Confidentiality Issue in Political Opinion Microdata: Bayesian Synthesis, Data Utility Analysis, & Disclosure Risk Evaluation

Yitong Wu

I. The Research Questions

Datasets consisting of microdata are subject to various confidentiality issues, and political opinion survey results is no exception. For purpose of analysis, research agencies often times collect respondents' preference in political party and ideology with their demographical and geographical information. Categories such as one's age, gender, and race expose survey respondents to potential identification and attribute disclosure risks. With knowledge of an individual's demographical information, the intruders could identify their political ideology, voting preferences, party affiliation, and stances upon certain socials issues and public policies, all of which could be private for some people. Therefore, in order to protect the confidentiality of the dataset, research agencies should address the potential disclosure risks in a dataset before its release.

On the other hand, data synthesis also faces the trade-off between utility and disclosure risks. Utility in a dataset lies in its ability to capture the true distribution of certain variables or the true relationship between different ones. However, as the process of synthesis inevitably changes the dataset, distributions and relationships would also experience distortions to a certain extent. In the context of a political opinion survey, researchers are mostly interested in the correlations between one's demographics and their opinions, some of which could be lost in the

synthesized dataset. That said, it is important for the statistical agency to decide which features to preserve before they synthesize the dataset. This paper explores issues like disclosure risks, synthesis methods, and data utility with a random sample of the AP VoteCast dataset.

II. Background & Significance of the Research

The motivation behind this research comes from the researcher's concern for a political opinion survey conducted within the Vassar student body. Due to the small size of the student body and the massive scale the questions have covered, several individuals could be easily identified along with their political affiliations. This kind of information could be particularly sensitive on a campus where the majority of the students are affiliated with one party than the other. Driven by this concern, the research hopes to explore the confidentiality issues in political datasets in general and narrows down to the AP VoteCast for a more in-depth analysis. Since the released AP VoteCast data has already been synthesized, the purpose of this research is not to address the confidentiality concern of this specific dataset but to rather provide an example of handling this kind of datasets in terms of data synthesis, utility measure, and disclosure risk evaluation.

III. Dataset Descriptions

AP VoteCast is a survey of the American electorate conducted in all 50 states by NORC at the University of Chicago for *The Associated Press* and *Fox News*. The survey of 138,929 registered voters was conducted from October 29 to November 6, 2018, concluding as polls closed on Election Day. Interviews were conducted via phone and web, with 11,059 completing by phone and 127,870 completing by web. AP VoteCast selected its respondents through a random sample of registered voters drawn from state voter files. The released AP VoteCast has

already been synthesized, and unfortunately the researcher does not have access to the original survey results. That said, this research only explores the remaining confidentiality issues in the synthesized dataset.

The AP VoteCast survey contains more than 200 questions, and the finalized dataset contains 220 variables with 138,929 observations in total. As this research only intends to showcase possible synthesis methods and risk and unity evaluation techniques, and is also greatly limited in the computation capacity, a random sample of 1,000 observations is drawn from the dataset for synthesis and evaluations. Furthermore, most of the questions concern one's attitude towards a specific social issue, public policy, or candidate, the answers of which are highly correlated with one's political ideology and party affiliation. These are also the questions with the highest non-response rate. Therefore, this research only selects the following variables for synthesis and analysis: state, age, race, sex, education, income, political ideology, and party preference. The resultant dataset consists of 1,000 observations and 8 columns.

IV. Synthesis Steps & Models

This research adopts sequential Bayesian synthesis as its technique, in which various Bayesian models are used to synthesize different variables. The basic idea is to create a universe made up of different people yet upholding the same relationship between one's demographics and their political attitudes. The main approach is to first independently synthesize the biological features including race, sex, and age. Other demographical variables like education and income are then synthesized in relation to the biological ones. Lastly, political ideology and party preference are synthesized based on all other synthesized variables in the dataset.

Below are the specific models used to synthesize the corresponding variables, with the

order of the list following that of the sequential synthesis.

① **Sex**

$$Y_i \sim Bernoulli(p)$$

$$p \sim Normal(0.5,1)$$

$$0 \leq p \leq 1$$

② **Race**

$$Y_i \sim Multinomial(n = 1, p_{i,White}, p_{i,African}, p_{i,Hispanic}, p_{i,Asian}, p_{i,Other})$$

$$p_{i,White}, p_{i,African}, p_{i,Hispanic}, p_{i,Asian}, p_{i,Other} \sim Dirichlet(\alpha_{White}, \alpha_{African}, \alpha_{Hispanic}, \alpha_{Asian}, \alpha_{Other})$$

$$\alpha_{White} = \alpha_{African} = \alpha_{Hispanic} = \alpha_{Asian} = \alpha_{Other} = 1$$

③ **Age**

$$Y_i \sim Multinomial(n = 1, p_{i,18-24}, p_{i,25-29}, p_{i,30-39}, p_{i,40-49}, p_{i,50-64}, p_{i,65+})$$

$$p_{i,18-24}, p_{i,25-29}, p_{i,30-39}, p_{i,40-49}, p_{i,50-64}, p_{i,65+} \sim Dirichlet(\alpha_{18-24}, \alpha_{25-29}, \alpha_{30-39}, \alpha_{40-49}, \alpha_{50-64}, \alpha_{65+})$$

$$\alpha_{18-24} = \alpha_{25-29} = \alpha_{30-39} = \alpha_{40-49} = \alpha_{50-64} = \alpha_{65+} = 1$$

④ **Education**

$$Y_i \sim Multinomial(n = 1, p_{i,HighSchool}, p_{i,SomeCollege}, p_{i,College}, p_{i,Postgraduate})$$

$$log(\frac{p_{ic}}{p_{i1}}) = \beta_{0,c} + \beta_{1,c} Race + \beta_{2,c} Age + \beta_{3,c} Sex$$

*Note: Race, age, and sex take on binary variables for different categories in application; the model is written here with simplification.

⑤ **Income**

$$Y_i \sim Multinomial(n = 1, p_{i,\leq \$25,000}, p_{i,\$25,000-\$49,999}, p_{i,\$50,000-\$74,999}, p_{i,\$75,000-\$99,999}, p_{i,\geq \$100,000})$$

$$log(\frac{p_{ic}}{p_{i1}}) = \beta_{0,c} + \beta_{1,c}Race + \beta_{2,c}Age + \beta_{3,c}Sex + \beta_{4,c}Education$$

*Note: Race, age, sex, and education take on binary variables for different categories in application.

⑥ Political Ideology

$$Y_i \sim Multinomial(n = 1, p_{i,VConservative}, p_{i,SConservative}, p_{i,Moderate}, p_{i,SLiberal}, p_{i,VLiberal})$$

$$log(\frac{p_{ic}}{p_{i1}}) = \beta_{0,c} + \beta_{1,c}Race + \beta_{2,c}Age + \beta_{3,c}Sex + \beta_{4,c}Education + \beta_{5,c}Income$$

*Note: Race, age, sex, education, and income take on binary variables for different categories in application.

Utility