The Social Data Environment

Mark ELLIOT and Elaine MACKEY
University of Manchester

Abstract. In previous work, we have introduced the data environment as a powerful explanatory concept in the realm of data privacy, with specific regard to the concepts of anonymisation and statistical disclosure. Here, we explain the concept more fully and examine how the socio technical system, that is titled social media, embodies and transforms the social data environment. We draw on both a social philosophy and sociological framework, but populate that framework with both agents and artefacts. We argue, that the data environment concept is an important organising principal, which has implications in epistemology and ontology, and for how we understand notions like disclosure, identity, society and perhaps most critically, privacy.

Keywords. data environment, social data environment, privacy, agency, identify, disclosure

Introduction

In previous work, we have introduced the data environment as an explanatory concept in the realm of data privacy, with specific regard to the concepts of anonymisation and statistical disclosure [1-4]. We describe the data environment as the context for any item of data. More precisely, we define it as: the set of formal and informal structures, processes, mechanisms and agents that either: (i) act on data; (ii) provide interpretable context for that data or (iii) define, control and/or interact with that data [4]. The argument, which we have put forward in this earlier work, is simply that one cannot decide whether or not an individual’s privacy is compromised by a piece of data, without knowing the context for that data, i.e. the data environment. If the data environment is unknown, with regards to what other data is available and to whom, and how and why it might be used to compromise another’s privacy, the protection of data privacy is made a great deal less precise. The practical upshot of this is that there is a potential for protection measures to be inadequate (as the Netflix and AOL cases referred to in the chapter attest to) or over cautious, leading to valuable data not being released or the value in the release data being unnecessarily reduced. The idea that you need to understand what is going on in the data environment in which you share or release data, in order to negate (privacy) risk more fully, is becoming more popular, but the stumbling block is that, at present, very little is known or even understood about it. To attend to this, the first step is to say what it is – to describe the data environment. This is what we do here, by way of identifying and explaining the important concepts related to it.

In this chapter, we explain the data environment concept more fully, and in doing so examine how the socio-technical milieu, known as social media, embodies and transforms
the social data environment. We draw on both a social philosophy and sociological framework to examine the relationship between agents, data and structure within the social data environment. We argue that the data environment is an important organising principal that has implications for how we understand notions such as identity, society, disclosure and indeed privacy.

The remainder of the chapter is divided into 5 sections. In Section 2, we provide an overview of the basic concept of the data environment. In Section 3, we discuss the importance of agency – the capacity of an agent to act in the world. Our previous discussions of the data environment have acknowledged the importance of agents in their constitution. However, we have been relatively circumspect on the role of agency in the formation and dynamics of data environments. In Section 4, we describe the set of structural, functional and emergent properties along which data environments vary. We use this framework to expand our previous use-limited notion of the data environment. In Section 5, we discuss how social networks are data environments. In the social media context, the data environment has some rather distinct properties, particularly the significant role played by agency (whereby the data subject and data generator are one and the same). In this section, we will describe social media/networks as data environments within the framework developed in Section 2, and then using this, describe how the global data environment is affected by their existence and use and the impact they have on identity and data choices. We close with some speculations about the future.

1. The Concept of Data Environment

The data environment is a relatively new concept. It was originally introduced, in order to formalise some very difficult problems, within a sub-field of data privacy known as statistical disclosure control, and we begin here by describing that genesis.

In a global sense, the data environment is becoming ever richer and more complex and concomitantly, concerns about data privacy have also increased. Within the data privacy field, when we talk about protecting privacy, we are in essence talking about ensuring that anonymised data remains anonymous once it is shared or released within or into a data environment. The de-anonymisation problem was graphically demonstrated by Narayanan and Shmatikov [5], who showed, that it was possible to identify individuals (and to infer sensitive information about them) within the Netflix prize dataset (which Netflix had released publically and claimed was anonymised).

All organisations will collect some information from their customers/clients/service users as part of their organisational activities. Almost always, this will include formal identifiers such as client’s names and addresses, but the information collected is often more extensive than this; demographic information, health data, attitudes, movement data and so on are all collected by different organisations for different purposes. This information will in all likelihood be stored in databases that hold very many individual level records of information.

These data are examples of what is termed personal data, which, as described by the European Data Protection Directive (95/46/EC) and similar legislation, as data that: relates to living individuals who are or can be identified from the data either alone or in combination with other data. Organisations that want or need to share and disseminate their data for secondary use (and who do not have explicit consent) are obliged 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 or altered in some way.

For researchers in the data privacy field, the first step to determining how best organisations can minimise the risk of de-anonymisation, is to assess how a de-anonymisation might actually happen. The term commonly used in the field to denote the process of de-anonymisation is ‘statistical disclosure’. 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 (see [6,7] for reviews of this area).

Determining how a statistical disclosure (data confidentiality breach) might actually occur and then play out is not straightforward. This is the crux of the problem: little is known about the factors, conditions and mechanisms involved in a statistical disclosure, largely because we know little about the data environment. The traditional perspective on statistical disclosure risk saw it as originating from, and therefore largely contained within, the data to be disseminated. This was largely because understanding how the data environment impacts risk is a very hard problem. As a consequence though, 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 models that have been built to assess disclosure risk, whilst statistically sophisticated, have been based on very crude assumptions about the context of the risk they were trying to model. The classic example of this was the release of search queries by AOL, which were pseudonymised, yet left people clearly identifiable with the addition of some common sense inferences; for example, if someone persistently searches for the name of a non-famous individual, it is likely to be the person themselves (see [8]).

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 orthodox position of asking how risky is this data?, towards a more critical position of asking how might a statistical disclosure occur? Some progress in addressing this latter question has 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 [9]). As a result, organisations have increasingly used simulation exercises, getting hackers to try and re-identify data under controlled conditions. This is an important step, but like all testing methodologies, it suffers from the contingent nature of the attack (how competent is the hacker? how realistic are his constraints?) and a lack of rigour, and for all intents and purposes, this is where the work has stalled. The aim of the data environment idea, then, is to support and enable further and more rigorous work in this area.

Most recently, we have extended this through the realisation that it was meaningless to attempt to assess whether data are anonymised or not, and therefore personal or non-personal, without knowing what other information is or could be co-present. This is explicit in the definition of personal data in law, but rarely do professionals talk about whether data is personal or not in anything other than an absolute sense. This, in turn, is based on the mistaken belief that anonymisation can be absolute. Unfortunately this
is not so, as Ohm states simply: you can have [absolutely] anonymised data or you can have useful data but not both [8]. The point, here, is that to ensure that nobody could be identified under any conceivable circumstances from within an anonymised dataset, one would have to mangle the data so that it bore no relation to the original data. So, for anonymisation to mean anything at all, one has to specify the circumstances and thus the only sensible answer to the question ‘is this personal data?’ is ‘in what context?’ or more specifically ‘in what data environment?’

This line of reasoning naturally leads to the general question of how we might formalise our notion of the data environment to allow such questions to be answered. Formally, we posit that any data environment is made up of four components: data, agency, governance processes and infrastructure.

(i) **Data** – what (other) data exists in the data environment? How does it overlap with or connect to the data in question? 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 that we have been pushing forward under the data environment framework.

(ii) **Agency** – we consider agents as capable of acting on as well as in the data environment. It may seem like an obvious point, but it is one worth emphasising – there is no risk of a privacy breach without human action. What is needed, is a way of putting the agent at the centre of the analysis, and one approach for this is Game Theory. We are currently using this approach within the data environment framework to develop greater insights into how agents might act and interact strategically, within specific contexts, to create a privacy breach [10].

(iii) **Governance processes** – We use the term here broadly to mean how the users’ relationships with the data are managed. This includes formal governance (e.g. laws, data access controls, licensing arrangements, and policies which prescribe and proscribe user behaviour) through *de facto* norms and practices to socio-cognitive user sets (e.g. risk aversion, prior tendency towards disclosure etc.).

(iv) **Infrastructure** – we use this term to consider how infrastructure shapes the data environment? Infrastructure can be best thought of as the set of interconnecting physical and technical structures and processes (organisational, managerial, contractual, legal) that frame and shape the data environment. This includes infrastructure such as information systems, storage systems, data security systems, authentication systems and platforms for data exchange.

For further discussion of this framework, and a description of one possible methodology for using the concept of the data environment and an understanding of its components to make inferences about whether data is anonymised (and therefore non-personal), see [11]. However, we will now move on and broaden our discussion of the data environment concept by introducing the notion of agency into the discussion.

---

2This is not the same as saying that anonymisation is not useful. The deadbolt on my door will not help if a burglar comes armed with a battering ram or simply smashes my living room window, but that doesn’t mean that my lock is useless; merely that it does not provide absolute security.

3We do not concern ourselves here with the content/topic of the data. This is a contestable exclusion. Data sensitivity, for example, is a feature which would arguably colour the environment that the data sits in. However, we are primarily concerned with defining the data’s context and then considering how the data relates to that context and, therefore, we do not consider the data themselves in defining those data’s environment.
2. The Importance of Agency

In the foregoing, we have acknowledged the importance of agents in the constitution of data environments. Here, we want to consider the role of agency in the formation and dynamics of data environments.

The question of agency is a complex and potentially troublesome one for the designers of intentional data environments. Intentional data environments tend to be based on tight governance and rigid infrastructure and, in that context, agency can appear like noise – and potentially damaging noise at that.

Clearly, agency is much more than just noise in an otherwise orderly data environment. It is the central component of the dynamics of any data environment. Data environments exist because we are informational beings inhabiting an informational society. To be human is to exchange information – to disclose. We agree with Webster [12] that the notion of the information society as a new development is misleading. All societies are axiomatically information societies. Accepting that, then the only meaningful questions are: ‘what sort of information society do we have and what sort do we want?’ Underlying this are conceptualizations of the relationships between (informational) agents and data environments, and key to this is privacy. The operation of informational privacy amounts to choices about disclosure, contact and engagement, and fundamentally it is about control of information flows between oneself and others. It is therefore critical to understanding the information society that we have a clear ontological position on privacy.

There is a conceptual interdependency between informational privacy, disclosure and identity. Without reference to identity, neither privacy nor disclosure are meaningful concepts. Disclosure is my representation of (information about) myself to others; privacy is – in part – the control of that representation. These are intentional operations in informational space or, in our terms, they are the operations of agents within data environments.

Whilst again acknowledging Webster’s fundamental point, we would argue that through the process that began with the invention of the printing press and continues today with the move online, these identity-privacy-disclosure processes are shifting from socio-cognitive to socio-technical. Our data now has technological substrate which both mirrors and shapes our offline identities. If one accepts (after Mead [13], Goffman [14], Berger and Luckman [15] etc.) that our selves do exist in our (symbolic) interactions with other selves, then we should also accept that those identities now extend to our formal data. ‘Who we are’ now has a socio-technical component: our cyber identities.

There is a clear relationship here with the notion of social machines coined by Berners-Lee and Fischetti [16] and now explored with some vigour within the burgeoning field of web science. Both there and here, the view is of an increasing connection between humans and their artefacts that is producing both a new humanity and new environments. Arguably, social machines are simply a description of the predictable outcome of agents interacting with data environments. The underlying point, here, is that without agency the data environment is simply an empty structure – a dead artefact. Agency gives data environments meaning and movement. Further, this dynamic interaction and the power of agents to disclose (or not) creates the possibility of differentiation of the data environment of variations in type and form.
3. Features and Types of Data Environments

In Section 1, we used ‘data environment’ in the definite singular; referring to the global information system. However, this global system has multiple layers/internal structures; it is partitioned and that partitioning creates a multitude of interrelated and potentially overlapping local environments. For example, a data lab can be termed a discrete data environment; it has context-specific physical, technical, organisational and managerial structures which determine what data goes in, how data is stored, processed and risk assessed, and in what format data comes out, as well as who the user community is and how they can interact with that data etc. By defining and regulating a local environment, the data owner/controller can render data anonymous, which ‘in the wild’ would be de-anonymised. So, when we share data, we are in effect moving it from one environment to another. Most data environments are looser in form and very much less controlled/controllable than the secure data lab. 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. On the other hand, it might be defined by less formal structures, such as group norms and shared practices. Hence, judgements about data sharing in a particular data environment need to take into account these definitions.

Thus data environments are not a homogenous class, but it seems likely that they vary across a defined set of properties which relate to the components described in Section 1: data, agents, governance and infrastructure. These components have two immediate implications:

Sub-Partitioning: Just as the global environment may be partitioned, so might local environments. 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 sub environments. Some of the properties of a sub-environment will be defined by the parent and some will be defined locally.

Boundary permeability: All data environments are – to some degree – permeable, if for no other reason than by virtue of the simple fact that users (agents) move in and out of them and users have knowledge of the external environment. Boundary permeability means that data, and indeed agents, will move in and out of the environment. The degree of boundary permeability will depend upon infrastructure and governance. We suggest that with both stronger governance and infrastructure, you will have less boundary permeability. In other words, in environments that have strong external controls, there will be less movement of agents/data in and out of that environment – the secure data lab is an example of this. Let us for a moment consider the role of infrastructure here. Infrastructure, we suggest, has a dual impact, in that it can both support the exchange of information, and also control/constrain it, as in the case of the secure lab scenario. So the converse – weak governance and infrastructure – is unlikely to lead to high boundary permeability, i.e. a more fluid and free exchange of information between environments, because although there may be few governance restrictions there will also be little infrastructure to support the exchange of information. Where you might get high boundary permeability is where you have low governance and high infrastructure; where the infrastructure supports the exchange of information and not its control.

We can use these features to describe the state of affairs, within a data environment, at any one time. Although described thus far as static features, data environments are
obviously not static; they will change over time as data and agents flow in and out, governance changes and new infrastructure is built. Two key questions that follow from this are: how do they start and how are they shaped over time?

**Emergent or Intentional:** A local environment may exist precisely because it is set up to be so, or it might emerge as a consequence of other processes. A research data lab, or a bank’s setting for its customer accounts, or an HR department’s record system are all examples of intentional data environments. At the other end of the scale, an informal network of friends or a family – essentially any set of agents with shared knowledge and understandings – are examples of an emergent data environment.

**Self-organising or Controlled:** The second dynamic feature is whether an environment is self-organising or controlled. Within a controlled data environment, its information flows and its development will be determined by one or more agents and/or imposed governance. Legal instruments are one form of control. Organisational policy or strategy is another. Self-organising data environments processes and development will not be determined, but rather arise from the (creative) interactions of agents.

These properties are clearly empirically related. An emergent data environment is more likely to be self-organising and an intentional one more likely to be controlled. However, it is likely that all environments will be a blend of properties, and this blend will itself change over time. The Internet itself has arguably moved from a self-organising environment into a more controlled one. Note here, that simply because an environment is self-organising and emergent, this does not mean that it has no impact on agents. However, the nature of the influence will be different and this will tend to be reflected in the blend of infrastructure.

### 3.1. Enabling and Restricting Infrastructure

Infrastructure, in effect, creates the medium of the data environment. In terms of impact on the manner in which agents operate within the environment, infrastructure, as we have previously suggested, comes in two distinct types. Both affect what happens within the data environment and thus the form that the data environment takes.

The first type, *enabling infrastructure*, comes in many forms: communication or storage media or the software for producing the desired functionality. Its primary purpose is to ensure that the environment delivers its intended or desired primary purpose. However, it does more than this: enabling infrastructure shapes the data environment. It impacts heavily on what can be done in the environment and how agents behave. The point here, is closely related to Marshall McLuhan’s reflection in his most famous work *Understanding Media*:

In a culture like ours, long accustomed to splitting and dividing all things as a means of control, it is sometimes a bit of a shock to be reminded that, in operational and practical fact, the medium is the message. This is merely to say that the personal and social consequences of any medium – that is, of any extension of ourselves – result from the new scale that is introduced into our affairs by each extension of ourselves, or by any new technology [17, p. 7].

---

4Indeed, as Feltz et al. [25] observe, the two concepts (emergence and self-organisation) are often dealt with as one.
The second type of infrastructure is *restrictive infrastructure*. It affects the data environment by controlling the behaviour of agents within the environment. It channels behaviour and places constraints on what agents can or cannot do. Security systems and access or functionality restrictions are all examples of this type of infrastructure. It is not concerned with the delivery of the environment’s primary purpose. A simple example is the requirement that a user enter a username and password in order to enter the environment. Governance, when enacted, most often reinforces the effects of restrictive infrastructure.

So, enabling infrastructure shapes, and restricting infrastructure controls.

**4. Social Networks as Data Environments**

Consider a coffee shop: we might think of this setting as the enabling infrastructure for a social network type of data environment. That is, it is designed to enable social interaction where disclosure of information between agents takes place. They have the shaping property that we introduced above – the layout will promote particular group dynamics; the style of seating and the lighting will lead to more or less relaxed discussion and so forth.

However, these environments lack something that might perhaps be an implicit feature in an orthodox understanding of what a data environment is: storage. True, the conversations will be remembered by the participants. Disclosed information will be retained to a greater or lesser degree in their memories. But when the agents leave the environment, the storage has walked out with them (and its contents are likely to have already begun a process of degrading). The coffee shop does not retain any information.\(^5\)

Let us examine this situation in a little more detail. Each agent, in the situation described, is themselves a data environment. When friends gather in the coffee shop, they are, in effect, (heavily bounded) sub environments of the coffee shop environment. On the other hand, the coffee shop without the agents is not a data environment at all; it is simply the enabling infrastructure for one. So the agents’ presence in the coffee shop turns what was simply some infrastructure into a data environment. When the agents are present, the group conversation and acts of disclosure are, in effect, the moving of data between (sub) environments.

Now consider social media. On the surface, they have many of the properties of the coffee shop. Without users, social media are simply the enabling infrastructures for a data environment, but not actual environments. Once they acquire users, then they become data environments. However, social media have something that the coffee shop lacks – storage. Unlike the coffee shop, the conversations and disclosures are retained. Another aspect of social media, that differentiates them from the coffee shop, is that they are much more intentional. This is a matter of purpose: the coffee shop’s *raison d’être* is the sale of coffee (and related foodstuffs), and its role as a transient data environment is very much secondary. Social media are intentional data environments: their primary function is processing personal information.

So what do social medial data environments look like? The first point of note is that this type of data environment covers a vast range of activities from social networking

---

\(^5\)A coffee shop could have CCTV of course, and some undoubtedly do, but for the sake of illustration we will say ours does not.
sites, blogs and wikis to photo and video sharing platforms. So this is not a single type of environment, but a class of environments with both shared and singular properties. However, the shared properties are noteworthy.

Firstly, like the coffee shop, the data subjects and the data generators are one and the same. Chatfield notes, ‘we are all authors today as well as audiences – not to mention our own part-time publicists, social secretaries, agents and ambassadors’, [18]. We argue that this is actually not new; we have always been engaged in all these activities. In the coffee shop, we could see individuals engaged in all of those roles in their social interactions. It is just that in the social media data environment, the medium in which we are doing our authoring is a less ephemeral one. Secondly and distinct from the coffee shop, the potential audience for our social media activity is vast – possibly even global.

The social media data environments highlight the issues of privacy and disclosure, perhaps more than any other. Due to the vast and unfettered exchange of personal information online, some have suggested that privacy is no longer relevant or important. Hyatt for example declared, ‘for all practical purposes, privacy is dead’, [19, p. 166]. Hyatt is not alone in this view; Zuckerberg also talks about the age of privacy being over. In an interview in 2010, Zuckerberg argued that the world had become less private and more public, and if he could design Facebook over again, users’ information would by default be public [20].

The argument underpinning this is: if there are concerns about privacy, it is not reflected in online behaviour or, more specifically, in the level of social media membership, which has not declined [21]. This is sometimes referred to as the privacy paradox. However, as Barnes argues, this is a rather simplistic reading of the situation. Rather than reflecting a lack of concern about privacy, social media participants do not fully understand the public nature of the Internet and its implications [22]. In short, some agents within the social media data environment behave as if they are in their local coffee shop – beguiled, perhaps, by the superficial similarity with that environment and not fully realizing that they are in a data environment that both remembers and broadcasts.

However, more importantly, the ‘privacy is dead’ argument represents a fundamental misunderstanding about the nature of both privacy and the historical role of social media. An analogy to illustrate this: when I step outside my front door, I enter a public space. If I live in a city, I might, within minutes, have been seen and possibly heard by thousands of other people. So, by stepping outside my door and allowing others to see me, am I giving my consent to them to do whatever they wish with my socio-physical form? Of course I am not. I would expect that certain social norms would be followed by anyone who wished to interact with me. By the same token, the assumption that posting something online is one and the same thing as saying, ‘I am giving everyone, everywhere, the right to do whatever they want with this’ is a crude one at best. The Internet is still very young, and the norms associated with the social media data environment are evolving and will continue to do so. Privacy, too, will be changed by the medium, but the question is not whether we have privacy or not, it is what sort of privacy we will have. Privacy is a process not a state, and it is a process that has manifested itself in very different forms in different cultures and at different points in the development of our species. However, it has always been present.
5. Concluding Remarks

In this chapter, we have demonstrated the importance of the data environment for analysing the nature and likely implications of the relationship between human agents and data (about humans). The concept, we believe, is a useful one for understanding human informational interactions, and also a medium for making sense of other complex ideas, like personal data and anonymisation. In other work [11] we describe a formulation for using the data environment concept to make practical judgments about these socio-legal concepts.

In a broader sociological sense, it should be clear that we regard the ontological separation of human agents and the data about them as increasingly problematic. Our socio-physical and digital selves are increasingly intertwined, and this process shows no sign of slowing. Indeed, one does not have to be a full-on singularitarian like Ray Kurzweil [23] to acknowledge that humanity appears to be on a trajectory that will see it merge, at least to some extent, with its digital artefacts. In describing the impact of the related phenomenon of self-quantification through digital technology, Swan remarks:

The individual body becomes a more knowable, calculable, and administrable object through QS activity, and individuals have an increasingly intimate relationship with data as it mediates the experience of reality [24, p. 85].

These processes will undoubtedly have implications beyond anything we can foresee. But one thing we do feel certain about is that the data environment – here a useful abstraction – will increasingly become as real as the socio-physical environment that human agents mostly inhabit today, and that what we put into the data environment, and what goes on in it, will be as important in determining the quality and meaning of our lives as the contents and activities of our socio-physical lives.

References