Exploring reidentfication risk: Is anonymization a promise we can keep?

Accepted for publication in the International Journal of Market Research

Dr. Daniel Nunan

Birkbeck, University of London

d.nunan@bbk.ac.uk

Professor. Marialaura Di Domenico

University of Surrey

Abstract

The anonymisation of personal data has multiple purposes within research: as a marker of ethical practice, a means of reducing regulation and as a safeguard for protecting respondent privacy. However, the growing capabilities of technology to gather and analyse data have raised concerns over the potential reidentification of anonymised data-sets. This has sparked a wide ranging debate amongst both academic researchers and policy makers as to whether anonymisation can continue to be relied upon. This debate has the potential to create important implications for market research. This paper analyses the key arguments both for and against anonymisation as an effective tool given the changing technological environment. We consider the future position of anonymisation and question whether anonymisation can remain its key role given the potential impact on both respondent trust and the nature of self-regulation within market research.

Keywords

de-identificaton, anonymisation, self-regulation, data protection, privacy

Introduction

> *"A survey of UK and US citizens suggests that 41% do not trust market research companies with their data…People appear to be more trusting of search engines, mobile phone companies and even national security agencies."* - *Tarran, 2014*

Trust has long been recognized as a key factor in facilitating the forms of relationships upon which market research relies (Moorman et al. 1993). It plays a key role in reducing the perception of risks in research (Morgan and Hunt, 1994) whilst having a positive impact on respondent engagement in research (Moorman et al. 1992). Any factors that have the potential to impact the trust that the public has in the research process are therefore significant. In a theoretical sense trust can be understood as a mechanism that serves to mitigate the risk of opportunism towards respondents in an exchange characterized by uncertainty (Pfeffer and Salancik, 1978). Within the context of the growing important of secondary online data the role of respondent trust has been recognized as playing an increasingly important role in the level of and quality of responses. A lack of trust has been associated with lower response rates, fabrication of personal information as well as other forms of obfuscation taken as acts to protect personal privacy (Lwin and Williams, 2003; Wirtz et al., 2007).

A key mechanism through which respondent data is protected, and trust maintained is by the effective anonymisation of personal data. In this context anonymisation is the process through which personal data is removed from datasets before they are shared more broadly, whether within organisations or externally. Despite its central role in discussions around the contemporary use of big data anonymisation has received relatively little coverage within the literature relating to market research. This may relate to more limited levels of systematic data sharing, use of open data or other forms of 'release and forget' data within commercial market research. This can be contrasted with

other research contexts, for example within health and social sector where sharing and use of open data is more widespread. However, due to the growing scope and availability of such datasets their use within commercial settings is already recognized as being of strategic importance, as highlighted by the examples given later in this paper. Although there has been broad coverage amongst scholars in legal and technology domains there remains a gap relating to the issues surrounding anonymisation in a social or market research context. This paper seeks to address this gap by recognising data anonymisation as not simply an issue of law or technology, but one that goes to the heart of the challenges around the wider social issue of trust in research.

By evaluating and reconciling the differing views relating to anonymisation, particularly in the light of changing patterns of data collection, we seek to build a greater understanding of the key role that anonymisation is likely to play going forward. This paper is structured as follows. We begin by exploring the concept of anonymisation and its role within contemporary research practice. Using examples, we discuss the often high profile debates amongst research relating to the risks in reidentifying anonymised data. A synthesis of this debate is presented from which we highlight the challenges and risks with maintaining the 'promise' of anonymity.

The Role and Importance of Anonymisation

Anonymisation is rooted in the defining principle of research ethics: that participants in research should not be harmed as a result of participation. Data collected during a research process could, if gathered in the wrong hands, cause harm to respondents by making public information that was not designed to be. Harm can be caused both directly and indirectly. In the direct case, the reidentification of personal identifiers such as name or address could lead to linking back personal details, such as financial or health information. In the indirect case reidentification can happen through the combination of multiple datasets even without any active or malicious attempt to de

reidentify the data. Examples of these scenarios are given later in this paper. Given the link between trust in the person or organisation carrying out research and response rates (Edwards, 2002) there is an impetus to ensure individuals know that their personal data could not, even theoretically, cause them harm. Whilst this sets the scene for anonymisation its importance within ethical and self-regulatory frameworks has emerged through legal drivers. The identification of the concept of personal data by the Council of Europe in 1981 created with it a regulatory necessary for researchers to understand whether what they could be dealing with was personal data and, by extension, an incentive to develop ways to avoid dealing with personal information to reduce the regulatory burden. Anonymity can be therefore characterised as playing a useful role in aligning the interests of researchers with participants.

From a UK perspective the Data Protection Act provided further specificity on where researchers could look to draw the line between personal and non-personal data. Crucially it applies not simply to reidentification from a single data set but other forms or combinations of data.

Figure 1. Goes about here

The importance of building understanding of anonymisation is three-fold. Firstly, working with anonymised data has been adopted by the research profession as a defining characteristic of the field. Ensuring that any data collected is correctly anonymised is a core feature of market research. For example, it features prominently in the both the MRS and ESOMAR codes (figure 2.).

Figure 2. goes about here

Secondly, anonymisation has also become, in legal and regulatory terms, a boundary between what might be considered as personal data and thus subject to data protection legislation. The

attractiveness of anonymisation has been partly been driven by a view that anonymised data is 'of no interest to regulators' (Aldhouse, 2014:405). Thus, in an environment where personal data is coming increasingly under regulatory scrutiny anonymisation provides a route through which to more easily protect the existence of self-regulation.

Thirdly, due to rising concern over the potential for reidentification, with scare stories appearing in the media on a near daily basis (Aldhouse, 2014), if the techniques that underly the principles of anonymisation are shown to be broken there are serious implications for those who rely upon it to maintain trust. The concept that anonymisation might be broken has become the subject of increasingly wide debate in academic circles with two alternative, and competing, views. On the one hand legal scholars argue for the importance of anonymisation to maintain the key legal underpinnings of research. Without anonymisation, it is argued, the utility of market research will be severely harmed. On the other hand are information systems and computer science academics who argue that the anonymisation as a concept cannot be guaranteed in a way that can be aligned with the patterns of data use that are seen within the contemporary 'big' data strategies (Nunan & Di Domenico, 2013). Whilst the importance of anonymisation is well embedded within research practice, the arguments against anonymisation should not be dismissed as a niche academic concern. For example, in the US the Presidents Council on Science and Technology has stated the following:

> "it is increasingly easy to defeat anonymization by the very techniques that are being developed for many legitimate applications of big data…PCAST does not see it as being a useful basis for policy." (PCAST, 2014).

The latter part of this statement is important given that anonymisation features heavily de facto in both policy and legislation.

The re-identification problem

Challenges to reidentification have emerged through the growing number of scenarios where supposedly anonymous data sets have been subject to reidentification. From the reidentification of the health details of Massachusetts Governor William Weld in 1997 following his 1996 collapse at a public event, hospitalisation and almost immediate recovery (Ohm, 2009) to more recent reidentifications these events, and the media coverage surrounding them, have tracked the growth in the internet. Whilst an old example, in internet terms at least, the story of Governor Weld highlights both the risks of reidentification, but also the danger in exaggeration from media hype. It was a graduate student named Latanya Sweeney, now a Professor at Harvard, who carried out the analysis[1]. The Massachusetts Group Insurance Commission had made available a set of, supposedly, anonymised data for researchers that contained information on treatment date, ZIP code and gender. At the same time Sweeney was able to purchase, for $20, an full set of electoral roll data containing ZIP code, birth date, and gender. Based upon these variables it is not difficult to see how the datasets could be easily combined, and from this a single match was identified and Sweeney able to send the correct set of re-identified health records to the Governor (Ohm, 2009). This example has since been used as a parable for the risks relating to data re-identification in the internet era - if the Governor of Massachusetts can have his health data reidentified so easily, couldn't anyone. However, there are a number of flaws in this argument, not least that the range of variables made available means that the data was not properly deidentified. The provision of ZIP code and gender narrows the field of potential individuals to such a great extent that the addition of even a small number of additional variables enables reidentification. In any case, the subsequent tightening of HIPAA privacy regulations in 2003 would have made the approaches used ineffective (Barth-Jones, 2012). In addition to the way that the data was anonymised there are the

[1]Whilst an illustrative summary is provided here a full analysis has been published by Barth-Jones (2012) and Ohm (2009).

characteristics of the data subjects themselves. This example included a high-profile individual for whom there existed a large set of existing public data. It was therefore possible to verify with a highly level of certainty that the individual was the one identified from the data even where, as in many cases, the public dataset was incomplete.

Despite the changing technological infrastructure since 1997 the issues of the effectiveness of anonymisation and the re-identification from a general population have much in common with a recent example of reidentification of NY Taxi data discussed later in this paper.

Figure 3. goes about here

A number of cases of reidentification and their causes are highlighted in figure 3. This is, by necessity, a non-representative sample of incidents of reidentification as it highlights only some of those that have reached prominence through the media. One trend that has can be observed from the data is reidentification shifting from being a complex, and perhaps difficult, laboratory project through to something that is accessible to those with more mainstream technical skills.

Concerns over the threats to anonymisation, perceived and actual, manifest themselves in two competing streams of debate. The first broadly argues that anonymisation *must* exist because through enabling research it provides significant value to the economy. The second argues that, given technological trends, the standards required for anonymisation keeps rising and we can therefore not guarantee that information remains anonymous. The paper now explores each of these perspectives in turn.

Perspective 1: Anonymisation Must Work - "the Tragedy of the Data Commons"

We label the first perspective "anonymisation must work" to reflect the focus on wider social and economic benefits of anonymisation whilst playing down the extent of the technical risks. Effectively, this argues that critics of anonymisation are highlighting purely theoretical and laboratory risks that matter far less in the real world.

Yakowitz (2011) frames this debate by introducing the concept of the data commons, defined as "the disparate and diffuse collections of data made broadly available to researchers with only minimal barriers to entry" (ibid:403). This recalls the analogy of the tragedy of the commons (Harden, 1968) used to highlight the tension between individual interests and the common good. The original example related to the grazing on common land, whereby individuals benefitted from its provision but over-exploitation of the resource harmed everyone. At the same time there was little incentive for an individual to reduce their own grazing. Harden did not interpret tragedy as a form of unhappiness but it was meant, in a philosophical sense, as a form of futility reflecting the "remorseless working of things" (Yakowitz, 2011:1245). The 'tragedy' in a research sense relates to a situation where individuals feel able to opt-out or remove their personal information from data collected but still profit from the benefits that are brought by the use of the data.

Yakowitz, and others supporting this argument, argue that the risks have been overstated and, when weighed against the benefits, are acceptable. This perspective does not question that anonymisation is foolproof, nor that the changing technology landscape leaves it unscathed. However, it suggests that it is a combination of media excitement and failures in the processes through which data was anonymised that have caused the issues. For anonymisation "the sky is not falling" (ibid:35) for three reasons:

1. **Anonymisation processes are defective.**

In many cases it is the ineffective use of anonymisation techniques that causes problems, not the concept of anonymisation itself. For example, the AOL search data case (figure 3.) relied on pseudonymization where a single numerical key - related to an individual - was attached to each search term. Given the personal nature of search information, for example that variables such as demographics and location can often by inferred from the types of search undertaken, it was not surprising that a number of individuals could be identified. From searches related to specific pets, health issues and homes for sale user 4417749 was identified as a 62 year old widow from Lilburn, Georgia (Barbaro & Zeller, 2006). The scenario for the New York Taxi case was enabled by a similar mistake. Whilst a common cryptographic algorithm, known as MD5, was used to encode the cab drivers medallion number (i.e. unique ID number) it was only effective when one is unaware of the original format of the number. Unfortunately, one doesn't need to live in New York to find out the format of a New York taxi medallion number as a search on Google images will quickly provide an example. The argument is therefore that anonymisation works if it is done properly and that the failure of anonymisation is a failure of the anonymisation process, not necessarily a failure of anonymisation itself. The answer is therefore better technical solutions to offer more effective approaches to anonymisation have become adopted, such as k anonymisation (Sweeney, 2002).

2. The low probability that adversaries exist.

Negative effects of reidentification assume that there is someone willing and able to exploit the misuse of data. As Yakowitz puts it (2011:34) "…the marginal value of the information in a public dataset is usually too low to justify the effort for an intruder". The point here is that it requires not only the intent to cause harm, but an acceptance that if someone wants to access personal data, and they are willing to ignore the law, there are far simpler mechanisms to gaining personal data than a complex reidentification process.

3. **The low level of risk posed by reidentification compared to tolerated risks.**

Once the effectiveness of anonymisation is characteristised as an exercise in calculating risk it requires an analysis of what is an acceptable level of risk. One of the widely cited examples of US health data being reidentified (Sweeney, 2011) found a reidentification rate of 0.04%, similar to the lifetime risk of being hit by lightening and considerably less risky than dying in an accident at home (Calman and Royson, 1997). Even as risks grow over time as the scope of data and power of technology increase these must be put into perspective in terms of other general risks within society.

Bringing the discussion full circle, the metaphor of the commons related specifically to the unregulated commons (Harden, 1998) and it could be argued that the issues anonymisation points towards greater regulation to prevent reidentification activity (Barbaro & Zeller, 2006). Overall, the argument is that the benefits of anonymised data for research, whether it commercial, social or scientific are very great whilst the risks have been overstated. Given that much of society functions on the basis of weighing up and evaluating difference forms of risk the types of risks created from anonymisation are manageable. In short, what is therefore needed is a regulatory solution to enforce the effectiveness of anonymisation.

Perspective 2: Anonymisation Can't Work - "The Database of Ruin"

> "For almost every person on earth, there is at least one fact about them stored in a computer database that an adversary could use to blackmail, discriminate against, harass, or steal the identity of him or her.… For almost every one of us, then, we can assume a hypothetical "database of ruin," the one containing this fact but until now splintered across dozens of databases on computers around the world, and thus disconnected from our identity." - Ohm, 2009: 41

A second perspective does not lay the problems of anonymisation with inconsistent implementation or management of risk, but argues changes in technology have left it ineffective in the context of contemporary characteristics of data generation. This argument has been highlighted by researchers within information systems and computer science, and taken forward by legal scholars in the context of the shifting debates on implications for privacy. It is necessary to differentiate this argument from reflection upon the media 'hype' that often accompanies examples of reidentification, such as those given earlier in this paper. Indeed, whilst the first perspective may be said to have a very practical foundation the second combines this with a theoretical base.

Whilst the question of challenges of reidentification have been widely debated amongst computer scientists it is law Professor Paul Ohm who has made the most widely cited arguments over the failure of anonymisation. The lessons provided by Ohm are notable for commercial researchers as they not only critique the failure of anonymisation from a technical perspective but consider how these failures might be remedied in a world where rights to privacy cannot be so easily dismissed.

Ohm argues that anonymisation is both important and flawed. It is important because many of the legal defences of privacy depend on the effective anonymisation of data and thus, indirectly, anonymisation plays a key role in the ordering of society. He argues that anonymisation worked well over a 15 year period in the early era of computing and the internet but, in the face of increasing computer power and available of datasets, it is no longer effective. The barriers to reidentification are lower than might be supposed with 87% of Americans being uniquely identifiable from a ZIP code, birth date and sex (Sweeney, 2000). The key point is not that data cannot be anonymised, but that doing effectively removes much of the utility that might be gained from analysing the data.

Researchers within computer science have provided a number of technical arguments that seek to highlight the limitations of anonymisation approaches[2]. Here, the promise of anonymisation is ineffective as we are unable to predict the direction of technologies in the future:

> "Due to the ad hoc de-identification methods applied to currently released datasets, the chances of re-identification depend highly on the progress of re-identification tools and the auxiliary datasets available to an adversary. The probability of a privacy violation in the future is essentially unknowable." (Narayanan et al. 2015)

At the root of this issue are difficulties in defining the concept of anonymisation itself. Whilst personal data is often considered as being a relatively narrow set of data directly relating to a persons identity, geography or demography, in the context of reidentification any variable that might be used to distinguish one person from another could be considered as personal (Narayanan and Shmatikov, 2010). Technical approaches to identification, such as use of k-anonymity, assume that some variables are non-identifying. Such approaches might be considered as being better, in providing an increase in the barriers to reidentification, but they are not a solution. Narayanan et al. (2015) argue that due to the difficulties in identifying future technology capabilities and the inability to delete data once made public it is not just that the risk is unknown - it is unknowable.

Ohm's emotively titled 'database of ruin' represents the sum of the risks contained through the potential for individual data to be identified. Ohm uses the phrase 'privacy theatre' to describe the current approach to anonymisation, implying that researchers are thoughtlessly going through the motions of privacy in procedures that are not effective. This alludes to the concept of 'security theatre' that has been used to describe forms of security protection that give the appearance of security without fully addressing the underlying risk factors, for example in airports (Schneier,

[2] see Narayanan et al. 2015 for a detailed explanation of the issues with anonymisation.

2009). The remedy is to abandon both the promise and language of perfect anonymisation, including terms such as anonymisation and deidentification, and present the process of protecting personal data in more nuanced terms. Both Ohm and Sweeney prefer the term ‘scrub’, although it is unclear whether this serves to add greater clarity. Beyond the basics of terminology there is a recognition that the existing regulatory regime in many countries, being so dependent on the concept of protecting personally identifiable information through anonymisation, will need to be rethought. Here the question or how to best regulate privacy, like the concept of privacy itself, becomes confused and divided between addressing privacy in a specific context rather than as a general framework. Arguably the most influential scholar in the field of privacy, Dan Solove, argues for regulations to be neither too general or too specific, but also that privacy issues can only be resolved in regulatory terms by considering both the general and the specific (Solove, 2008).

As an example we return to the case of the New York taxi data being reidentified.. Whilst the media coverage related to this scenario focused upon identifying the tips, or lack of tips, that various celebrities gave to the taxi driver, other potential sources of personal information such as a home address could also be inferred from the data. The point is that the type of anonymisation required to have prevented such reidentification taking place would also have so severely restricted the value of the data so as to be of very limited use. It also highlights that anonymisation can only be effective in the light of an analysis of what other data sources and variables it can be combined with. Of wider interest is the implication that the range of variables from which personal data can be inferred is broader than had been anticipated, particularly when location can be inferred from data (Krumm, 2007).

Despite the critiques of Ohm’s argument discussed earlier he draws two, more modest, conclusions that from which wider agreement might be possible (Ohm, 2009:5). The first is that the changes in technology relating to data collection and analysis will continue to increase the risks of

reidentification. The second is that further regulations with the goal of strengthening requirements for anonymised data have the potential for significant negative impacts upon the utility of information available to researchers. This argument can be summarised as a longer term view - it is not necessarily the current risks that are important, but the direction of travel and the techniques that enable future reidentification using data collected today.

Anonymisation and market research: where do we go from here?

Whilst the two arguments outlined in this paper are not aligned, they are not necessarily opposed. Despite the very different approaches and philosophies underlying these two perspectives they both reach similar conclusions: a combination of regulatory and technical solutions are required to protect the public from the risks of reidenfication of personal data and determine levels of acceptable risk. Thus, even where we are clear on how technology works the wider social consequences only become clear over time (Coates, 1982). If anonymisation cannot be guaranteed in strictly technical terms a case can be made that the risks are outweighed by the wider social or economic benefits.

For market research the issue is less straightforward. As with the wider research sector there is a desire for the status quo not only in terms of anonymisation but also in terms of regulation, with bodies such as the Wellcome trust characterising the choice as between 'privacy and possibility' (Wellcome Trust, 2015). However, with proposed updates to EU privacy regulations drawing a distinction, for example, between public health research and broader commercial research there is also the issue of maintaining of self-regulation. These regulatory drivers are not limited to Europe with updating and tightening of data protection legislation on the agenda around the world, not least in the US. The key limit of self-regulation is that it can only apply to the actions of those organisations that have chosen to be part of a particular regulatory regime. As the 'data sector' has

grown to be much broader in scope that the traditionally defined market research sector so has the potential for externalities that cannot be controlled through self-regulation.

With the growth of big data and the subsequent trend towards collection and analysis of secondary data such as social media or sensor data, many areas of research involve a shift from researchers being 'creators' of data to users of data created by others. To appreciate the extent of this, one only has to look at how many leading firms that used to define themselves as being in market research now describe their business as 'data science'. In this shift to analysis a dependency is created upon the norms of organisations driven by a commercial imperative to work with individual level rather than anonymised data. An approach that pushes for statutory legislation to protect the effectiveness of anonymisation therefore carries with it an implicit acceptance of the limits of self-regulation.

The alternative is an acceptance that blanket anonymity is a promise that cannot be kept. The promise in this context is part of the overall argument that researchers can, and should, be trusted with data. It is this promise that is, at the least, undermined by the technical factors outlined in this paper. Whilst at this point in the paper it would be cleaner to provide a clear resolution and way forward, the reality is more messy. The changing technological landscape creates a level of uncertainty that makes predictions impractical. Rather than be prescriptive our goal in this paper is to draw attention to an important side-effect of the changing nature of research technology and data use.

However, in a more practical sense this paper underlines the importance of the management of current and future potential risks related to anonymisation and of communicating them to research participants. These risks should not be seen as being entirely, or even wholly related to the potential for data reidentification. To reemphasise the point made earlier: given the large volumes of data being collected the risk of accidental reidentification remains comparatively low. Rather, the

potential risk is that participants in research *perceive* the concept of anonymisation to be misleading, or ineffective, and that this perception impacts upon the types of data that they are willing to share with researchers. The collection of data carries with it risk, and researchers need to share these risks with the public. The question is how researchers can ensure that participants are better aware of the risks whilst still maintaining the principle of anonymity from a data protection perspective. With this in mind we offer two directions for further debate amongst the research community.

- Firstly, if we are dealing with research participants rather than respondents (as the new MRS code implies) then researchers have a duty to provide individuals with the information they need to enable them to participate. This means that consideration must be given to being less equivocal over the types of guarantees offered to participants. A follow-on from this is that it creates a risk around respondents being less willing to partake in research. However, at the same time these respondents are likely to be making their decisions in a more informed way, and thus this would partly fulfill the need for adapting informed consent to fit contemporary technology.

- Secondly, there is the more strategic view of the effectiveness of the concept of anonymised data. Researchers have to consider both the potential direction of legislation as well as the development of social norms with regards to privacy. How might market research cope with a world where anonymisation of data is legally mandated, technically difficult and treated with indifference by participants?

Conclusion

This paper has reviewed the key debates surrounding the issues of anonymisation. Whilst this has specific relevance to the use of anonymisation as a key characteristic of 'legitimate' market

research, it also serves as a microcosm of the wider tensions around legislation, technology and participants that have the potential to pull market research in multiple directions. As the size of the data economy increases it also highlights the increasing influence of external actors upon the research world and with it the limits of self-regulation.

Although in the context of the methods described in this paper it refers to the future of research, as much as the present, the central importance of the concept of anonymisation remains. We have argued that anonymisation has a unique place within market research, as opposed to other uses of commercial data, as it enables respondent trust and the maintenance of a self-regulatory regime. The ambiguity over the impact of new technologies together with a changing legal climate threatens this. Whilst informing respondents over the risks related to anonymisation may result in reduced response rates, or access to more limited datasets, researchers would do well to remember that trust does not come free.

References

Aldhouse, F. (2014) Anonymisation of personal data: A missed opportunity for the European Commission, Computer Law & Security Review, 30(4), pp.403-418.

Barbaro M., and Zeller, T. (2006) A Face Is Exposed for AOL Searcher No. 4417749. New York Times, August 9th. Available online at: http://www.nytimes.com/2006/08/09/technology/09aol.html?pagewanted=all (accessed 2nd June 2015).

Barth-Jones, D, (2012) The 'Re-Identification' of Governor William Weld's Medical Information: A Critical Re-Examination of Health Data Identification Risks and Privacy Protections, Then and Now. Working paper. Available online at: SSRN: http://ssrn.com/abstract=2076397 (accessed 2nd June 2015).

Calman K, and Royston G. (1997) Risk language and dialects. British Medical Journal 315, pp.939-42.

Coates, J. (1982) Computers and business ? A case of ethical overload. Journal of Business Ethics 1(3), pp.239-248.

Edwards, P. (2002) Increasing Response Rates To Postal Questionnaires: Systematic Review. British Medical Journal, pp.1183-1183.

Gymrek, M., McGuire, A., Golan, D., Halperin, E., and Erlich, Y. (2013) Identifying Personal Genomes by Surname Inference. Science, 339(6117), pp.321-324.

Hafner, K. (2006) Tempting Data, Privacy Concerns; Researchers Yearn To Use AOL Logs, But They Hesitate, The New York Times. August 23rd

Hardin, G. (1968) The Tragedy of the Commons. Science, 162(3859) pp.1243-1248.

Hardin, G. 1998. Essays on Science and Society: Extensions of "The Tragedy of the Commons". Science, 280(5364), pp.682-683.

Krumm, J. (2007) Inference Attacks on Location Tracks. Fifth International Conference on Pervasive Computing (Pervasive 2007), May 13-16, Toronto, Ontario, Canada

Lwin, M. O., Wirtz, J. and Williams, J. (2007), "Consumer responses to personal informational privacy concerns on the Internet: a power-responsibility equilibrium perspective", Journal of the Academy of Marketing Science, Vol. 35 No. 4, pp. 572-585.

Morgan, R. M. and Hunt, S. D. (1994), "The commitment-trust theory of relationship marketing", Journal of Marketing, Vol. 58 July, pp. 20-38.

Moorman, C., Zaltman, G. and Deshpandé, R. (1992) Relationships Between Providers and Users of Market Research: The Dynamics of Trust Within and Between Organizations, Journal of Marketing Research, 29 (August), pp.314-29.

Moorman, C., Deshpande, R., & Zaltman, G (1993) Factors Affecting Trust in Market Research Relationships, Journal of Marketing, 57(1).

Narayanan, A. and Shmatikov, V. (2008) Robust De-anonymization of Large Sparse Datasets (How To Break Anonymity of the Netflix Prize Dataset), IEEE Symposium on Security and Privacy.

Narayanan, A., and Shmatikov, V. (2010). Myths and fallacies of "personally identifiable information". Communications of the ACM, 53(6).

Narayanan, A., Huey, J. and Felten, E. (2015) A Precautionary Approach to Big Data Privacy. Computers, Privacy & Data Protection Conference.

Nunan, D. and Di Domenicom M. (2013) Market research and the ethics of big data. International Journal of Market Research 55 (4).

Ohm, P. (2009) Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization, UCLA Law Review, 97(1701).

Pandurangan, V. (2014) On Taxis and Rainbows: Lessons from NYC's improperly anonymized taxi logs. Available online at: https://medium.com/@vijayp/of-taxis-and-rainbows-f6bc289679a1 (accessed 2nd June 2015)

PCAST (2014) President's Council of Advisors on Science and Technology. Report to the President: Big Data and Privacy: A Technological Perspective, at 38-39 (May 2014).

Salancik, G. and Pfeffer, J. (1978). A Social Information Processing Approach to Job Attitudes and Task Design. Administrative Science Quarterly, 23(2), p.224-241

Schneier, B. (2009) Beyond Security Theater, Schneier on Security. Available online at: https://www.schneier.com/essays/archives/2009/11/beyond_security_thea.html (accessed 2nd June 2015)

Solove, D. (2008) Understanding privacy. Cambridge, Mass: Harvard University Press.

Sweeney, L. (2000) Simple Demographics Often Identify People Uniquely. Carnegie Mellon University Privacy Working Paper 3, Pittsburgh.

Sweeney, L. (2002) k-anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 10(05) pp.557-570.

Sweeney, L. (2011) Patient Identifiability in Pharmaceutical Marketing Data. Data Privacy Lab Working Paper 1015. Cambridge, MA.

Tarran, B. (2014) 41% do not trust market research companies with their data. Research Live. Available online at: http://www.research-live.com/news/news-headlines/41-do-not-trust-market-research-companies-with-their-data/4011288.article (accessed 2nd June 2015).

Wellcome Trust (2015) Safeguarding the Status Quo: Privacy and Possibility in Research. Available online at: http://blog.wellcome.ac.uk/2015/01/22/safeguarding-the-status-quo-privacy-and-possibility-in-research/ (accessed 2nd June 2015).

Wirtz, J. and Lwin, M. O. (2009), “Regulatory Focus Theory, Trust, and Privacy Concern”, Journal of Service Research, Vol. 12 No. 2, pp. 190-207.

Yakowitz, J. (2011) Tragedy of the Data Commons. Harvard Journal of Law & Technology. 25(1) Fall.