UNIVERSITY OF CAMBRIDGE

First Year Report

Research as a Service

A tunable and adaptive platform for privacy-preserving data sharing

PhD proposal

Jovan Powar

Trinity College

October 2018

First year report submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

Contents

1 Introduction **2**

2 Summary of first year projects **4**

2.1 ADINT 4

2.1.1 Threat modelling experiments 5

2.1.2 Business vs. ethics 6

2.2 Device Analyzer 7

2.2.1 How the dataset is managed 7

2.2.2 Parquet 8

2.2.3 Lessons for the next generation of research data sharing 9

2.3 Research-as-a-Service: proposing a code-to-data platform 11

2.3.1 Thinking broadly about data sharing models 12

2.4 Other projects 13

2.4.1 Teaching 13

2.4.2 IVIR summer course on Privacy Law and Policy 13

3 SoK: Sharing of sensitive datasets **14**

3.1 Aims 14

3.2 Survey methodology 15

3.2.1 Datasets vs. Platforms 15

3.2.2 Defining personal and sensitive data 16

3.2.3 Requirements for dataset consideration 18

3.3 Classification methodology 18

3.4 Evaluating linkability by constructing attacks 23

3.4.1 Linkability is qualitative, not quantitative 23

3.4.2 A qualitative attack-based survey 24

4 Thesis plan **29**

4.1 Research question 29

4.2 Chapter One 30
4.3 Chapter Two 31
4.4 Chapter Three 31
4.5 Chapter Four 32
4.6 Timeline 33

A ADINT **37**
A.1 Investigation plan 37
A.1.1 Outline 37
A.1.2 Background 37
A.1.3 Methodology 39
A.1.4 Hypotheses 40
A.1.5 Risks of data collection and precautions to be taken 41
A.2 Ethics committee application 41
A.2.1 Description 42
A.2.2 Precautions 43

B SoK **44**
B.1 Dataset features 44
B.2 Initial survey 50

Chapter 1

Introduction

A significant focus of both academic and public interest recently has been the privacy of big data analytics. The ubiquity of sensor-packed smartphones and internet tracking techniques has enabled the collection of increasing quantities of data on users. This is enabling remarkable insights and the development of powerful new technologies, but has also raised widespread concern over a person's control—or simply awareness—of how their data is being used. 'Big Data' today is established on a notion of a trade-off between the utility of data and the privacy preserved for the subject. Put simply: to achieve insight and new features, the developers must have as unrestricted a view of the data as possible.

In academia, there has been a significant body of work attempting to enable **privacy-preserving** data analytics and publishing. Techniques such as differential privacy [1, 2, 3, 4] and randomised response [5] allow people to contribute data under some mathematically constructed guarantees of anonymity. (Wang et al. [6] present a general overview of anonymisation techniques.) However, these techniques are hard to apply outside theory, and choosing how and where to apply them remains an open problem. The key issue is that the context of any data collection exercise is hard to fully characterise. To adequately apply differential privacy, for example, we must choose our privacy budgets and noise parameters, which requires balancing the amount of noise added to data against its utility. In a vacuum, we can choose these parameters, but in real-world and continuous-data applications the risks associated with this trade-off are hard to properly quantify [7, 8, 9, 2]—it entails fully understanding correlations and relationships not only within a dataset, but with potential sources of side information. This problem has proven to be non-trivial even when we have side information to reason about! [10, 4]

Meanwhile, from the public perspective it appears (often long after the fact) that any and all data collected on us is being swiftly proliferated—traded, linked to other datasets, and used for profiling capabilities, conferring technical [11] and social powers [12, 13]

heretofore unimagined.

This narrative has ignited a huge debate in governance: it is now clear that seemingly 'runaway' technical capabilities must be tackled not only technically but through procedure and regulation. As such, future work in privacy must take a holistic view to tackle the challenges that data-driven applications now pose:

- Giving data subjects control of and visibility over how their data is used;
- balancing the utility of 'Big Data' against the amount of information recoverable from the dataset on any given subject;
- and understanding what the risks to privacy are, given the huge dimensionality of—and heterogeneity of access to—side information.

In this report I outline the reading and projects I undertook with the aim to better understand these challenges, and the beginnings of my work in the area of planning and building systems to enable data sharing with strong privacy guarantees.

My first year of work has concentrated on two examples of data sharing and privacy concerns: research and advertising. Through my involvement with the Device Analyzer project, I formed an understanding of the benefits and processes of a research data collection and publishing project. I also gained a perspective on the challenges in providing utility to third party researchers and privacy to data subjects (challenges that other projects have notably failed to address).

Privacy concerns have been particularly topical recently in advertising, and this gave me a valuable perspective on the deanonymisation capabilities available in industry today, and the capabilities of real-world adversaries, bad actors, or (more commonly) simply third parties with vastly differing incentives to data subjects.

This report will present the work I have begun on systematising knowledge on data sharing in research and my efforts to properly characterise the salient risks of handling potentially sensitive datasets. The aim of this work is to produce a process for designing a system to handle sharing of sensitive datasets that uses technical, procedural, and legal mitigations to address the wide range of risks encountered. The ultimate aim of my PhD will then be to produce such a system, or a basis for a class of such systems, and evaluate its application to real research problems.

Chapter 2

Summary of first year projects

2.1 ADINT

Privacy concerns in the advertising industry have been a focus of academic research for some time. The techniques that advertising companies use to track web users have been and continue to be well documented, as does the structure of the ecosystem [14, 15, 16, 17, 18, 11], and a large industry has emerged around constructing user profiles and selling analysis and targeting tools. However, investigating privacy in the advertising industry is also an economic problem. To understand the real privacy risks, it has become necessary to investigate the features and usability of the systems that data companies (such as Facebook or Google) provide [19, 20, 21].

A recent work by Vines et al. [22] showed that the purchasing of mobile ads could be used to track the location of a user on whose device the ads were shown, without the need for dynamic content in the ad creatives, or a click-through on the ad. This class of attacks, dubbed ADINT, leverages the rich data collected on a user by advertising networks and assembled into a profile, and the many targeting options available to an advertiser—including the retargeting of specific user profiles.

The question I set out to investigate was: how realistic are these attacks? Are these toy examples that would be an infeasible process for anyone but a highly motivated adversary? Or does this show that user profiles can be exfiltrated from Google, Facebook et al. at a small cost? If the advertising identifier is unknown, unlike Vines et al. assume, are there other related attacks that would allow a malicious advertiser to construct a proxy profile, or extract the identifier?

The mobile advertising ecosystem

In the investigation proposal for my experiments (Appendix A.1), I described the mobile advertising ecosystem:

> Briefly, the mobile advertising ecosystem operates as follows: a **User** installs an app or visits a page provided by a **Publisher**. This Publisher embeds in their app or website an ad library, provided by a **Service-Side Provider (SSP)**. The role of the SSP is to manage Publishers and auction off this advertising inventory to parties interested in advertising to the user. These auctions are contested by **Demand-Side Providers (DSPs)**, who manage and bid on behalf of **Advertisers**, the parties who provide the content for the ads.
>
> The SSP constructs profiles of Users, to allow targeted advertising, which is more valuable to an Advertiser. It does this by tracking user behaviour using browser cookies, or through a **Mobile Advertising Identifier (MAID)**. This is a unique and device-specific identifier present on iOS and Android devices, which provides a means to associate all (in-app, not including in-browser) behaviour on a device with the same profile—this cannot be done via cookies as apps will store cookies separately, or often not at all. In current versions of these operating systems, a User may reset this identifier to a new value.
>
> The DSP exposes targeting functionality to an Advertiser through a web interface or API. Upon a successful impression (an instance of the ad being loaded on a User's device), the DSP reports this to the Advertiser, along with information such as the MAID. The DSP may provide user information as forwarded by the SSP, or maintain its own profiles on users based on previous campaigns.

The nature of this ecosystem presents an interesting problem: where do we draw the line between business model and attack?

2.1.1 Threat modelling experiments

I outlined a set of experiments designed to investigate the attack surface of mobile advertising. My complete investigation plan is given in Appendix A.1. The hypotheses I intended to investigate were:

1. Ads can be reliably directed at a specific device or person, and only at that person.
2. Ads purchased through a DSP can be tailored—through targeting and bidding—such that they are shown frequently in natural use by regular users.
3. Fine localisation of an impression can be achieved using ad campaigns with overlapping location targeting.

2.1.2 Business vs. ethics

To begin, I started an account with a DSP, *Go2mobi*. Though the ad creatives uploaded to the exchanges passed all programmatic checks, shortly before the campaign was due to go live I received the following email from a *Go2mobi* representative:

> We are unsure if this type of campaign is allowed on our platform and our platform traffic source partners.
>
> For us to investigate, we would need to do a legal review and we are not a large enough company to be able to do this review in a timely manner. Unfortunately we will not be able to approve your campaign at this time.
>
> We will get in contact with you if we do a legal review and figure out if this type of campaign is acceptable on our platform.

This was a clear example of the difficulties I had expected to encounter in this study: as part of our ethical obligation to be transparent to the users whose data would be handled, our ad campaign explicitly stated "We're looking for participants to help with our research on privacy in the mobile ad ecosystem." However, once this was noticed by a person at the (relatively tiny) DSP I had chosen to use, my account was disabled for fear of retribution from the SSPs (e.g. Google). As a result, I moved my experiments to another DSP, *Splicky*, whose process was entirely automated.

A preliminary evaluation of Facebook's advertising platform showed similar functionality to that offered by the SSP-DSP model, but due to the integrated nature of the bidding and distribution platforms, the capabilities exposed to the advertiser change at a rapid pace—during my experiments, the audience estimation functionality was removed after an attack was published [19].

From inspection of APIs provided by SSPs Google AdX and Facebook, and related 'retargeting' services (Kochava), I saw that the possible capabilities a DSP can offer are very powerful—retargeting lists can be built (providing similar functionality to MAIDs) and the data collected on a user can be extremely specific (e.g. eye colour). However, to gain access to this functionality, one must use significantly more expensive DSPs than *Go2mobi* or *Splicky*, which usually establish a relationship with businesses through a designated salesperson. Given my experience with *Go2mobi*, I suspected that the overhead of establishing a business relationship would require more thought, and so I have paused work on this project.

Though a setback, this obstacle did provide insight into data brokerage markets: establishing a business arrangement with background checking and minimum spends raises the barrier to entry for the most powerful tools, and thus serves as a deterrent and shield from potential bad actors. However, this mechanism is also rather opaque—it is not clear

at this point what criteria the advertising companies look for to judge an advertiser safe to provide access to these tools, if indeed safety is a consideration in the onboarding process at all. Put simply, advertising companies put procedural protections in place, but this tells us little about the legality or privacy risks of the activity.

2.2 Device Analyzer

I have become involved in the Device Analyzer project here in the DTG, taking over third-party researcher onboarding. Device Analyzer is a project started by Daniel Wagner that offers volunteers an Android app, which logs a wide range of usage data from their phone. This data is then sent to us, where we produce a dataset and make it available under license to other researchers. The dataset has been used for research in a wide range of academic disciplines.

I performed some analysis on the dataset to gain an understanding of the practicalities of working with large, rich datasets, and the challenges faced by researchers attempting to share potentially sensitive datasets.

2.2.1 How the dataset is managed

The job of managing the Device Analyzer project is split into two parts: onboarding and supporting third-party researchers; and maintaining the infrastructure and mobile app.

The onboarding process for third parties is as follows:

1. Third parties email us requesting a sample of the dataset, with a brief description of their research project.

2. We verify that they seem to be a member in good standing at their stated research institution (usually a university but often an industrial research lab). As a general rule, we give access to PhD students or staff at an institution—we do not usually provide access for Masters students, as the lengthy legal process for full licensing would usually be prohibitive within the timeframe of their course, and we also require that third parties' work is performed with intent to publish the results.

3. We provide a sample of the dataset—the logs from around five devices—so that the third party may perform a preliminary investigation to see whether the full dataset would be of use to them.

4. If the third party wishes to apply for access to the full dataset, they must provide us with a license agreement signed by a representative of their institution (legally, licenses must be made between our University and their institution itself), along with a detailed research proposal for the intended work.

5. The license agreement is completed, with any amendments requested by either party. The agreement lists individual researchers who will be allowed to access the data, with the stipulation that the data will only be used by those persons and that no effort will be made to deanonymise or identify the subjects. This agreement is time-limited to three years.

Maintaining the mobile app is for mostly an exercise in stability. The app was first launched in 2010 and we have aim to maintain the same functionality; this is difficult to achieve given the Android upgrade cycle. The permission system has shifted priority from install-time prompts to run-time, which has caused bugs and will likely necessitate significant engineering work to keep Device Analyzer running on future versions of Android. The APIs that apps have access to have also changed and so functionality has been lost. For example, Device Analyzer can no longer read from the /proc directory, and so cannot log rich app usage data.

Earlier this year we pushed an update to Device Analyzer, adding calendar logging functionality. Written by Ricardo Mendes, I helped test the new version of the app and its integration with our server-side infrastructure. This entailed standing up a testing server and testing the incoming data to ensure schema compliance and to check that the increase in bandwidth needed was reasonable.

Logs are pre-processed on the subject's device, pseudonymising any identifiers by hashing them, such as MAC addresses, phone numbers, and WiFi SSIDs. On the server side, we store these in a database, and perform some more pre-processing. First, as described in the Device Analyzer paper [23] timestamps are corrected for consistency. This produces the dataset that members of the University can access. Then a second round of pre-processing is performed to further strip out location data (by hashing) and replace any app names with the category of the app. This pseudonymised dataset is provided to third party researchers. These pre-processed datasets are 'released' infrequently—until recently the dataset in use was produced in 2016. Pre-processing is very computationally expensive and takes days to complete.

2.2.2 Parquet

I attempted to address one particular pain point that I encountered when attempting to analyse the data: the speed of analysing data at such scale. Due to the flat-schema .csv format used by device logs, any job performed on the dataset had to serially iterate over every line of each log file, as a string. This meant that even the most basic aggregation queries had to be left overnight if performed on even a 200GB subset of the ∼18TB dataset.

At Martin Kleppman's recommendation, I investigated data formats used in existing big data applications. Of the commonly used schemas I judged Parquet (an Apache standard developed for Hadoop distributed architectures which Martin had recommended) to

be a good fit for Device Analyzer log files. Parquet uses a columnar format, where each column in a table is stored individually, with regular markers to allow quick row lookups. By storing each column individually, a different encryption or dictionary table can be used for each, saving on space and search time—numerous tools are available to query Parquet datastores with SQL. Translation of the .csv format to Parquet ran at roughly 1000 lines per second, and produced files of similar size to bzip2 compression, which is used on Device Analyzer's servers. Simple one-device SQL-based queries execute in times on the order of thousandths of a second, rather than previous Python scripts which would take times on the order of minutes.

Parquet was clearly a more flexible and easily analysed format for Device Analyzer logs, so I have created a workspace on the University's high performance computing cluster, to which I intend to migrate the Device Analyzer log database, perform translation into Parquet, and set up tools to efficiently query the whole dataset. This should significantly speed up common tasks performed by members of the DTG on the dataset.

2.2.3 Lessons for the next generation of research data sharing

The object of my involvement with the Device Analyzer project was to use it as a case study, using its history and my practical experience as a frame of reference for sharing sensitive data. By gaining insight into the successes, challenges, and failures of data sharing in research, I aimed to come away with a direction for a new model for research data sharing. Below I will summarise my key takeaways:

Applicant audit: Device Analyzer operates a strict audit process (Section 2.2.1) for third-party researchers looking to access the dataset. This allows us to ensure that whatever work third parties do with the dataset has been well thought through. However, it can be a lengthy process, usually taking around three months to complete the licensing process. This limits the number of projects that can make use of the dataset—students with less time, such as Masters students, or PhD candidates at a late stage in their project cannot get access in a timely fashion, no matter how simple the task they wish to perform. Additionally, relatively few of the sample datasets that I have provided to inquiring researchers have resulted in a full application for a license. Most of these cases are probably due to the third parties' inspection of the sample data not proving fruitful or applicable, but some applicants may have been turned off by the friction of the process—we do not know to what extent this may be the case.

Liability: The necessity of the license agreement is due to the sensitive nature of the data we collect. Applications that exploit such rich data have the capacity to do harm, and so the University assumes some liability in the case of Device Analyzer. The license agreement

is a legal measure to mitigate this liability. To weaken the barrier to entry of the dataset by weakening the license agreement would entail the legal question of how to manage the increased risk this creates.

Use cases: The utility of the dataset ended up being broader than the original authors had expected—instead of simply being a smartphone usage tool for computer scientists, the dataset has been of use in other academic fields.

Pre-processing: The process we use for pre-processing is rather expensive, and requires a large window of time for correcting timestamps. This means that we batch process the dataset in 'releases', which prevents researchers (internal and third parties) from using the most recent data. If recent data is required, a release must be generated, which takes on the order of a week to complete.

Scale of analysis: Analysing the data can also be extremely slow. Re-running code after a small tweak is an overnight job, and slows down development of an analysis and iterative investigation. In one recent case, another member of the DTG could not use a new result in a paper because the code, written days beforehand, would not finish before the deadline. This is even more of a problem for third parties, as downloading the dataset is a significant bandwidth requirement, and up until recently was made manageable using a download tool, *Picky*, which is no longer functional.

Establishing trust: The license agreement is our main way of establishing trust in a third party. After vetting them to convince ourselves that they are a member in good standing of a trustworthy institution, we have them legally pledge to preserve the anonymity of data subjects and to handle the data under appropriate security measures. However, we do no audit of the institution's security provisions (such as requiring Cyber Essentials compliance). Once the data is in the third party's hands, we trust that they will handle it securely and preserve subjects' privacy in perpetuity.

Android engineering challenges: The Android update cycle is yearly, often with significant changes to APIs that limit or change the basic functionality of the Device Analyzer app. Keeping up is a significant engineering challenge, and I expect we are approaching the point at which significant refactoring is needed to keep Device Analyzer running on new devices—probably with reduced utility.

With these considerations in mind, I began thinking of what the next generation of Device Analyzer, or a Device Analyzer-like project would look like. Enabling easier access and more powerful tools increases the privacy risks associated with the data, so new methods of establishing trust must be applied. Alternatively, certain capabilities could retain or

increase the strictness of access in the current model, while the barrier to entry could be lowered for 'safer' analyses. The legal and procedural hurdles in the licensing process—as in the case of the advertising ecosystem—serve as privacy risk mitigations to just as appreciable an extent as our data sanitisation processes. Clearly, any rigorous formulation of protections in a 'next generation of data sharing' ought to consider technical, procedural, and legal hurdles in combination from the outset of its design.

Once trust is established under the current model, full download access to the sanitised dataset is given. This provides us with no strong technical or procedural security guarantees. Furthermore, providing access to such a large dataset is difficult, with high bandwidth and storage costs for the third party. To kill both of these birds with one stone and possibly provide benefits beyond, a future system might, instead of giving the data to third parties, allow them to bring their code to the data. The next section describes this proposed model.

2.3 Research-as-a-Service: proposing a code-to-data platform

Although the Device Analyzer sharing model does offer some legal protections against data misuse, it cannot provide any strong guarantees about misuse by technical or procedural means, leaving the dataset prone to risks from human error or misconfiguration. This has been shown to be a problem in data sharing for research even recently [24]. In Section 2.2.1 I noted that once the data is in the hands of a third party, our only assurance that it will be treated securely and attempts to deanonymise subjects will not be made is the legal agreement. However, there are many technical and procedural means that might be employed to gain stronger assurances. I believe an appropriate way to go about this would be to use a 'code-to-data' model, in which third parties provide the operator (i.e. Device Analyzer developers) with their analysis code, and the operator runs it and sends back the results.

This would allow the operator to have a full view of all work done with the dataset, and ensure that the raw data is never in the hands of a third party. To be useful without significant burden for the operator, such a system would need to provide the operator with tools to analyse the code and enforce policies in an automated way. This might be done with fine-grained and dynamic access control to both data and analysis tools.

A system using this model might also be able to incorporate privacy-preserving sanitisation techniques such as randomised response and query-time differential privacy adaptively based on the code, thereby providing some strong information-theoretic guarantees about the privacy risks of the input and output of the third party's analysis. The produced results could also be checked over by a human, if necessary.

2.3.1 Thinking broadly about data sharing models

Device Analyzer's model is a comprehensive solution to a specific combination of expected use-case, datatype, and users. It has even served as a model for a similar sharing project at the Cambridge Cybercrime Centre. However, rather than applying a clear and well-established framework to that use case and data, its protections and processes were reasoned about from first principles.

Furthermore, the technical landscape of data research has changed significantly in the nine years of Device Analyzer. Android capabilities have changed significantly—many of the capabilities given to the Device Analyzer app have been withdrawn for security or usability reasons, and new 'digital health' operating system features might provide more utility to the user than Device Analyzer could. The ways in which people use their devices have changed, and the statistics someone might wish to examine about phone usage might not be covered by Device Analyzer; and of course new technologies have been developed, such as differential privacy (a big deal for future privacy-preserving efforts) or containerisation (which could help not only sharing but security of analysis).

If the goal is to find some sharing model that provides better utility, privacy guarantees, or both—thus allowing it to serve more use-cases—it is necessary to have an idea of the greater design space we are venturing out into.

This space is multivariate both in the risks and challenges associated with datasets and in protection mechanisms that the sharing model could implement. The first step in addressing this problem, then, is to establish a baseline of what approaches are available, which are in use, and how well-applied they have proved to be.

This led me to my first major piece of work (presented in its current state in Chapter 3): an effort to systematise knowledge on data sharing, with a focus on privacy protection. The contributions I intend to make with the work are the answers to the following questions:

How should we describe a dataset? — what data is being shared, and what features present in those datasets impact which protections are required?

How should we describe a sharing infrastructure? — what privacy-preserving processes are available when sharing data?

How do we choose the latter given the former? — is there a definable framework or process that can map our characterisation of a dataset to a characterisation of the most suitable platform for sharing it?

2.4 Other projects

2.4.1 Teaching

This year I have also been supervising undergraduates in Economics, Law and Ethics (for Queens and Trinity) and Paper 1 (for Trinity), and ticking for 1A ML, Java, and Algorithms.

2.4.2 IVIR summer course on Privacy Law and Policy

In July I attended a week-long course at the University of Amsterdam's Institute for Information Law, on privacy law and policy. This gave me a strong understanding of contemporary privacy law and its application, as well as the perspectives of lawyers and regulators.

Chapter 3

SoK: Sharing of sensitive datasets

In this chapter I present the current draft of my paper surveying data sharing and systematising knowledge on protecting sensitive datasets.

3.1 Aims

This survey aims to systematise current practices for sharing sensitive datasets. This space is highly heterogeneous, with many existing practices being some hybrid of technical, procedural, and legal protection techniques. Sectoral guidelines exist for handling sensitive data [25, 26, 27], which focus on procedure and are based on many years of sectoral experience (making them difficult to draw good general lessons from). In research, consideration of possible guidelines is gaining increasing attention [28]. While there is a wealth of literature on privacy-preserving algorithms, we lack an understanding of the real-world deployment of such systems—especially when combined with procedural and legal techniques—and an ability to judge their suitability and feasibility when social, economic, and legal factors come into play.

By constructing a framework for classifying these systems by in across these aspects, we intend to provide a method for assessing the suitability of sharing techniques for a given dataset/use case, and formulate principles and a decision process that can be used in the future for designing sensitive data sharing systems.

This work is expected to consist of four phases:

A – A survey of instances of sensitive data sharing, drawing from a wide range of disciplines, and attempting to give representative examples across a range of features, such as sensitivity of underlying data and protection methods employed. (We elaborate on this in Section 3.2.)

B – For each dataset and sharing methodology identified in the survey, we describe them according to a set of features, which is iteratively arrived at as more examples are added, and important common considerations of their features or designs become clear.

C – From this set of dataset features, we identify the most important considerations to serve as the axes for a design space. We consider the surveyed sharing methodologies in this space to gain insights into common themes and concerns when designing sharing platforms, and to identify risks and possible blind spots in existing designs.

D – With these two parametrisations of sensitive datasets and privacy-preserving data sharing in place, we then consider the risks found across the space. This will allow us to formulate a decision process for choosing a suitable sharing platform design for a given dataset and use case.

3.2 Survey methodology

The survey consists of a stratified sample of systems and case studies, covering the range of each feature investigated in the survey. We proceed iteratively: first collecting a set of samples and noting a maximal set of possibly pertinent features; then narrowing this down by grouping, or redefining features; then re-surveying to ensure a representative sample for each feature, or to re-inspect datasets to answer questions that arose from the previous review. We repeat until a clear parametrisation arises and presents a useful model for reasoning about datasets and platforms. Section 3.3 presents the state of the set of features at our current survey iteration.

3.2.1 Datasets vs. Platforms

Existing instances of sensitive data sharing should be differentiated into two categories: ***datasets*** and ***platforms***. While two instances might be different in terms of datatype, sensitivity, and intended use cases, both might use similar or identical means to manage sharing of data to third parties. Similarly, two instances may be similar in datatype and intent, but use entirely different processes for sharing. Thus, in our survey we separate datasets from the means of sharing (platforms).

Each dataset surveyed will necessarily have been shared via a platform. While each dataset contributes uniquely to our dataset survey, they may be shared via an already observed platform. This might either be an *explicit platform*—e.g. a data brokerage platform—or an *implicit platform*—e.g. a publicly linked download over HTTP.

The relationship between datasets and platforms is thus many-to-one. This enables us to identify classes of datasets that use the same sharing mechanisms, and where similar

datasets use different platforms we are able to compare them, to see why such decisions may have been made, and evaluate. This evaluation is an 'adequacy test', which we expect would then form the basis for the decision process that is the object of Phase D—a function that maps a dataset to a platform.

3.2.2 Defining personal and sensitive data

In Section 3.1, we defined our scope of datasets to data 'about behaviour of individuals'. Here, we will motivate and provide a more concrete definition of personal data, which extends existing legal definitions to better facilitate the characterisation of the risks and technical challenges a dataset presents.

Our definition of personal data aligns broadly with the definition given by the GDPR:

> 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;
>
> — *GDPR Article 4(1)*

This definition couples the notion of personal data with the task of identification, and so is able to capture the wide semantic breadth of personal data; personal data is information about an identified individual, or about an individual *that may be identified*. Therefore data can be considered personal not only when it includes 'direct' means of identification—identifiers or quasi-identifiers—but also when it includes data that can be used 'indirectly' to identify a natural person.

This process of indirectly identifying a person by reference to various factors presents the fundamental challenge behind the collection of data on individuals: anonymising sufficiently rich data is difficult. Through combination with other data sources, subjects may be re-identified, or more personal information may be inferred about them. While legal regulation might (for the purposes of judgement) cast this risk as a binary—whether subjects can or cannot be re-identified indirectly from a given dataset—the risk of indirect identification is in fact a matter of degrees.

For example, some datasets may be trivially combined with external sources to re-identify all data subjects, whereas others may require a significant amount of work and access to privileged data to re-identify a single subject. The nature of risk in each of these datasets is different, and so each requires a different approach to provide adequate protection. While there exist best practices such as pseudonymisation for mitigating the risk of

direct identification, if we intend to also address the risk of indirect identification, we must set out some concrete methods of identifying and evaluating this risk.

To achieve this, we give a more rigorous footing to the definition of personal data by introducing the notion of **linkability**—put simply, a multivariate description of a dataset's potential to be combined with other information. By formalising linkability not only as a description of the level to which indirect re-identification is possible, but as the effort required to do so, we can better characterise the full scope of re-identification risks. This will then allow us to better outline the necessary practical considerations for sharing a given dataset. A detailed definition of linkability is given in Appendix B.1.

This brings us to our definition of personal data as: *information relating to an individual who is directly or indirectly identifiable; indirect identification being the reduction of anonymity by exploiting the **linkability** of data.*

The GDPR also addresses the notion of **sensitivity**—the existence of particular categories of data that warrant extra protections—with its definition of special categories:

> Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.
>
> — *GDPR Article 9(1)*

It is important to note that this is not an *absolute* prohibition on the processing of these categories of data; derogations are defined in Article 9 for data controllers and processors that allow for the processing of these categories. For example, processing of health data by the NHS is permitted where 'necessary for reasons of public interest in the area of public health'. We believe this attribution of special status to types of data that require especially strong protection to be a useful distinction, and so adopt a definition of sensitive data that aligns with it.

We group the given special categories into two supercategories: data that reveals membership of a vulnerable group[1], and extraordinarily powerful identifiers (genetic[2] and biometric data).

It should be noted that like the GDPR definition of personal data, this definition also intentionally leaves open to interpretation a question of indirect identification—whether data

[1]The term 'vulnerable group' is here used to mean a class of individuals who may be made victims of harmful discrimination or some other violation of fundamental rights, for example in the European Union their dignity.

[2]Genetic data can also, under certain use cases, reveal membership of vulnerable groups such as ethnic minorities, or provide a basis for health-based discrimination by indicating an subject's predisposition to congenital disease.

reveals a subject's membership of a vulnerable group. Therefore data might be 'indirectly sensitive' or 'directly sensitive'.

3.2.3 Requirements for dataset consideration

To ensure a limited and 'interesting' scope for our survey, we require datasets to

- be intended for academic study;
- and include potentially personal data (potentially PII, pseudonymous, or linkable).

We currently have 29 individual datasets, and 6 groups of datasets under consideration in the survey. Appendix B.2 shows the initial survey of datasets and platforms that contributed the candidates, and Appendix B.1 presents the features currently of interest, with descriptions of how the iterative process arrived at them.

3.3 Classification methodology

We describe each dataset or platform example by a set of features, to allow comparison and highlight common considerations and decisions made when working with sensitive data.

As noted, we proceed iteratively, beginning with a set of features chosen according to our existing understanding of sensitive data, and when new common features are identified or when notable differences are found, we add to the feature set; if existing features appear inadequately descriptive, they are redefined.

So far, the survey has been conducted in three phases. The initial survey (Appendix B.2) built an initial list of samples, leading to the formulation of the distinction between dataset and platform. The second phase focuses on datasets, building a table of features.

This dataset survey encountered difficulties in assessing some features: mostly the question of how to describe the linkability of a dataset, even given our feature definition (below). We needed a mechanical or directed decision process to evaluate linkability. This led to the third phase (Section 3.4): building a framework for evaluating the scope for linkage attacks on a given dataset.

Appendix B.1 shows the definitions for all current features under consideration, with a brief description for each possible value, and justifications drawn from the survey. Figures 3.1 to 3.4 show an excerpt from the dataset survey for four of the datasets surveyed.

Dataset	Sensitivity	Linkability				Datatypes
		Availability	Difficulty	Cardinality	Example	
Device Analyzer	vulnerable group (indirect)	internal	?	majority	coarse location trace	smartphone logs, location traces
23andMe	superidentifier	private		individual		genome
India Human Development Survey-II (IHDS-II), 2011-12	vulnerable group	internal	hard	individual	if you know they're in the dataset, you may be able to extract sensitive information	questionnaire
Youth Development Study, 1988-2011	vulnerable group	private	mild	subset	trawl social media for people at those schools, and match	questionnaire
Maternal Lifestyle Study in Four Sites in the United States, 1993-2011	vulnerable group	private	mild	individual	one in forty people born at this particular place had mothers who had taken cocaine, indentifying them within the dataset using perhaps social media reveals developmental information about them and sensitive information about their mothers	questionnaire
Census of Juveniles in Residential Placement	vulnerable group	private		individual		

Figure 3.1: This figure shows sensitivity, linkability, and datatypes collected on this subset of datasets.

Dataset	Source			Intent		
	Discipline	Motivation	Organisation	Discipline	Motivation	Organisation
Device Analyzer	CS	academic	University of Cambridge	CS	academic	
23andMe	health	commercial	23andMe	health	academic, industrial > commercial	
India Human Development Survey-II (IHDS-II), 2011-12	economics/sociology	academic	University of Maryland, and National Council of Applied Economic Research	economics/sociology, governance	academic	
Youth Development Study, 1988-2011	economics/sociology	academic	University of Minnesota	economics/sociology/education	academic	
Maternal Lifestyle Study in Four Sites in the United States, 1993-2011	health	academic	Barry Lester, Women and Infants Hospital of Rhode Island; Henrietta Bada, University of Kentucky; Charles Bauer, University of Miami; Seetha Shankaran, Wayne State University; Toni Whitaker, University of Tennessee; Linda LaGasse, Women and Infants Hospital of Rhode Island; Jane Hammond, RTI International	health/governance	academic	
Census of Juveniles in Residential Placement						

Figure 3.2: This figure shows source and intent information collected on this subset of datasets.

Dataset	Scale		Subject			Age	Collector
	N	k	Permission	Active	?		
Device Analyzer			volunteer	no		recent	first-party
23andMe	5,000,000	exome	volunteer	yes		recent	first-party
India Human Development Survey-II (IHDS-II), 2011-12	204,569	huge	volunteer	yes		pertinent	first-party
Youth Development Study, 1988-2011	2,573	huge	volunteer	yes		pertinent	first-party
Maternal Lifestyle Study in Four Sites in the United States, 1993-2011	1,388	huge	volunteer	yes		pertinent	first-party
Census of Juveniles in Residential Placement							

Figure 3.3: This figure shows scale, subject, age, and collection information collected on this subset of datasets.

Dataset	Collection		Data quality	Format	Pathetic dot	
	Temporal scope	Continuing			Market	Norms
Device Analyzer	long-term	yes	good?		only similar are academic	similar things done by OS companies, with opt-in
23andMe	snapshot	yes	good	genome, reports	lots of other sequencing companies, personal and otherwise	concordat on insurance, continuing questions about the ethics of genomics, recent guidance
India Human Development Survey-II (IHDS-II), 2011-12	snapshot	no	good	a bunch	none	a frequent thing for governments to do, but rarely of this scale
Youth Development Study, 1988-2011	long-term	no	?	a bunch	none	governmental
Maternal Lifestyle Study in Four Sites in the United States, 1993-2011	long-term	no	?	?		
Census of Juveniles in Residential Placement						

Figure 3.4: This figure shows more collection and market and norm information collected on this subset of datasets, as well as some auxiliary data on data quality and format.

3.4 Evaluating linkability by constructing attacks

As noted in Section 3.3, describing the risk of linkage with another dataset is difficult to achieve by inspection. In this section we present the basis and justification for a qualitative attack-based approach to describing linkability. Understanding the risks associated with a dataset through its ability to be linked to other datasets is inherently a qualitative task; the fact of vulnerability to a certain attack, and the extent of the harm it could cause are influenced not only by technical and mathematical factors, but procedural and economic ones.

3.4.1 Linkability is qualitative, not quantitative

Mathematically, one could evaluate these potential attacks and quantify the associated risk by analysing factors such as the number of data subjects, the nature of the data points collected, the distribution of those data points both in the collected data and 'in the wild'. There has been a great deal of information-theoretic work that attempts to quantify the individuality of a row within a dataset, with a view to preventing recovery of unique data by altering the distribution, such as in the example of differential privacy where specially crafted noise is applied to the data. However, this purely mathematical approach to risk is not globally applicable—noise cannot be applied to every dataset while preserving utility—and even where mitigations derived from this process can be deployed, taking a more holistic view of risk may show that efforts could be better applied elsewhere.

We must also consider the factors that affect the capabilities of the attacker, and the context of any potential harm. To illustrate the complexity of this task, and the range of risk responses a rational publisher might consider, consider a large dataset of geotagged photos used for computer vision research. The presence of individuals in these photos creates the risk of identifying a person (if you have a picture of them), and finding out where they were at a particular time. However, if the dataset is very large, it would require a great deal of computational power to find a person by facial recognition. This attack would be made easier if more information can be linked to, such as the city the target lives in, which allows a narrowing of the search space. Even then, the information gained may not be particularly harmful. Knowing where a person works may be personal data, but the inference could not be considered a reduction in anonymity, or an increased risk of potential harm, if the information is also available on a public LinkedIn profile, and the attacker has learned that they often go there between 9am and 5pm on weekdays. We could also consider another attacker with access to this dataset, who does not have a particular target in mind. A generic attack aiming to reidentify many people, perhaps indiscriminately would require a great deal more computational power and side information—in this case an analysis might conclude that this attack model would only be cost-rational for a state-level attacker. Should a

computer vision researcher apply protections to their dataset against foreign governments?

We see that the risks associated with a dataset cannot be fully captured by considering the potential uniqueness of a user and the shape of the distribution across collected features. Instead, we must also consider the intent of an attacker, the kinds of data they have access to, and in what norms or other context the potentially inferable data lie.

3.4.2 A qualitative attack-based survey

Therefore, we present a set of threat models, parametrised in two dimensions by the intent of the attacker and their access to side information (following the definition of availability of linkage datasets given in Appendix B.1).

We consider four distinct attacker intents:

- targeted attacks (assuming some identifying information is already known about an individual);
- class attacks, where an attacker aims to deanonymise data subjects who satisfy some criterion (that is not a piece of identifying information);
- 'any one subject' attacks, where the attacker aims to deanonymise a single subject, indiscriminately, and learn as much as possible about them;
- and general mass attacks, where the attacker aims to deanonymise as many subjects as possible.

We further break down targeted attacks into 'blind' and 'nosy'. This is due to the importance of the membership inference problem—the problem of determining whether a target is present in a dataset. Not knowing this information significantly changes the feasibility of an attack, and forces the consideration of different factors, such as the global uniqueness of data. Therefore we consider both the general 'blind' targeted attack, where the attacker must solve the membership inference problem) and the 'nosy' attack, where the attacker already knows their target is a subject of the dataset.

This parametrisation aligns with our subcategorisation of the linkability feature given in Appendix B.1—attacker intents with cardinality, availability directly, with effort given by the description given for each combination of the former. Figures 3.5, 3.6, and 3.7 present the first results of this attack survey. From these preliminary results, we see some trends start to emerge:

- As expected, the internal case (given in the tables as 'Extra' for readability) is usually of little interest.
- Mass attacks are for the most part a repetition of the 'any one' attack.

- In many cases, private access can enable an attack but cause no increase in harm, given the preconditions for access to the private dataset.
- The membership inference problem is a significant factor in determining both the feasibility and potential harm of an attack. This perhaps should lead to a rethinking of the definition of linkability to include a special case for the 'nosy' scenario.

This survey methodology has proved promising. The next steps we intend to take are to expand the survey with a focus on comparing similar datasets by datatype.

Name	Availability	Target		Members of class	Any one subject	General mass attack
		Blind	Nosy			
Device Analyzer	Intrinsic	none	none	None	None	None
	Public	none	none	None	None	None
	Private	Difficult: the nosy attack can be used but the profile built has to be much more detailed to satisfy global uniqueness	Detailed profile of behavioural data can be linked against to reidentify target in database (detail has to satisfy local uniqueness)	None	Very very difficult: a detailed (globally unique) profile can be made up of movements and usage patterns, which would have to be linked to an extraordinarily large dataset of behavioural data, probably impossible to satisfy cohort membership problem)	None (the one-subject investigative attack does not scale)
	Extra	Collectors could link against app install data or IP addresses to reidentify, plus use pre-sanitised data				
Strava Metro	Intrinsic	None (nosy attack doesn't work because we assume we don't know more about the user than identifiers, so we don't know where they go, just where they live)	We can use heuristics to construct user profiles of location traces, assuming they have many logged exercises. Knowing where they live, we can apply these and traffic analysis to the roads around them and the aggregates for the bin in which they live. This increases degree of confidence in new routes starting from their address and may allow reconstruction of some full runs, including points of interest.	Nosy intrinsic attack could be used centred instead of on home locations but on places of interest such as schools or workplaces or religious places	Areas of low density can be trivially watched to construct an individual's profile	Route reconstruction and heuristic modelling can be applied to whole dataset but this will only recover a small subset of subjects, mostly those who run at odd times
	Public	Corresponding nosy attack can be used, cohort membership problem is made easier in the presence of social media posts about running etc.	Intrinsic nosy attack can be further improved and made to work with fewer data points and better modelling, by using publicly available check-in events (four square, facebook etc)	Intrinsic attack can be strengthened as in targeted nosy	Intrinsic attack can be linked to public data such as four square to find more identifiers on a person, or information	Census data on home-work pairs could increase the power of the intrinsic attack but doubtful as to achieving majority
	Private	Corresponding nosy attack can be used, as above	Datasets with behavioural data on fitness or visited locations, or place of work, give even stronger data to improve models than public as above	Stronger strengthening as with targeted nosy	Stronger version of public	Stronger location datasets than census data with appropriate size to broaden the attack beyond subset would be state-level or health data with addresses in them
	Extra	Strava have trivially global access to underlying dataset and user information				
SNAP ego sets	Intrinsic	None (the targeted threat model does not assume access to the target's Facebook profile and data is pseudonymised)		None (ego networks, political affilliations etc. are not sufficiently globally unique)		
	Public	Nosy attack is very unlikely to work because globally unique feature combinations for a subject will by definition be difficult to extract from pseudonymised datasets	Reidentify by comparing public Facebook data to correspond to anonymised features, after some whole-dataset distribution analysis of feature set. Private features about the subjects can be inferred.			
	Private	Private access to a user's Facebook network allows reconstruction of their ego network, which can be matched against the dataset, which is probably quite globally unique and locally unique. This probably doesn't gain any information over existing facebook access though, unless through combination of other ego networks in the dataset, features can be inferred about friends. That's not really a compelling attack.		Access to the subjects Facebook account as in the private targeted case might occasionally provide a globally unique identifier, but a global analysis is too costly and there again is no benefit once you have access to account details		
	Extra	No extra capabilities beyond access to raw data				

Figure 3.5: A sample of the datasets surveyed for potential linkage attacks. 'Extra' is used in place of *internal* for readability.

Name	Availability	Target		Members of class	Any one subject	General mass attack
		Blind	Nosy			
SNAP email networks	Intrinsic	None				
	Public					
	Private					
	Extra	No extra capabilities beyond access to raw data				
SNAP location sets	Intrinsic	Small chance of retrieval. Home/work pairs may be included in the dataset, but unlikely as they are usually not places of interest to be checked into. Checkins nearby each of these could be used as a proxy.	Same as blind, but with higher probability of constructing PPOI proxies through nearby checkins.	Mobility profiles of users who visit sensitive points of interest can be created, and through constructing quasi-home-work pairs they become identifiers, and this attribute can be propagated through their friend graph to attempt recovery of social circles or communities with sensitive attributes, such as political or religious affiliations, or health conditions	A user who regularly checks in at a home location or a workplace provides a trivial one-subject recovery, and is easily found by frequency of checkins	Small chance of retrieval by constructing home-work pairs, so probably only a very very small subset could be reidentified.
	Public	Linking against other checkins or similar posts such as Foursquare, Facebook, and more recently than these datasets Instagram strengthens intrinsic attack substantially		Performing the intrinsic attack with other checkin information and perhaps subsequently social network information allows more accurate reconstruction of the class, as well as names and other stronger identifiers than fuzzy home-work pairs	If the home can be found, this can be linked to public census or voter roll records to find names	In cases where users check in on both the service in question and another such as Foursquare and Facebook, these could be matched and the intersection identified. Assuming migration to one of theese servers after shutting down of Gowalla and Brightkite, profiles could be matched against similar Foursquare or Facebook users at great technical cost. Unlikely to recover many.
	Private	Linking against card purchase information, search history, or any other proxies for behaviours such as rest, home, work, coffee, can be matched against timestamps and GPS, in the absence of other 'checkins'		Similarly to the targeted case, the accuracy of linking extra identifiers to a mobiity graph is improved with private datasets	No more capability than public	Purchasing information could be used to match people to traces. A Google could match web traffic to locations and extract friendships
	Extra	None				
Enron emails	Intrinsic	The same attack as the untargeted case could be performed but narrowing down by searching for the user's name in the dataset.	The same attack as the untargeted case would be used, removing the uncertainty posed by global uniqueness of any identifying features derived.	NLP analysis and graph of email addresses can retrieve a small, outdated, and incomplete set of features on each person such as dates of birth, names of family and friends, religious interests etc, home addresses. These may be so outdated that name-DOB pairs will be the only identifiers available (enron emails and even addresses would likely not be current), which may not be sufficiently globally unique to provide any utility to an attacker. Therefore the maximum impact of this attack is to infer old personal, political or religious affiliations of a few users (or, much less likely, juicy personal life stories). This attack is identical for each of the 150 subjects.		
	Public	Matching names and other details against Facebook or old census data allows a limited increase in information to be attributed to each user, and stronger chances of reidentification. The most informative outcome would be to tie outdated behavioural data to current facebook user accounts, which is probably not very useful in any attack scenario				
	Private	Connecting to phone records from the time could identify the phone numbers and addresses (at the time) of business contacts				
	Extra	None				
Netflix prize	Intrinsic	None				
	Public	Connecting to IMDB or other ratings can solve membership inference. This means a blind attacker has a reduced but significant probability of accomplishing the nosy attack.	Linking to IMDB or other ratings can expose the user within the dataset. This can reveal significant details about their personal life based on preference, such as sexuality, race, etc.	The general attack can be performed, specifically on those users who have rated or watched particular films	Assuming sufficient computing power to scrape IMDB and link, the blind targeted attack could be performed on the IMDB users. If these accounts are linked to Facebook accounts, this can provide more information for matching as well as contribute more identifying information.	

Figure 3.6: A sample of the datasets surveyed for potential linkage attacks. 'Extra' is used in place of *internal* for readability.

<table>
<tr><th rowspan="2">Name</th><th rowspan="2">Availability</th><th colspan="2">Target</th><th rowspan="2">Members of class</th><th rowspan="2">Any one subject</th><th rowspan="2">General mass attack</th></tr>
<tr><th>Blind</th><th>Nosy</th></tr>
<tr><td rowspan="2"></td><td>Private</td><td colspan="2">If the IP address of a user can be matched against their IMDB or Netflix usage patterns, an ISP or other global passive adversary could learn sensitive data about the user</td><td>The global passive adversary from the targeted case could find the IP addresses of a sensitive class of subjects</td><td colspan="2">Passive adversary with traffic data</td></tr>
<tr><td>Extra</td><td colspan="5">Trivial access to underlying data</td></tr>
<tr><td rowspan="4">Yahoo music ratings</td><td>Intrinsic</td><td colspan="5">None</td></tr>
<tr><td>Public</td><td colspan="5">None (nothing to link against March 2004 track ratings)</td></tr>
<tr><td>Private</td><td colspan="2">Requires very specific side information, any adversary who had this would gain nothing from an attack</td><td colspan="3">No (known) dataset to link against</td></tr>
<tr><td>Extra</td><td colspan="5">Trivial access to underlying data</td></tr>
<tr><td rowspan="4">Flickr YFCC100M</td><td>Intrinsic</td><td>Location metadata can be used to reidentify based on home location, with membership inference performed using photo contents and facial recognition.</td><td rowspan="3">Possible attacks are the same as blind, but no further risk created by this dataset because all media was already public under creative commons</td><td rowspan="3">General attack (mass and one subject) can be concentrated on photos at certain locations or at certain tags</td><td colspan="2">Location data and facial recognition can be used to construct behavioural profiles attached to faces or photographers</td></tr>
<tr><td>Public</td><td>Linking to Facebook, Instagram, or Foursquare checkins allows reidentification of other people in photos. Because dataset was already Creative Commons the behavioural information conferred by the dataset is mostly only notable where names attributed to Flickr profiles have low global uniqueness or where pseudonyms are used</td><td>Profiles from intrinsic attack can then be linked against social media checkins to link names and other information to faces or profiles</td><td rowspan="2">Same as any subject but unlikely to be a majority attack because only half of images have geolocation, and not all photos are of people.</td></tr>
<tr><td>Private</td><td>Any knowledge of where a person was at a given time (or was likely to be) would allow the public attack but without relying on users actively checking in</td><td>Private records of attendance at a particular location can be cross-referenced against photos taken there, reducing the reliance on active social media checkins from the public attack</td></tr>
<tr><td>Extra</td><td colspan="5">Trivial access to user accounts</td></tr>
</table>

Figure 3.7: A sample of the datasets surveyed for potential linkage attacks. 'Extra' is used in place of *internal* for readability.

Chapter 4

Thesis plan

4.1 Research question

I intend to produce a dissertation investigating the privacy-utility trade-off in data sharing, with a particular focus on health and mobile applications. This trade-off has an analogue in security engineering in the consideration of usability: it is often thought of as a zero-sum problem, where an increase in security (*privacy protection*) for the user (*data subject*) necessarily reduces the usability (*utility*) of the system. In the case of usable security, it has been shown that this characterisation is insufficient; in fact, human behaviour is inherent to the security problem, rather than strictly being an obstacle, and successful security design takes usability into account. Similarly, I aim to achieve a win-win, where the application of privacy-preserving techniques can in fact increase the utility of sensitive data in real-world applications. The research question the dissertation will aim to answer is:

> *Can we enable privacy-preserving data sharing in real-world situations, while preserving or increasing data utility?*

This scope—'real-world'—is broad and multivariate; earlier in this report and in the SoK paper I am currently working on, I argue that the problems of privacy protection and risk analysis are not simply a function of the richness and nature of a dataset, but of the threat models, markets, laws, and norms that make up the context of the sharing activity. This also results in a variety of ways to evaluate utility, and thus many different trade-offs to make and optimise for: the availability of data, speed of analysis, data portability, or reproducibility of analysis to name a few. To ensure a workable scope, I thus will limit my consideration of data sharing use cases to **research on sensitive datasets**—such as the Device Analyzer project.

To tackle this question, I expect to break down the problem into three sub-questions:

1. Given a potentially sensitive dataset, what are the best practices for sharing it, and can we construct a decision process for arriving at them?

2. Can we construct a code-to-data platform for sensitive data that can be tuned to provide adequate protection for the use case, under a range of threat models?

3. What is gained and lost when we apply this new approach to a real research project?

The first question establishes a baseline and my proposed framework for data sharing. The second question applies this framework in practice, building a system that aligns. The third and final question then evaluates this system, to validate the proposed approach. I intend to arrange the chapters of the dissertation as follows: the first question will be the focus of the first two chapters, and the second and third questions will be the focus of Chapters Three and Four. Below I present a brief description of each chapter.

The second and third questions will be addressed by Chapters Three and Four respectively. I intend to build a system that can address many different models of privacy risk identified in the first chapter. I envision this as a code-to-data platform, where third parties wishing to perform some analysis on a sensitive dataset can upload their analysis code, and appropriate data sanitisation and oversight can be applied adaptively according to the trust placed in the third party, as can the capabilities with which they are presented. The production of a 'core' platform of this kind and its application in the instance of a 'Device Analyzer 2.0' will be the focus of Chapter Three, and its evaluation and any lessons learned from the approach will be given in Chapter Four.

Below I give a brief description of each chapter:

4.2 Chapter One

The first question establishes a baseline for data sharing—i.e. what does the privacy-utility trade-off actually look like in practice today? This chapter will investigate to what extent actual guidance, best practices, and systems in use align with the state of the art in privacy-preserving data publishing.

Working from this baseline, as with earlier efforts to establish 'Privacy by Design', this chapter will present a process that will allow researchers to better understand the privacy risks associated with their dataset, and then to map the dataset to an appropriate platform/set of protection techniques.

These are the aims of the SoK given in Chapter 3, which will form the basis for this chapter.

4.3 Chapter Two

This chapter will consist of applying and refining the framework from Chapter One, using Device Analyzer as a case study. This case study will consist of a retrospective on Device Analyzer from the point of view of the development team, and a users study of researchers who licensed the dataset.

I expect this case study to illuminate the pain points associated with protecting the data, the risks assumed and addressed, and utility pain points and preferences on the part of the third party researchers.

Applying the framework from Chapter One should present a new model (or set of potential models) for Device Analyzer, and cross-referencing this with the learnings from the case study should verify, or help refine, the framework.

4.4 Chapter Three

I propose to build a platform with which analysis can be performed on sensitive (or potentially sensitive) datasets, while providing strong guarantees of privacy for data subjects. A model for this that I intend to investigate is a code-to-data platform, which provides privacy assurances by never directly allowing third parties to view data, and can provide different levels of protection through technical means (pre-sanitisation, query-time sanitisation, reduced library support, etc), procedural means (rate-limiting, standard report formats for output), and legal means (clickwrap for low-risk access, Device Analyzer-style legal agreement for high-risk).

The platform built should be a 'core', which can then be easily tunable to address various risk profiles—some subspace of the platforms that the process in Chapter One laid out. This approach must be evaluated, so I intend to port existing projects to work atop this core, probably Device Analyzer, and possibly also cybercrime or health datasets.

There is a large space of possible capabilities for this system, both in privacy-preserving algorithms and features for research utility. Some possible avenues to explore are as follows:

- Containerisation: creating an individual environment for each researcher, such as a Docker container, with carefully chosen data access and packages available.
- Static analysis: applying taint analysis or other data provenance techniques to construct an audit trail for any analysis code uploaded to a code-to-data platform.
- Access control-based differential privacy parametrisation: applying differing levels of noise to a dataset, or to different rows in the table, based on the trust profile associated with a user.

- Differentially-private stream processing: using a system such as PeGaSus [29] to allow safe stream processing, eliminating the utility overhead in the current Device Analyzer created by producing 'releases'.
- Pan-private randomised response: collecting a set of publicly-available analytics on a dataset that uses an approach similar to RAPPOR [5], with infrequent updates to preserve pan-privacy.
- *Secret-shared non-interactive proofs*: adapting the Prio system [30] which allows the computation of aggregate statistics across multiple servers, while preventing servers from learning any of the underlying personal data as long as one server remains honest. This might allow a halfway point between data release and code-to-data, where third parties can operate a server which works in concert with our 'trustworthy' server, whose presence in the system alone would provide strong privacy guarantees.

Provisionally, I expect to start by outlining a stratification of trust placed in third party researchers, and investigate combining the above avenues for each.

4.5 Chapter Four

Once the system in Chapter Three has been built, it must be compared against earlier efforts to share sensitive data. In the case of Device Analyzer, the user study performed for Chapter Two will provide a baseline for researcher utility. Some of these users could then be asked to use the new system, and evaluate its utility. This would show where usability has been improved or worsened, where new capabilities have been added, and generally a notion of whether the system is improved.

A security analysis of the system will also be performed to show via multiple threat models and examples that the system provides better assurances than the original Device Analyzer, and verify that in some cases the strong legal constraints of the original project can be relaxed.

4.6 Timeline

Table 4.6 shows my planned timeline of work, by date, pieces of work outlined in this chapter, and deliverables.

<table>
<tr><th>Month</th><th colspan="2">Work</th><th>Deliverable</th></tr>
<tr><td>Rest of 2018</td><td rowspan="2">Complete SoK</td><td></td><td rowspan="13">SoK paper (PETS)</td></tr>
<tr><td>Jan 2019</td><td rowspan="4">User study of existing users and retrospective of Device Analyzer</td></tr>
<tr><td>Feb 2019</td><td rowspan="3"></td></tr>
<tr><td>Mar 2019</td></tr>
<tr><td>Apr 2019</td></tr>
<tr><td>May 2019</td><td rowspan="5">Designing platform</td><td></td></tr>
<tr><td>Jun 2019</td><td rowspan="8">Building platform</td></tr>
<tr><td>Jul 2019</td></tr>
<tr><td>Aug 2019</td></tr>
<tr><td>Sep 2019</td></tr>
<tr><td>Oct 2019</td><td rowspan="4">Design paper (HotMobile)</td></tr>
<tr><td>Nov 2019</td><td rowspan="4"></td></tr>
<tr><td>Dec 2019</td></tr>
<tr><td>Jan 2020</td></tr>
<tr><td>Feb 2020</td><td rowspan="2">Completed system</td></tr>
<tr><td>Mar 2020</td><td colspan="2" rowspan="4">User study & security analysis</td></tr>
<tr><td>Apr 2020</td></tr>
<tr><td>May 2020</td></tr>
<tr><td>Jun 2020</td><td rowspan="2">System paper</td></tr>
<tr><td>Jul 2020</td><td colspan="2" rowspan="6">Writing up</td></tr>
<tr><td>Aug 2020</td></tr>
<tr><td>Sep 2020</td></tr>
<tr><td>Oct 2020</td></tr>
<tr><td>Nov 2020</td></tr>
<tr><td>Dec 2020</td><td>Dissertation</td></tr>
<tr><td>Jan 2021</td><td colspan="2"></td><td>Submit</td></tr>
</table>

Table 4.1: My proposed PhD timeline, broken down by month. Italics in the deliverable column denote the venues to which I intend to submit the deliverable.

Bibliography

[1] C. Dwork, "Differential Privacy: A Survey of Results," in *Theory and Applications of Models of Computation*, pp. 1–19, Berlin, Heidelberg: Springer, Berlin, Heidelberg, Apr. 2008.

[2] C. Dwork, M. Naor, T. Pitassi, and G. N. Rothblum, *Differential privacy under continual observation*. New York, New York, USA: ACM, June 2010.

[3] N. E. Bordenabe, C. Palamidessi, M. E. Andrés, N. E. Bordenabe, and K. Chatzikokolakis, *Geo-indistinguishability: differential privacy for location-based systems*. differential privacy for location-based systems, New York, New York, USA: ACM, Nov. 2013.

[4] C. Liu, S. Chakraborty, and P. Mittal, "Dependence Makes You Vulnerable: Differential Privacy Under Dependent Tuples," in *Network and Distributed System Security Symposium*, (Reston, VA), pp. 1–15, Internet Society, July 2016.

[5] Ú. Erlingsson, V. Pihur, and A. Korolova, "RAPPOR: Randomized Aggregatable Privacy-Preserving Ordinal Response," *arXiv.org*, pp. 1054–1067, July 2014.

[6] K. Wang, R. Chen, B. C. Fung, P. Y. A. C. Surveys, and 2010, "Privacy-preserving data publishing: A survey on recent developments," Dec. 2017.

[7] Y. Xiao and L. Xiong, *Protecting Locations with Differential Privacy under Temporal Correlations*. New York, New York, USA: ACM, Oct. 2015.

[8] R. Al-Dhubhani and J. M. Cazalas, "An adaptive geo-indistinguishability mechanism for continuous LBS queries," *Wireless Networks*, vol. 36, pp. 1–19, May 2017.

[9] R. Schlegel, C.-Y. Chow, Q. Huang, and D. S. Wong, "User-Defined Privacy Grid System for Continuous Location-Based Services," *IEEE Transactions on Mobile Computing*, vol. 14, no. 10, pp. 2158–2172, 2015.

[10] A. Olteanu, K. Huguenin, R. Shokri, M. Humbert, and J. Hubaux, "Quantifying interdependent privacy risks with location data," *IEEE Transactions on Mobile Computing*, vol. 16, pp. 829–842, March 2017.

[11] A. Razaghpanah, R. Nithyanand, and N. Vallina-Rodriguez, "Apps, Trackers, Privacy, and Regulators: A Global Study of the Mobile Tracking Ecosystem," *Network and Distributed System Security Symposium*, 2018.

[12] E. Graham-Harrison and C. Cadwalladr, "Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach," Mar 2018, last accessed 2018-11-03. *https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election*.

[13] S. Salinas, "Zuckerberg says he's seen evidence russia tried to influence election using social media," Jul 2018, last accessed 2018-11-03. *https://www.cnbc.com/2018/07/18/facebooks-mark-zuckerberg-no-reason-to-doubt-russian-election-medd.html*.

[14] C. E. Wills and C. Tatar, *Understanding what they do with what they know*. New York, New York, USA: ACM, Oct. 2012.

[15] G. Acar, C. Eubank, S. Englehardt, M. Juarez, A. Narayanan, and C. Diaz, *The Web Never Forgets: Persistent Tracking Mechanisms in the Wild*. Persistent Tracking Mechanisms in the Wild, New York, New York, USA: ACM, Nov. 2014.

[16] S. Englehardt, D. Reisman, C. Eubank, P. Zimmerman, J. Mayer, A. Narayanan, and E. W. Felten, *Cookies That Give You Away: The Surveillance Implications of Web Tracking*. The Surveillance Implications of Web Tracking, New York, New York, USA: International World Wide Web Conferences Steering Committee, May 2015.

[17] S. Englehardt and A. Narayanan, *Online Tracking: A 1-million-site Measurement and Analysis*. A 1-million-site Measurement and Analysis, New York, New York, USA: ACM, Oct. 2016.

[18] S. Son, D. Kim, and V. Shmatikov, "What mobile ads know about mobile users.," in *NDSS*, 2016.

[19] G. Venkatadri, A. Andreou, Y. Liu, and A. M. I. S. on, "Privacy Risks with Facebook's PII-based Targeting: Auditing a Data Broker's Advertising Interface," *mislove.org*, 2018.

[20] A. Datta, M. C. Tschantz, and A. Datta, "Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination," *arXiv.org*, Aug. 2014.

[21] A. Andreou, G. Venkatadri, O. Goga, and K. P. Gummadi, "Investigating ad transparency mechanisms in social media: A case study of Facebook's explanations," *Network and Distributed System Security Symposium*, 2018.

[22] P. Vines, F. Roesner, and T. Kohno, "Exploring adint: Using ad targeting for surveillance on a budget-or-how alice can buy ads to track bob," in *Proceedings of the 2017 on Workshop on Privacy in the Electronic Society*, pp. 153–164, ACM, 2017.

[23] D. T. Wagner, A. Rice, and A. R. Beresford, "Device analyzer: Large-scale mobile data collection," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 4, pp. 53–56, 2014.

[24] P. Waterfield and T. Revell, "Huge new facebook data leak exposed intimate details of 3m users," May 2018.

[25] UK Data Service, "Legal and ethical challenges surrounding big data: energy data," pp. 1–3, Nov. 2017.

[26] Future of Privacy Forum, "Privacy Best Practices for Consumer Genetic Testing Services," tech. rep., July 2018.

[27] C. Culnane, B. Rubinstein, and V. Teague, "Health Data in an Open World," pp. 1–25, Dec. 2017.

[28] I. Lundberg, A. Narayanan, and K. L. a. p. arXiv, "Privacy, ethics, and data access: A case study of the Fragile Families Challenge," *arxiv.org*.

[29] Y. Chen, A. Machanavajjhala, M. H. P. o. t. 2017, and 2017, "PeGaSus: Data-Adaptive Differentially Private Stream Processing," *dl.acm.org*, pp. 1375–1388, 2017.

[30] H. Corrigan-Gibbs and D. Boneh, "Prio: Private, Robust, and Scalable Computation of Aggregate Statistics," *arXiv.org*, Mar. 2017.

Appendix A

ADINT

A.1 Investigation plan

A.1.1 Outline

This study intends to investigate the extent to which the purchase of mobile advertising inventory can be used to surveil and track the location of users—introduced in recent work as ADINT. We intend to evaluate the attack surface under realistic and concrete threat models.

We also propose to develop a method of establishing a covert channel through mobile ads (StegAd). This channel requires no installation of native code on the part of the 'undercover' party, hides communication in naturalistic phone usage, and does not require the 'undercover' party to share an identifier prior to communication.

A.1.2 Background

A recent work by Vines et al. [22] showed that the purchasing of mobile ads could be used to track the location of a user on whose device the ads were shown, without the need for dynamic content in the ad creatives, or a click-through on the ad. This class of attacks, dubbed ADINT, leverages the rich data collected on a user by advertising networks and assembled into a profile, and the many targeting options available to an advertiser—including the retargeting of specific user profiles.

The mobile advertising ecosystem

Briefly, the mobile advertising ecosystem operates as follows: a **User** installs an app or visits a page provided by a **Publisher**. This Publisher embeds in their app or website an ad library, provided by a **Service-Side Provider (SSP)**. The role of the SSP is to manage

Publishers and auction off this advertising inventory to parties interested in advertising to the user. These auctions are contested by **Demand-Side Providers (DSPs)**, who manage and bid on behalf of **Advertisers**, the parties who provide the content for the ads.

The SSP constructs profiles of Users, to allow targeted advertising, which is more valuable to an Advertiser. It does this by tracking user behaviour using browser cookies, or through a **Mobile Advertising Identifier (MAID)**. This is a unique and device-specific identifier present on iOS and Android devices, which provides a means to associate all behaviour on a device with the same profile—this cannot be done via cookies as apps will store cookies separately, or often not at all. In current versions of these operating systems, a User may reset this identifier to a new value.

The DSP exposes targeting functionality to an Advertiser through a web interface or API. Upon a successful impression (an instance of the ad being loaded on a User's device), the DSP reports this to the Advertiser, along with information such as the MAID. The DSP may provide user information as forwarded by the SSP, or maintain its own profiles on users based on previous campaigns.

Advertising privacy attacks

Much previous work on advertising privacy has centred around malicious ad content, usually using JavaScript to exploit poor permission design choices or vulnerabilities in the web-views provided by ad libraries, to break out of the sandbox of the library and extract sensitive information from a device.

The Advertiser provides the ad content ("creatives") to the DSP, either by upload or as a link to a URL. These can be plain text, HTML (with or without JavaScript), or video. Regardless of the format of the creatives, the targeting and impression information provided by the DSP are the same. This underlies the contribution from Vines et al. that targeting can be performed without active content.

The work by Vines et al. requires the knowledge of a user's MAID prior to tracking, and presents attacks that assume the regular use of certain apps in locations of interest. It remains unclear whether these attacks will generalise to ads served in webpages, and to what extent the impressions actually occur in naturalistic usage patterns.

This investigation intends to evaluate the potential of mobile ads as an attack surface under realistic threat models—this will begin with a characterisation of the frequency of ad impressions and their contexts, and as a result attacks can be made with the consideration of any restrictions or tradeoffs discovered. Methods of learning information on a user without prior knowledge of the MAID will also be explored.

A.1.3 Methodology

The investigation will explore the attack surface of mobile advertising through the running of ad campaigns through DSPs. These campaigns will have varying targeting options—targeting users by location, interest, MAID, or any other differentiators, to understand the varying reliability and frequency of data collection.

As detailed in the previous section, the DSP provides statistics on impressions (reporting an impression around ten minutes after the fact, per Vines et al.), and the MAIDs of devices on which impressions occurred.

The general statistics on impressions and their times will provide a general picture of the total volume of data provided to advertisers, while drilling down on the impression data associated with specific MAIDs will provide a characterisation of the ability to track or learn about specific users. Such analysis will be performed using the MAIDs of users that have consented to the study and provided us with the MAIDs of their devices (see Consenting Participants below).

To verify the reliability of the data returned to the advertiser by a DSP, we will require some degree of self-reporting from consenting participants. This reporting of impressions will be compared to the DSP's reporting. Self-reporting will also provide a ground truth for user location when testing location targeting.

Ad campaigns

Multiple ad campaigns will be created with the DSP, with varying targeting parameters. The creatives for these campaigns will also vary, to test the impact of creatives on impressions, and also to test the capabilities of the creatives. Examples of these different creatives are: - creatives containing JavaScript to consenting users to attempt to breach sandboxes as detailed in previous work, - HTML ads with multiple buttons, each of which sends a different request to the advertiser's server, expanding the return bandwidth for StegAd communication.

As advertising networks will often reject creatives if the campaign does not seem to be a *bona fide* attempt to advertise a brand or product, the creatives will advertise this study itself, inviting users who click-through to learn about the study, and if desired participate further through providing their MAIDs.

Individuals subject to data collection

There are two classes of individuals to whom ads in this investigation may be shown to, and thus whose impressions will be recorded by the DSP. These classes, and their data and its utility, are detailed below.

Consenting participants We will be recruiting participants, who will give consent to our storing of MAIDs associated with their devices, and our performing specific analysis on the data we collect associated with those MAIDs. We will be able to target their devices directly, constructing patterns of impressions from our campaigns.

The consenting participants may also be asked to participate in self-reporting, keeping records of when they have seen our ads, and reporting their location at the time of impression. They will also be asked for certain demographic information, such as age, sexual orientation, and interests.

The consenting participants may be asked to participate in controlled experiments, such as visiting a certain webpage or using a certain app in a specified location, and at a specified time.

At any point, participants may withdraw from the study, and request that we delete any data associated with them. At this point we will delete all analysis on our systems (and, wherever possible, on the DSP's system) pertaining to their MAIDs, as well as the data collected from self-reporting.

"Drive-by" participants Some of the ad campaigns we will run will use very coarse-grained targeting, and as such ads will be shown to members of the public who have not consented to participate in the study.

The data from these users will be of use as part of statistics of impressions, to understand in a general sense the magnitude and distribution of impression events. We will not individually analyse any of the data associated with an MAID from such users who have not explicitly given consent. A more thorough discussion of the risks of data collection on these users and the mitigations we intend to enforce can be found later in this document.

A.1.4 Hypotheses

The hypotheses we intend to test are as follows:

ADINT: 1. Ads can be reliably directed at a specific device or person, and only at that person. 2. Ads purchased through a DSP can be tailored—through targeting and bidding—such that they are shown frequently in natural use by regular users. 3. Fine localisation of an impression can be achieved using ad campaigns with overlapping location targeting. The three above hypotheses will be tested by analysis of impression data from ad campaigns targeted directly at consenting users. The targeting criterion will be the user's device MAID, and their location (for hypothesis 3).

StegAd: 1. A user can transmit information beyond the one-bit impression by interacting with an ad. 2. A user exhibiting a pre-agreed behavioural pattern can be found using ADINT without prior knowledge of their MAID.

These two hypotheses will be tested by controlled experiments. Testing the first hypothesis will require experimentation with different creatives, to understand the scope of possible user interactions with the ads. Testing the second hypothesis will require asking a user to exhibit various patterns of behaviour and self-reporting the ads shown.

A.1.5 Risks of data collection and precautions to be taken

As noted previously, where experiments require coarse-grained targeting, ads will inevitably be served to members of the public. The ads will advertise the study itself, so users will be informed of the campaign through the content of the creatives. As much information on the nature of the data collection effort will be conveyed in the creatives as possible. However even in this case the user will not be explicitly giving consent, and we cannot know that they have read whatever information is included in the creative. As a result, information will be returned via the DSP on impressions associated with non-consenting users.

To minimise the incidence of such impressions, ad campaigns will be targeted directly at consenting participants wherever possible. For the campaigns that are not directed only at consenting users, any information beyond the fact of impression will not be stored on our systems, and no individual analysis of "drive-by" users will be performed. Where patterns of user behaviour across ad campaigns are analysed, these will always be done in the aggregate, and only data associated with consenting users will be extracted for individual inspection.

On the consenting participants, we will be collecting sensitive data, either from self-reporting or inferred through ad campaigns. Such sensitive data will always be stored on secure, encrypted systems managed by our group, with access restricted only to us. Participants will be able to request the deletion of this data at any time.

A.2 Ethics committee application

Title: Constructing ADINT attacks: evaluating ad-based surveillance
Applicants: Jovan Powar, Alastair Beresford, Stan Zhang, Diana Vasile
Email: jsp50@cam.ac.uk, arb33@cam.ac.uk
Dates: 23/02/2018 - 31/03/2021
Study type: controlled experiment
Funding body: EPSRC

A.2.1 Description

This study intends to investigate the extent to which the purchase of mobile advertising inventory can be used to surveil and track the location of users?introduced in recent work as ADINT. We intend to evaluate the attack surface under realistic and concrete threat models. We hope that understanding the capability of this attack surface will enable further work on advertising privacy.

A recent work by Vines et al. showed that the purchasing of mobile ads could be used to track the location of a user on whose device the ads were shown, without the need for dynamic content in the ad creatives, or a click-through on the ad. This class of attacks, dubbed ADINT, leverages the rich data collected on a user by advertising networks and assembled into a profile, and the many targeting options available to an advertiser?including the retargeting of specific user profiles.

The work by Vines et al. requires the knowledge of a user?s unique Mobile Advertising Identifier (MAID) prior to tracking, and presents attacks that assume the regular use of certain apps in locations of interest. It remains unclear whether these attacks will generalise to ads served in webpages, and to what extent the impressions actually occur in naturalistic usage patterns.

Our evaluation of ADINT under realistic threat models will begin with a characterisation of the frequency of ad impressions and their contexts?by considering any restrictions or tradeoffs discovered, the class of feasible attacks will be better understood.

The investigation will explore this attack surface through the running of ad campaigns through Demand-Side Providers (DSPs). These campaigns will employ varying targeting options?targeting users by location, interest, MAID, or any other differentiators, to understand the varying reliability and frequency of data collection.

The general statistics on impressions and their times will provide a picture of the total volume of data provided to advertisers, while drilling down on the impression data associated with specific MAIDs will provide a characterisation of the ability to track or learn about specific users. Such analysis will be performed using the MAIDs of users that have consented to active participation in the study and provided us with the MAIDs of their devices.

To verify the reliability of the data returned to the advertiser by a DSP, we will require some degree of self-reporting from consenting participants. This reporting of impressions will be compared to the DSP?s reporting. Self-reporting will also provide a ground truth for user location when testing location targeting. This self-reporting will be in the form of making records of when they have seen our ads, and reporting their location at the time of impression. Participants will also be asked for certain demographic information, such as age, sexual orientation, and interests.

The consenting participants may be asked to participate in controlled experiments, such

as visiting a certain webpage or using a certain app in a specified location, and at a specified time.

Participants will be recruited both personally and via the advertisements themselves, after being presented with the aims of the study, its background, the methodology used for data collection and analysis, and the precautions taken to protect their privacy.

A.2.2 Precautions

Where testing our hypotheses requires the running of campaigns using coarse-grained targeting, ads will inevitably be served to members of the public. As a result, information will be returned via the DSP on impressions associated with non-consenting users.

To minimise the incidence of such impressions, ad campaigns will be targeted directly at consenting participants wherever possible. For the campaigns that are not directed only at consenting users, any information beyond the fact of impression will not be stored on our systems, and no individual analysis of ?drive-by? users will be performed. Where patterns of user behaviour across ad campaigns are analysed, these will always be done in the aggregate, and only data associated with consenting users will be extracted for individual inspection.

The campaigns will advertise this study itself, inviting users who click through to understand more about the study, its intentions and background, its methodology, and the precautions we will have taken when dealing with sensitive data. Users may be invited to participate further by registering their MAID with us.

On the consenting participants, we will be collecting sensitive data, either from self-reporting or inferred through ad campaigns. Such sensitive data will always be stored on secure, encrypted systems managed by our group, with access restricted only to us. At any point, participants may withdraw from the study, and request that we delete any data associated with them. At this point we will delete all analysis data on our systems (and, wherever possible, on the DSP?s system) pertaining to their MAIDs, as well as the data collected from self-reporting.

Appendix B

SoK

B.1 Dataset features

Sensitivity

Values: {*vulnerable group, superidentifiers, not sensitive*}
We describe the sensitivity of a dataset according to the definition given in Section 3.2.2. In that definition we noted that a dataset may be considered directly or indirectly sensitive. However, the problem of deciding whether a dataset might indirectly leak sensitive information is hard, due to the broad scope of possible risks.

Therefore we note whether a dataset explicitly contains sensitive data, and further subcategorise this to illustrate the nature of the sensitive data: *vulnerable group* or *superidentifiers* (as in Section 3.2.2).

Linkability

In Section 3.4 we present a linkage risk analysis to be applied to a dataset to characterise the scope of potential linkage attacks on it under a range of threat models. This allows us to pick out the most successful, harmful, or likely attacks on a dataset, and so derive the general approaches to linkage attacks particular to that dataset that would be important to consider. This should inform a description of its linkability:

For each approach we note three features: the availability of the dataset linked against (the linkage dataset); the effort required to perform the linkage, ranging from purely mechanical—e.g. direct matching of identifiers or generic computation of statistical correlations—to manual investigation; and the numerical extent to which the data subjects can be identified.

For example, we might say that a dataset is vulnerable to *public-mechanical-subset* link-

age attacks—i.e. there exists an attack that links the dataset to a publicly available dataset, performs a simple mechanical inference on the resulting combined dataset, and successfully re-identifies a subset of data subjects. A rigorous definition of these features follows:

Linkability – *Availability of linkage dataset*
Values: *intrinsic, public, private, internal*

The procedural ease of obtaining the linkage dataset: encapsulates the data landscape surrounding the dataset and so the capabilities of an adversary. *Intrinsic* availability describes the case where access to the dataset trivially provides access to linking data; i.e. where applying inference techniques to an anonymised or pseudonymised dataset without using external data allows the re-identification of subjects.

Deciding whether a linkage dataset is *public*ly or *private*ly available is a question of technical, procedural, and legal restrictions which disqualify parties from access. If data can be obtained by any party, we consider it public, but if access is subject to significant restrictions, we consider it private. For example, most Twitter profiles would be considered public, as any party with internet access may view it. However, if a profile were restricted such that only other users that the owner follows can view it, we would consider it private.

Note that we only consider a barrier to operation a 'significant hurdle' when it is not generally achievable by all parties. For example, we consider data on a person's Facebook page that has been made visible to any Facebook user to be *public*, despite the fact that a party wishing to access the data is required to have a Facebook account. While this requirement is indeed a procedural and legal hurdle, we do not consider it a restriction as (by and large) any party is capable of fulfilling this requirement (anyone can create a Facebook account).

At the other end of the spectrum is *internal* availability. This is an extremely strong case of *private* linkage, where the linkage dataset is available to the data collectors only, e.g. through leveraging institutional resources. In many cases, this would be trivial and of no interest to a discussion of data privacy. A subject may usually assume that the conductors of the data study will log their IP address, but have a reasonable expectation that the IP address will not be included in the published dataset. But there may be cases where the user might expect (or the collector might like to argue) that the act of collection itself carries no privacy risk; in these cases the presence of possible *internal* attacks is a risk worth identifying.

Linkability – *Linkage difficulty*
Values: *mechanical, investigative*

Difficulty here describes the technical complexity of the linkage task, and can be thought of as a proxy for the difficulty of producing an algorithm that links the datasets. If a linkage is *mechanical*, it can be performed as a direct cross-referencing task, or an unsupervised statistical inference process. By contrast, an *investigative* linkage would be entirely hands-on, with a human operator tweaking and crafting matching algorithms to fit the data exactly.

There is, of course, considerable grey area between the extreme cases of mechanical and investigative linkages. This could be thought of as a spectrum, where the more intensively an algorithm must be tuned by an attacker, the closer the linkage would be to *investigative*. However, it is not clear what a useful parametrisation of this space would be, and so for now we choose the closer of the poles, *mechanical* or *investigative*, by a qualitative assessment.

Linkability – *Cardinality of re-identification*
Values: *full, majority, subset, individual*

The greatest set of subjects that can be re-identified. The strongest such re-identification (*full*) completely re-identifies all data subjects. *Majority* re-identification re-identifies more than half of the data subjects, usually indiscriminately, whereas *subset* re-identification reduces the anonymity of a particular class of subjects, those that match some criteria via a targeted inference. *Individual* re-identification describes the case in which analysis can pick out only a few individuals, either by specially-crafted analysis or due to the existence of certain high-entropy subjects.

Datatypes

Values: string
This feature is a qualitative description of the nature of the data, e.g. medical records, genomic data, smartphone logs, or user study questionnaires.

Source & Intent

We distinguish between the work that originally produced the dataset (the **source**) and the works expected to make use of the dataset (the **intent**). This describes the research context of the dataset—important to note as one would expect datasets from different disciplines to fall under different technical and ethical rules. We note intent separately from source as these may differ (health produced by a psychologist could be intended for use by neuroscientists or computer scientists), as might the eventual uses from intent (Device Analyzer is a prime example of a dataset whose uses went far beyond those envisaged). Both features are subcategorised as follows:

Research motivation — ***Values:*** {*academic, industrial*}

Organisation — ***Values:*** string
Discipline — ***Values:*** string

To ensure a workable scope we focus mainly on datasets produced by or intended for the field of computer science, but to preserve generality we attempt to capture a range as wide as possible of datasets pertaining to other disciplines.

Scale

(N, k), where the dataset consists of at least N data items on at least k subjects. It is unclear at this point in the survey whether lower bounds on these parameters should be imposed.

Permission given

Values: *volunteer, opt-in, opt-out, implied/derogation permission, none*

Describes the subject's perceived relationship with the collection. This ranges from entirely voluntary (*volunteers*), to no knowledge of or control over the collection (*implied/derogation permission* or *none*). Between these two extremes lies: *opt-in*, where the subject is already engaged in some way with the mechanism of collection, but their data is not contributed until they so choose; and *opt-out*, in cases where data is collected on a subject without their explicit permission—and perhaps not even with their knowledge—but they have the capability to withdraw consent and oblige the collector to delete their data.

The importance of this feature is due to legal and ethical considerations—if a dataset is obtained by opt-in or volunteer collection, terms of the contract may place restrictions on the sharing of the dataset with third parties. Conversely, as we found in the case of many health record-derived datasets, data is drawn from hospital records legally on the basis of legal derogations for health projects. This may impose legal and ethical constraints on analysis, as well as forcing us to consider certain threat models differently—in health datasets, private linkages (to other health datasets) would be a highly pertinent risk (as access to one NHS dataset implies access to others).

Age

Values: *historical, outdated, pertinent, recent, real-time*

We describe age not as a strict period of time but with respect to a notion of 'freshness'; we stratify datasets according to their power to provide personal data, so the groupings reflect the decay in 'informativeness' of a dataset conferred by its age.

Historical data is collected long enough ago that it would no longer be considered personal. *Outdated* data does confer some information about an individual that is true of them

in the past, and would be considered personal, but that is not necessarily indicative of current behaviour, such as a person's working hours from two decades ago. By contrast, *pertinent* data is not current, but confers personal information that is indicative of current behaviour—for example, drinking habits from two decades ago might confer information about alcoholism, and from this we may assume that this past personal information is also true today. *Recent* data directly confers pertinent information that is current, with *real-time* data a special case of this, in which the dataset may be updated during the time frame in which a third party uses it.

The motivation for these particular distinctions were driven by two main examples in our survey: one longitudinal study of working families in a particular village in the 1950s, and an 18 year longitudinal maternal health study at four hospitals in the USA. The questions posed by the first are: do we consider the subjects to be alive, and if so, is the data personal or indicative of current behaviour? Given the workers' ages we expected some to be still alive, and certainly their children. While certainly not indicative of current behaviour, the data may include workers' sensitive information (e.g. sexual or health history), though this would be highly unlikely. In this case, we realised the need to distinguish between outdated and pertinent, and if the questions not related to health were present, we would not even consider it sensitive, and the dataset would be most usefully classified as historical.

In the maternal health study, the questions about mothers' health, drug use, and highly detailed sexual history definitely contribute sensitive data, and would certainly be strong indicators for current behaviour—the dataset is certainly not outdated. The dataset was stratified into phases, where some subjects were studied longitudinally and some only at the time of the birth. There would thus be a distinction between longitudinal subjects, on whom very recent data was collected which could reasonably be expected to be current, while the 18 year time-gap for snapshot subjects would reduce this data to an indicator. Thus we came to the distinction between pertinent (snapshot subjects) and recent (longitudinal subjects).

Collection

Values: *first-party, third-party*

The methodology used to collect the data in the dataset. A distinction between *first-party* and *third-party* shows whether the data collector is the 'owner' of the means of data collection from the user's perspective, or obtains data from that first party—such as through a data sharing agreement, purchase, or the first party integrating a third-party logging SDK into their app.

Temporal scope

Values: *snapshot, short-term, long-term*

This feature describes whether the work that generated the dataset collected data at a certain point in time—*snapshot*—or collected data as part of a longitudinal study, either for a short period of time or a long period of time (*short-term* or *long-term*). The reasons for the distinction between longitudinal and snapshot data collection are clear—the richness and time-variance of longitudinal data hugely increases the risks of linkage—but through inspection of location datasets, we expect that short-term and long-term longitudinal study should be considered differently as well. Consider running data, such as Strava exercise logs. A single continuous-time run log would confer information about the subject's home location (and perhaps health), while a month of their run logs would be significantly more informative, giving information about their daily routine and multiple locations of interest. Thus we currently make this distinction, although it may later prove to be irrelevant when considering risks.

Continuing

Values: *continuing, closed*

If the collection is longitudinal (*short-term* or *long-term*), we note whether data is still being collected.

Market, governance, and norms

This feature is a qualitative one, and consists of a description of any existing markets that deal with similar data to the dataset, or data that serves one of the purposes of the dataset, as well as the norms and regulations surrounding the collection and usage of data of this kind. These are the non-technical factors used in the 'pathetic dot' (New Chicago School) model of regulation. This feature is useful in that it can capture economic incentives in play—this is clearly an important factor, as evidenced by the previous work on advertising. Capturing ethical norms can inform the choice of sharing mechanisms through user expectations. It also contributes information on linkage datasets—if there is a thriving brokerage market for this kind of data, we might consider similar datasets to be public rather than private.

B.2 Initial survey

The following table shows the initial datasets identified in the first phase of the survey and what data was collected on them, with entries in bold being groups of datasets, usually found via a dataset portal.

	Requirements			
Name	**Purpose**	**Dataset, tool, or platform**	**Nature of sensitivity**	**Linkability**
Device Analyzer	Academic	Dataset	Behavioural	Direct (weak)
AOL Query Logs 2006	Academic	Dataset	Behavioural	Derived
SNAP ego sets	Academic	Dataset	Social structure	Derived
SNAP email sets	Academic	Dataset	Social structure	Direct (weak)
SNAP location sets	Academic	Dataset	Behavioural	Direct
Netflix Prize	Industrial competition	Dataset	Behavioural	Derived
Flickr YFCC100M	Mixed	Dataset		
MIRFLICKR sets	Mixed	Dataset		
NUS-WIDE	Mixed	Dataset		
Yahoo! Music ratings	Industrial research	Dataset		
Yahoo! Messenger network flows	Mixed	Dataset		
Yahoo! Messenger User communication pattern	Mixed	Dataset		
Yahoo! IM Friend graph	Mixed	Dataset		
Yahoo! Property + IM usage	Mixed	Dataset		
Yahoo! IM protocol-level events	Mixed	Dataset		
Google RAPPOR	Mixed	Tool		
Strava Metro	Mixed	Dataset	Behavioural	Derived
National Health and Nutrition Examination Survey				
Open Humans	Academic	Platform	PII	Direct (weak)
23andme	Academic	Platform	PII	Direct
NY Taxis	Mixed	Dataset	Behavioural	Derived
US Census				
National Pupil Database	Public good	Dataset	PII	Direct
National Office of Statistics				
UCL JDIRL	Academic	Platform	?	?
UK Data Service (open and safeguarded)	Mixed	Dataset(s)	PII	Derived
UK Data Service (controlled)	Mixed	Dataset(s)	PII	Direct, derived
Nokia Mobile Challenge				
IMPACT datasets	Mixed	Both		
metadac	Academic	Dataset	PII (genetic)	Direct