The confounding problem of private data release

Graham Cormode[1]

1 **University of Warwick**
Coventry, UK
G.Cormode@Warwick.ac.uk

Abstract

The demands to make data available are growing ever louder, including open data initiatives and "data monetization". But the problem of doing so without disclosing confidential information is a subtle and difficult one. Is "private data release" an oxymoron? This paper (accompanying an invited talk) aims to delve into the motivations of data release, explore the challenges, and outline some of the current statistical approaches developed in response to this confounding problem.

1998 ACM Subject Classification H.1 [Models and Principles]: Miscellaneous—Privacy

Keywords and phrases privacy, anonymization, data release

Digital Object Identifier 10.4230/LIPIcs.ICDT.2015.1

1 Introduction

One can scarcely glance at the Internet these days without being overwhelmed with articles on the great promise of, and excitement surrounding, humanity's ability to collect, store and analyse data. Whether under the banner of "big data", "data science" or some other catchy phrase, data is the new sliced bread, and computer scientists are the new master bakers, in this half-baked metaphor. The list of activities that will benefit from this data-based perspective is lengthy, and has been enumerated enthusiastically in other venues.

The starting point for this article is that for this revolution to succeed, a vital component is the data itself, and in many cases this data derives from the activities of individuals who may be unaware of the manifold uses to which it is being put. Indeed, the cynic can view the era of big data as merely the second coming of data mining, rebranded due to the tarnish that this term carries with it. The original association of data mining, or data dredging as its detractors termed it, was of interrogating data in the pursuit of a purpose for which it was not originally collected.

Putting aside the statistical concerns around torturing the data long enough until it confesses, there are ethical and legal concerns that arise from this use of data. Data protection regulations mandate the ways in which data can and cannot be used, and specifically preclude some reuses of data beyond its original purpose. The individuals who have contributed to the data are naturally concerned about what can be learned about them from this process: either what information is revealed about them directly, or what can be inferred about them from the various metrics and measurements. Laws are increasingly being drawn to protect the data of an individual. The notion of 'privacy' is becoming recognized as a human right[1].

This appears to create a tension between the bright future of a data-driven society, and the dark Orwellian nightmare of a world without privacy. Since it was computer scientists and statisticians who were most guilty of creating the hype, they should feel some responsibility for resolving this

[1] http://www.un.org/ga/search/view_doc.asp?symbol=A/C.3/68/L.45/Rev.1

18th International Conference on Database Theory (ICDT 2015).
Editors: Marcelo Arenas and Martin Ugarte; pp. 1–11

Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

tension. This focuses attention on a solution that is primarily technical in nature, as opposed to one that is legal, social, or moral.

The solution that is most commonly suggested is seemingly simple: just anonymize the data before it is used. That is, make it impossible to determine the identity of individuals contributing to the data. Then it will be fine to go ahead with any and every subsequent piece of data mangling, since the association between people and their data has been broken.

The annoying flaw in this prescription is that the act of anonymization is far more intricate than one would ever imagine. The awkward history of attempts to anonymize data is littered with anecdotes of failure. These are sufficiently illuminating that they bear retelling, albeit with only light regard for historical accuracy.

I know what you tipped last summer

In 2014, the New York City Taxi and Limousine Commission released a dataset comprising information about taxi trips taken the preceding year[2]. This was in response to a Freedom of Information request. The data had the identifying information of the cabs "masked", so that the same cab had the same masked identifier throughout the data. But with a little effort, the masking function could be reversed, since it was performed by applying a standard hash function (MD5) to the number. A dictionary attack iterating over the moderate ($< 10^7$) number of possibilities was sufficient to reidentify the cabs. From this, various privacy breaches were possible: combining the pick-up time and location with the cab number obtained from press photographs, it was possible to identify where celebrities had traveled and how much they had tipped [3]; and by finding establishments of ill-repute frequented late at night, find trips from these to specific locations to identify the homes of their clients. ◀

Mass Data Release

When Massachusetts Group Insurance Commission released data on hospital visits of state employees, they performed due diligence by removing the names, addresses and social security numbers from the data. However, Latanya Sweeney was able reidentify a large fraction of individuals in the data based on other signals: specifically, date of birth, sex, and postal (ZIP) code. These three attributes—none of them identifying in isolation–turn out to be uniquely identifying for a majority of Americans [9]. Moreover, data linking individuals to their demographics is quite widely available. The result is that the supposedly private detailed medical data for many people could be recovered from the released information. ◀

Many other notable examples follow a similar pattern: the release of Internet search histories by AOL in 2006 [4]; the extraction of individual movie viewing histories from Netflix data in 2008 [5].

From these horror stories to chill the spines of researchers, certain patterns and themes can be derived:

- Attempts to release data are usually done with the best of intentions. These can include attempts to create useful data sets for researchers and the public, response to freedom of information requests, and business purposes (attempts to "monetize" rich data sets).

---

[2] https://archive.org/details/nycTaxiTripData2013
[3] http://research.neustar.biz/2014/09/15/
[4] http://www.nytimes.com/2006/08/09/technology/09aol.html
[5] http://www.nytimes.com/2009/10/18/business/18stream.html

- Those releasing data are not oblivious to the sensitivity of the data they are sharing: they make some efforts to remove or mask identifying data. However, these fail in what in retrospect appear to be obvious ways: the free availability of external information with which to join to reidentify, or trivial attacks on the masking function.
- The consequences vary: in some cases, a large fraction of individuals in the data can be reidentified, in others it is just an unlucky few. The current consensus seems to be that these are equally undesirable outcomes. Similarly, the nature of data does not affect the perceived severity of the breach. Even seemingly innocuous data sets (taxi trips or movie viewings) can inform on people's activities or beliefs that they would consider private.

It is worth noting that there is selection bias in these examples, and that there are many more releases of data which do not expose private information.

In response to these "surprising failures of anonymization" [7], there are a variety of possible responses. One is to despair of the difficulty of private data release, and to resolve to oppose any further releases. However, the various pressures, including the clarion calls from Governments and advocate groups to make data open, mean that data releases will continue to happen and grow as more data becomes available.

Equally pessimistic is to begin with the same premise, and instead to declare that privacy is an artifact of the past, which can no longer be attained in a world of Google and Facebook [6]. However, thus far society seems not to have abandoned its need for privacy.

Legal responses are a valid option, but mostly seem to provide some attempt at recompense after the fact rather than prevent, and at best may provide a sufficient penalty that those releasing data do so with more caution and control over its spread. The scepticism with which the computing community has viewed efforts such as the "Right to be forgotten" [7] to put the genie back into the bottle show that information spreads too widely and too quickly for the law to be an effective information removal implement.

Thus, providing tools and mechanisms to understand and assist the release of private data remains the main viable option to respond to these challenges. The computer science and statistical research community has risen to this challenge over the past few decades, by providing a vast literature on the topic. Nevertheless, the problem remains a difficult and confounding one, that will occupy researchers for years to come.

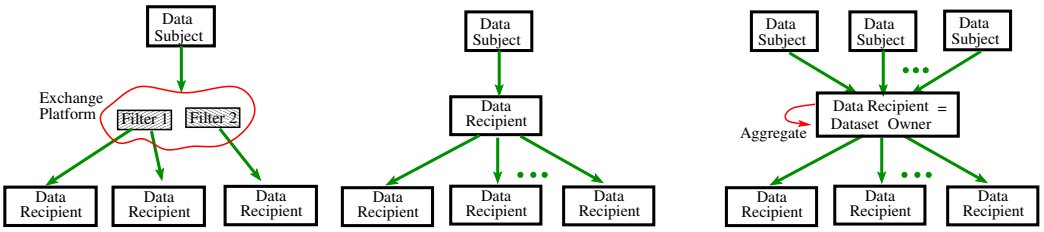
1.1 Outline

The remainder of this article attempts to touch on some of the technical issues surrounding privacy and data release. Section 2 delineates various privacy problems and their connection to other areas. Section 3 outlines some basic principles for working with private data and technical directions. Section 4 identifies some of the most interesting areas for research, and makes some initial suggestions for efforts here.

2 Privacy Preliminaries

In a typical privacy setting, there are a number of players, with possibly competing interests. There are *data subjects*, whose information is the subject of privacy concern. A data subject can be an individual or user, but it can also be a device, e.g., a computer or cell phone. The subject's information

[6] https://www.eff.org/deeplinks/2009/12/google-ceo-eric-schmidt-dismisses-privacy
[7] http://www.stanfordlawreview.org/online/privacy-paradox/right-to-be-forgotten



(a) First-person privacy. **(b)** Second-person privacy. **(c)** Third-person privacy.

Figure 1 Classification of scenarios raising privacy concerns.

may be released to specific *data recipients*. In some cases, this involves an exchange platform that typically provides both sharing and filtering capabilities. Examples include social networks and photo sharing sites. In other cases, it involves an intermediate entity that either re-shares the data directly (e.g., advertisement-supported sites), or aggregates it first and may then reshare it (e.g., stores, credit card companies, phone companies). The aggregator can end up collating large amounts of data about the demographics, habits, activities and interests of the data subjects, and becomes the *dataset owner* of this collection. Intermediate entities may choose to perform a data release and make some amount of information available to a data recipient, with or without notification to individual data subjects. However, such a release is still governed by legal obligations to data subjects (e.g., opt-in vs. opt-out), as well as good practice (a data release should not harm or upset the data subjects involved). Note that the information disclosed is based on the collected data, but may be modified in a number of ways. A data release can be addressed to a particular recipient, rather than being made generally available to all (a public release).

Within this scenario, there are three broad classes of privacy concerns. The classification is based on how close a data subject is to the final data recipient, and is illustrated in Figure 1. We distinguish between intermediate entities that re-share data at an individual level (Figure 1b) and those that aggregate it first (Figure 1c) because the privacy concerns are significantly different.

2.1 First-person privacy: Users (over)sharing their own data

First-person privacy issues concern the private data that a user shares with other entities (e.g., websites, social networks), and the rules and policies that determine who is able to see this information. This is illustrated schematically in Figure 1a. There have been a large number of headlines about privacy in recent years which ultimately derive from the difficulty of ordinary users to appreciate the consequences of their data sharing. These include stories about people sacked for inadvertently sharing their feelings about their manager directly with their management chain[8]; and a fugitive captured after revealing his whereabouts to his 'friends' who included a justice department official[9]. Examples like this arise because users imagine that their information is being shared with their "true friends", but in fact it may be shared with all people they have marked as "friends" through a system. When these two groups differ, such unintended consequences can occur.

There are several opportunities for research into ways to help users cope with these privacy problems. Efforts to date have included the provision of tools and warnings to users (from service providers directly or from third-party plugins) about the extent of their (over)sharing, and encouraging

[8] http://www.dailymail.co.uk/news/article-1206491/Woman-sacked-Facebook-boss-insult-forgetting-a
html
[9] http://www.guardian.co.uk/technology/2009/oct/14/mexico-fugitive-facebook-arrest

them to check before they post information.

2.2 Second-person privacy: information spreads fast

Second-person privacy issues arise when the data that a user has shared with one entity is then passed on to other entities, as shown in Figure 1b. Oversharing of this data can lead to privacy concerns. For example, AT&T researchers identified that when MySpace was connecting to external advertisers to place ads on a page, they were passing detailed private data about the user viewing the page direct to the advertiser[10]. This ultimately led to a binding settlement with the FTC monitoring MySpace's activities for twenty years[11].

There is much research potential here. Focus so far has primarily been on detecting when this happens and tracking the flow of information. A natural approach is to provide languages for users to express their privacy preferences about how and with whom their information can be shared; P3P can be seen as an effort in this direction from the start of the century [12], and "do-not-track" a more recent example on the web. Adoption of such methods have been limited thus far, due to implicit opposition from users (lacking interest in expressing their privacy requirements and difficulty in doing so), and service providers (whose business interests may rely on allowing as much sharing of data as possible). Future directions may be to more actively track information sharing, and the development of "peer-to-peer" data networks where encryption tools are used to control access to information, such as the diaspora social network [13].

2.3 Third-person privacy: private data release

Third-person privacy issues surround the practice of collecting large amounts of data about many individuals together, and sharing this with other entities, illustrated in Figure 1c. This can be in the form of a static data set, a live data feed, or via exposing an API for interactive interrogation. The goal of this activity is to provide general information about a large user base without revealing detailed information about any one individual. However, there is the potential for such data sets to inadvertently reveal private information. The high-profile failures of private data release outlined in the introduction all fall under the heading of third-person privacy.

2.4 Privacy, Utility and Trust

Guaranteeing the privacy and maintaining the utility of the released data are fundamentally opposing objectives. There are many possible compromise approaches, some favoring privacy over utility, and others the reverse. Where in this continuum one chooses to perform a data release is governed by the level of *trust* in the data recipient. Assessing trust is easier in some cases than others: For first-person privacy, data subjects assess the trustworthiness of their friends, who are the intended data recipients (however, pitfalls exist, as discussed above). For third-person privacy, data owners have various control levers over the data recipients (e.g., if the recipients are analysts employed by the data owner, they are subject to internal rules and regulations; if they are external partners, they are subject to contractual obligations). However, second-person privacy is less amenable to trust analysis, partly because of a lack of transparency in the data release process (see Section 4). Under all three models, understanding the data flow is essential to assessing trust.

[10] http://online.wsj.com/article/SB10001424052748704513104575256701215465596.html
[11] http://ftc.gov/opa/2012/05/myspace.shtm
[12] http://www.w3.org/P3P/
[13] https://github.com/diaspora/diaspora

2.5 Privacy Versus Security

Philosophically, privacy and security (as understood within computer science) have many tenets in common. However, there are essential differences. Security is primarily concerned with a binary decision: is access granted to a particular resource? For example, if a user has the right key they can decrypt a file and access its contents in full, otherwise they learn nothing. In privacy, the issues are more subtle: a data owner has detailed information about a collection of users, and must decide which data items, and in what form, they should be revealed to another entity. As such, the foundations of privacy technology are less mature, and less widely deployed than security technologies such as encryption.

3 Privacy Principles

Some necessary (but not necessarily sufficient) conditions for ensuring private data release include the following:

3.1 Protecting Personally Identifiable Information

When sharing data, it is important to remove information which can uniquely identify the data subject, unless this is absolutely needed. Examples of such "personally identifiable information" (PII) are well-documented, and include names, account numbers, license plate number, and social security numbers. In the communications domain, attributes such as IP address, MAC address, and telephone number are also considered PII.

A US FTC report[14] advocates that data be "de-identified". That is, all PII is removed prior to sharing. However, in some cases, it is necessary to "join" two data sets to collate data on individuals from different sources. If this cannot be done prior to release (e.g., if the two data sets are owned by different organizations), then more complex technical solutions are needed (see below).

3.2 Quasi-identifiers

A major subtlety of data release is that information which does not obviously qualify as PII may nevertheless be sufficient to identify an individual. For example, learning the zip (postal) code of an individual does not typically identify that person[15]: most zip codes contain around 10,000 households. However, if taken in combination with other attributes, this can become identifying, as discussed in the Massachusetts Group Insurance commission example.

There has been much work in the research world on how to anonymize data so that quasi-identifiers are not identifying. The work on k-anonymity tries to delete and reduce precision of information so that each individual matches against at least k entries in the released data [8]. The database community happily occupied itself for many years in generating new variations on k anonymity; the enthusiasm for this has waned since it was observed that k-anonymity/diversity does not necessarily provide very useful protection [5].

A different policy approach is to remove the "obviously" identifying fields (PII), and to ensure that what remains cannot be "reasonably linked" back to a specific individual or device[16]. The FTC's current standard for reasonability is qualitative, and lacking in examples, and so provides little actionable guidance for data release.

---

[14] `http://ftc.gov/os/2012/03/120326privacyreport.pdf`
[15] With some exceptions, e.g., 20252
[16] `http://ftc.gov/os/2012/03/120326privacyreport.pdf`

3.3 The data minimization principle

A basic concept in data sharing is the data minimization principle: it should reveal no more information than is needed for the task at hand. Clearly, this can guide how much information to delete or mask prior to release. However, putting this into practice can still be challenging. In particular, it is hard to fully anticipate what information could be of use for the data recipient, and it is too easy to fall into the trap of including data "just in case" it is needed. Moreover, the default option is often to allow data to remain in place: it takes an active decision to remove or modify the values. In these cases, it is helpful to remember the value of data to its collator: great effort is often expended in collecting and curating rich data. It is therefore incumbent on organizations to avoid freely giving such wealth to others without good reason.

3.4 Data Correlations

An additional risk to privacy arises from the accretion of data about individuals. That is, if a data set includes a lot of information about someone's activity collected over a long period of time, this can build up into a picture that is unique, and identifying. In the AOL example, it was the large collection of search terms that helped to identify certain users, and hence learn about other searches that they had made. Modern communications and internet applications are a particularly rich source of information about individual's interests and views. Relevant data can include internet browsing history, phone call activities, and set-top box TV data. Even if attempts are made to mask these (e.g., by suppressing or hashing phone numbers), these patterns can quickly become unique for most users; moreover, a determined entity could observe a targeted individual in order to find their entries in the data. As with quasi-identifiers, there may be no reasonable technical provision which can preclude such a determined effort to re-identify while still providing the desired functionality. However, it is possible to ensure that such efforts become costly to enact, and that casual inspection of the released data does not allow easy identification: a far weaker standard than the goal of perfect privacy, but perhaps a more reasonable one for data that is released to a single party rather than to the world at large.

3.5 Aggregations and Differential Privacy: Safety in Numbers

A natural way to improve the privacy of data is to provide it in aggregated form. That is, instead of reporting the raw data, just provide statistics on groups of the data. For example, instead of releasing full lists of phone calls made on a mobile network, one could compute just the number of calls and average call length (etc.) per account. Or, customer information could be aggregated up to the neighborhood level, rather than at the household level. Great care and thought is still required: information still leaks when some groups are allowed to be small, or when the behavior within a group is uniform.

Within the privacy research community, the concept of "Differential Privacy" is close to such aggregation in spirit [4]. In its most common form, differential privacy typically computes aggregate statistics over grouped data, and adds statistical noise to further perturb the result. In practice, it may be sufficient to rely on aggregation alone, but augmented with suppression of small groups: the uncertainty in which individuals contribute to the data is sufficient to provide the perturbation necessary. Use of aggregation is one technique specifically mentioned in the FTC report (see footnote 14).

3.6 Privacy Checklist

The following checklist is an attempt to articulate a set of questions that should be answered about any planned use of private data:

- What is the data that will be used?
- What are the different fields in the data?
- Which are the uniquely identifying fields?
- How was the data obtained? What control did users have over the inclusion of their data?
- How much data is there? At what rate is it produced?
- Who is the intended recipient of the data?
- What is their intended use for the data?
- What other (linkable) data sets would they have access to?
- What contractual obligations will the data recipient be placed under?
- What transformations will be applied to the data to ensure privacy? Who will perform each step?
- Can each of the data elements be justified or can the list be shortened?
- Will all PII be removed from the data prior to release?
- What partially identifying information will remain in the data? How easy would it be to identify an individual from this?
- How much data will there be on each individual? Will this allow correlation attacks?
- What are the consequences of re-identification? What are the possible harms that could result?
- What are the benefits of the use of the data contrasted against the potential risk of re-identification?
- What are the benefits to the user to share the data?
- Who will ensure that the privacy procedure will be adhered to? Can this procedure be audited?

4 Privacy Pinch-points

The discussion thus far has ranged broadly over different types of data and different data release settings. The subsequent sections consider some specific applications and techniques in more detail. The cited work here is heavily biased to the author's recent research.

4.1 Location and Mobility Data

Data about people's location, gathered from GPS devices and mobile phones, is increasingly available. This gives insight into the distribution of people, but also their movements. The possible applications suggested for mobility data are many and varied: urban planning, dynamic advertising, road traffic analysis, emergency management and more. At the same time, it is understood that the detailed location of an individual is very sensitive: their presence at a particular medical facility, say, may be very private. Even coarse location data is sensitive: an individual may not wish it to be learned that they were far from where they said they would be. Longitudinal location data is also particularly susceptible to correlation attacks of the kind described above: observing someone's location late at night typically identifies a "home" location, while location in the middle of the day identifies a "work" location. This can isolate an individual, and then reveal everywhere else they go.

Consequently, great care is required in releasing location data. Raw trajectories of movements over extended periods reveal too much. Instead, different approaches are needed. These can include: (1) Demographic snapshots. Describe the demographic occupancy of grid-cells of sufficient size, e.g., the (approximate) number of people there; the gender and age breakdowns, etc. [2]. (2) Short trajectories. Describe the detailed movements of (anonymous) individuals for short periods of time. It must be made difficult or impossible to "sew these back together". (3) Density maps. The approximate

locations of an identified sub-population can be revealed at regular (e.g., hourly) intervals. Each of these brief outlines needs further research to refine into a specific, robust, procedure.

4.2 Joining private data sets

Much value in working with data comes from the ability to join together multiple data sets, and hence to learn from the combination. For example, a telecoms provider might wish to study the impact of call drops on customers' usage by joining logging data on call drops with billing data. This becomes problematic to achieve under privacy, since such joins are best performed making use of a unique identifier to isolate records corresponding to a particular entity ("the key"); however, such unique identifiers are typically considered PII. Moreover, several attacks on privacy have occurred due to the possibility of joining a private data set with a public one.

There are several natural approaches to dealing with joins over private data: (1) The (trusted) data owner performs the linkage, and then drops the uniquely identifying attributes from the resulting joined data set, before releasing it. (2) Appropriate "hashing" (using a secret 'salt' value[17]) to replace occurrences of the key in both data sets. Then they can be joined using the hashed key value, rather than the true key value. (3) Both data sets can be entrusted to a trusted third party, who will perform the linkage, and return the results to the data user. It is important that the data recipient cannot easily compare the joined output to the input and so reidentify the source of some data items. Each one of these approaches necessitates some amount of trust between the parties. There are cryptographic protocols for performing joins without revealing which items matched, but these are considered slow and costly to put into practice.

4.3 Synthetic data sets

One approach that can significantly enhance the privacy of a dataset is to generate a synthetic dataset that mimics certain properties of the original, but contains made-up entries that are generated according to some model [6]. A synthetic dataset is designed such that specific tasks can be performed over it with sufficient accuracy (e.g., analyzing traffic patterns), but will most likely introduce large errors in other, unrelated types of analysis.

Generating synthetic data may seem very different in nature to anonymization of data, as they start from opposite extremes. Data anonymization is often viewed as starting with the original private data in full, and chipping away at it by removal and coarsening of information, whereas synthetic data generation may be seen as starting with nothing, and creating a new data set by sampling from an appropriate statistical model whose parameters are derived from the full data. However, this can also be viewed as a spectrum. One perspective on private data release is that it should be viewed as designing an appropriate model for the data, the parameters of which are learned from the data, and which is rich enough to generate faithful data.

This approach is of particular value when combined with models such as differential privacy. Applying differential privacy to the function $f(x) = x$, i.e. trying to simply release the input data in full, can be seen as a trivially complex model, where the parameters describe the data in full. The effect of differential privacy in this setting is simply to add noise that drowns out all signal in the data. At the other extreme, a simple model of the data, described in terms of sums and averages across all individuals (say) can be obtained very accurately through differential privacy, but may only describe the data poorly. Abstracting from these extremes, the difference between the input data and the released data can be broken into two pieces: the model error (the noise introduced by fitting the

[17] It was a lack of salt that made the NYC taxi data so easy to reidentify.

data to a model) and the privacy error (the additional noise added to the parameters of the model to provide a privacy guarantee). Much recent work in private data release can then be viewed as trying to find appropriate models for data so that the model error and noise error can both be contained [3].

4.4 Graph Structured Data

One aspect of "big data" is the variety of forms that the data can arrive in. Different types of data require different approaches to allow private data release. An important class of data is that which can be represented in the form of a graph, such as the pattern of interactions between individuals. This provides a suitable target: a problem simple enough to state, yet complex enough to give pause, and flexible enough to model a number of different settings. The reasons that graph data presents a challenge for data release hinges on the fact that typically an individual will correspond to a node in the graph, and the associated information (edges) can be quite substantial. Finding a suitable representation of the graph data so that appropriate statistical noise (say) can be added to mask the presence of an individual while preserving properties of the graph has so far eluded researchers.

4.5 Inference and Privacy

One of the reasons that private data release remains a confounding problem is the difficulty in pinning down a suitable definition. Lacking a precise definition of what properties a private data release should satisfy, it is possible to be fooled into believing that stronger guarantees result. A case in point is the ability to draw strong conclusions about individuals from the released data. One might assume that if data is released under an appropriate privacy model, then it should not be possible to infer supposedly private information about individuals in the data. However, this is often the case, under a variety of privacy models [5, 1].

The reason is that effective classifiers can be built for data where the parameters of the classifier depend not on the behaviour of any one individual, but collectively on large groups within the population. Data released under privacy often preserves statistics on large groups – indeed, this is very much a requirement for utility. Consequently, it is possible to build accurate classifiers for seemingly private information. Applying the classifier to individuals (either from within the data or drawn from a similar population) leads to accurate inferences about them.

4.6 Data-as-a-Service

The concept of data-as-a-service (DaaS) is a powerful one: since companies have access to much data of interest, it should be possible to monetize this, and license access to other organizations who would like to make use of it. Here, privacy concerns come to the fore. It is vital to ensure that detailed customer data is not released as part of DaaS: revealing who a business's customers are, let alone what they are doing, would be viewed as a serious privacy breach. Multiple privacy techniques may be needed to ensure that such transactions can proceed effectively. Specifically, all identifiers should be removed, and it should be ensured that there are no shortcuts that would allow identifiers to be restored. The minimal amount of information should be included, and the contribution of one individual to the data should be limited, to reduce the chances of exploiting data correlations to re-identify a person. Ideally, data would be provided in an aggregate form, possibly with some random noise added and small counts suppressed. Even then, the privacy analysis should take into account the possibility of linking the data to other external sources, and thus establish the potential risks of this.

5 Conclusion

Enabling the release of private data remains a fundamental challenge. Guaranteeing privacy and being able to share useful data stand in fundamental opposition: the only way to provide perfect privacy is to entirely prevent all access to data, and the only way to ensure full use of the data is to make no attempt to address privacy concerns. Nevertheless, there can be workable compromises that provide a reasonable level of privacy against re-identification while enabling legitimate data uses. This article has attempted to outline the different ways in which privacy can be at risk, and discussed principles and ongoing efforts to find workable solutions. Despite the many horror stories and conceptual challenges, there remains optimism that suitable technical solutions can be found to all the promise of big data to be realized while providing strong and effective privacy protections for all.

Acknowledgments

I am indebted to Balachander Krishnamurthy, Magda Procopiuc and Divesh Srivastava for many lengthy and detailed discussions at AT&T Labs–Research around the topic of privacy. These discussions developed many of the perspectives on privacy promulgated in this paper, and led to notes on which this article is based (perhaps explaining the telecoms bias in some of the examples chosen). I similarly thank my many other collaborators with whom I have worked on topics in privacy over the years. These include Smriti Bhagat, Xi Gong, Xi He, Zach Jorgensen, Ninghui Li, Tiancheng Li, Ashwin Machanavajjhala, Entong Shen, Tanh Tran, Xiaokui Xiao, Ting Yu, and Jun Zhang. I also thank the many other researchers in the privacy community with whom I have discussed ideas and algorithms over the years.

My work is supported by a Royal Society Wolfson Research Merit Award, and European Commission Marie Curie Integration Grant PCIG13-GA-2013-618202.

References

1 Graham Cormode. Personal privacy vs population privacy: Learning to attack anonymization. In *ACM SIGKDD*, August 2011.

2 Graham Cormode, Magda Procopiuc, Divesh Srivastava, and Thanh Tran. Differentially private publication of sparse data. In *International Conference on Database Theory*, 2012.

3 Graham Cormode, Magda Procopiuc, Divesh Srivastava, Xiaokui Xiao, and Jun Zhang. Privbayes: Private data release via bayesian networks. In *ACM SIGMOD International Conference on Management of Data (SIGMOD)*, 2014.

4 Cynthia Dwork. A firm foundation for private data analysis. *Communications of the ACM*, 54(1):86–95, 2011.

5 Dan Kifer. Attacks on privacy and deFinetti's theorem. In *ACM SIGMOD International Conference on Management of Data*, 2009.

6 Ashwin Machanavajjhala, Daniel Kifer, John M. Abowd, Johannes Gehrke, and Lars Vilhuber. Privacy: Theory meets practice on the map. In *IEEE International Conference on Data Engineering*, 2008.

7 Paul Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57(6):1701–1778, August 2010.

8 Pierangela Samarati and Latanya Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI, 1998.

9 L. Sweeney. Simple demographics often identify people uniquely. Technical Report Data Privacy Working Paper 3, Carnegie Mellon University, 2000.