Personal data privacy

Krumm, John [1]

[1] *Université de Californie du Sud, US*

> ABSTRACT
>
> *Conférence de John Krumm, chercheur en résidence d'écriture, dans le cadre du cycle "Paris IAS Ideas"*

A person may willingly reveal their age, gender, and home city to a third party company. However, this same person may be uncomfortable with the inferences that come from this data, such as their income, political preferences, income, and education level.

Regular people should understand what can be inferred about them from revelations of seemingly innocuous personal data. There is likely a simple, underlying theoretical foundation that makes this clear. For instance, we can look at data to understand how a person's age can be used to infer their income. This is based on a simple joint probability distribution of age and income. In theory, there is a larger joint distribution, often approximated with a deep neural network, that gives the probabilistic relationship between tens of different, personal variables. Examining a distribution like this will help understand:

- What can be inferred from a revelation of a few personal details?

- How does this change if a person gives fuzzy answers, such as an age range instead of a specific age?

- How are the inferences affected if a person lies about their personal data? What are the best lies in order to confuse the inference?

- Some inferences are more sensitive than others, e.g. an inference about income may be more sensitive than an inference about sports teams preferences. How does the

sensitivity of the inferences affect the sensitivity of the personal data revelations?

- A company receiving personal data will likely not reveal the trade secrets of what it can infer. How can an individual still make smart choices about what to reveal given the individual's uncertainty about what else the company can infer?

There is a simple, underlying theory that can answer these questions, using joint probabilities, probabilistic expectation, and possibly information theory. This can be easy to illustrate with census data, which will demonstrate answers to the questions above. This project can lead to practical, real-life guidelines on the consequences and best practices for revealing personal data.

Personal data privacy

Bibliography

Banovic, N., & Krumm, J. (2018). Warming Up to Cold Start Personalization. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 1, 4*, 1–13.

Krumm, J. (2022). Sensitivity Analysis of Personal Location Disclosure. *2022 23rd IEEE International Conference on Mobile Data Management (MDM*, 73–82.

Krumm, J. (2022). Maximum entropy bridgelets for trajectory completion. *Proceedings of the 30th International Conference on Advances in Geographic Information Systems*, 1–8.