Understanding the Data Environment

DOI:
10.1145/2508973

Document Version
Accepted author manuscript

Understanding the Data Environment

A better understanding of the conditions and mechanisms under which data privacy and anonymity may be threatened is key to protecting data privacy and anonymity.

Elaine Mackey and Mark Elliot

The *data environment* is a new concept in the field of data confidentiality. Although there have been references to its various aspects, manifestations, and impacts, it is only now that it has become a focus of enquiry in its own right. It is a focus, we would argue, that is long overdue and rather urgent given the manner and pace in which the data landscape is evolving. The huge amounts of data being generated, combined with the economic drivers and political will to share it more widely, means that concerns about data privacy and anonymity are ever more founded. Here, we explain why we need to understand the data environment in order to minimize threats to data privacy and anonymity.

When we talk about protecting data privacy and maintaining anonymity in the data confidentiality field, we are in essence talking about ensuring that anonymised data remains anonymous once it is shared, disseminated, and released in the data environment. So what does this actually mean in practice? To answer this, we will first discuss data and anonymisation, as this will set the scene for what we really want to discuss, the data environment.

Data and Anonymisation

All organisations will collect some information from their customers/clients/service users as part and parcel of their organisational activities. Almost always, this will include classic identifiers such as client's names, addresses, and contact details. However, the information that is collected is often more extensive than this; for example, schools also collect data on their pupils' exam scores, special educational needs, and health, the police also collect data on crime and anti-social behaviour, and retailers also collect data on shopping and leisure habits, finance, employment status, and occupation. This information will in all likelihood be stored in databases that hold very many individual level records of information.

This data is termed *personal data* which, as described by the UK Data Protection Act (DPA, 1998), is "data that relates to living individuals who are or can be identified from the data". Organisations that want or need to share and disseminate their data for secondary use are obliged under the DPA (1998) to process the data in such a way as to render it anonymous and therefore no longer personal. The transforming of data from personal to anonymous requires that identifiers are removed, obscured, aggregated, and or altered in some way. There are two types of identifiers that organisations need to think about when processing data: formal identifiers and complex identifiers. Formal identifiers are relatively easy to spot and deal with and include data such as a subject's name, address, and unique reference numbers (e.g. their social security number or National Health Service number). Complex identifiers are less easy to spot and deal with. They could in principle include any piece of

information (or combination of pieces of information). For example, take age and marital status; considered in the abstract, they are not immediately obvious identifiers. But, if we consider the case of an eighteen year old widow, our implicit demographic knowledge tells us that this is a rare combination (at least in peace time). This means that such an individual could potentially be re-identified by, for example, someone spontaneously recognising that this record corresponded to their friend/neighbour/colleague/family member.

Just this data complexity problem alone means that anonymising data is not straight forward. To complicate matters further, organisations preparing data for dissemination don't just have to think about sufficiently anonymising their data, but also about retaining data utility. After all, there is little point in sharing and disseminating data that doesn't represent whatever it is that it is meant to represent (because it has been altered during the anonymisation process).

Because anonymisation is difficult and has to be balanced against data utility, the risk that a re-identification will happen will never be zero. In other words, there will be a risk (although extremely small) of de-anonymisation present in all useful anonymised data. The only way to remove this risk entirely is not to share any data at all, which is obviously undesirable if we are to exploit the undoubtedly huge social and economic value locked up in the data.

Statistical Disclosure

For researchers in the data confidentiality field, the first step to determining how best organisations can minimise the risk of de-anonymisation and optimise the trade-off they must make between anonymisation and data utility is to assess how the process of de-anonymisation might actually occur. The term commonly used in the field to denote the process of de-anonymisation (and one that we will use from here on in) is '*statistical disclosure*'. A statistical disclosure, we should point out, incorporates not just the idea of de-anonymisation (or re-identification), but also captures the idea that confidential information is revealed (or disclosed). See Duncan et al. (2011) or Hundepool et al. (2012) for recent reviews of the *statistical disclosure control* field.

Formally, we describe a statistical disclosure as a form of data confidentiality breach that occurs when, through statistical matching, an individual population unit is identified within an anonymised dataset and/or confidential information about them is revealed. However, determining how a statistical disclosure might actual occur and then play out is not straight forward. This is the crux of the problem: as it stands, we know little about the factors, conditions, and mechanisms involved in a statistical disclosure largely because we know little about the *data environment*. We will give a technical description of this term shortly; for now, consider it as the context for any piece of data, without which that data has no meaning.

You may wonder why it is only now that attention is being directed towards the data environment. After all, it would seem like an obvious point of focus given the task in hand. The explanation for this lies with: (i) the particular perspectives that have underpinned and

informed data confidentiality work and (ii) the intractability of understanding and gathering data from the data environment.

The traditional perspective was one where statistical disclosure risk was seen as originating from, and therefore largely contained within, the data to be disseminated, released, or shared. It meant that data researchers and practitioners rarely looked beyond the statistical properties of the data in question. More precisely, it meant that they did not concern themselves with issues such as how or why a data intruder might make a disclosure attempt, or with what skills, knowledge, or access to other data they would require to ensure their attempt was a success. As a consequence, the statistical models they built to assess disclosure risk, whilst statistically sophisticated, were based on very crude assumptions about the context of the risk that they were trying to model. To address these failings, there has been a broadening of perspective in the last twenty years which has seen attempts to incorporate some context beyond the data itself. This has usually taken the form of *intruder scenario analysis* which has shifted attention away from the traditional position of asking *how risky is the data for release* to a more critical position of asking *how a statistical disclosure might actually occur.* Some inroads in addressing this latter question have been made, most notably the: (i) development of a framework for identifying plausible intrusion scenarios and (ii) identification of sets of key variables, i.e., information that can be used for statistically matching one dataset with another (see Elliot and Dale, 1999). But, for all intents and purposes, this is where the work has stalled not least because much of it is theoretical.

It is certainly true that we lack a real world view of statistical disclosure and have relatively little direct data on it. This may be because an act of statistical disclosure is a rare event and or is one which the key protagonists (i.e., the *data intruder* and the organisation releasing data) are both incentivised to conceal (albeit for differing reasons). It is difficult to speculate productively on this and we do not do so here. The important point we wish to make is that whilst there is little direct data in the form of cases of disclosure, it does not mean that there isn't any (key) data; the data environment can potentially tell us all we need to know about how a statistical disclosure might actually happen.

The Data Environment

The data environment is made up of a small number of components: (i) data, (ii) agents, and (iii) infrastructure. It is these components that we need to look at in order to ascertain how a statistical disclosure might occur and play out.

Data

What (other) data exists in the data environment? This is what we need to know in order to identify what data (key variables) are risky, i.e., can be used for statistically matching one dataset with another thereby providing (some of the) conditions for statistical disclosure. This is still a developing area which, at Manchester University, we have been pushing forward. Our *Data Environment Analysis Service* (a bespoke service for the Office for National Statistics) has involved developing a methodology for investigating, cataloguing,

categorising, and documenting data in the data environment. However, this methodology is operationalized manually and is therefore constrained in its scope and ability to deal with the complexity of the problem. The next step is to develop a methodology for automating these processes in order to enable comprehensive capturing of the global data environment.

Agents

Who are the key protagonists, and how might they act and interact to bring about a disclosure event? What might the consequences of this be? There is no risk of statistical disclosure without human action. This may seem an obvious point, but it is one worth emphasising. As of yet, we know little about the key protagonists and how they might interact. We are currently working on an approach using a game theoretic reasoning to develop greater insights into how agents might act and interact strategically, within specific contexts, to create a statistical disclosure (see Mackey and Elliot, 2009).

Infrastructure

How does infrastructure and wider social and economic structures shape the data environment? Infrastructure can be best thought of as the set of interconnecting structures (physical and technical) and processes (organisational, managerial, contractual, and legal) that frame and shape the data environment. It provides the context to data and agents, so, for example, it will influence what data is shared, to whom it is given, and how that process takes place. It will also influence key agents, such as National Statistical Institutes, any organisation releasing data, data users, specialist interest groups, the general public, and the media in terms of their possible actions, interactions, and counter responses. Infrastructure includes storage systems, information systems, data security systems, governance structures, and national and international legislation.

Thus far, we have talked about the data environment in the definite singular; in other words, we are referring to the global system. But the global system has internal structure; it is partitioned and that partitioning creates many local environments. For example, a secure data centre can be termed a discrete data environment: it has context-specific physical, technical, organisational, and managerial structures that determine what data goes in, how data is stored, processed, and risk assessed, and in what format data comes out, who the user community is, and how they can interact with that data. By defining and regulating a local environment, the data owner/controller can render data anonymous that would, "in the wild", not be. So when we share data we are in effect moving it from one environment to another. Data environments can be looser in form than the secure data centre. An environment might be defined purely by regulation and licensing. For example, a community of allowed users might have access to data, and that community and the instruments that define what can and cannot be done with the data comprise the environment. Such an environment cannot be as tightly controlled as the secure data centre environment, but it does allow for some control of the data environment not (currently) present when for example data is published on the internet.

All data environments will contain the features outlined above but are likely to differ in form depending on how they are made up and how they are operationalized. A local data environment may in turn contain sub-environments. For example, an organisation may have multiple servers with differential access. Individual machines themselves are then sub-environments.

Note that all sub-environments are, to some degree, permeable since users move in and out of them with knowledge of the external environment. So, even if I am a bona fide user of an environment acting in a legal and compliant fashion, my knowledge of the external environment might cause spontaneous de-anonymisation (e.g., if I recognise the eighteen year-old widow as my neighbour).

In conclusion, understanding and building models of the data environment is of paramount importance if are to continue to protect data privacy. The reason for this is that we can only effectively guard against the threat to data privacy and anonymity when we have a clear idea of what it is we are guarding against. Whilst we have some understanding of the factors, mechanisms, and conditions under which data privacy and anonymity may be threatened, the long and the short of it is, at this present time, we do not know enough about them. Further work on this topic is both necessary and urgent.

References

Duncan, G., Elliot, M. J. and Salazar, J. J. (2011) *Statistical Confidentiality*. Springer, New York.

Hundepool , A., Domingo-Ferrer J., Franconi L., Giessing S., Schulte Nordholt, E., Spicer K., and de Wolf, P-P. (2012) *Statistical Disclosure Control.* Wiley, London.

Elliot M. J. and Dale, A. (1999) 'Scenarios of Attack: The Data Intruder's Perspective on Statistical Disclosure Risk.' *Netherlands Official Statistics*. Spring 1999. pp 6-10.

Mackey, E. M. and Elliot, M. J. (2010) 'The application of Game theory to disclosure events.' *Proceedings of UNECE worksession on Statistical Confidentiality, Bilboa, December 2009*.

Authors

Elaine **Mackey** is a well established researcher into the broader aspects of statistical confidentiality where the statistical, data management and social policy meet. Her PhD demonstrated the value of using game theory to map disclosure attack scenarios. She has recently worked as part of the Data Environment Analysis Service mapping the data that an attacker might feasibly use to identify individuals in anonymised datasets.

Mark **Elliot** has an international reputation in the field of data privacy. He has led numerous interdisciplinary projects in the field and his special uniques methods are at the centre of the SUDA system for anonymisation decision support developed at Manchester and used in statistical agencies across the world. He has a long track record of relevant stakeholder

engagement most recently through his work on the Administrative Data Liaison Service (www.adls.ac.uk) and as lead for the UK Anonymisation Network (www.ukanon.net).