $P(B_1|A, B_2)$ or $P(B_2|A, B_1)$. We consider a setting in which D_s is representative on $P(B_1|A, B_2)$ but not on $P(B_2|A, B_1)$, constructed as follows. For $\{A_g, A_a\}$, we oversample women and older individuals by keeping all observations with $A_g = 2$ or $A_a > 4$, and sample each of the remaining observations with probability 0.5. This results in $n_s = 2837$ auxiliary cases. We sample each record's B_2 from $\{1, 2, 3\}$ with probabilities $(0.7, 0.15, 0.15)$. This is highly nonrepresentative, as the true marginal probabilities are $(0.41, 0.32, 0.27)$. We sample each record's B_1 from $\{1, 2\}$ with probabilities given by the empirical $P(B_1|A, B_2)$ from the original data. Thus, D_s is representative in terms of $P(B_1|A, B_2)$, but not on $P(B_2|A, B_1)$ or any marginal distributions. We fit the DPMPM model to D_s to estimate $P(B_1|A, B_2)$ and $P(B_2|A, B_1)$, as described in step (1), and construct D_s^* as described in step (2). The resulting marginal distribution for the imputed B_1 is extremely close to the empirical distribution of B_1 from D_1, with differences of only 0.01. The marginal distribution for imputed B_2 is $(0.57, 0.23, 0.20)$, quite far from the original data values. The diagnostic suggests that $P(B_2|A, B_1)$ is not representative, whereas it may be reasonable to rely on $P(B_1|A, B_2)$.

4. HarperCollins data fusion with CivicScience glue. We now turn to the HarperCollins data fusion. We seek to combine information from two surveys. The dataset D_1 contains $n_1 = 2000$ respondents who answered questions related to the discovery of new authors, for example, "Do you become aware of an author by [medium]?" for different media. Although this first survey was administered to 4001 respondents, we restrict ourselves to the half that was asked about author discovery. In D_2, HarperCollins asked $n_2 = 5015$ different people about their interest in various authors. Each person was asked about different subsets of authors, and so D_2 includes many missing values. We let B_1 represent author discovery via the media Best Seller List, Facebook, the library, online, recommendations and the bookstore. We let B_2 represent interest in the authors Shel Silverstein, Agatha Christie, Suzanne Collins, Stephanie Meyer and Lisa Kleypas. Each B_1 variable is recorded as yes or no. Each B_2 variable is recorded as one of three categories, namely, read, interested or not interested. Both D_1 and D_2 contain the demographic variables age, gender and income, all of which are of strong interest to HarperCollins for market segmentation. Our goal is inference on relationships between discovery media and author interest, in particular on the distributions $P(B_1|B_2)$, $P(B_1, B_2)$ and $P(B_1, B_2|A)$.

We provided CivicScience with a list of questions to ask in one of their surveys, with the goal of procuring glue. CivicScience collected $n_s = 2730$ simultaneous observations on author discovery and interest, along with age and gender for many (but not all) respondents. There are some key differences between the data collected by CivicScience and those in the original HarperCollins surveys. In particular, the CivicScience respondents tend to be older; over 60% are 55+ years old compared to only 30% of HarperCollins respondents (see Figure 5). We conjecture that is a consequence of the voluntary nature of the internet data collection done by CivicScience. We note that the distributions of A variables in D_1 and D_2 are very similar.

As discussed in Section 3.3, it is not prudent to proceed with data fusion by appending the nonrepresentative sample from the CivicScience survey to (D_1, D_2). We therefore construct D_s^* that reflects the marginal distribution of $\{A, B_2\}$ in

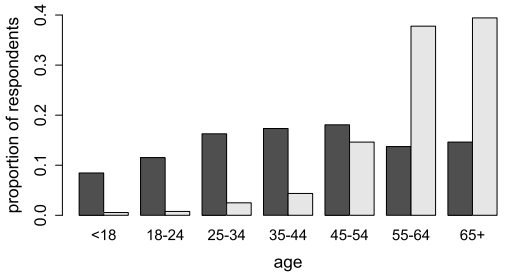

FIG. 5. *Age distributions from the HarperCollins* (*dark gray*) *and CivicScience* (*light gray*) *surveys.*

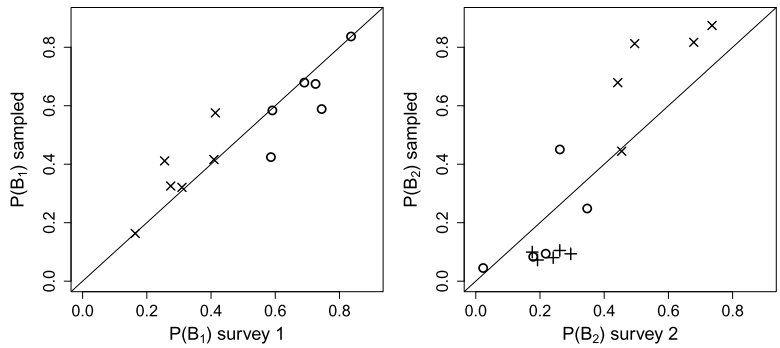

FIG. 6. *Left*: *Empirical probabilities assigned to* no (*"o" symbol*) *and* yes (*"×" symbol*) *for each of* 6 *discovery questions by sampling* B_1 *as implied by inference for* $P(B_1|A, B_2)$ *from the CivicScience data versus marginal distributions of* B_1 *from the survey data. Right*: *Empirical probabilities assigned to* read (*"o" symbol*), interested (*"+" symbol*) *and* not interested (*"×" symbol*) *for each of* 6 *author interest questions by sampling* B_2 *as implied by inference for* $P(B_2|A, B_1)$ *from the CivicScience data versus* B_2 *from the survey data.*

D_2 and the conditional distribution $P(B_1|A, B_2)$ estimated from the collected CivicScience data, following the procedure for nonrepresentative glue described in Section 3.3.3. We first duplicate $\{A, B_2\}$ from D_2, and then sample values of $\{B_1|A, B_2\}$ for these duplicated records using a DPMPM applied to the CivicScience data. As evident in Figure 6, the empirical probability distributions for the observed values of B_1 in D_1 and the sampled values of B_1 from $P(B_1|A, B_2)$ are similar, suggesting that it is not unreasonable to use the CivicScience data to estimate $P(B_1|A, B_2)$. We also considered creating D_s^* by duplicating $\{A, B_1\}$ from D_1 and sampling $\{B_2|A, B_1\}$ for the duplicated records. However, as shown in Figure 6, the sampled marginal distributions for B_2 do not closely match the empirical distributions in D_2. We therefore do not assume $\{B_2|A, B_1\}$ in the CivicScience data is representative, and construct D_s^* only from the duplicated $\{A, B_2\}$ sample from D_2. Due to the large sample size, there is little uncertainty in the estimates in Figure 6. However, in cases where the size of the original survey is smaller, there can be a fair amount of variability in the empirical probabilities, and this procedure could be repeated to quantify this uncertainty.

After appending the constructed D_s^* to (D_1, D_2), we estimate the DPMPM model on the concatenated data. In the process we impute all missing values in D_1 and D_2. As in the simulation studies, we keep $m = 50$ of these completed datasets, spacing them far apart among the 100,000 MCMC iterations post burn-in to ensure approximate independence. We use the completed versions of D_1 and D_2 for multiple imputation inferences. Standard MCMC diagnostics, such as examination of trace plots, did not suggest problems with convergence or inadequate mixing. The appropriateness of the DPMPM was evaluated using posterior predictive checks, which did not reveal substantial inadequacies in the latent class

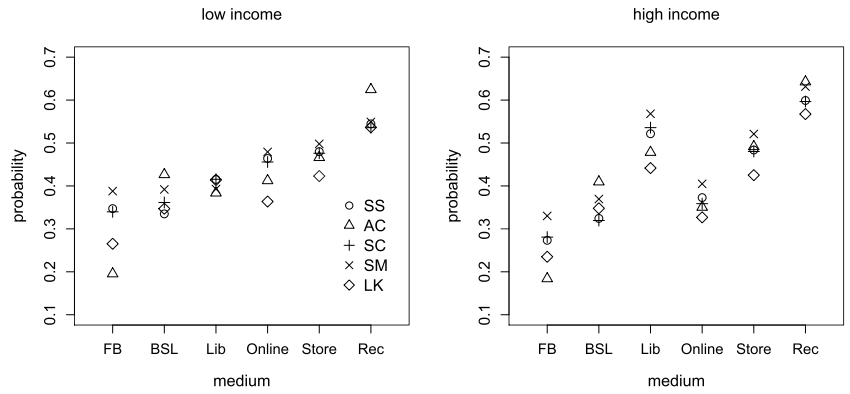

FIG. 7. *Multiple imputation point estimates for* $P(B_1 = \text{yes}|B_2 = \text{read, income})$ *for low- and high--income groups, all media* B_1 *and all authors* B_2.

model fit. Details on the model checking and MCMC diagnostics are provided in the online supplemental file [Fosdick, DeYoreo and Reiter (2016)].

As a first data fusion inference relevant for marketing strategies, we estimate probabilities of discovery via a given medium for those who have read or are interested in reading a particular author. As evident in Figure 7, high-income individuals appear very likely to discover books via recommendations regardless of author. Low-income individuals are also likely to discover books through recommendations, but the extent to which this is the case is more variable by author; for instance, low-income individuals who have read Agatha Christie are more likely to discover new books via recommendations than those who have read other authors. Probability of discovery by recommendations is the highest for both groups and all authors. Among individuals who have read Meyer, those with high incomes are very likely to discover books at the library, whereas those with low incomes are not as likely. Variability in probability of discovery via library is much greater across authors for high-income individuals than for low-income individuals. Furthermore, low-income individuals are more likely to discover books via the Internet than high-income individuals for readers of all authors; however, this difference is minimal for Kleypas. In fact, low- and high-income individuals who have read Kleypas do not appear to differ in terms of discovery.

We also look at author discovery conditional on reading interest and age, as opposed to income. Figure 8 displays inference for $P(B_1 = \text{yes}|B_2 = \text{read, age})$ across age groups for three different combinations of discovery media B_1 and authors B_2. There appears to be an increasing trend in discovery via Best Seller List for those who have read Meyer. In other words, older individuals who have read Meyer are more likely to discover new books through the Best Seller List than younger individuals. Quadratic trends are present for discovery via the Internet for those who have read Silverstein and for discovery via Bookstores for those who have read Collins. As evidence of the impact of glue, Figure 8 also displays the

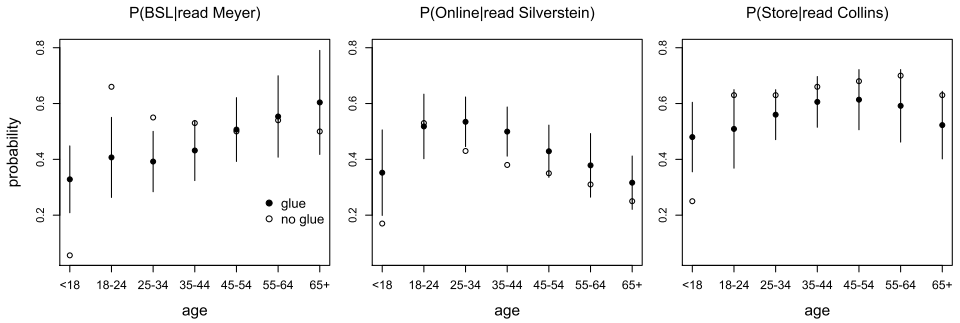

FIG. 8. *Multiple imputation point estimates and* 95% *confidence intervals for* $P(B_1 = \text{yes}|B_2 = \text{read}, \text{age})$ *across age groups for three different combinations of medium* B_1 *and authors* B_2. *Open circles refer to the estimates under the DPMPM model applied without any glue. Left*: *Probability of discovery via Best Seller List given one has read Meyers. Middle*: *Probability of discovery Online given one has read Silverstein. Right*: *Probability of discovery via Bookstores given one has read Collins.*

multiple imputation point estimates obtained from the DPMPM model fit without using the CivicScience data. In some cases these estimates agree in terms of the trends they suggest (e.g., the middle figure), but sometimes there are fairly stark differences, such as in the leftmost figure.

Finally, we estimate the conditional distributions $P(B_1|B_2)$ for particular discovery media and authors. Figure 9 displays these probability distributions for authors Silverstein and Christie, under models applied with and without glue. It appears that fans of Silverstein's books use Facebook to find out about new books more frequently than fans of Christie's books; however, both readerships rely on recommendations equally. We note that the glue impacts inference for even these

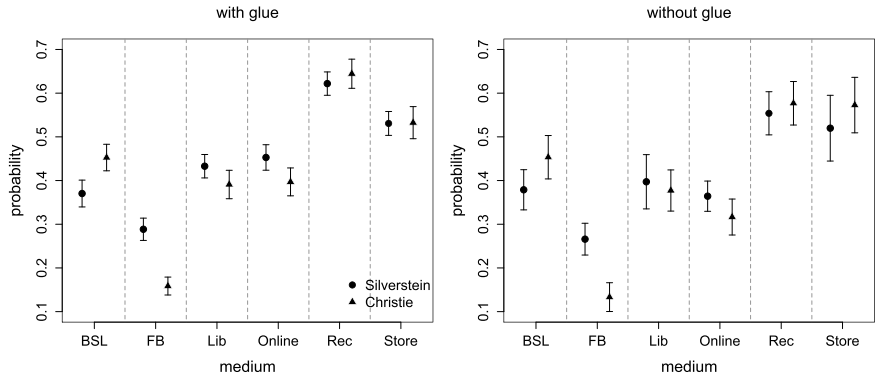

FIG. 9. *Multiple imputation point estimates and* 95% *confidence intervals for* $P(B_1 = \text{yes}|B_2 = \text{read})$ *for* B_1 *representing each of* 5 *media and* B_2 *representing Silverstein and Christie under the model applied with glue* (*left*) *and without glue* (*right*).

marginal probabilities. For instance, without glue, one cannot conclude that readers of Christie are more likely to discover via Best Seller List than readers of Silverstein. However, with glue, the 95% confidence intervals are reduced in width and no longer overlap; hence this conclusion can be made with more confidence.

5. Concluding remarks. The results of the simulation studies offer a number of general lessons about data fusion. The most important lesson is the value of collecting and using glue: compared to assuming conditional independence, using glue can improve the quality of inferences from data fusion substantially. The approach presented here—concatenating the samples and auxiliary data, and estimating models with the concatenated data—offers a principled way to take advantage of such auxiliary information and enhance data fusion, even when the additional data include only portions of the full joint distribution of interest.

Our experiences with the HarperCollins and CivicScience data fusion also provide insights about integrating online and traditional survey data. The results suggest that data from online polling companies like CivicScience, not surprisingly, are likely to not be representative on some dimensions. Obviously, one should be very cautious in making population inferences based on the online data. Less obviously, one should not naively merge online and traditional survey data and proceed with a data fusion (or other data integration) analysis; selection biases in the online data can degrade the quality of the resulting inferences. However, while joint distributions are prone to bias from selection into the online poll, it may be plausible to believe that conditional distributions in the polling data are reliable. When this is the case, our methodology provides a simple, yet general approach to leveraging the information in the conditional distributions. It prescribes a means of generating representative glue, thereby avoiding some of the consequences of selection bias in online polls. With this approach, we believe that analysts can take better advantage of timely, inexpensive online data collection to supplement traditional surveys.

Finally, the simulations with the HarperCollins data point to interesting directions for future research. In those simulations, adding gender to glue already containing age does not noticeably improve the inferences. In practice, one would expect the cost of collecting glue to increase with the number of variables; hence, in this simulated fusion context, it may not be cost effective to collect gender as part of the glue. This suggests a benefit for research on methods for selecting the variables that most improve the accuracy of data fusion, taking into account the cost of obtaining those variables.

APPENDIX: POSTERIOR COMPUTATION

In order to obtain inference under the hierarchical model, we use a Gibbs sampler to simulate from the posterior distribution $P(\{\phi^{(j)}\}, Z, V, \alpha, Y^{(\text{mis})}|\text{data})$, where $Y^{(\text{mis})}$ refers to all missing values in $Y_i = (A_i, B_{i,1}, B_{i,2})$ from D_1 and D_2,

and data refers to all observations of $(A_i, B_{i,1}, B_{i,2})$ in D_1, D_2 and D_s. For computational expediency, we need not impute missing values for D_s, as we are simply using this data to inform nonidentifiable relationships. However, it would be straightforward to impute these missing values just like we impute missing values in D_1 and D_2. We now describe the posterior full conditionals for all model parameters.

Full conditional for Z. The mixture allocation variables Z_i, for $i = 1, \ldots, n$, are updated from categorical distributions with probabilities given by

$$P(Z_i = h | Y_i, \pi, \phi) = \frac{\pi_h \prod_{j=1}^{p} \phi_{hY_{ij}}^{(j)}}{\sum_{k=1}^{N} \pi_k \prod_{j=1}^{p} \phi_{kY_{ij}}^{(j)}} \tag{A.1}$$

for $h = 1, \ldots, N$. For the glue cases $i = n + 1, \ldots, n + n_s$, let J_i represent the variables in $\{1, \ldots, p\}$ that are observed for observation i. The variable Z_i, $i = n + 1, \ldots, n + n_s$, is updated from a categorical distribution with

$$P(Z_i = h | Y_i, \pi, \phi) = \frac{\pi_h \prod_{j \in J_i} \phi_{hY_{ij}}^{(j)}}{\sum_{k=1}^{N} \pi_k \prod_{j \in J_i} \phi_{kY_{ij}}^{(j)}} \tag{A.2}$$

for $h = 1, \ldots, N$.

Full conditional for $\{\phi^{(j)}\}$. Let J_i be defined as above for $i = n + 1, \ldots, n + n_s$, and define $J_1 = \cdots = J_n = \{1, \ldots, p\}$. To update $\phi_h^{(j)}$, for $h = 1, \ldots, N$, and $j = 1, \ldots, p$, sample from a Dirichlet distribution. The full conditional $P(\phi_h^{(j)} | Y^{(\text{mis})}, \text{data}, Z)$ is proportional to

$$\text{Dirichlet}\Bigg(\phi_h^{(j)}; a_1 + \sum_{\substack{\{i: Z_i = h,\\ j \in J_i\}}} 1(Y_{ij} = 1), \ldots, a_{d_j} + \sum_{\substack{\{i: Z_i = h,\\ j \in J_i\}}} 1(Y_{ij} = d_j)\Bigg), \tag{A.3}$$

where the summations are over all survey and glue cases, $i \in \{1, \ldots, n + n_s\}$.

Full conditional for V. The stick-breaking proportions V_h, for $h = 1, \ldots, N - 1$, can be sampled from Beta distributions:

$$P(V_h | \alpha, Z) \propto \text{beta}\Bigg(V_h; M_h + 1, \alpha + \sum_{j=h+1}^{N} M_j\Bigg), \tag{A.4}$$

where $M_h = \sum_{i=1}^{n+n_s} 1(Z_i = h)$. Fixing $V_N = 1$, the probabilities π are given by $\pi_1 = V_1$ and $\pi_h = V_h \prod_{j=1}^{h-1}(1 - V_j)$ for $h = 1, \ldots, N$.

Full conditional for α. The DP precision parameter α can be sampled from a Gamma distribution:

$$P(\alpha | V) \propto \text{gamma}\big(\alpha; N + a_\alpha - 1, b_\alpha - \log(\pi_N)\big). \tag{A.5}$$

Imputing $Y^{(\text{mis})}$. Missing Y_{ij} in D_1 and D_2 can be imputed by sampling from categorical distributions with the form given in equation (3.1).

Acknowledgments. We thank members of the working group from the SAMSI program on Computational Methods in the Social Sciences, especially Nicole Dalzell, Elena Erosheva, Monika Hu, Tracy Schifeling, Joe Sedransk and Aleksandra Slavkovic. We also thank David Boyle and Zach Sharek for providing the data from HarperCollins and CivicScience. Finally, we thank the Associate Editor and two anonymous reviewers whose comments were extremely constructive and helpful in improving the paper.

SUPPLEMENTARY MATERIAL

Model checking and MCMC diagnostics (DOI: 10.1214/16-AOAS925SUPP; .pdf). Model goodness-of-fit checks to the HarperCollins and CivicScience data and MCMC convergence diagnostics results.

REFERENCES

D'ORAZIO, M., DI ZIO, M. and SCANU, M. (2006). *Statistical Matching*: *Theory and Practice*. Wiley, Chichester. MR2268833

D'ORAZIO, M., DI ZIO, M. and SCANU, M. (2002). Statistical matching and official statistics. *Rivista di Statistica Ufficiale* **1** 5–24.

DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.* **104** 1042–1051. MR2562004

FOSDICK, B., DEYOREO, M. and REITER, J. (2016). Supplement to "Categorical data fusion using auxiliary information." DOI:10.1214/16-AOAS925SUPP.

GIBBS, A. and SU, F. (2002). On choosing and bounding probability metrics. *Int. Stat. Rev.* **70** 419–435.

GILULA, Z. and MCCULLOCH, R. (2013). Multi level categorical data fusion using partially fused data. *Quantitative Marketing and Economics* **11** 353–377.

GILULA, Z., MCCULLOCH, R. and ROSSI, P. (2006). A direct approach to data fusion. *Journal of Marketing Research* **43** 73–83.

GOODMAN, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika* **61** 215–231. MR0370936

ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96** 161–173. MR1952729

ISHWARAN, H. and ZAREPOUR, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87** 371–390. MR1782485

KADANE, J. B. (2001). Some statistical problems in merging data files. *Journal of Official Statistics* **17** 423–433.

KAMAKURA, W. and WEDEL, M. (1997). Statistical data fusion for cross tabulation. *Journal of Marketing Research* **34** 485–498.

KAMAKURA, W., WEDEL, M., DE ROSA, F. and MAZZON, J. A. (2003). Cross-selling through database marketing: A mixed data factor analyzer for data augmentation and prediction. *International Journal of Research in Marketing* **20** 45–65.

KIESL, H. and RÄSSLER, S. (2006). How valid can data fusion be? IAB Discussion Paper, 15.

MORIARITY, C. and SCHEUREN, F. (2003). A note on Rubin's statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econom. Statist.* **21** 65–73. MR1973805

MORIARTY, C. and SCHEUREN, F. (2001). Statistical matching: A paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics* **17** 407–422.

POLLARD (2002). *A User's Guide to Measure Theoretic Probability*. Cambridge Univ. Press, Cambridge.

RÄSSLER, S. (2002). *Statistical Matching*: *A Frequentist Theory*, *Practical Applications*, *and Alternative Bayesian Approaches*. *Lecture Notes in Statistics* **168** 60–63. Springer, New York. MR1996879

RÄSSLER, S. (2004). Data fusion: Identification problems, validity, and multiple imputation. *Austrian Journal of Statistics* **33** 153–171.

REITER, J. P. (2012). Bayesian finite population imputation for data fusion. *Statist. Sinica* **22** 795–811. MR2954362

RODGERS, W. L. (1994). An evaluation of statistical matching. *J. Bus. Econom. Statist.* **2** 91–102.

RUBIN, D. B. (1976). Inference and missing data. *Biometrika* **63** 581–592. MR0455196

RUBIN, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *J. Bus. Econom. Statist.* **4** 87–94.

RUBIN, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. MR0899519

SCHIFELING, T. A. and REITER, J. P. (2016). Incorporating marginal prior information in latent class models. *Bayesian Anal.* **11** 499–518. MR3472000

SETHURAMAN, J. (1994). A constructive definition of Dirichlet priors. *Statist. Sinica* **4** 639–650. MR1309433

SI, Y. and REITER, J. P. (2013). Nonparametric Bayesian multiple imputation for incomplete categorical variables in large-scale assessment surveys. *Journal of Educational and Behavioral Statistics* **38** 499–521.

VAN HATTUM, P. and HOIJTINK, H. (2008). The proof of the pudding is in the eating. Data fusion: An application in marketing. *Journal of Database Marketing & Customer Strategy Management* **15** 267–284.

VAN DER PUTTEN, P., KOK, J. N. and GUPTA, A. (2002). Data fusion through statistical matching. Working paper 4342-02. MIT Sloan School of Management, Cambridge, MA.

VERMUNT, J., GINKEL, J., DER ARK, L. and SIJTSMA, K. (2008). Multiple imputation of incomplete categorical data using latent class analysis. *Sociological Methodology* **38** 369–397.

WICKEN, G. and ELMS, S. (2009). Demystifying data fusion—The "why?", the "how?" and the "wow!" Technical report, Advertising Research Foundation Week of Workshops, New York.

B. K. FOSDICK
DEPARTMENT OF STATISTICS
COLORADO STATE UNIVERSITY
102 STATISTICS BUILDING
FORT COLLINS, COLORADO 80523-1877
USA
E-MAIL: bailey@stat.colostate.edu

M. DEYOREO
J. P. REITER
DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY
BOX 90251
DURHAM, NORTH CAROLINA 27708-0251
USA