XAMAX
Consultancy

Submission
to
Productivity Commission

Data Availability and Use

Xamax Consultancy Pty Ltd
78 Sidaway St
Chapman ACT 2611
AUSTRALIA
http://www.xamax.com.au

5 May 2016

Productivity Commission Inquiry
Data Availability and Use

Submission

5 May 2016

Executive Summary

The attached Submission provides evidence and argument in support of one general statement and further statements on six key issues.

The general statement is that many issues arise in relation to personal data that are of limited relevance to other categories of data. There is a serious risk of positive recommendations in respect of most categories of data being undermined by inevitably contentious debates about personal data.

1. The 'Big Data' Meme

The 'big data' meme is being accompanied by wildly excessive enthusiasm. Its proponents have failed to factor in issues of data quality, inferencing quality, decision quality, large volumes of false positives, and incompatibility of data drawn from multiple sources and applied in ways very different from the original purposes of collection.

2. Public Trust and Distrust

Data quality is in many circumstances highly dependent on trust by the data subject in the organisations that the data subject deals with. The abuses of trust involved in secondary use of data, re-purposing of data, and disclosure of data, result in greatly lowered data quality.

Distrust is becoming a very serious issue in Australia, because the abuses are increasingly evident to the public. Current examples include Census 2016, the PCEHR / MyHR, telecommunications data retention, myGov and so-called 'positive' credit reporting.

3. Government Policy

Although the Australian Government has paid lip-service to transparency and open data, many actions by the Government and by individual government agencies have been thoroughly inconsistent with those values.

4. Data Propertisation vs. Data Openness

The ongoing unbalancing of copyright law in favour of copyright-owners and against users of copyright works has created very substantial economic incentives for corporations to develop and sustain data monopolies. There has been a strong tendency away from business models based on services towards the extraction of rents through data propertisation.

5. 'Data Rights' not 'Data Ownership'

The 'data ownership' meme was invented by US business, for US business. Its purpose is to reduce privacy from a human right, by treating it as a mere economic right. That would enable corporations to buy consumers out of their privacy rights for very low costs.

The notion of ownership is not applicable to data, because there are always many parties that have interests in data. An appropriate legal regime needs to establish rights relating to data that vest in different parties, some of which are inalienable, and to enable balances to be achieved through constructive tension among those parties and rights.

6. Re-Identifiability vs. Anonymisation

Rich data-records are largely incapable of being effectively anonymised, because they offer so many opportunities for re-identification. Irreversible 'data falsification' – a strong form of the technique of 'data perturbation' – is the only approach that can enable data that originally related to specific individuals to be used and disclosed without breaching both individual privacy and public trust.

1. Introduction

The focus of Xamax's consultancy practice is on the strategic and policy impacts and implications of advanced information technologies. My professional and disciplinary background is in Information Systems, although my current Visiting Professorships are in Computer Science (at ANU) and Cyberspace Law (at UNSW). I have long involvement in public interest advocacy, including as a Board member of Electronic Frontiers Australia (EFA, 2000-05), as a Director of Internet Australia (IA, 2010-15, incl. 3 years as Secretary), and as a Board member of the Australian Privacy Foundation (APF, 1987-, incl. 8 years as Chair).

This Submission draws on my experience across consultancy, research and advocacy. Brief summaries of material are provided in relation to each point. Sources are identified, to enable drill-down for greater detail. I have discussed various of these points with colleagues in EFA, IA, APF, the Financial Rights Legal Centre (FRLC) and the Australian Communications Consumer Action Network (ACCAN).

The Submission commences by distinguishing personal data from all other categories of data, and identifying the enormous risks that the Commission faces because of the inclusion of both within a single Terms of Reference. It then identifies and addresses key issues that need to be confronted. The contentions made are not currently mainstream. Proponents of 'big data' have naively and/or willingly blinded themselves to a number of critical realities associated with data and its use.

2. Distinct Categories of Data

The Inquiry's Terms of Reference are very broad-ranging, draw on several prior reports, and relate to data of many different kinds. There are undoubtedly considerable benefits to be gained from increased openness of many kinds of data, and from measures to improve quality and consistency of the data-sets, data-definitions and data-formats. The niches in which opportunities arise range across data categories as diverse as geographic, geological, geophysical, meteorological, oceanographic, biological and transportation data.

One kind of data, however, presents a cluster of challenges that seldom arise with those categories. When data is collected about soil, bed-rock, magnetic fields, bodies of water, low-level organisms and traffic congestion, those entities are not recognised as having interests in that data's collection, processing, storage, use, disclosure and retention. Human beings, on the other hand, do not accept that they are passive data subjects that are to be measured as governments and corporations see fit, and to have data about them trafficked in whatever means suits the effectiveness and efficiency of government and business. Rightly or wrongly, people consider that the institutions of economies and societies are there to serve people, not the other way around.

The Commission faces the very serious prospect of its research and recommendations in relation to data about many inanimate entities and processes being submerged in a tsunami of debate about issues arising with personal data. This will be driven by the desires of some segments of business and government for far greater freedom to abuse personal data, and to roll back the limited data protection laws that they see as an unwelcome constraint on their freedom of action.

For example, the current pitching of the notion of 'benchmarking' Australian data protection laws against those elsewhere is readily perceived as a trojan horse for the drastically watered-down US FTC and US-driven APEC frameworks, whose purpose is to export the USA's anti-privacy framework to the world. A related notion peddled by business and government is that 'data silos' need to be broken down. This completely misses the critical point that data silos are, and always have been, a far greater protection for the privacy of Australians than the country's weak data protection laws.

I submit that the Commission should divide the Inquiry into two segments, one addressing personal data specifically, and the other addressing all other categories of data. This would enable a variety of stimulative and expansive measures to be discussed and proposed in one context, without them being poisoned by association with positions adopted by business and government in relation to personal data that are highly unpalatable to the public.

3. Key Issues

This section identifies and discusses six key issues arising from the Inquiry's Terms of Reference.

3.1 The 'Big Data' Meme

Many inferences from 'big data' are currently being accorded greater credibility than they actually warrant. Proponents of big data analytics pay very little attention to a considerable number of important factors that affect the appropriateness and reliability of their techniques. The papers cited below examine these factors in some depth.

1. Many elements contribute to data quality and information quality. See Table 1 in Clarke (2016)
2. Achieving quality is expensive
3. When an organisation collects data, it spends only what it needs to in order to satisfy its own requirements. It seldom considers possible future uses, still less those of other organisations that may gain access to the data
4. Data acquired from multiple sources is of variable quality, and often of low quality
5. The meaning of each data-item, and of the data-content of that data-item in each record, is often difficult to determine, or ambiguous, or subject to varying interpretations across individuals, organisations and time
6. When an organisation collects data, it defines the data-items in accordance with its own perceived needs. It seldom considers possible future needs, still less those of other organisations that may gain access to the data
7. Data acquired from multiple sources has varying meanings, even where data-items and data-content appear to be the same
8. Data matching is based on data-items that have variable quality and variable meaning
9. Data matching has moderate to high false-positive and false-negative measures
10. Data scrubbing, when it is performed, is seldom undertaken against an external standard, but is instead usually based on heuristics or on inferences drawn from within the data-set
11. Data scrubbing has moderate to high false-positive and false-negative measures
12. Inferencing processes are commonly only applicable to some categories of data – most commonly to data that is on a ratio scale, or perhaps on a cardinal scale, although more commonly only on an ordinal or a nominal scale
13. A great deal of data is not on a ratio scale, and application of powerful statistical processes to it produces results that are unreliable, and quite probably distinctly misleading
14. Many inferencing techniques are 'inscrutable', in the sense that it is not feasible to answer the question 'How did you reach this conclusion?'. This applies in particular to neural net and learning algorithms, but also to many applications of expert systems techniques
15. The results of inferencing processes are seldom audited against reality
16. Decision processes seldom take account of the above issues
17. Decisions are frequently not communicated transparently to people affected by them, and hence a great many inappropriate decisions are never discovered to be so
18. There is little evidence of risk assessment and risk management techniques being applied to big data activities

I submit that the Commission needs to draw attention to the challenges confronting big data analytics, to the strong tendency for those challenges to be overlooked, and to the probably considerable degree of 'over-claiming' by the proponents of big data analytics.

I further submit that the Commission needs to draw to attention the care necessary to identify attractive opportunities, and to evaluate whether the potential is realisable.

These matters are considered in the following recent papers:

Wigan M.R. & Clarke R. (2013) 'Big Data's Big Unintended Consequences'
IEEE Computer 46, 6 (June 2013) 46 - 53, PrePrint at
http://www.rogerclarke.com/DV/BigData-1303.html

Clarke R. (2014) 'Quality Factors in Big Data and Big Data Analytics'
Working Paper, Xamax Consultancy Pty Ltd, at http://www.rogerclarke.com/EC/BDQF.html

Clarke R. (2015) 'Quality Assurance for Security Applications of Big Data'
Presented at Redefining R&D Needs for Australian Cyber Security, at the Australian Centre for Cyber Security (ACCS) at the Australian Defence Force Academy (ADFA), 16 November 2015,
Revised version accepted for Euro. Intelligence & Security Informatics Conf., Uppsala, August 2016
at http://www.rogerclarke.com/EC/BDQAS.html

Clarke R. (2016) 'Big Data, Big Risks'
Information Systems Journal 26, 1 (January 2016) 77-90,
PrePrint at http://www.rogerclarke.com/EC/BDBR.html

3.2 Public Trust and Distrust

In principle, it is possible to gather data about business activities, and about people and their behaviour, without their involvement or even knowledge. On the other hand, activities of this kind represent covert mass surveillance, and are deprecated as a massive threat to economic, social, cultural and political freedoms.

In practice, a great deal of data collection depends on the more or less willing participation of individuals, and on a sufficient degree of accuracy and honesty on their part when providing the data. It is in many cases feasible to perform authentication of the quality of the data captured from individuals; but such processes are subject to limitations, and are very costly. In short, public trust is essential if organisations are to successfully capture much of the data that they need. Deeper analysis of the concepts of trust, lack of trust and distrust is in section 3 of Clarke (2014).

Unfortunately, the behaviour of both government and business over the last 50 years has taught people to be vague, to be obstructive, to be constructively misleading, and in some cases to habitually lie. This is necessary in order to avoid floods of consumer marketing, to avoid the construction of consumer profiles that enable marketers to manipulate consumer behaviour, and to frustrate government agencies that expropriate data from multiple contexts and infer from inconsistencies among the various sources that individuals are 'up to no good'.

At any particular point in time, there is a vast array of instances of government and business initiatives whose purposes – and in many cases primary purposes – are to disadvantage individuals and favour organisations. Current examples include:

- public sector:
  - Census 2016. See Clarke (2016)
  - PCEHR / MyHR. See APF (2016)
  - [telecommunications] data retention. See APF (2015a, 2015b)
  - Telecommunications Act s.313. See Clarke (2015)
  - mass traffic surveillance through Automated Number Plate Recognition (ANPR). See Clarke (2009)
  - the Document Verification System (DVS) and National Trusted Identity Framework (NTIF)
  - myGov
  - Austrac's gross impositions on relationships between consumers and service-providers
- private sector:
  - 'positive' (i.e. comprehensive) credit reporting. See the Submission from FRLC
  - online behaviour tracking

- person location and tracking.
  See Clarke (2009), Clarke & WIgan (2011), Michael & Clarke (2013)
- vehicle tracking, e.g. by insurance companies
- identified public transport tickets and identified use of toll-roads
- many mooted applications of 'the Internet of Things'

The significance for the Inquiry of public distrust is at its greatest in the context of personal data. However, there are doubtless areas in which distrust by SMEs will influence data quality and even preparedness to disclose data to government. Large enterprises, meanwhile, are straightforward in their expressions of concern to, for example, the ABS.

I submit that the Commission needs to present an analysis of the impact of open data initiatives on public trust, on preparedness to disclose, and on the accuracy of data that is collected in circumstances in which the discloser knows, expects or even suspects that the data will be used and disclosed, to multiple organisations, for multiple purposes.

I further submit that the Commission needs to identify and consider some additional themes, in particular the question of over-regulation, and the importance of freedoms and self-determination as the appropriate basis for a vibrant, innovative and adaptive economy and society, rather than a surveillance society with nanny-state central planning.

APF (2015a) 'Telecommunications (Interception and Access) Amendment (Data Retention) Bill 2014, Submission to Parliamentary Joint Committee on Intelligence & Security (PJCIS)', Australian Privacy Foundation, 19 Jan 2015, at http://www.privacy.org.au/Papers/PJCIS-DataRetention-150119.pdf

APF (2015b) 'Telecommunications (Interception and Access) Amendment (Data Retention) Bill 2014, Supplementary Submission to Parliamentary Joint Committee on Intelligence and Security' Australian Privacy Foundation, 31 Jan 2015, at http://www.privacy.org.au/Papers/PJCIS-DataRet-Supp-150131.pdf

APF (2016) 'My Health Record – Campaign Page' Australian Privacy Foundation, March 2016, at http://www.privacy.org.au/Campaigns/MyHR/

Clarke R. (1999) 'Person-Location and Person-Tracking: Technologies, Risks and Policy Implications' Proc. 21st International Conference on Privacy and Personal Data Protection, pp.131-150, September 1999. Re-published in Information Technology & People 14, 2 (Summer 2001) 206-231, PrePrint at http://www.rogerclarke.com/DV/PLT.html

Clarke R. (2009) 'The Covert Implementation of Mass Vehicle Surveillance in Australia' Proc. Fourth Workshop on the Social Implications of National Security: Covert Policing, April 2009, at http://www.rogerclarke.com/DV/ANPR-Surv.html

Clarke R. (2014) 'Privacy and Social Media: An Analytical Framework' Journal of Law, Information and Science 23,1 (April 2014) 1-23, PrePrint at http://www.rogerclarke.com/DV/SMTD.html

Clarke R. (2015) 'Telecommunications Act s.313 – Notes for the Standing Committee on Infrastructure and Communications' Xamax Consultancy Pty Ltd, March 2015, at http://www.rogerclarke.com/DV/TA313.html

Clarke R. (2016) 'Census 2016 – Information Sheet' Xamax Consultancy Pty Ltd, March 2016, at http://www.rogerclarke.com/DV/Census-2016.html

Clarke R. & WIgan M.R. (2011) 'You Are Where You've Been: The Privacy Implications of Location and Tracking Technologies' Journal of Location Based Services 5, 3-4 (December 2011) 138-155, PrePrint at http://www.rogerclarke.com/DV/YAWYB-CWP.html

Michael K. & Clarke R. (2013) 'Location and Tracking of Mobile Devices: Überveillance Stalks the Streets' Computer Law & Security Review 29, 3 (June 2013) 216-228, PrePrint at http://www.rogerclarke.com/DV/LTMD.html

3.3 Government Policy

The Inquiry's Terms of Reference suggest that Government policy is strongly in favour of open data and transparency, in both the public and private sectors.

However, the Government's behaviour is very different from that rhetoric. It sought to disestablish the OAIC. When this was frustrated by the Senate, it breached constitutional norms by failing to make effective appointments to two of the three Commissioner positions and by reducing the agency's funding. It currently has the remaining Commissioner on 3-monthly drip-feed contracts that are renewed about the time the previous contract expires, requiring him to divide his time across three functions, and to cope with the lack the resources to do any of them properly.

In addition, government agencies continue to take every opportunity to avoid auto-publication, and to resist FOI applications. One small element in the mix that gives rise to a great deal of waste is the refusal of Ministers and government agencies to make publicly available the legal opinions that they use taxpayers' funds to acquire.

I submit that the Commission needs to draw attention to the importance of open access and transparency across the board, and the vital need for all Governments and all government agencies to adapt their philosophies and their *modi operandi* to reflect the need for openness and transparency.

A useful source summarising the appalling behaviour in relation to the OAIC is here:

Farrell P. (2015) 'Governments do not like freedom of information: the war on Australia's privacy and information watchdog' The Guardian, Thursday 1 October 2015 13.09 AEST, at http://www.theguardian.com/australia-news/2015/oct/01/governments-do-not-like-freedom-of-information-the-war-on-australias-privacy-and-information-watchdog

3.4 Data Propertisation vs. Data Openness

There are very strong disincentives against data openness. During recent decades, there has been a massive shift in balance away from the interests of consumer users and towards copyright-owners. This has reinforced the position of many supra-nationals, and encouraged many more corporations to presage their business models not on value-added services but on the exploitation of monopoly rights.

A current instance of this problem arises in the environmental space. A PhD candidate I'm supervising in ANU RegNet is examining the factors militating against open access to such sources as atmospheric data, oceanographic data, and river pollution measures. His intention is to identify stimulative and regulatory mechanisms that can achieve open data across international borders, despite the all-too-apparent scope for corporations to protect the data and extract rents from it.

The Commission has already identified this problem in its draft report 'CopyNotRight', which addresses aspects of this important cluster of problems.

I submit that it is vital that the Commission sustain its position in the face of the inevitable backlash by large, copyright-dependent corporations, and by the government agencies that participate in the public-private melange that sustains the present system.

The sources below adopt an information economics approach to innovation, and point to the very different conception of copyright that needs to be applied in digital contexts, including software, music and research papers.

Clarke R. & Dempsey G. (2004) 'The Economics of Innovation in the Information Industries' Xamax Consultancy Pty Ltd, April 2004, at http://www.rogerclarke.com/EC/EcInnInfInd.html

Clarke R. & Kingsley D. (2008) 'e-Publishing's Impacts on Journals and Journal Articles' Journal of Internet Commerce 7,1 (March 2008) 120-151, PrePrint at http://www.rogerclarke.com/EC/ePublAc.html

Dempsey G.C. (1998) 'Knowledge and Innovation in Intellectual Property: The Case of Computer Program Copyright' PhD Thesis, Australian National University, March 1998

Dempsey G.C. (1999) 'Revisiting Intellectual Property Policy: Information Economics for the Information Age' Prometheus 17, 1 (1999) 33-40

3.5 'Data Rights' not 'Data Ownership'

Since privacy emerged as a public policy issue in the 1960s, US business interests have waged war against it. One particular thread of the attacks has been the attempt to reduce privacy from the human right that it is to a mere economic right. The human rights aspects are discussed in Clarke (2000, 2014, 2016a).

Individuals are well-known to sell rights to their data very cheaply. In addition to the anecdotal evidence of the scant return offered by 'rewards' and 'loyalty' schemes, many laboratory experiments have thrown light on the ease with which people are duped into not appreciating the value of data.

Corporations perceive opportunities to gain access to vast quantities of personal data very cheaply, by getting the 'data ownership' meme to proliferate. That would enable corporations to enveigle people into signing away their human rights, which would in effect perform an end-run around pitifully weak data protection laws.

A deeper consideration of the treatment of privacy and 'personal data markets' in relevant literatures is in section 5 of Clarke (2016). This identifies the conventional Westin / Posner / Laudon / Varian line of analysis as being utterly committed to the corporate perspective, and utterly hostile to the perspective of the individuals whose data the corporations want to traffick. This is the philosophical basis for the corporate behaviour that is deepening consumer distrust.

A further consideration in this context is the nature of the consent granted by individuals to the use of data. To be meaningful, consent must have the characteristics of being informed, freely-given, variable and revocable. The desires of corporations for 'data ownership' to enable cheap purchase from unwitting consumers, are an abuse of the notion of consent. For deeper treatment of the notion of consent, and necessary characteristics of consent, see section 3 of Clarke (2002).

A particular device that has been dreamt up to achieve proliferation of the 'data ownership' meme is the UK 'midata' scheme (Midata 2011). Despite high-sounding claims about enabling consumers to compare alternative sources of services, the essential purpose of the midata initiative is to make personal data readily available to corporations. Government interest in it extends beyond its role as an economic stimulatory measure. This is because consumers' data would thereby become increasingly accessible by a government that is largely outsourced, and monolithic rather than being constrained to agency silos.

If the corporate sector were to succeed in reducing data privacy to a mere economic right, the results would be institutionalisation of consumer behaviour manipulation, monolithic government ousourced to the private sector, and a public whose political and economic behaviour was chilled, and whose cultural and social behaviour was driven by the private sector. It would be a more exciting existence than East Germany in the 1970s, but just as un-free.

I submit that the Commission needs to present an analysis of the proposition that data privacy be reduced from a human right to a mere economic right, and to resoundingly reject it.

The notion of 'data ownership' is not only repugnant to a free society. It is also an intellectual nonsense. Data is not an entity to which the concept of 'ownership' is appropriately applied.

Multiple parties have interests in data. Those interests need to be reflected in a variety of rights, some of which are entrenched and inalienable. Mechanisms need to exist to enable balance and reconciliation to be achieved in the vast array of circumstances in which interests conflict. Examples

of these rights include the right to collect, and the right to deny or preclude collection, the right to store and the right to deny or preclude storage, the right to use and the right to deny or preclude use, the right to disclose and the right to deny or preclude disclosure, and the right to retain and the right to deny or preclude retention.

None of the baskets of rights that the various parties can reasonably claim correspond to the notion of 'ownership', whether modelled on the law of real estate or of chattels, nor even any of the numerous concepts sometimes referred to generically as 'intellectual property'.

For a deeper examination of the 'data ownership, data control, data rights' issues, see section 4 of Wigan & Clarke (2013). For a deeper examination of the sources of value in data, see section 4.1 of Clarke (2013).

I submit that the Commission needs to present an analysis of the proposition that 'data ownership' is an appropriate approach, and to resoundingly reject it, in favour of the recognition of multiple rights, some inalienable, and processes for resolving conflicts among rights.

Clarke R. (2000) 'Beyond the OECD Guidelines: Privacy Protection for the 21st Century' Xamax Consultancy Pty Ltd, January 2000, at http://www.rogerclarke.com/DV/PP21C.html

Clarke R. (2002) 'e-Consent: A Critical Element of Trust in e-Business' Proc. 15th Bled Electronic Commerce Conference, Bled, Slovenia, 17-19 June 2002, at http://www.rogerclarke.com/EC/eConsent.html

Clarke R. (2013) 'Data Risks in the Cloud' Journal of Theoretical and Applied Electronic Commerce Research 8, 3 (December 2013) 59-73, at http://www.rogerclarke.com/II/DRC.html

Clarke R. (2014) 'Privacy and Free Speech' Invited Presentation at the Australian Human Rights Commission Symposium on Free Speech, Sydney, 7 August 2014, at http://www.rogerclarke.com/DV/PFS-1408.html

Clarke R. (2016a) 'A Framework for Analysing Technology's Negative and Positive Impacts on Freedom and Privacy' Datenschutz und Datensicherheit 40, 1 (January 2016) 79-83, PrePrint at http://www.rogerclarke.com/DV/Biel15-DuD.html

Clarke R. (2016b) 'Personal Data Markets: A Matter of Perspective' Xamax Consultancy Pty Ltd, February 2016, at http://www.rogerclarke.com/EC/PDMP.html

Midata (2011) 'The midata vision of consumer empowerment' Department for Business, Innovation & Skills and The Rt Hon Edward Davey, 3 November 2011, at https://www.gov.uk/government/news/the-midata-vision-of-consumer-empowerment

Wigan M.R. & Clarke R. (2013) 'Big Data's Big Unintended Consequences' IEEE Computer 46, 6 (June 2013) 46 - 53, PrePrint at http://www.rogerclarke.com/DV/BigData-1303.html

3.6 Re-Identifiability vs. Anonymisation

The suggestion is frequently made by proponents of open data that personal data can be rendered harmless to the people it refers to, through processes referred to variously as anonymisation or de-identification.

A few sources of guidance are available in relation to techniques for anonymisation / de-identification (UKICO 2012. See also Slee 2011, DHHS 2012). The techniques include deletion of specific rows and columns, generalisation or suppression of particular values and value-ranges, and what is politely referred to as 'data perturbation' - including micro-aggregation, swapping, adding noise and randomisation.

In rich data collections, however, it is generally feasible for a significant proportion of the records involved to be able to be later re-identified (Sweeney 2002, Acquisti & Gross 2009, Ohm 2010).

The literature contains various demonstrations and analyses, but few reports on actual abuses. This is because there are powerful factors at work which militate against the collection and publication of such data.

Such uses are perceived even by their users to be underhand, by their subjects to be abuses, and by regulators to be at least 'sailing close to the wind', dubiously legal, and even illegal. To the extent that the techniques are effective, they have competitive overtones. Added to that, it has become increasingly common for research that involves access to data from corporations and government agencies to be subject to non-publication constraints. Universities – which lack market and institutional power and which are dependent on external funding – are now accepting those serious compromises to the once-respected principles of open research data and auditability.

The only way in which rich data sets can be processed in order to prevent subsequent re-identification and consequential breaches of privacy is by the application of comprehensive 'data falsification'. This goes well beyond the limited notion of 'data perturbation'. It requires active measures to ensure that individual records no longer relate to any individual and hence will be entirely useless for any inferencing about an individual instance, and will be known to be so.

Depending on how data falsification is done, it may have a significant negative impact on the utility of the data-set as whole, not only for inferences about individuals – as is intended – but also for population inferencing. There are prospects that data falsification routines may be able to be developed that retain statistical value for specific purposes, and hence it may prove feasible to use large sets of personal data records to generate falsified data-sets whose usability is suitable for particular applications. However, at this stage, this remains an active area of research.

I submit that the Commission needs to present an analysis of the propositions relating to de-identification and anonymisation, debunk the claims that it is an established and reliable mechanism, and call for accelerated research into data falsification methods that destroy the utility of individual records while sustaining the overall value of data-sets.

Acquisti A. & Gross R. (2009) `Predicting Social Security Numbers from Public Data'
Proc. National Academy of Science 106, 27 (2009) 10975-10980

DHHS (2012) 'Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule' Department of Health & Human Services, November 2012, at http://www.hhs.gov/ocr/privacy/hipaa/understanding/coveredentities/De-identification/guidance.html

Ohm P. (2010) 'Broken Promises of Privacy: Responding to the Surprising Failure of Anonymization' 57 UCLA LAW REVIEW 1701 (2010) 1701-1711, at http://www.patents.gov.il/NR/rdonlyres/E1685C34-19FF-47F0-B460-9D3DC9D89103/26389/UCLAOhmFailureofAnonymity5763.pdf

Slee T. (2011) 'Data Anonymization and Re-identification: Some Basics Of Data Privacy: Why Personally Identifiable Information is irrelevant' Whimsley, September 2011, at http://tomslee.net/2011/09/data-anonymization-and-re-identification-some-basics-of-data-privacy.html

Sweeney L. (2002) 'k-anonymity: a model for protecting privacy' International Journal on Uncertainty, Fuzziness and Knowledge-based Systems 10, 5 (2002) 557-570, at http://arbor.ee.ntu.edu.tw/archive/ppdm/Anonymity/SweeneyKA02.pdf

UKICO (2012) 'Anonymisation: managing data protection risk: code of practice' Information Commissioners Office, November 2012, at http://ico.org.uk/for_organisations/data_protection/topic_guides/~/media/documents/library/Data_Protection/Practical_application/anonymisation-codev2.pdf