Protecting identities of panelists in market research

October 10 2022, by Tom Fleischman

Credit: Pixabay/CC0 Public Domain

News alert: Just because a marketing research company tells survey participants that their personal information will remain anonymous doesn't mean it's true.

No, this is not a big secret. But it's not just possible that personal information could be compromised: According to research by a Cornell SC Johnson College of Business professor and colleagues, it's highly likely that a survey participant's identity and other sensitive information can, in fact, be traced back to the individual.

"When organizations release or share data, they are complying with privacy regulations, which means that they're suppressing or anonymizing personally identifiable information," said Sachin Gupta, Ph.D., the Henrietta Johnson Louis Professor of Management in the Samuel Curtis Johnson Graduate School of Management, in the SC Johnson College.

"And they think that they have now protected the privacy of the individuals that they're sharing the data about," he said. "But that, in fact, may not be true, because data can always be linked with other data."

Nearly all market research panel participants are at risk of becoming de-anonymized, Gupta and colleagues say in a new paper, "Reidentification Risk in Panel Data: Protecting for k-Anonymity," published Oct. 7 in *Information Systems Research*.

Co-authors are Matthew Schneider, M.S., Ph.D., associate professor of decision sciences and management information systems at Drexel University; Yan Yu, Ph.D., the Joseph S. Stern Professor of Business Analytics at the University of Cincinnati; and Shaobo Li, assistant professor at the University of Kansas School of Business.

It's no secret that personal data—name, date of birth, email address and other identifiers—are floating in the ether, ripe for the taking by a highly motivated person or company. This has been proven countless times; Gupta and colleagues referenced a 2008 paper by a pair of researchers from the University of Texas, Austin, who developed a de-anonymization algorithm, Scoreboard-RH, that was able to identify up to 99% of Netflix subscribers by using anonymized information from a 2006 competition, aimed at improving its recommendation service, coupled with publicly available info on Internet Movie Database.

That research, as well as Gupta's, relies on "quasi-identifiers" or QIDs, which are attributes that are common in both an anonymized dataset and a publicly available dataset, which can be used to link them. The conventional measure of disclosure risk, termed unicity, is the proportion of individuals with unique QIDs in a given dataset; k-anonymity is a popular data privacy model aimed to protect against disclosure risk by reducing the degree of uniqueness of QIDs (i.e., any individual's QID information should be the same as at least k-1 other's QID information).

"Unicity was developed for cross-sectional data, where you have one observation per individual," Gupta said. "But in many of these datasets, you have longitudinal data—the same individual is observed over time. And now the reidentification risk changes, because of the availability of multiple observations."

Gupta and his colleagues have developed what they term "sno-unicity"—as in snowballing unicity—which is basically the worst-case-scenario reidentification risk, as it iteratively collects individuals who can be uniquely reidentified by at least one of their multiple records.

In their research, Gupta and colleagues studied market research data across 15 frequently bought consumer goods categories, as well as physician prescription-writing. They found that based on unicity alone (just one observation per panelist), the reidentification risk in panel data is very high—up to 64% for carbonated beverage purchases, for example.

However, when employing sno-unicity (multiple observations per panelist), that number soars to 94%, and is higher in all 15 categories. In other words, people's data isn't as secure as marketing researchers might have them believe. "We demonstrate," Gupta said, "that the risk of reidentification in such data is vastly understated by the conventional unicity measure."

An example of the risk: The researchers' analysis found that among households that were re-identifiable based on their purchases of salty snacks in a given store, 20% bought beer and 2% bought cigarettes from a different store. Even if this information is never used, just the fact that it can be obtained is a compromise of data privacy.

The researchers' new approach, called graph-based minimum movement k-anonymization (k-MM), was especially designed to preserve the usefulness of panel data with minimal loss of information. Distortion is used to protect panelists' identities—by slightly modifying a panelist's brand choices, for example—but it negatively affects the value of the data.

"The consumers of this panel data are paying for this information, so we don't want to lose too much of it," Gupta said. "And yet, we want to protect privacy, so you want to find that point on the curve where you are guaranteeing some threshold of privacy—in our case, k-anonymity—while minimizing information loss."

Although privacy laws are being enacted in the U.S. and elsewhere that will make it tougher for information to be nefariously obtained, Gupta said this research is still vital. Market researchers will still collect and store data, meaning protecting privacy will still be a challenge.

"The nature of the problem will probably reduce and change," he said, "but I don't think it's going away."

More information: Shaobo Li et al, Reidentification Risk in Panel Data: Protecting for k-Anonymity, *Information Systems Research* (2022). DOI: 10.1287/isre.2022.1169

Provided by Cornell University

Citation: Protecting identities of panelists in market research (2022, October 10) retrieved 23 April 2024 from https://techxplore.com/news/2022-10-identities-panelists.html