Data release and privacy: An algebraic theory for combining and linking data

Liang-Ting Chen, Markus Roggenbach, and John V. Tucker

Department of Computer Science, Swansea University

The vast stores of data built up by governments, agencies, institutions and companies in the course of their operations hold information of value in diverse and unexpected situations. Some governments have launched initiatives to encourage bodies to share their data with other organisations and the public. For example, in the UK, there are several national and local registers and a plethora of statistical data that are now widely shared. The UK's Open Data Initiative demonstrates the ambition to publish internal government data as open data sets. There are many patterns of data sharing, of which three are particularly important: *a)* making data public—data release into the wild; *b)* data sharing by contract with a data analysis organisation; and *c)* data sharing with delegation to a new data controller for further onward sharing. However, data custodians have a legal duty, and a social duty of care, to ensure that privacy is not breached by the release of open data sets.

The technical question arises: What information is revealed by, or can be inferred from, the data? Naturally, prior to its release, a data set can be filtered and anonymised but *a)* anonymisation is difficult and often flawed; and *b)* data from various other sources can be combined with a given data set to reveal much more. There are many data sources to call upon, and many unknown unintended consequences in making data publicly available.

An early example is Sweeney's finding [2] that 97% of voters in Cambridge, Massachusetts, USA, can be uniquely identified by combinations of birth dates and postcodes, and these can be further joined with a hospital discharge database to discover individuals' medical history, for example, the governor of Massachusetts, William Weld, at that time [3].

Lately, Narayanan and Shmatikov [1] devised an algorithm exploiting sparsity to combine datasets. As a case study they analysed the Netflix prize dataset and found '84% of (Netflix) subscribers present in the dataset can be uniquely identified if the adversary knows six out of eight movies outside the top 500' that the subscriber rated. Such source of film ratings may come from social engineering or the Internet Movie Database (IMDb). In response to these privacy concerns, Netflix decided to withdraw the datasets. Unfortunately, they are still available to download using BitTorrent or archive.org.

In this paper we take a fresh look at the challenge of combining data sets and linking pieces of data. Our aim is to develop tools to analyse formally the abstract structure of data sharing, and technical issues of policy specification and compliance. To this end, we develop the notion of a *data representation algebra*, whose operations combine two or more pieces of data from potentially different sources to form data with higher information content.

Data ordering. The domains from which the data is released often come with a hierarchy. Take for example the domain of postcodes. In the UK they consist of four components (e.g., the postcode of our university, SA2␣8PP, is composed of area SA, district 2, sector 8, and unit PP). For the sake of anonymization, often only a prefix postcode is released. This results naturally in a preorder representing the amount of information that was released: $\epsilon \preceq$ SA $\preceq$ SA2 $\preceq$ SA2␣8 $\preceq$ SA2␣8PP. Note that in a number of relevant examples this *data ordering* fails to be antisymmetric. The relation $x \cong y \iff x \preceq y$ and $y \preceq x$ expresses that x and y hold the same information; it is an equivalence, so we can speak about the *information classes* of a data domain. This allows us to distinguish between representation and information content of data.

Data representation algebra. Data from the same domain can be *combined* for the purpose of gaining more information, e.g., for de-anonymization. Often, such data originates from different releases. Combination is a partial operation as not all information is consistent, e.g., a person's main address can't be both SA1 and SA2; however, the information SA**PP and *2** (where * represents 'hiding' a part of a postcode) can be combined to SA2*PP. This holds more information than the single pieces of data in the information order consisting of postcodes P with anonymizing stars, where $[\![P]\!] \coloneqq \{\, p \in \mathbf{Post}_{\mathrm{UK}} \mid P \text{ results from hiding parts of } p \,\}$ and $P \preceq Q \iff [\![Q]\!] \subseteq [\![P]\!]$. Note that the domain of postcodes P with anonymizing stars includes the element **** which holds no information at all. We call the structure $(M, \preceq, \oplus, 0)$ a *data representation algebra*, where M is the data domain equipped with an information order $\preceq$, a partial combination operator $\oplus$ that is associative, commutative, and compatible with $\preceq$, i.e. $x_1 \preceq x_2 \Longrightarrow x_1 \oplus y \preceq x_2 \oplus y$, and a unit element 0. The operations lift to the information classes to make an *information algebra*. Both algebras turn out to be ordered partial commutative monoids.

Data linkage. Data of different kind can be *linked*, again for the purpose of gaining more information. Consider, e.g., a non-empty set of suspects with their hiding places $U \subseteq \mathbf{Pop}_{\mathrm{UK}} \times \mathbf{Addr}_{\mathrm{UK}}$ and a non-empty set of house addresses and their owners $V \subseteq \mathbf{Addr}_{\mathrm{UK}} \times \mathbf{Pop}_{\mathrm{UK}}$. Their combined information may represent pairs of suspects, addresses, and house owners who possibly provide shelters to suspects. In order to capture such data combination, we develop the formal definition of a *linkage passage*.

In our paper we develop a comprehensive algebraic theory of data and information algebras by providing extensive motivational examples for the chosen axioms and studying their properties.

References

1. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: 2008 IEEE Symposium on Security and Privacy. pp. 111–125. IEEE (May 2008)
2. Sweeney, L.: Weaving technology and policy together to maintain confidentiality. The Journal of Law, Medicine & Ethics 25(2-3), 98–110 (Jun 1997)
3. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(05), 557–570 (2002)