Version 2.2

Draft Guidelines for Third-Party Use of Big Data Generated by Mobile Network Operators[1]

Rohan Samarajiva, LIRNE*asia* (rohan@lirneasia.net)

August 2014

This work was carried out with the aid of a grant from the International Development Research Centre (IDRC), Canada and the Department for International Development (DFID), UK

[1] The contributions of Sriganesh Lokanathan, Senior Research Manager leading the big data work at LIRNEasia, are gratefully acknowledged.

Contents

List of acronyms

BTS	Base Transceiver Station
CALEA	Communications Assistance to Law Enforcement Act
CDR	Call Detail Records
CPNI	Customer Proprietary Network Information
FCC	Federal Communication Commission
FISA	Foreign Intelligence Surveillance Act
GPS	Global Positioning System
HA	Historical Anonymized
HI	Historical Identifiable
MNBD	Mobile Network Big Data
MNO	Mobile Network Operator
NRRI	National Regulatory Research Institute
RBOC	Regional Bell Operating Company
RTA	Real Time Anonymized
RTI	Real Time Identifiable
SIM	Subscriber Identification Module
SMS	Short Message Service
VLR	Visitor Location Register

Introduction

Big data generated by Mobile Network Operators (MNOs) in the course of business has potential to make their operations more efficient and improve the management of customer relationships, including marketing. Big data also has immense, and at this point unique, potential to bring forth a qualitative transformation of urban design including resilience, improve transportation and government-service delivery, and enhance management of the economy, among others. This is especially true in developing economies where datafication[2] is less common. While "datafied" data sets with broad population coverage are plentiful in developed-market economies, the only "born digital" data sets that cover most of the populations of developing economies are generated by MNOs. In addition, the richness of the data which also includes evidence of physical location and movement makes mobile network big data (MNBD) uniquely valuable in the development field at this time.

As with any phenomenon that can do good, big data also can do harm. This document is the first step of a series of actions that will hopefully lead to the adoption of voluntary Guidelines by MNOs which will minimize the likelihood of harm by embedding safeguards into standard procedures while reducing the transaction costs of making the data available to third parties who will use them for public purposes. The potential harms have been identified through the literature (Annex 1) and through engagement with ongoing analysis of MNBD at LIRNE*asia*.

Definitions and classification

"Big data" is defined as data that are very large in volume, diverse in variety and/or moving with such velocity (3 Vs), that traditional modes of data capture and analysis are insufficient and the insights generated therefrom are qualitatively more valuable. The declining cost of collection, storage, and processing of data, combined with new sources of data like sensors, cameras, geospatial and other observational technologies, have resulted in a proliferation of big data and a qualitative change in what can be analyzed and by whom. No longer are the fruits of big data analysis available only to those with super computers.[3]

MNBD include, but are not limited to, Call Detail Records of CDRs (denoting not only those associated with voice calls, but with Short Message Service (SMS) and data use), Visitor Location Register (VLR) data used to locate mobile devices in use (collected but rarely stored) and reload data applicable to prepaid customers who constitute the great majority of customers in developing economies. In the case of smartphones, tablets and similar advanced mobile devices, not yet in the hands of a majority of customers in developing economies, additional data such as GPS [Global Positioning System] and some data generated by mobile applications would be included. MNOs do not have access to some big data generated on devices connected to their networks by Internet companies such as Google and Facebook.

---

[2] "Datafication" is defined as transforming a phenomenon into a quantified format that allows it to be measured and analyzed: Mayer-Schonberger, V.; Cukier, K. (2013). *Big data*. London: John Murray. Pp. 78-86.

[3] While the current excitement about big data goes back only a few years, big data analysis was being conducted by organizations such as credit-card companies and security agencies many decades back: Samarajiva, R. (1996). Surveillance by design: Public networks and the control of consumption, in *Communication by Design: The Politics of Information and Communication Technologies*, eds. R. Mansell & R. Silverstone, pp. 129-156. Oxford: Oxford University Press.

MNBD may be analyzed in real time or in batch mode depending on purpose. If objectives are behavioral change (e.g., marketing), actionable prediction or proving culpability (e.g., surveillance), it will be necessary to retain at least the address, even if the complete identity of the user of the address is not known.[4] The ability to communicate to a specific individual based on the results of data analytics may be seen as a basic requirement of data-based behavior modification efforts, e.g., targeted marketing or political campaigning. But it is possible to communicate with audiences in other ways, for example, through cell broadcasting (where all mobiles within the coverage of a base transceiver station (BTS) will receive a message)[5] or where those who had signed up for some location-based service would receive messages.

For some development purposes such as understanding the location or movements of large numbers of people, individual identification or even the retention of the address is not necessary. However, retaining address (and associated identity data) may have value when, as is often the case with development-related knowledge, the MNBD or analyses are combined with data from other sources. It is important to note that considerable insights can be gained from historical, anonymized[6] big data. Of course, some development activities do require the insights of data analytics as well as the ability to communicate with the data subjects.

For the purposes of a Guideline, it was thought useful to develop a classification of MNBD. In actual fact, different kinds of MNBD may be situated on different continuums, rather than in discrete categories. MNBD may be classified in various ways. In practical terms, there is value in classifying based on actions related to processing.

Anonymization and real-time versus historical processing are the most relevant criteria in the present instance because the focus is on third-party use. For real-time processing to be done outside the premises of an operator, as is the case with big data from stock exchanges, significant additional expenditures would have to be incurred. These additional expenditures may be made if and when MNOs choose to make their big data available for third-party commercial use, most likely through joint ventures or subsidiaries. It is unlikely that entities seeking to harvest development insights will benefit from real-time access in the near term.

Anonymization requires additional resources; so much so that it is at present one of the most significant barriers to third parties gaining access to MNBD. Again, as in the case of making data available in real time, anonymization can be built into routine processing. But commercial incentives would be needed

---

[4] Given the preponderance of prepaid SIMs in developing economies and imperfections in subscriber registration procedures, the reliability of identifying information varies significantly from country to country and is never 100 percent. Even with regard to post-paid SIMs, it may not always be possible to identify the actual individual user since the billing information may refer to a corporate entity.

[5] An individual address is not needed to communicate via cell broadcasting.

[6] Given concerns about individual privacy, any identity-related data, e.g., phones numbers, can be replaced by unique number randomly generated number that will allow patterns to be observed, but will not allow re-identification. Safeguards against de-anonymization are necessary. Encrypting data, removing unique identifiers, or perturbing data so it no longer identifies individuals are some of the current technological solutions. But "too much" de-identification may strip the data of both its utility and the ability to ensure its provenance and accountability. It is difficult to predict how technologies to re-identify seemingly anonymized data may evolve.

for these kinds of additional heavy expenditures. In terms of non-commercial, third-party use in the short term, it is safe to assume that anonymization will be available at most in near real time, and for the most part as historical data, anonymized when computer resources are available.

Granular MNBD such as CDRs and VLRs are "born" as real-time identifiable data. Storage converts them to historical identifiable data. Storage costs which have decreased significantly in recent times (but are not negligible) have been a factor in the rise of data analytics. A conscious act is required to store data, which is a precondition for the analysis of historical data. On the other hand, analysis of historical data may be less of a strain on the MNO's resources since it can be done at off-peak times.

Intermediate data such as load factors of BTSes may not include individually identifiable data. Here, the software in the BTS does some processing of real-time, identifiable data to produce real-time data that does not include identifying elements.

Some things that can be done with real-time data cannot be done with historical, and vice versa. For example, it is possible to predict traffic congestion using deviations from the norm of VLRs generated by the mobile devices of people in vehicles passing roadside BTSes. If the VLR shows longer-than-normal stays by SIMs at a BTS, traffic congestion in that location may be inferred. The value is higher if these findings can be generated in real time, because they may then be communicated to vehicles that have yet to join the traffic jam along with suggestions of alternative routes. This is not to say that these data are worthless if analyzed later. They can be used for transportation planning.

The distinction between real-time and historical is also important in disaster-related applications. If evacuation managers want to ensure smooth evacuations, they need real-time data on people movements.

Criminal investigations occur after a crime is committed.[7] Nowadays, mobile data, mostly small (in the sense that the interest is in specific persons), are a valuable resource for investigators. However, in certain instances, security agencies may wish to predict and preclude certain phenomena such as large gatherings. Here, the interest would be in real-time data on movements of SIM carrying individuals and possibly even of their identities.[8] This could be in the context of political protests or in relation to public safety.

Table 1: Categories of big data generated by mobile operators

	Real time	**Historical**
Identifiable	RTI	HI
Anonymized	RTA	HA

[7] Except in science fiction such as *Minority Report*, the film, and *Person of Interest*, the TV series. The plots of both include big data.

[8] Murphy, H. (2014 January 22). Ominous Text Message Sent to Protesters in Kiev Sends Chills Around the Internet, *New York Times*. http://thelede.blogs.nytimes.com/2014/01/22/ominous-text-message-sent-to-protesters-in-kiev-sends-chills-around-the-internet/?module=Search&mabReward=relbias%3Aw

However, further analysis did not reveal significant difference in the harms or remedies applicable to real-time versus historical MNBD. Therefore, the draft Guidelines are based solely on the basis of the two categories of identifiable and anonymized data.

Analysis of MNBD may not necessarily yield causal models. For the most part, the current state of data analytics yields illuminating correlations, rather than causal explanations. Therefore, it is important to be aware of various assumptions that are built into the analysis. The normal scientific safeguards of replicability and data being available for review are absent in big data analytics causing a need for careful assessment of results and for the development of alternative safeguards.[9]

Big data generated by MNOs in the course of business has potential to make their operations more efficient and improve the management of customer relationships, including marketing. Big data also has immense, and at this point unique, potential to bring forth a qualitative transformation of urban design including resilience, improve transportation and government-service delivery, and enhance management of the economy, among others. This is especially true in developing economies where datafication is less common.

As with any phenomenon that can do good, big data also can do harm. These Guidelines seek to minimize the likelihood of harm by embedding safeguards into standard procedures while reducing the transaction costs of making the data available to third parties who will use them for public purposes. The harms that are to be addressed have been identified through the literature, described in detail in Annex 1. Ideally, the elements of the Guidelines will be incorporated into agreements between mobile operators and third-party data users.

[9] Markoff, J. (2012 May 21). Troves of personal data, forbidden to researchers. *New York Times*. http://www.nytimes.com/2012/05/22/science/big-data-troves-stay-forbidden-to-social-scientists.html?nl=todaysheadlines&emc=edit_th_20120522

Elements of draft Guidelines

Out of the dozen selected harms discussed in Annex 1, several harms were identified as being applicable to MNBD. Remedies were then developed. The harms are described in detail in Annex 1. Once there is rough consensus on the elements, LIRNE*asia* is willing to work up model Guidelines that may be adapted for use by MNOs.

Remedy	Identified potential harm	Include in agreements transferring identifiable MNBD	Include in agreements transferring anonymized MNBD
Mobile Network Operators (MNOs) will not engage in active surveillance of their customers, except as required by applicable law. MNOs will desist from collecting more data than are needed for the efficient operation of the networks and the supply of good service to customers. To the extent feasible, data collection practices will be transparent.	***Active surveillance***	No. Applying only to MNOs, this need not be included in agreements. However, active surveillance is a root cause of problems that could be manifested in other forms at the subsequent information processing and dissemination phases.	No. Applying only to MNOs, this need not be included in agreements. However, active surveillance is a root cause of problems that could be manifested in other forms at the subsequent information processing and dissemination phases.
Best efforts will be made to prevent de-anonymization. Working groups may be formed with data users to monitor the state of knowledge in techniques of anonymization and de-anonymization.	***De-anonymization***	No	Yes
Individually identifiable data will not be released to third parties, unless the purposes are specified in the agreement and have been approved by an ethics review committee, if one is available. If an ethics review committee is not available, an equivalent third-party review should be sought.	***Individual identification***	Yes	No
Any agreement transferring identifiable data to a third party will also transfer responsibility to maintain safeguards to ensure security of individually identifiable data.	***Insecurity***	Yes	No
The agreement governing the transfer will include provisions to minimize risks	***Increased accessibility***	Yes	Yes

posed by increased accessibility when data are released to third parties.			
The principle of non-discrimination shall govern the release of MNBD to third parties who do not compete with the MNO. Those in the same class will be treated equally, subject to reasonable accommodation for resource constraints.	***Anti-competitive effects***	Yes	Yes

Annex 1: Harms that are sought to be avoided

The report entitled "Big data: Seizing opportunities, preserving values" issued by the US President's Executive Office in May 2014 quotes Harvard Professor of Science & Technology Studies Sheila Jasanoff as saying that framing the policy implications of big data is difficult because it manifests in multiple contexts that each call up different operative concerns, including big data as property (who owns it); big data as common pool resources (who manages it and on what principles); and big data as identity (it is us ourselves, and thus its management raises constitutional questions about rights).

If MNBD are property, one has to ask who owns the data. The answer depends on who produced it. It could be said that the customer who moved around with a charged-up mobile device "produced" a VLR. But was the VLR actually produced solely by the customer, who in most cases did not know or barely knew her mobile device was communicating with the BTS? It may be more accurate to say it was jointly produced by the mobile operator and by the customer. Therefore, it is the joint property of the company and the customer. Minimally, the joint property may be used by either party, but in a way that does not cause harm to the other. Maximally, nothing can be done with the joint property without agreement of both. In any case, big data is more about extracting value from the raw data, than it is about the raw data itself. After all, MNBD has been in existence for several decades now with no value extracted from it. Approaching the problem from the perspective of property does not appear to be very productive, except as the basis for esoteric debates on how the co-creators can be monetarily compensated.[10]

The common-pool approach is central to making optimal use of MNBD. However, it has limited value in identifying and remedying harms. In the case of physical common-pool resources, the central question is whether use by one detracts from use by another. Big data is infinitely replicable (though not without cost) and therefore, the Lockean principle that one must leave "enough and as good" for others is not violated.

Harms in the areas of privacy, competition and marginalization are examined below.

Privacy problems

Mobile data centrally involves identity.[11] Many aspects of identity and self are covered by the common-sense understanding of privacy. Privacy, as commonly understood, "is a sweeping concept, encompassing (among other things) freedom of thought, control over one's body, solitude in one's home, control over personal information, freedom from surveillance, protection of one's reputation, and protection from searches and interrogations."[12] Attempts to define it in terms of boundary control

---

[10] E.g., O'Malley, J.P. (2013 March 22). Interview with a writer: Jaron Lanier, *The Spectator*. http://blogs.spectator.co.uk/books/2013/03/interview-with-a-writer-jaron-lanier/

[11] The US Supreme Court in *Riley v California and US v Wurie* (13-132, 25 June 2014) http://www.supremecourt.gov/opinions/13pdf/13-132_8l9c.pdf described the "small data" such as call histories, search histories and so on found on mobiles, as being "such a pervasive and insistent part of daily life that the proverbial visitor from Mars might conclude they were an important feature of human anatomy." It is not a stretch to attribute similar qualities to MNBD held by operators.

[12] Solove, D.J. (2008). *Understanding privacy*. Cambridge MA: Harvard University Press, p. 1.

by individuals[13] are difficult to translate into practical policy. For example, where are the bright-line boundaries of what an individual has authority over in the case of data generated as a by-product of a transaction?

Solove argues that privacy as an abstract concept is difficult to pin down, since it "involves a cluster of protections against a group of different but related problems."[14] He concludes correctly, that the focus should be shifted away from defining privacy, to addressing privacy problems (or harms). He proposes 16 privacy problems, grouped into four general types: Information collection (comprising surveillance and interrogation); information processing (comprising aggregation, identification, insecurity, secondary use and exclusion); information dissemination (comprising breach of confidentiality, disclosure, exposure, increased accessibility, blackmail, appropriation and distortion); and invasion (intrusion and decisional interference). MNBD harms are primarily located in the second of the clusters, information processing, and secondarily in information collection, the first cluster, and information dissemination, the third cluster.[15] Only nine relevant problems are discussed below.

There may be concern about the over-representation of US statutes and case law in the discussion below. It is partly because the US has produced most of the relevant case law. Daniel J. Solove, an American teaching law at the George Washington Law School in Washington DC, has made a genuine effort to survey non-US law, including references to India, Hungary, Japan and Sri Lanka among others, but the end result is still heavily biased to the US. This is natural because the US has been at the forefront of market and technology developments that have brought big data to the fore of legal, policy and media agendas.

Surveillance

Within the first cluster proposed by Solove, the most relevant problem is surveillance. In the context of big data, it is useful to distinguish between active and passive surveillance. Installation of devices such as a GPS tracker is active surveillance.[16] Active surveillance, where the activity is undertaken for the primary purpose of collecting data, is normally associated with law enforcement and espionage and is, for the most part, a "small data" problem. What is relevant in the context of big data is "data exhaust," or passive surveillance in the form of data that are a by-product of some activity.[17] Where systems are explicitly engineered to collect more data than are needed for normal operations, the line between passive and active is blurred.[18]

The harms are the gathering of information about a person through active or passive surveillance. The former may be prohibited or constrained. But the latter is difficult to control without negative effects on the activity that generates the data as by-product. If the base activity is one that benefits the data

---

[13] E.g., Samarajiva, R. (1994). Privacy in electronic public space, *Canadian Journal of Communication*, 19(1): 90.

[14] Solove, D.J. (2008). *Understanding privacy*. Cambridge MA: Harvard University Press, p. 174.

[15] Solove, D.J. (2008). *Understanding privacy*. Cambridge MA: Harvard University Press, ch. 5.

[16] *United States v. Jones*, 132 S. Ct. 945, 565 U.S. ___ (2012).

[17] Mundie, Craig (2014). Privacy pragmatism. *Foreign Affairs*. March/April. http://www.foreignaffairs.com/articles/140741/craig-mundie/privacy-pragmatism

[18] The US Communications Assistance to Law Enforcement Act (CALEA) of 1994 is one of the earliest examples involving electronic technology. http://itlaw.wikia.com/wiki/CALEA

subject and is one that he/she engages in willingly, there may be merit in not prohibiting collection, and instead focusing remediation on subsequent processing, as suggested by Mundie.[19]

Harm may also be caused by over-engineering systems to collect data that are not needed for normal operations.

Aggregation

Aggregation can take two principal forms, in relation to MNBD. First it is the aggregation of discrete data elements related to a single individual within one dataset, e.g., not just the datum that A called B, but the pattern of A's calls to B and vice versa. Second is the aggregation of data from different sources, e.g., from mobile CDRs and from surveys or from payment terminals in shops. Anonymization is not necessarily a barrier to the former.[20] Anonymization makes the latter form of aggregation much harder. However, it is also possible that de-anonymization can be achieved if one of the datasets has identity information.[21]

It is widely accepted that aggregated data yields a richer picture than non-aggregated data. But aggregation may also reduce the potential for wrong conclusions being drawn from the partial picture presented by non-aggregated data.[22]

Therefore, the first set of potential harms comprises of errors caused by aggregation or lack thereof. The second is about "true" insights drawn through aggregation, when the "truth" is not intended to be disclosed. The third is about the dangers of identification through de-anonymization made possible because of aggregation. At the individual level, the third is the most critical.

One may ask what harm is caused by erroneous or "truthful" information generated through aggregation as long as the data subject is anonymous. So for example, one may conclude through aggregation that a particular data subject has undergone an illegal/morally questionable medical procedure. This may be true, or may be false because the aggregation was incomplete and missed some significant data (the data subject may be visiting the medical facility for a different reason). As long as the data subject's identity is not known, it is difficult to discern the harm.

Therefore, the true harm is in the likelihood that aggregation may permit identification through de-anonymization.

Identification, individual and group

Identification is a central concept. According to Solove, identification "is connecting information to individuals. . . . Aggregation creates . . . a portrait composed of combined information fragments. Identification goes a step further—it links the digital person directly to a person in realspace."[23]

---

[19] Mundie, Craig (2014). Privacy pragmatism. *Foreign Affairs*. March/April. http://www.foreignaffairs.com/articles/140741/craig-mundie/privacy-pragmatism

[20] As long as each identity-related element is replaced by the same number string.

[21] De Montjoye, Y-A; Hidalgo, C.A.; Verleysen, M.; Blondel, V.D. (2013) Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports* 3, Article number: 1376 doi:10.1038/srep01376 http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html

[22] Recognizing, of course, that all data are partial representations of "reality." The debate is not about fully accurate versus inaccurate, but about the relative veracity of partial representations.

It is clear that identification is an essential element of the postulated harms at the individual level, where much, if not all, of the privacy discussions focus. But it is also the essential element in harms at the collective or group level.

It is widely believed that there is greater consumption of adult or pornographic entertainment when conventions attended large numbers of Christian Evangelicals are held at US hotels.[24] Whether true or false, this perception is harmful to the image of Christian Evangelicals.

Let us assume that there is hard evidence to substantiate the above claim. For that, it would not be necessary for the hotels to release the video viewing records of individuals, violating the provisions of the US Video Privacy Protection Act of 1988. Instead, they could simply provide the aggregate use records by title or category of videos. With this information, one could observe the peaks and valleys of adult entertainment use by date in a specific hotel or hotels. It may be mildly interesting, but not newsworthy.

However, if the information on daily consumption of adult entertainment is correlated with the numbers of hotel guests attending specific conventions, be they atheist or evangelical, the story begins to become interesting. If one can establish a consistent correlation between increased consumption of adult entertainment over a period of time and with particular kinds of conventions, especially if the viewing of such movies is contrary to the public positions taken by the convention organizers, the story becomes truly interesting. It could, ostensibly, be damaging to a group that is hostile to depictions of sexual behavior in entertainment.

This is an example of a breach of collective or group privacy, as commonly understood. The simple aggregation of individual video rental records does not constitute the breach; it is the combination of that data with data identifying the group. Here too, the conceived harm is connected to identification of the group.

It is critically important, however, to recognize the dangers associated with safeguarding "collective privacy" or "group privacy" of the type discussed above. First, these concepts have been rarely discussed in the scholarly literature. Second, a prejudice against group attributes would pretty much put an end to social science and to efforts to improve the functioning of society in systematic, evidence-based ways. For example, it is routine to associate various characteristics or behaviors with persons living in geographical areas (e.g., rural areas, cities), by age group and gender and so on. It is not only routine but considered desirable to "target" various policy measures to specific groups. Without group identification it will be impossible for modern societies to function. This is possibly the reason why safeguards against group identification do not exist.

Insecurity

"Glitches, security lapses, abuses and illicit uses of personal information all fall into this category [of] insecurity, . . . a problem caused by the way our information is handled and protected."[25] As the volume

---

[23] Solove, D.J. (2008). *Understanding privacy*. Cambridge MA: Harvard University Press, pp. 122-25.

[24] http://gospeldrivenchurch.blogspot.com/2011/03/what-you-do-in-your-hotel-room-gives.html. This site is sympathetic to Christians and hostile to adult entertainment.

and value of aggregated data increases (becoming big data), the harms that can be caused by the data falling into wrong hands or being distorted increase. Here too, the harm is tied to identity; it is difficult to imagine what harm could be caused by anonymized data lacking any connection to an individual or group falling into the hands of a criminal.

Secondary use

"'Secondary use' is the use of data for purposes unrelated to the purposes for which the data was initially collected without the data subject's consent."[26] As can be seen from the definition, it is an artifact of law developed in the 1970s. It is anchored in practices such as individuals filling out forms and ticking boxes indicating consent that have little relation to the passive and pervasive surveillance that is the norm today. When one makes a phone call, one generates a CDR. Was the data given or collected, or was it jointly generated in the course of completing the call? How and when could consent be given? Is it possible to maintain an effective mobile network without aggregating and analyzing different elements of data within the CDR such as the loading of the BTSes? Is the use of the data for network optimization a secondary use?

Secondary-use absolutism poses the danger that uses by all but the entity co-generating the data will be prohibited. As senior Microsoft official Craig Mundie states "Today, there is simply so much data being collected, in so many ways, that it is practically impossible to give people a meaningful way to keep track of all the information about them that exists out there, much less to consent to its collection in the first place."[27]

The only way this problem can be managed is through omnibus consent forms that have to be signed at the moment of obtaining service. Depending on the skill of the lawyers drafting the documents, one would have to give consent to all imaginable uses by the service provider or make do without the service.[28] Since this particular subterfuge will not be effective in the case of third parties, the practical result will be exclusion of all third parties from the benefits of data analytics of data co-generated by others. In the case of for-profit entities, the loss will be to innovation and competition. The use of big data for public purposes will also suffer.

Exclusion

Solove proposes the term "exclusion"[29] for failure to provide individuals with notice and input about their records. He states that the harm is created by the data subject being shut out from participating in the use of the data, from not being informed about how it is used, and by not being able to affect how it

---

[25] Solove, D.J. (2008). *Understanding privacy*. Cambridge MA: Harvard University Press, pp. 127.

[26] Solove, D.J. (2008). *Understanding privacy*. Cambridge MA: Harvard University Press, pp. 131.

[27] Mundie, Craig (2014). Privacy pragmatism. *Foreign Affairs*. March/April. http://www.foreignaffairs.com/articles/140741/craig-mundie/privacy-pragmatism, See also, Viktor Mayer-Schonberger's views at https://privacyassociation.org/news/a/keynote-forget-notice-and-choice-lets-regulate-use/

[28] "Because privacy notices under the 1980 Guidelines constrain future data uses, notices have become increasingly broad and permissive. The result has been the increasing erosion of information privacy." –Cate, F.; Cullen, P.; Mayer-Schonberger, V. (2013, December). *Data protection principles for the 21st century: Revising the OECD guidelines*. http://nova.ilsole24ore.com/wordpress/wp-content/uploads/2014/01/Data_Protection_Principles_for_the_21st_Century.pdf

[29] Perhaps the least felicitous of the set.

is used.[30] While it is present in Fair Information Practices, Solove states that "for the most part, tort law has not recognized exclusion as a harm."[31]

The Kafka quotation used by Solove[32] illustrates the possible harm: "For in general the proceedings are kept secret not only from the public but from the accused as well." When benefits/harms are decided on the basis of data sets, the argument is that they must be known and subject to correction. While this is intuitively correct for credit reports, the "ground zero" of modern privacy remedies. But the harms are small compared to the massive transaction costs that would be associated with notifying all data subjects whose data are in big data sets and permitting them rights to examine and correct them. For example, every BTS in a mobile network contains data on thousands of "data subjects" including ephemeral data as such as what is recorded on the VLR on when they moved within the range of the BTS and when they moved out. It would serve little purpose to notify them of this. The transaction costs would be very high. Allowing access to commercially sensitive data sets would also not be practical.

Exclusion, therefore, poses no harm in relation to MNBD. It could, however, be the cause of considerable problems in the form of high transaction costs if attempts were made to apply remedies that may have been appropriate in the days of credit reports.

Breach of confidentiality

Most privacy problems sought to be addressed by the tort of breach of confidentiality are not relevant to MNBD. It requires consideration because of the "third-party doctrine" exemplified by the *United States v. Miller* and *Smith v. Maryland* decisions which govern government access to MNBD of individuals (small data).[33] In the former, the Supreme Court held that no breach occurred when a person's bank records were released to government because "all of the documents obtained, including financial statements and deposit slips, contain only information voluntarily conveyed to the banks and exposed to their employees in the ordinary course of business."[34] In *Smith v Maryland*, the logic was extended to call details (not the content of the call), on the basis that people "know that they must convey numerical information to the phone company," and, cannot "harbor any general expectation that the numbers they dial will remain secret."[35]

The US government's justification for the collection and use of telephone metadata pertaining to US citizens by the NSA exposed by Snowden was based on the third-party doctrine, derived from the above judgments.[36] A 2013 decision from the District Court of the District of Columbia (perhaps the most important, because Washington, DC is in the District) attracted a lot of attention because it explicitly

---

30 Solove, D.J. (2008). *Understanding privacy*. Cambridge MA: Harvard University Press, pp. 134.

31 Solove, D.J. (2008). *Understanding privacy*. Cambridge MA: Harvard University Press, pp. 135.

32 Solove, D.J. (2008). *Understanding privacy*. Cambridge MA: Harvard University Press, pp. 133.

33 425 U.S.435 (1976) and 442 U.S. 735, respectively.

34 425 U.S. 435 (1976), at 442-43.

35 442 U.S. 735, at 743.

36 Savage, C. (2013 October 18). NSA plan to log calls is renewed by court. *New York Times*, http://www.nytimes.com/2013/10/19/us/nsa-plan-to-log-calls-is-renewed-by-court.html?module=Search&mabReward=relbias%3Aw

contradicted the Smith rationale.[37] However, a subsequent decision by a District Judge from the Foreign Intelligence Surveillance Act (FISA) Court responsible for oversight of the National Security Agency's surveillance activities reaffirmed the third-party doctrine.[38] Until the various appeals work their way up to the Supreme Court, *Smith v Maryland* will continue as the ruling precedent in the US. As stated by the FISA judge: "The Supreme Court may someday revisit the third-party disclosure principle in the context of 21st-century communications technology, but that day has not arrived."[39]

It must be noted that there is no question in either Miller or in Smith about whether the bank and the telephone company can use the data. The only question at issue was whether the data could be given to a third party, the government, without the data subject's authorization. Since the focus here is on use of MNBD by third parties, the privacy problem or harm may be restated as one of harms cause by aggregation and identification at the individual or collective levels, as discussed above.

Disclosure

Disclosure refers to disclosure of true information about a person. In many countries, there are laws restricting the disclosure of data from educational institutions, video rental companies, health services, etc. The harm caused by disclosure is damage to reputation. Reputation being tied to identity, anonymization can avoid the harm.

Increased accessibility

Here, the information is public, but is difficult to get to. This is an important issue in the context of the Internet, with its easy search capabilities, and the increasing trend toward open data and open government. It primarily applies to public records held by government and not to data held by private entities where there is no presumption of openness.

But the issue may become relevant if and when MNBD in raw or semi-processed form are made available on the web, especially if these actions are a result of government direction.

Anti-competitive effects

Competition is seen as a good. It is seen as requiring a metaphorical "level playing field" that gives all competitors equal opportunities, though not identical or equal outcomes.

In infrastructure industries, it is recognized that the competition that exists deviates from the ideal to a greater or lesser extent. It is recognized that certain elements such as "essential facilities" have monopoly characteristics and that policy and regulatory safeguards have been set in place to prevent the extension of market power into other market segments that are otherwise conducive to competition.

---

[37] Klayman v Obama, Civil Action 13-0851(RJL). http://www.nytimes.com/interactive/2013/12/17/us/politics/17nsa-ruling.html?ref=politics&_r=0

[38] Savage, C. (2014 April 25). Phone company bid to keep data from NSA is rejected. *New York Times*, http://www.nytimes.com/2014/04/26/us/phone-company-bid-to-keep-data-from-nsa-is-rejected.html?emc=edit_th_20140426&nl=todaysheadlines&nlid=9770121

[39] Savage, C. (2013 October 18). NSA plan to log calls is renewed by court. *New York Times*, http://www.nytimes.com/2013/10/19/us/nsa-plan-to-log-calls-is-renewed-by-court.html?module=Search&mabReward=relbias%3Aw

The 1982 Consent Decree[40] that divested AT&T into seven Regional Bell Operating Companies (RBOCs) and a long-distance and information services company that retained the AT&T name was a pivotal event with significance not limited to the US borders. The Consent Decree sought to provide a structural remedy for the alleged anti-competitive actions of AT&T by separating the potentially competitive segments (new AT&T) and the monopolistic segments (RBOCs) into structurally independent companies. When an RBOC wished to offer a new service, it had to obtain prior approval from the District Court Judge who maintained authority over the Decree.

The approval was based on competitive implications for firms that were offering services that depended on the monopoly segment controlled by the RBOC. In some cases, there were additional conditions imposed by the relevant regulatory authorities. For example, when the courts permitted RBOCs to offer enhanced services, the Federal Communication Commission (FCC) mandated that the RBOC obtain prior authorization from business customers with more than 20 lines before permitting RBOC marketing personnel to access Customer Proprietary Network Information (CPNI).[41]

The Consent Decree's design to control AT&T's anti-competitive conduct through structural separation and the policing of the monopolistic-competitive boundary did not last very long in the face of pressure from the companies and rapid technological and market changes. It is referred to here to illustrate the fact that policy and regulatory authorities have accepted that data generated in the course of providing services in one market segment can have implications for the "level playing field" in another related market.

The difference between the fact-pattern examined in the 1992 NRRI Report and that existing at present is that the RBOC then had almost total coverage and thus had a unique informational advantage; whereas today's mobile operators in most countries do not.

Today, there are few barriers to mobile operators offering services that ride on their networks in competition with other firms. On one hand, it is in the interest of the mobile operators to have others provide services over their networks thus generating additional data consumption and greater utilization and revenues. On the other, it may be in their interest to "tilt" the playing field in favor of services they offer and to the disadvantage of non-vertically integrated competitors.

In this context, it will be necessary to consider anti-competitive implications alongside consideration of the privacy problems discussed above.

Marginalization

The former President of Brazil and eminent scholar, Fernando Henrique Cardoso, said that countries of the South faced two dangers, the first being that of exploitation and the second ending up in "the 'worst

[40] 552 F.Supp. 131 (DDC 1982).

[41] Burns, R.E.; Samarajiva, R.; Mukherjee, R. (1992 September). *Utility customer information: Privacy and competitive implications*. NRRI 92-11. Columbus OH: National Regulatory Research Institute, p. 121.

of all possible worlds.' They will not even be considered worth the trouble of exploitation; they will become inconsequential . . ."[42]

Anyone who has moved to a new country and tried to obtain services that require a credit rating will understand the consequences of exclusion. The earliest data protection laws were enacted in response to the rise of credit cards and credit reporting.[43] They were intended to safeguard against problems of data aggregation, identification, insecurity, secondary use, etc. But an unintended consequence of rules against credit reports being sent across national borders was that individuals who crossed those borders were denied services or had to make do with more expensive options until they built up credit histories in the new country. Being excluded from the purview of data analytics based on MNBD could have a similar effect on large segments of populations, especially those in the developing countries.

[42] Cardoso, F.H. (1993). "North-South relations in the present context: A new dependency?" in *The new global economy in the information age*, p. 156. College Park PA: Pennsylvania University Press.

[43] Rule, J.; McAdam, D.; Stearns, L.; Uglow, D. (1980). *The politics of privacy: Planning for personal data systems as powerful technologies*. New York: Elsevier, ch. 6.