dataprotectionlaw&policy

FEATURED ARTICLE

cecile park publishing

Head Office UK Cecile Park Publishing Limited, 17 The Timber Yard, Drysdale Street, London N1 6ND
tel +44 (0)20 7012 1380 fax +44 (0)20 7729 6093 info@e-comlaw.com
www.e-comlaw.com

The complexities of defining personal data: anonymisation

Privacy professionals have long debated the complexities of anonymisation when assessing the boundaries of personal data. Now, with the review of the EU Data Protection Directive and of US privacy practices, the debate has been ignited again. Omer Tene, Associate Professor at the Israeli College of Management School of Law, examines how the conundrum arose and practical solutions that can be adopted.

The EU Commission, the Council of Europe, the Organisation for Economic Co-operation and Development and the US Government are currently reviewing the legal framework for data protection and privacy, which dates back to the 1980s and 1990s. One of the fundamental issues concerns the definition of the most basic building block of the data protection framework - that of 'personal data' (in Europe) or 'personally identified information' or PII (in the US). A narrow definition of personal data may fail to account for the increasingly sophisticated means of re-identifying apparently anonymised or pseudonymised data sets. At the same time, an overly broad definition would expand the framework, making it unworkable.

Article 2(a) of the EU Data Protection Directive (95/46/EC) defines personal data as 'any information relating to an identified or identifiable natural person', where 'an identifiable person is one who can be identified, directly or indirectly'. Recital 26 of the Directive states that 'to determine whether a person is identifiable, account should be taken of all the means likely reasonably to be used either by the controller or by any other person to identify the said person' and that 'the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable'. Individuals are considered 'identifiable' even though they have not yet been identified but it is possible to do so.

In the past, data that were rendered anonymous or pseudonymous, encrypted or key-coded, were considered exempt from the scope of the Directive as long as the data subjects were 'no longer identifiable'. However, technological developments including advances in analytics, once the domain of national security agencies but increasingly available off-the-shelf for use by individuals and small businesses, and the massive increase in computing power and data storage capacity have also undermined the effectiveness of de-identification techniques. Moreover, individuals are now sharing increasing amounts of personal data online, thus facilitating the linkage of data across multiple sources and reducing the prospects of stable de-identification.

In an influential 2010 law review article, Paul Ohm observed that 'clever adversaries can often re-identify or de-anonymize the people hidden in an anonymized database...Re-identification science disrupts the privacy policy landscape by undermining the faith that we have placed in anonymization'[1]. In computer science circles, the prospect of re-identification has been known for years. In 2000, Latanya Sweeney demonstrated that merely three pieces of information - ZIP code, birth date and gender - are sufficient to uniquely identify 87% of the US population[2]. De-anonymisation of seemingly anonymous databases was more recently shown by researchers who were able to identify a large proportion of anonymised Netflix subscribers by matching data in their movie ratings against an additional online database[3].

In another case, two *New York Times* reporters were able to sparse out the identity of an AOL user, whose online search queries were anonymised and posted on an AOL research website[4]. The Netflix researchers, Narayanan and Shmatikov, argued that 'the amount of perturbation that must be applied to the data to defeat our algorithm will completely destroy their utility for collaborative filtering.' Hence, they contend that data are either useful or truly anonymous - never both.

These findings have significant implications for policymakers. On the one hand, critics claim that any legal framework which exempts data that are anonymised or pseudonymised, encrypted or sharded (broken up into fragments stored on different equipment in different locations) is faulty of emphasising form over substance. Surely, legal protection must depend on the risk of privacy harm, not on whether data are directly or indirectly linked to a name. Indeed, behavioral targeting companies or pharmaceutical manufacturers typically do not care or need to know the name of a data subject whose data they collect; instead they seek to find and identify specific profiles which are then targeted with relevant treatments, content or ads. In a 1980 *Yale Law Journal* article, Ruth Gavison characterised precisely such activities, which commodify individuals treating them as 'profiles' or 'numbers', as an infringement of privacy and human dignity. In its 2007 working paper on the concept of personal data, the Article 29 Working Party (WP29) echoed this approach, stating that 'while identification

through the name is the most common occurrence in practice, a name may itself not be necessary in all cases to identify an individual.' On the other hand, protecting data based on a remote risk of re-identification could lead to gross over-expansion of the legal framework. It means that information, ostensibly not about individuals, would come under full remit of data protection law based on a possibility of it being linked to individuals at some point in time.

We have clearly entered an age of data ubiquity, a 'data deluge' where organisations seek innovative ways to manage data being accumulated through various business processes[5]. De-identification has become a key component of numerous business models, most notably in the context of health data, online behavioral advertising and cloud computing. Health data, for example, while extremely sensitive, could be harnessed for secondary uses yielding immense value to society. A strict interpretation of data protection laws applying the legal framework to de-identified data would threaten the advancement of anonymisation and encryption as practical concepts[6]. This, in turn, would increase, not alleviate, privacy and data security risks. Ontario Privacy Commissioner Ann Cavoukian and Professor Khaled El Emam recently observed: 'As long as proper de-identification techniques, combined with re-identification risk measurement procedures, are used, de-identification remains a crucial tool [for] privacy'[7].

There are several possible solutions to the re-identification quandary, besides the 'all or nothing' approaches of tagging any de-identified data as 'personal' or not. First, the nature of data as personal or not could be viewed as a continuum, as opposed to the current dichotomy. This means that data which are only identifiable at great cost would remain within the legal framework, yet be subject to only a subset of fair information principles. Hence, for example, it makes little sense to provide individuals with a right of access and rectification to data that are not readily identifiable, as this would require data controllers to proactively re-identify data, infringing the privacy of individuals requesting access and others.

'Clever adversaries can often re-identify or de-anonymize the people hidden in an anonymized database...Re-identification science disrupts the privacy policy landscape by undermining the faith that we have placed in anonymization'

Another approach would be to restrict the scope of the term personal data based on the likelihood of identification. Such a solution confirms to the spirit of Recital 26 of the Directive. It would facilitate the processing of data about de-identified individuals and encourage organisations to anonymise personal information wherever possible. For example, key coded data would not be regarded as personal data in the hands of a research institute if it has no means to access the key, which is held by a sponsoring entity. Likewise, a user profile compiled and analysed by a service provider in the online advertising ecosystem would not be considered as personal data unless such company is able to link the profile to an individual user. The test for re-identification risk must be context specific depending on factors such as the type of data, the duration of use and the techniques used for de-identification. This approach relies on a realisation that while re-identification attacks have been demonstrated by researchers seeking to prove their theoretical possibility, they may be difficult to effect in practice.

There is no single right way to achieve de-identification, much less to introduce it into regulation. In the US, the Health Insurance Portability and Accountability Act excludes from its scope de-identified health data based on one of two standards: a 'safe harbor standard', which specifies 18 data elements that must be removed including patient names, full dates, and ZIP codes, and a 'statistical standard' requiring that an expert perform the de-identification, that re-identification remain 'very low', and that the de-identification method is documented. Both standards reduce the risk of re-identification, though not to zero.

Policymakers currently grappling with the concept of personal data should not dispose of the important tool of de-identification. Instead, they should define best practices setting forth technical and organisational measures for robust de-identification and seek pragmatic, practical solutions rather than expansive formula that might yield adverse results.

Omer Tene Associate Professor
Israeli College of Management School of Law
omer.tene@bezeqint.net

1. Paul Ohm, 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization', 57 *UCLA Law Review* 1701 (2010).
2. Latanya Sweeney, 'Uniqueness of Simple Demographics in the US Population', Laboratory for International Data Privacy Working Paper, LIDAP-WP4 (2000).
3. Arvind Narayanan & Vitaly Shmatikov, 'Robust De-anonymization of Large Sparse Datasets', 2008 IEEE Symposium on Security & Privacy 111.
4. Michael Barbaro & Tom Zeller, Jr., 'A Face is Exposed for AOL Searcher No. 4417749', *NY Times*, 9 August 2006.
5. 'A special report on managing information: Data, data everywhere', *The Economist*, 27 February 2010.
6. A tangential issue is that since the process of anonymization is considered to be 'processing' of personal data, controllers are typically required to seek individuals' consent to the anonymization of their data.
7. Ann Cavoukian & Khaled El Emam, Dispelling the Myths Surrounding De-identification: Anonymization Remains a Strong Tool for Protecting Privacy, Discussion Paper, 16 June 2011.