



© 1997–2004, Millennium Mathematics Project, University of Cambridge.

Permission is granted to print and copy this page on paper for non-commercial use. For other uses, including electronic redistribution, please contact us.

September 2002

News

Random privacy



How old are you? How much do you earn?

What would you answer if asked these questions at website when you were buying your next TV or ordering groceries online? A lot of us would lie, and for a very good reason – to protect our privacy.

But the companies posing these questions also think they have a good reason. Information about customer profiles is becoming increasingly important in business, both for marketing and for development and improvement of services.



Time to tell the truth? Image from www.freeimages.co.uk

"Right now, the rate of falsification on Web surveys is extremely high," says Dr Ann Coavoukian, the commissioner of information and privacy in Ontario, U.S.A. "People are lying and vendors don't know what is false [or what is] accurate, so the information is useless."

But researchers at IBM think they have the solution. They have developed an ingenious method to protect our privacy, while still giving companies the information they crave. Rakesh Agrawal and Ramakrishnan Srikant, computer scientists working for IBM in California, realised that companies are often interested in aggregate

Random privacy

data, rather than in being able to track individual responses. The two researchers have developed software that never records original answers, instead randomising them in such a way that companies can reconstruct the information they need from all the responses.

For example, if you are asked your age, instead of recording exactly what you entered, the software records your answer plus or minus some randomly generated number within a known range, say ± 20 years. So if you enter 28, it might record anything between 8 and 48, with this process of randomization occurring independently for each response. Since your original answer and the random number masking it are never recorded, there is no way to reconstruct information about, say your next birthday.

But what remains is still useful for companies. They know not only what was recorded for all the responses, but also the probability distribution used to generate the random numbers masking the answers. This is not enough to reconstruct individual answers, but enough to deduce something about the whole group that responded. Using a version of Bayes Theorem applied to probability distributions, they can deduce the distribution of answers over the whole group. For example, they can get good estimates of how many respondent were aged 15–24, 25–34, and so on. (See [Ye Banks and Bayes](#) in issue 9 of *Plus*, or [Beyond reasonable doubt](#) in this issue for more on Bayes theorem.)

The inaccuracy resulting from this method may be a small price to pay if it persuades customers to be more truthful in their responses. "The beauty of this research is that retailers and other Web businesses are able to extract the valuable demographic information they need without necessarily knowing the underlying personal consumer data", says Harriet P. Pearson, IBM's Chief Privacy Officer.

Companies using such privacy software may have a real advantage over competitors, as many web users say that privacy is a major concern. However, for the system to work companies are going to have to convince consumers to trust them enough to stop fibbing about their answers.

Rachel Thomas



Plus is part of the family of activities in the Millennium Mathematics Project, which also includes the [NRICH](#) and [MOTIVATE](#) sites.