Releasing data as a risk management problem

In this presentation, I present the motivations for reframing privacy-preserving data publishing as a risk management problem, and describe the concept of *linkability*—the basis for our novel threat modelling approach. I then go on to discuss our learnings from running Device Analyzer, a popular smartphone-based data collection project which provided data to researchers, as well as broader lessons drawn from a wide-ranging investigation of the current data publishing landscape. Finally, I discuss how to design and build data collection and sharing systems that address these identified risks, and present possible future architectures to handle sharing of rich data streams from pervasive and ubiquitous systems.

Risk, not quantitative metrics

Releasing datasets that contain information on real people presents difficult challenges. Researchers are required—by convention and by law—to preserve the privacy of our data subjects, to make them as "anonymous" as possible or to minimise the data that is included about them.

The field of Privacy-Preserving Data Publishing has presented many techniques for tackling these challenges that are now well known—such as k-anonymity, differential privacy, and randomised response. Each of these techniques presents a technical, mathematical approach to quantifying the expected potential information gain on data subjects from a dataset, and attempts to bound this gain.

There are two problems with this approach. First, these techniques often require the application of noise, and so reduce the specificity of data. While this may not affect aggregate analysis, it hampers analysis in which we want to pick out particular subjects, such as health record analysis. Second is the spectre of **side information**. Threat models adopted in current approaches to data publishing rarely make a detailed account of the potential for an adversary to use some external dataset in efforts to infer information about a subject; this is usually termed *linking* or *joining* datasets.

Threat modelling the impossible

The difficulty in modelling side information attacks is in anticipating what side information is available, and the nature of the attacker with access to it. In our work, we present a new way to model threats due to these *linkage attacks*. Because the possible space of linkage-based threats is so large and difficult to anticipate, our approach attempts to first narrow this space using qualitative analysis.

A rational approach to linkage attacks must first rely on an evaluation that identifies where tradeoffs between privacy risk and utility could lie. Considering the maximal case of side information would entail locking down data as much as possible, constricting utility, but this may be irrational as the maximal dataset might only be accessible by a nation-state level attacker. Thus, we present the notion of *linkability*, a qualitative-quantitative description of risk, which relies on first performing a guided qualitative assessment of potential threats, and then a quantitative risk assessment for each identified threat.

Alternatives to technical mitigation

By reframing data protection as a risk mitigation exercise, we see that technical protections are only one tool in our risk mitigation toolbox. From our experience running a data collection project and a further study of the data publishing landscape, we show that legal and procedural techniques can act as equal citizens to technical ones, allowing us to make tradeoffs not previously possible.

In the context of data collection from mobile systems, this enables us to present a design for a "Device Analyzer 2.0", in which rich continuous data can be responsibly made available for research purposes while achieving a better privacy-utility tradeoff than was possible with our original system. By bringing researchers' code to the data, rather than vice versa, we argue that a lower barrier to entry can be achieved, with stronger guarantees about subject privacy—by leaning on procedural techniques, both privacy **and** utility can be strengthened.

Jovan Powar (jsp50@cam.ac.uk) & Alastair Beresford (arb33@cam.ac.uk), Department of Computer Science and Technology, University of Cambridge