Big data for development for the Global South: A research and policy agenda

Rohan Samarajiva, Ph.D.

March 2017

LIRNE*asia*
info@lirneasia.net | www.lirneasia.net

LIRNE*asia* is a pro-poor, pro-market think tank whose mission is *Catalyzing policy change through research to improve people's lives in the emerging Asia Pacific by facilitating their use of hard and soft infrastructures through the use of knowledge, information and technology.*

Contact: 12 Balcombe Place, Colombo 00800, Sri Lanka. +94 11 267 1160. info@lirneasia.net
www.lirneasia.net

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

Table of Contents

Introduction

Business analytics is all the rage in the private sector. For the most part the buzz phrase "big data" is associated with companies crunching massive volumes of data to figure out what to sell more effectively. Strawberry flavored pop tarts are said to be much in demand just before hurricanes hit the East Coast of the United States. Walmart is supposed to have gleaned this valuable insight by analyzing big data (Hayes, 2004). How Target figured out someone was pregnant from their purchasing patterns is also part of the conversation (Duhigg, 2012).

Big data applications for public purposes began to be researched a few years after the new millennium (e.g., Weslowski & Eagle, 2009). More than in the case of business analytics, big data for development (BD4D) requires greater attention to representivity, validation and attention to the minimization of harm. The fact that much of the relevant datafied records (Mayer-Schonberger & Cukier, 2013) happened to be in the hands of private firms because of the slower pace of computerization within government added a layer of complexity.

BD4D research is gathering momentum. Efforts are being made by actors ranging from the World Economic Forum to the Data Pop Alliance to establish frameworks for the conduct of BD4D research. With a few exceptions researchers from the Global South are absent at these deliberations. This is not simply about representation. The conception of BD4D research that sees it as some kind of magic bullet that would displace all prior research methods is not defensible. It can provide highly valuable insights but it will not displace prior methods completely. The insights have to be contextualized and validated. For that, it is important have local voices at the table. It is also more likely that government and other actors will use the insights when they are communicated by local researchers.

The focus of this discussion paper is to begin the process of involving organizations from the global south in shaping the BD4D research and policy agenda. Given the heavy emphasis being placed on possible harms in the ongoing conversations, the paper also gives extra weight to the potential harms caused, not only in terms of privacy, but also in terms of marginalization as when large swaths of the populations are excluded from the analysis because their data are not analyzable. In addition, group harms and harms to competition and innovation are also discussed.

Other issues such as the multi-disciplinary challenges, addressing gender and the opportunities afforded by the metric-challenged sustainable development goals are briefly set out for discussion. Finally a proposal for an articulated set of actions that would develop capacity for BD4D research and policy in the global south and create conditions for mutual learning and policy impact is presented as a basis for deliberation on future actions.

Why big data? Why now?

Information and control are closely connected. Beniger (1986: 7-8) states that the twin activities of information processing and reciprocal communication (or feedback) are inseparable from the concept of control. Control is defined in the broadest sense as "purposive influence toward a predetermined goal." Even though he wrote well before the current democratization of big-data

analytics,[1] Beniger provides possibly the best answer to the questions "why big data?" and "why now?"

He postulates that the roots of present-day developments associated with information and communication technologies (ICT) lie in the industrial revolution, which saw an unprecedented speeding up of the entire material processing system. This, he states, precipitated a series of crises of control, wherein innovations in information processing and communication technologies lag behind those of energy and its applications in manufacturing and transportation. Each crisis is resolved by a burst of innovation associated with ICTs, described as a control revolution (p. vii).

The revolution relevant to the emergence of big-data analytics is associated primarily with the control of consumption.

Until the 19th Century, the dominant mode of production was craft production. The distance between producer and consumer was not big. To some extent, production took into account the preferences of consumers. Volumes were low; quality was variable.

Mass production is qualitatively different. Volumes are high; quality is uniform. What the consumer needs has to be imagined. But this mode of production is prone to crises of under-consumption. Beniger (1986) describes some of the innovations that sought to control the system to prevent it from going into crisis. But the core problem is consumer behavior. The solutions are periodic improvements in its control and alignment with production. Advertising is a critical element in this effort.

Advertising plays a dual role in market processes. On one side, it helps buyers "discover" what is available to satisfy their needs. On the other, it shapes and even creates needs; in Beniger's terms, it controls. Conventional advertising sent identical messages to an imagined homogenous audience, with little feedback received. But the ideal solution has always been to understand individual consumers and target them in a customized manner. If the product or service can also be customized, the targeting is likely to be even more effective.

Now, finally, ICTs are beginning to make that possible. Ability to generate feedback and process it into actionable insights are now possible. Better feedback is provided by the subset of behavioral big data that can be described as transaction-generated data (TGD), than error-prone survey results. Survey results use up respondents' time and attention. TGD is a by-product of their activities and imposes no costs on the data subjects.

Analytics using TGD is pulled into existence to meet the requirements of control not just of marketers of consumer goods and services, but of urban planners, designers of social programs, political campaign managers, etc. On the surface big data has achieved prominence at this time because of lower costs of computer processing, memory and software. But the fact that analytics

[1] Big data analytics is not new. Back in the 1980s, the National Security Agency, the Central Intelligence Agency, and their counterparts in developed economies were using super computers for this purpose. In the private sector, only very large enterprises such as American Express were doing analytics and were listed as purchasers of super computers. See, Samarajiva (1996).

were being conducted by the National Security Agency (NSA) and by companies such as American Express using supercomputers shows that the underlying causes run deeper (Samarajiva, 1996).

As with previous bursts of innovations in control, this will not bring the system into a state of equilibrium. But as with prior bursts of control innovations, it will make life easier in the short term for entities that get in front of the phenomenon.

Placing the present developments in analytics in historical context is helpful. The introduction of zip or PIN codes in postal addresses was a control innovation. They were the basis of forms of audience segmentation and targeting used in marketing. The collection, analysis and use of TGD are driven by an underlying economic logic that could be described as inexorable. Such forces can be shaped,[2] but it is doubtful whether they can be stopped or reversed.

Among big data, the most important and sensitive subset right now comprises the above described TGD. However, there are other kinds of high volume and variable data that are of relevance. These include data generated from sensors, placed in various locations. For example, insights are being generated from earth observation satellites. As with TGD, data from sensors on satellites have been used for many decades. The difference is that data from multiple satellites are now available to third-party users and the costs of hardware and software have come down. The control logic applies to big data generated by sensors, including but not limited to satellites.

Public policy is necessarily intertwined with issues of control that can range from "hard' or "soft" control of behavior to the control of undesirable forms of control of one group in society by another. Much of present-day concerns about the negative effects on privacy are based on the perceived increase in the gathering of data that could lead to greater control (generally understood as hard, and therefore undesirable, control, not the Benigerian form); about more aspects of citizen's lives being made visible to governments or to corporations. But this is not the sole issue worthy of attention.

As public-policy decisions become increasingly based on analytics, it is necessary to address issues of marginalization, wherein some subsets of the population are excluded from consideration because they are not represented in the data or not legible to the state. On one side this is about the quality of the findings produced by analytics. On the other, it is about equity.

Some, such as Scott (1998: 183-84) inspired by anarchist writers such as Proudhon, see legibility as intrinsically tied to undesirable aims such as manipulation. For them, it is good for the citizens to be not fully known because what is not known cannot be manipulated. Those who do not share the anarchistic conception of the state and who instead see public policy as an instrument, albeit imperfect, of addressing issues of effectiveness, efficiency and equity cannot share this conception of legibility. For them, what is undesirable is marginalization, the opposite of legibility. Unless all citizens (and in some cases, all residents in a territory) are visible, the data are incomplete or erroneous. Therefore, the public policies based on such data are ineffective. Since in their view public policies are well intentioned, lack of legibility would also harm the goals, especially of equity. Those who are invisible would be denied the benefits. Another form of this concern is about

[2] Or controlled, in the broad sense used by Beniger.

misrepresentation. If data subjects are inaccurately represented that can lead to flawed public policies and detriment to all or some of the data subjects.

Another important public-policy issue is that of competition. In a market economy, economic activity and innovation occurs in a decentralized manner. Consumer welfare is optimized by the checks and balances provided by the existence of supplier competing on a level playing ground. For this reason, public policy has sought to break up concentrations of market power caused by actions defined as being anti-competitive. When certain actors in economic value chains accumulate vast amounts of behavioral big data, there is a legitimate question about the impact on other economic actors and market segments.

In contrast to the static view that privileges consumer welfare only, the dynamic perspective gives weight to effects on innovation. Here, the principal concerns would be the effects on the ability of startups to innovate.

Marginalization

Lerman (2013) sketches out two archetypes relevant to big data analytics, and extends to a third:

> The first is a thirty-year-old white-collar resident of Manhattan. She participates in modern life in all the ways typical of her demographic: smartphone, Google, Gmail, Netflix, Spotify, Amazon. She uses Facebook, with its default privacy settings, to keep in touch with friends. She dates through the website OkCupid. She travels frequently, tweeting and posting geotagged photos to Flickr and Instagram. Her wallet holds a debit card, credit cards, and a MetroCard for the subway and bus system. On her keychain are plastic barcoded cards for the "customer rewards" programs of her grocery and drugstore. In her car, a GPS sits on the dash, and an E-ZPass transponder (for bridge, tunnel, and highway tolls) hangs from the windshield.
>
> . . . .
>
> Now consider a second person. He lives two hours southwest of Manhattan, in Camden, New Jersey, America's poorest city. He is underemployed, working part-time at a restaurant, paid under the table in cash. He has no cell phone, no computer, no cable. He rarely travels and has no passport, car, or GPS. He uses the Internet, but only at the local library on public terminals. When he rides the bus, he pays the fare in cash.
>
> Today, many of big data's tools are calibrated for our Manhattanite and people like her—those who routinely generate large amounts of electronically harvestable information. A world shaped by big data will take into account her habits and preferences; it will look like her world. But big data currently overlooks our Camden subject almost entirely. (And even he, simply by living in a U.S. city, has a much larger data footprint than someone in Eritrea [the third archetype], for example.)

Lerman's short piece on exclusion is an exception to the general emphasis on problems attributed to inclusion. In many fields of public policy, practitioners are well aware of the problem of exclusion, such as that of those who administer sample surveys oversampling roadside communities and thus

marginalizing those who are more difficult to reach, and post-disaster aid not reaching those in less visible locations.

Box 1: Boston's Street Bump app

The City of Boston makes available an app called Street Bump that can be downloaded to smartphones. Any citizen can place the smartphone in a holder in a car and press one button to start the app at the beginning of a journey. No calls would be taken during the journey. The accelerometer of the smartphone collects data that has been proven to be effective in identifying pot holes and speed bumps. At the end, another button is pressed and the collected data including the GPS coordinates of the starting and ending points are sent to City Hall. Algorithms differentiate between the bumps that should be there and those that should not be. Roads with an excess of the latter get routed into the work order system for repairs.

The assumption is that smartphones are ubiquitous in Boston. What if a similar crowdsourced big-data application is deployed in a city which has less than 10 percent smartphone users? Even if penetration is higher, if smartphones and cars have significantly lower penetration is certain parts of the city? Because resources are always limited, will this result in greater resource allocations to areas with more wealth?

The issue has to be situated within the larger problem of representivity (Miller, et al., 2015; Samarajiva, 2014). Miller, et al. propose an approach that would require researchers to explicitly address the representivity of a particular data set in the hope that over-broad claims will not be made for it and biased policy prescriptions that would not be derived from the findings. Samarajiva, et al. (2015) argue for reliance on the less rich data generated by mobile networks (as against smartphones) in developing countries to avoid marginalizing the poor. The outcomes of marginalization may be optimal in terms of privacy because none of the privacy harms are caused by marginalization. Indeed, marginalization may well describe the aspiration of privacy absolutists.

It must be noted that marginalization is not a binary condition, but that there is a continuum of conditions. Certain groups such as the homeless or illegal immigrants are marginalized by conventional surveys and censuses. Mobile network big data (MNBD) cover more people than data collected from smartphones or from Twitter, but do not cover every person. Every method has its biases. The best that can be done is to be aware of biases and to use the method best suited for the purpose.

Privacy harms

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" (Solove, 2008, p. 1). Attempts to define it in terms of boundary control by individuals (e.g., Samarajiva, 1994: 90) are difficult to translate into practical policy. For example, it is difficult to clearly demarcate what an individual has authority over in the case of data generated as a by-product of a transaction, where the data are co-produced and held by one party.

Solove (2008, p. 174) contends that privacy as an abstract concept is difficult to pin down, because it "involves a cluster of protections against a group of different but related problems." 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) (Solove, 2008, Ch. 5). Harms that may be caused by behavioral big data or transaction-generated data that fall within the scope 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.

1.1. Surveillance

Within the information-collection cluster proposed by Solove, the most relevant problem is surveillance. In the context of behavioral big data, it is useful to distinguish between active and passive surveillance. Installation of a device such as a GPS tracker constitutes active surveillance.[3] Active surveillance, where the activity is undertaken for the primary purpose of collecting data on a specific individual is normally associated with law enforcement and espionage and was in the past a "small data" problem. However, efforts to emulate e-commerce sites by tracking customers as they move through brick and mortar shops have moved active surveillance into "realspace" (Clifford & Hardy, 2013).

What is relevant in the context of big data is passive surveillance in the form of data that are a by-product of some activity (Mundie, 2014). Where systems are explicitly engineered to collect more data than are needed for normal operations, the line between passive and active is blurred.[4]

The harms are the gathering of information about a person through active or passive surveillance. The former may be prohibited, constrained or subject to notice requirements. But the latter is difficult to control without stifling the activity that generates the data as by-product. If the base activity is one that benefits the data 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 (2014).

1.2. Aggregation[5]

Aggregation, as defined by Solove (2008), can take two principal forms in relation to behavioral big data. 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 interacted with B, but the pattern of A's interactions with B and vice versa. Second is the aggregation of data from different sources, e.g., from mobile networks and from surveys or from payment terminals in shops. Pseudonymization is not a barrier to the

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

[4] 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

[5] The term aggregation is here used not as a tool for obscuring identity as it is sometimes understood, but exactly in the sense used by Solove (2008). It is a technical term that is central to his analysis.

aggregation of data regarding a person within a dataset, though the resulting insights about the digital person will not be connected to the person in "realspace." Pseudonymization makes aggregation across multiple data sets more difficult.

Aggregated data yields a richer picture than non-aggregated data. Aggregation may also reduce the potential for wrong conclusions being drawn from the partial picture presented by non-aggregated data.[6]

Therefore, the first set of potential harms comprises 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 significant.

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 cannot be identified, it is difficult to discern the harm at the individual level.

However, harm may occur to an organization or a group using that organization's services. It may be possible to infer the location of an illegal service provider using aggregated anonymized/pseudonymized data sets even if the identities of individuals using the services continue to be effectively masked. While the specific persons included in the data sets may escape prosecution, the organization providing the service and future users may suffer the consequences of engaging in actions illegal under that country's laws. Increasingly, law enforcement authorities are using analytics for purposes such as predictive policing (Perry, et al., 2013). Whether we describe the consequences of such actions as harmful or not depends on the purpose. If against criminals or those engaging in socially undesirable actions, it is unlikely that it will fall within the definition of harm as discussed here.

At the individual level, the harm is in the likelihood that aggregation may permit identification through de-anonymization. At the group level, it is possible that harm may result if techniques used in law enforcement are used against non-criminal actors.

1.3. Identification, individual and group

Identification is a central concept. According to Solove (2008: 122-25), 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."

[6] 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 privacy discussions focus. Group identification is also the essential element in harms at the collective or group level (discussed below under group harms).

1.4. 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" (Solove, 2008, p. 127). As the volume 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 at the individual level is tied to identity. Effectively anonymized data falling into the hands of an ill-meaning or unintended person or organization is unlikely to cause a person whose data are included within the data set harm.

However, some scholars such as Taylor (2015) contend harms may be caused to groups from anonymized data falling into the hands of unintended persons.

The harms caused by insecurity are increasingly common. The standard privacy remedies anchored on inform-and-consent play no role in alleviating this harm.

1.5. 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" (Solove, 2008: p. 131). The definition hints that it is an artifact of law developed in the 1970s 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 Call Detail Record (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 a Base Transceiver Station (BTS)? 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 Mundie (2014) 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."

One way this problem may be managed is through omnibus consent forms that may be obtained at the moment of establishing the commercial relationship. Depending on the skill of the lawyers drafting the documents, one would have to give consent to all imaginable uses by the provider of goods or services, or make do without the service.[7] Since this particular subterfuge will not be

[7] "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, Cullen & Mayer-Schonberger (2013).

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 privately held big data for public purposes will also suffer.

1.6. Exclusion

Solove (2008, pp. 134-35) proposes the term "exclusion"[8] 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 is used. While it is present in Fair Information Practices, Solove (2008, p. 207) states that "for the most part, tort law has not recognized exclusion as a harm,"

The Kafka quotation used by Solove (2008, p. 133) 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 not only the data but the algorithms that are used to extract insights from them must be known and subject to correction (Pasquale, 2015; Tufekci, 2014).

Concern about exclusion or opacity is intuitively correct for credit reports, the starting point 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 transaction-generated 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 Visitor Location Registry (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 so high that use of the data would be all but impossible. Allowing access to commercially sensitive data sets would also not be practical.

The algorithms applied to the data to produce insights pose difficulties of a higher order of magnitude. Even if the data were understandable, there are few realistic solutions to the problem of eliminating the opacity of the algorithms (Pasquale, 2015, ch. 6).

Exclusion, therefore, poses no harm in relation to many forms of transaction-generated big data. 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.

1.7. Breach of confidentiality

Most privacy problems sought to be addressed by the tort of breach of confidentiality are not relevant to the TGD subset of big data. 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 transaction-generated data of individuals (small data).[9] In the former, the US

[8] Perhaps the least felicitous of the set.

[9] 425 U.S.435 (1976) and 442 U.S. 735, respectively.

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."[10] 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."[11]

The US government's justification for the collection and use of telephone metadata pertaining to US citizens by the National Security Agency (NSA) exposed by Snowden was based on the third-party doctrine, derived from the above judgments (Savage, 2013). A 2013 decision from the District Court of the District of Columbia (perhaps the most important, because Washington DC is within the District) attracted significant attention because it explicitly contradicted the Smith rationale, stating that the surveillance of meta-data in 2013 was qualitatively different from that which was decided in 1979.[12] 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. 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" (Savage, 2013)."

It must be noted that there is no question in either *Miller* or in *Smith* about whether the bank and the telephone company could 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 transaction-generated data 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.

1.8. Disclosure

Disclosure refers to disclosure of true information about a person. In some 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/pseudonymization can avoid the harm at the individual level. There may be circumstances under which groups suffer harm, but they have to be dealt with on a case-by-case basis, outside the realm of privacy.

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

10 425 U.S. 435 (1976), at 442-43.

11 442 U.S. 735 (1979), at 743.

12 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

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 data such as MNBD in raw or semi-processed form are made available on the web, especially if these actions are a result of government direction.[13]

Group harms

Identification is central to all discussions of privacy. 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" (Solove, 2008, pp. 122-25).

Identification is an essential, if not the core, element of the postulated harms at the individual level, where much of the conventional privacy discussions are focused upon. But even absent identification at the individual level, it may contribute to postulated harms at the collective or group level.

Group or collective harms may be illustrated thus. It is widely believed that there is greater consumption of adult or pornographic entertainment when conventions attended by large numbers of Christian Evangelicals are held at US hotels.[14] Whether true or false, this perception harms the collective image of Christian Evangelicals in the United States by showing them up as hypocrites.

To substantiate the above claim, it would not be necessary for hotels to release the video viewing records of individuals, an act that would violate the provisions of the US Video Privacy Protection Act of 1988. Instead, the hotels could simply provide the aggregate use temporal records by title or category of videos together with the numbers of guests attending Evangelical and other conventions and when. With this information, it would be possible to correlate the consumption of adult entertainment in hotels with Evangelical and other conventions.

This is an example of a group harm, described by some as a breach of collective privacy. 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. The harm is connected to identification of the group not the individuals.

It is critically important, however, to recognize the dangers associated with attempting to build safeguards against collective harms of the type discussed above.

Rights are usually understood to belong to individuals, not to groups. The only group or collective right recognized in international law is that of peoples having the right of self-determination.[15] Even

---

13 For a discussion in the context of open government, see Borgesius, van Eechoud, & Gray (2015).

14 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.

15 The United Nations, *International Covenant on Civil and Political Rights*, Article 1.

with this right, the value and operationalization of group rights are highly contested in the literature.[16]

Furthermore, a prejudice against actions based on group attributes would pretty much put an end to efforts of the state to improve the functioning of society in systematic, evidence-based ways through public-policy instruments. For example, it is routine to associate various characteristics or behaviors with persons living in geographical areas (e.g., in poverty mapping), by age group and gender and so on. It is considered desirable to "target" various policy measures to specific groups and indeed to improve the targeting by various means. Without group identification it will be impossible for the modern state to function. This is possibly the reason why safeguards against group identification do not currently exist and are not likely to exist in the future.

A contrary position is advanced in a book entitled Group privacy which seeks to extend privacy into a group right is to be published in 2017 by Luciano Floridi, Bart van der Sloot and Linnet Taylor of the University of Amsterdam.[17]

Harms to competition

> *The traditional belief is that we want firms to compete to provide the best mix of products and services. But if the critical resource in many multi-sided markets is data (not merely to target advertising, but also to optimize the products and services themselves), then the firms with a competitive advantage in the four Vs of data are not merely in the best position to dominate their own sectors—they are also poised to take over adjacent fields* (Stucke & Grunes, 2016: 335).

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.

The 1982 Consent Decree[18] 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

---

16 Group rights, *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu/entries/rights-group/

17 The central argument is in Taylor (2015) and a response by Floridi (https://www.academia.edu/14389367/Open_Data_Data_Protection_and_Group_Privacy). Book information at http://www.springer.com/us/book/9783319466064.

18 552 F.Supp. 131 (DDC 1982).

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).[19]

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 many entities that possess TGD in most countries do not. Exceptions are monopoly suppliers of services such as energy and water distribution companies, some public-transport providers and, of course, government. For example, the French firm GDF Suez used data collected through its regulated monopoly in electricity to unfairly compete in the competitive gas-supply market. In 2014, the French competition authority found GDF Suez's conduct to be improper (Stucke & Grunes, 2016: 152-530).

Even when the entities controlling data are not regulated monopolies, they may approach monopoly status is certain markets (Rosoff, 2014). In such instances as well as in cases of mergers or acquisitions that would increase market share, it is likely that competition authorities will pay attention to the effects of data in addition to conventional competition issues. This appears to be at least part of the justification for the attention being paid to Google by European competition authorities. The issue here are whether traditional conceptions of market definitions continue to be relevant.

Increasingly, electronic networks and services are being seen as platforms where upon suppliers of products and services offer their services. If the platform is withdrawn or its quality is degraded (e.g., a change in a search-engine algorithm), these suppliers may find it difficult or even impossible to function. The entity controlling the platform possesses big data of significant value to the design of goods, services and advertising. Using this data to unfairly compete with suppliers who use the platform but do not own it is the problem that was addressed in the GDF Suez case. The other instance of potentially unfair competition would be when entities operate solely atop the platform are subject to undue discrimination, in terms of access to big data insights from the platform. Strict privacy safeguards that hinder the sharing of TGD with third parties are likely to have anti-competitive implications.

Stucke and Grunes (2016) advance the thesis that looser privacy safeguards constitute a degradation of quality in services offered. Given many of the Internet services under discussion in relation to big data are offered at a zero price, their argument is that it is meaningless to look for negative effects of market power in the raising of prices. Instead, the focus must shift to quality. Their contention is

[19] 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.

that privacy is the most significant element of quality in these services. Others would disagree, claiming that the convenience afforded by greater responsiveness to a user's needs made possible by data analytics is the most significant element.

For example, loyalty programs operated by various service suppliers such as airlines rest on voluntary diminution of privacy, as commonly understood. Because of customers' actions of enrolling in loyalty programs and of providing the membership information with transactions, the service supplier is able to aggregate the transactions and extract insights. This appears to be an instance of customers giving greater weight to the rewards and convenience offered by the loyalty program than the ability to shield patterns of behavior from the service supplier, or safeguard privacy as commonly understood. If privacy is the preeminent element of quality, loyalty programs should not exist. Identifying the appropriate weights to be given to convenience and privacy in different contexts would be useful to advance the discussion.

Aggregation of data from disparate sources is among the potential harms associated with big data, as discussed above. In terms of competition, the relevant issues are the effects of mergers and acquisitions across distinctly different markets on aggregation of data. For example, it has been argued that Google's acquisition of Nest, a supplier of home thermostats and Carbon Monoxide detectors, operating in a distinctly different market, should still have attracted greater regulatory scrutiny because of the potential of data aggregation (Stucke & Grunes, 2016: 89-92).

Challenges of melding big data research and policy studies

Even at the present formative phase of big data research, most research is being with applications in mind. Partly because of difficulties of obtaining the data, researchers are compelled to be sensitive to privacy concerns, even if not issues of marginalization, group harms and competition. On the other hand, there are those who engage solely with policy and social implications without engaging in analytics themselves. The needed skill sets are different: data science and domain knowledge for the former and law and economics for the latter (though economics tends to get neglected frequently).

There is value in conversations between the two sides. Big data research is not an individual activity. It has to be undertaken by teams, if only because of the inherent multi-disciplinary nature of the work. It does not take much to add those with expertise in the social and economic implications to teams that undertake analytics research. Including data scientists in teams that specialize in social and economic aspects is not as easy, though it would be of value. First, it's possible that the policy analysis is done by individuals, not teams. Second, data scientists engaged in actual big-data research are hard to find. They tend to be much in demand and unlikely to be attracted to sit in meetings discussing hypothetical negative implications of their work.

However, it is important that those engaged in policy analysis make the effort to understand what data is available, in what formats and what is being done with it. For example, the mobile networks in developing countries are different from those in developed economies. Given the imperative to keep costs down because revenues per user and minute are considerably lower, operators do not install non-essential network management modules. For example, the ability to geo-locate users is cruder in these networks. Networks that include researchers from both orientations would be a

pragmatic solution to the problem of ensuring that policy analysis is conducted on a sound factual basis.

Data analytics research is best done by multi-disciplinary teams. Data science and statistics knowledge must be complemented by domain knowledge. Because the same data sets can be used to obtain insights of value to a range of domains, this requires the assembling of different multi-disciplinary teams. It is one thing to talk about multi-disciplinary work in the abstract. Actually doing it is quite challenging. It requires hard work to bridge the "languages" used in different disciplines. Unless this work is done, it is unlikely that the results will be fully absorbed by policy actors in the relevant domains. For example, insights on the spread of infectious disease must be presented in language that is understandable to epidemiologists and must satisfy their criteria of quality.

Contextual knowledge matters. Much of the analytics work is still based on correlations and assumptions. It is important that adequate attention is paid to the need to ensure that the assumptions are defensible. Multi-disciplinary teams are useful in this regard. Having an organic connection to the country from where the data originates and where the insights are likely to be used is also useful.

BD4D and Sustainable Development Goals (SDGs)

Measuring performance of countries in relation to the Millennium Development Goals (MDGs) was a challenge. It is said that performance was better on the MDGs that were more easily measured. The reaction was not to narrow the scope of UN priorities to measurable goals, but to expand the range even further including a whole new set of goals with no established metrics. Apparently, the hope was that "the data revolution" would solve the problem along with the pressure created by the adoption of the SDGs.[20] Essentially, the UN decided to double down. It is too early to tell whether the strategy will be effective or not, but there are early positive signs, such as various data holding entities such as Twitter entering into agreements with UN.[21]

The efficacy of the SDGs and the strategy for keeping track of them is not relevant to the present discussion. What is relevant is the fact that there will be strong demand for insights on metrics to measure progress on the 17 goals and the associated 169 targets. To the extent that BD4D researchers can contribute to the herculean task of developing metrics at the national level, there should be demand for their insights and, hopefully, resources to produce the insights.

Addressing gender implications of BD4D research

Pseudonymization intended to address privacy concerns often wipes out gender and other demographic attributes from datasets. However, it is possible to overcome this problem by close

[20] Comments of Alex Pentland, a co-author of the Data Revolution report commissioned by the UN Secretary General (http://www.undatarevolution.org/) at Symposium on Big Data and Human Development, Oxford Internet Institute, 16 September 2016.

[21] http://www.un.org/sustainabledevelopment/blog/2016/09/twitter-and-un-global-pulse-announce-data-partnership/

collaboration with government-run large-sample surveys or through the conduct of customized surveys to calibrate and train the models to be run on the big data such as mobile network big data. Behavior attributes that most accurately predict gender can be found (Blumenstock, 2012; Frias-Martinez, Frias-Martinez, & Oliver, 2010). Such methods can then provide for gender disaggregation of ongoing efforts to understand socio-economic wellbeing cell-phone users. On going efforts (some supported by Data2X) are also exploring the potential of inferring gender characteristics from born-public data such twitter feeds and other social media content. Similar efforts are also underway to explore the use of satellite imagery to to increase the spatial resolution of existing information from standard surveys, such as the DHS, on key indicators of relevance to women's welfare.[22]

However these efforts are still very much in its embryonic stages. Its level of consistent reliability especially in the case of the use of satellite data and social media are not yet known. With social media represenatitivty will be even harder to establish.

Big data research will complement, not replace previous qualitative and quantitative research methods. For example, the best way to find out *why* people do certain things is to ask them. Data analytics can help overcome problems of recall and transaction costs with regard to *what* people have done or are doing. If a complete understanding requires answers to both what and why questions, it is best to do both.

Modalities of conducting BD4D research in the Global South

The systematic mapping of big data for development actors conducted as part of this research shows that there are only a few organizations based in the Global South that are active in BD4D research, from conceptualization, analytics, communication to policy audiences and work on policy/regulatory aspects, despite the significant contributions to development that could be made by such research and the potential harms that should be considered. The objective of the present exercise is to increase the number of organizations based in the Global South engaged in all, or as many as possible, of the components of the BD4D research value chain. It is also to develop capacity to undertake high-quality, multi-disciplinary BD4D research. Below is a proposal for developing this capacity based on our experience in forming multidisciplinary teams for Systematic Research (SR) in 2014.

BD4D research is ideally undertaken by multidisciplinary teams and supported by resources in terms of skilled analysts, data, hardware and software. This leads to a focus on organizations, not individuals. The work, if it is to be of good quality, will have to be funded. The proposed capacity development will therefore differ from LIRNEasia's previous capacity development initiatives which were focused on individuals not organizations (CPRsouth and SR research). It will also be anchored on the funding of at least a few projects that emerge from the capacity development initiative and the encouragement to seek funding from additional sources. Establishment of organic relationships among the participating organizations is an objective. A small "rapid-response" fund will enable the organizations to seek assistance when policy windows open in their areas of operation.

[22] For more information on going efforts supported by Data2X please refer to http://data2x.org/wp-content/uploads/2014/08/Big-Data-Projects.pdf

The central element is a five-day training program with four days of tutorials on different aspects of BD4D research and policy engagement that will be held annually for five years in the first instance. A core group of individuals from organizations engaged in BD4D will provide the instruction and mentoring; make the selections; supervise the rapid-response fund and expand the network.

Participants will be selected through a single-blind selection process from among applications solicited through the core group's networks as well as through open advertising. Factors such as institutional affiliation and potential for gaining access to data and processing capability in the case of data analytics projects and potential to work in multidisciplinary teams in the case of policy projects will be among the selection criteria. Seed research funds will be made available to a subset of teams. All will be mentored and assisted in fund raising.

The tutorial program will conclude with teams of participants preparing and presenting proposals for BD4D research from which the recipients of seed funds will be selected. The fifth day will feature presentations on completed research. In the first year, the showcased research will be from existing outputs from the core group. In subsequent years, successful projects from previous rounds will be showcased on the fifth day.

The experience with the SR research pointed to the importance of the qualified and motivated team leaders. It also demonstrated the difficulty of effective team performance when members are geographically dispersed. Therefore, the participation of pre-formed teams with designated leaders will be the norm. Exceptions may be possible, especially for policy-focused teams.

The ultimate objective of BD4D research is policy impact, broadly defined. Policy impact is best achieved by effectively communicating relevant research within policy windows. Research takes time. Therefore, the first-best solution is difficult to implement. A second-best solution would be to quickly adapt or add a "bridging" section to relevant work done in a different country and communicate it to the decision makers during the policy window. This is the rationale for the building of organic relationships among organizations engaged in BD4D research in the Global South and the Rapid Response Fund.

Above outlined is one modality for developing capacity in the Global South to extract insights relevant to development from big data and to effectively communicate those insights to decision makers while actively engaging in the shaping of global big data practices in ways that would minimize harms. It requires a donor or donors to make a multi-year commitment of resources to organize annual training events, provide seed funding and support rapid response. It also requires a core group of organizations to agree to contribute their stretched resources to build and expand the network. Based on the experiences of establishing a regional think tank (LIRNEasia), a capacity-development conference and associated activities (CPRsouth) and conducting research through multi-country teams (SR Project), we believe this is the optimal solution. We hope this will provide a good starting point for a wide-ranging discussion of options.

References

Beniger, J. R. (1986). The Control Revolution: Technological and Economic Origins of the. *Information Society*.

Blumenstock, J., & Eagle, N. (2012). Divided We Call : Disparities in Access and Use of Mobile Phones in Rwanda. *Information Technologies & International Development*, *8*(2), 1–16.

Blumenstock, J., Cadamuro, G., & On, R. (2015). Predicting poverty and wealth from mobile phone metadata. *Science*, *350*(6264), 1073-1076.

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.

Cate, F. H., Cullen, P., & Mayer-Schonberger, V. (2013). Data Protection Principles for the 21st Century.

Clifford, S., & Hard, Q. (2013, July 14). Attention, Shoppers: Store Is Tracking Your Cell. Retrieved September 25, 2016, from http://www.nytimes.com/2013/07/15/business/attention-shopper-stores-are-tracking-your-cell.html

Duhigg, C. (2012). *The power of habit: Why we do what we do in life and business* (Vol. 34, No. 10). Random House.

Eagle, N.; Macy, M. and Claxton, R. (2010). "Network Diversity and Economic Development." *Science* 328: 1029–31.

Floridi, L. (2014). Open data, data protection, and group privacy. *Philosophy & Technology*, *27*(1), 1.

Frias-martinez, V., Frias-martinez, E., & Oliver, N. (2010). A Gender-centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records. In *AAAI Spring Symposium on Artificial Intelligence for Development (AI-D)*.

Hayes, Constance L. (2004). What Wal-Mart knows about customers' habits, *New York Times*.

Hayes, C. L. (2004, November 14). What Wal-Mart Knows About Customers' Habits. Retrieved September 25, 2016, from http://www.nytimes.com/2004/11/14/business/yourmoney/what-walmart-knows-about-customers-habits.html

Lerman, J. (2013). Big data and its exclusions. *Stanford Law Review Online*,*66*.

Mayer-Schonberger, V., & Cukier, K. (2013). Big data: A revolution that will change how we live, work and think. *London: John Murray. ISBN*, *978*, 0544227750.

Mundie, C. (2014). Privacy Pragmatism; Focus on Data Use, Not Data Collection. *Foreign Aff.*, *93*, 28.

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press.

Perry, W.L.; McInnis, B.; Price, C.C.; Smith, S.C.; Hollywood, J.S. (2013). *Predictive policing*. Santa Monica CA: Rand.

Rosoff, M. (2014, November 24). Here's how dominant Google is in Europe. Retrieved September 25, 2016, from http://www.businessinsider.com/heres-how-dominant-google-is-in-europe-2014-11

Samarajiva, R. (1994). Privacy in electronic public space: Emerging issues. *Canadian Journal of Communication*, *19*(1), 87.

Samarajiva, R. (1996). Surveillance by design: Public networks and the control of consumption. *Mansell R. and Silverstone R.(eds.)*, 129-156.

Samarajiva, R., Lokanathan, S., Madhawa, K., Kreindler, G., & Maldeniya, D. (2015). Big Data to Improve Urban Planning. *Economic & Political Weekly*,*50*(22), 43.

Scott, J. C. (1998). *Seeing like a state: How certain schemes to improve the human condition have failed*. Yale University Press.

Solove, D. J. 2008. Understanding Privacy, Cambridge, MA: Harvard University Press

Soto, V., Frias-Martinez, V., Virseda, J., & Frias-Martinez, E. (2011, July). Prediction of socioeconomic levels using cell phone records. In *International Conference on User Modeling, Adaptation, and Personalization* (pp. 377-388). Springer Berlin Heidelberg.

Stucke, M E.; Grunes, A. P. (2016). *Big data and competition policy*. Oxford: Oxford University Press.

Taylor, L. (2015). No place to hide? The ethics and analytics of tracking mobility using mobile phone data. *Environment and Planning D: Society and Space*, 0263775815608851.

Wesolowski, A. P., & Eagle, N. (2009). Inferring human dynamics in slums using mobile phone data. *Santa Fe, NM: Santa Fe Institute*.

Zuiderveen Borgesius, F. J., Van Eechoud, M., & Gray, J. (2015). Open Data, Privacy, and Fair Information Principles: Towards a Balancing Framework.*Berkeley Technology Law Journal, Forthcoming*.