# Beyond encryption: Protecting consumer privacy while keeping survey results accurate

April 17 2020, by Niki Gianakaris

Credit: CC0 Public Domain

It comes as no surprise that consumer data is continuously being collected by various organizations, including local governments, marketing agencies and social media companies. These organizations assure anonymity and confidentiality when collecting this data, however, existing data privacy laws don't guarantee that data breaches won't occur.

According to a recent report, more than 2,000 confirmed data breaches occurred in 2019 alone, with 34% of those executed by internal actors such as employees. To add to that, city and state agencies collect sensitive data that they are required by law to share with the public—courtesy of Open Data movements and the Freedom of Information Act.

Data privacy laws require encryption and, in some cases, transforming the original data to "protected data" before it's released to external parties. But for researchers like Matthew Schneider, Ph.D., an assistant professor of Decision Sciences and Management Information Systems at Drexel University's LeBow College of Business, this isn't adequate.

"Encryption definitely helps, but it does not prevent a data breach," he said. "It's similar to safeguarding your email password. An internal actor with access to the encryption key could easily cause a data breach. It's more conservative from a risk perspective to assume that all data will eventually get out and should be transformed prior to sharing anywhere within the organization."

In a recent paper published in the *Journal of Marketing Analytics*, Schneider and Dawn Iacobucci, Ph.D., of Vanderbilt University, proposed a new methodology that permanently alters survey datasets to protect consumers' privacy —when data is shared— while still preserving a level of reasonable accuracy for these datasets.

According to the authors, survey data is often held within organizations and used for purposes beyond the original reason for collecting the data. "Databases and customer information have become a contemporary asset that makes one business attractive to another when forging alliances," Schneider said. "Even firms with high standards of data security can find it challenging to protect the privacy of consumer data."

Another less common, but all-too-real, threat, according to the authors, are cases where employees have illegally taken data from their former companies to a position with a new employer—for reasons ranging from gaining a favorable impression with the new company, to harming the old company, to even having to provide the data as a condition of the job offer.

For Schneider, the solution to fulfilling data privacy promises turns out to be a technological one.

"Survey data are increasingly used for respondent-level analytics, such as in linkage to other proprietary datasets, and promises of privacy may not be guaranteed in the myriad of subsequent uses of the data," said Schneider. "Confidentiality does not guarantee anonymity. It takes about three or four carefully posed questions in a survey to uniquely identify anyone."

In the paper, the authors analyzed a survey data set that was collected in 2015 by the city of Austin, Texas and released to the public following an Open Data movement. Other cities have similar movements, including New York and Philadelphia.

"There are lots of privacy risks in Open Data since they don't do privacy as well as the federal government that has the large budget and resources to hire statisticians, economists or computer scientists to address this technological problem," said Schneider. "Protection often depends on how the data is used."

The city of Austin administered a survey to 2,614 Asian Americans living in the city to explore the health and service needs of one of the city's fastest growing populations aiming to create higher levels of community engagement, policies and to identify resources to address the needs of the Asian American community. Officials in Austin posted

their data sets, as required, to make them readily available for users.

In one survey dataset, each respondent was asked their ethnic origin, which had 32 categories; age, which had 77 categories; zip code, which had 61 categories; and gender.

"Nearly everyone is identifiable with these four variables—some more so than others," said Schneider. "Once you identify them, this survey revealed other sensitive responses such as employment status, religious affiliation, household income, housing affordability and many attitudinal questions. "

Similarly, New York City experienced an Open Data problem with the New York City Taxi and Limousine Commission where 124 million driving routes could be traced to a driver's home address.

One major challenge when considering methodologies to alter participant data effectively is to do this in a way that doesn't greatly change the accuracy of the survey results. The methodology proposed by the authors, was built upon a technique found in genomic sequencing applications that was able to disguise the identity of consumers while maintaining the accuracy of insights within 5%.

"Our method would essentially 'shuffle' the demographic data in a survey dataset," said Schneider. "But, unlike previous methods, ours only shuffles data when it maintains the correlations between important variables that are essential to analysts. The protected data is simulated on a consumer level but still valuable to the end user. If this dataset got out, then only the organization's insights would be known."

The paper, "Protecting Survey Data on a Consumer Level," was published in the *Journal of Marketing Analytics* and is available at this link. Details about the new methodology are included in the paper.

**More information:** Matthew J. Schneider et al, Protecting survey data on a consumer level, *Journal of Marketing Analytics* (2020). DOI: 10.1057/s41270-020-00068-6