DATA POP ALLIANCE

HARVARD
HUMANITARIAN
INITIATIVE

WORLD BANK GROUP
Leadership, Learning and Innovation

DATA-POP ALLIANCE WHITE PAPERS SERIES

The Politics and Ethics of CDR Analytics

Emmanuel Letouzé

Patrick Vinck

Draft for discussion*

This version: Dec 10th, 2014

About this document

This paper is part of Data-Pop Alliance's 'White Papers Series'. It originated in discussions between members of the 'D4D team' involved in the first Orange D4D challenge in the summer of 2013—including representatives of Orange France Telecom Group, World Economic Forum USA, Université Catholique de Louvain, MIT, UN Global Pulse, GSMA, and the authors of this paper.

Its completion benefited from financial support from the World Bank Leadership, Learning and innovation group, which is gratefully acknowledged.

Data-Pop Alliance is the first global think-tank on Big Data and Development, jointly created by the Harvard Humanitarian Initiative (HHI), the MIT Media Lab, and the Overseas Development Institute (ODI) to promote a people-centered Big Data revolution.

Emmanuel Letouzé is the Director and co-Founder of Data-Pop Alliance. He is a Fellow at the Harvard Humanitarian Initiative, a Visiting Scholar at MIT Media Lab, a Senior Research Associate at ODI, and PhD Candidate at UC Berkeley.

Patrick Vinck is the co-Director and co-Founder of Data-Pop Alliance. He is Director of the Program on Vulnerable Populations at the Harvard Humanitarian Initiative, and A Research Scientist at Harvard School of Public Health. He also serves on the Committee on Scientific Freedom and Responsibility of the American Association for the Advancement of Science (AAAS).

The views presented in this paper are those of the authors and do not represent those of their institutions. This version benefited from suggestions and ideas from the Data-Pop Alliance team, members of the D4D team and others, notably William Hoffman, Nicolas de Cordes, Kyla Reid, Ed Barker, Alex 'Sandy' Pentland, Isabelle Huynh and Nuria Oliver. The standard disclaimers apply.

Table of Contents

Introduction

The 2014 Ebola crisis may well be a game changer in the emerging use and sharing of Call Detail Records (CDRs) for social good—in other words, the use to inform humanitarian and development action of meta-data that telecom operators (thereafter 'Telcos') record primarily to bill customers based on their cell phone usage, but also to understand and target their customers based on their cell-phone consumption patterns.[1]

Newspapers announced with bold headlines that such data would help stop the spread of the Ebola epidemics.[2] Since then, academics and practitioners have called for the release of more data to ever growing numbers of researchers to leverage the technology and show the value that can be extracted from it.[3] Yet despite the anticipated benefits of CDR analytics, their potential remains largely locked and under-explored for reasons ranging from technical issues to ethical and commercial considerations.

As the volume of CDRs continues to increase exponentially, (on par with the spreads of hand-held mobile communication devices serving a growing number of people and functions), this white paper explores ways to enhance the responsible use of CDRs for social good. CDRs contain rich information, including on the point of origin and destination of a call (or text message) and its duration (or length); GPS enabled phone will also constantly record movement between calls. The content of these exchanges is unknown, but a wide range of variables can be created from CDRs, which, especially combined with other data sets, creates promising avenues for research and policy.

The analysis of CDRs has gained significant attention in the past few years along with a broader interest in extracting social value from the analysis of *"traces of human actions picked up by digital devices"* – or Big Data.[4] Telcos themselves have contributed to the phenomenon. In 2012, Orange, a multinational telecommunications company operating mainly in Europe and Africa, and its partners organized the first 'Data for Development Challenge' (D4D), and, as discussed in greater details below, were overwhelmed by the response. A second challenge was organized in 2014-15. Another example is the effort of Telefónica—operating mainly in Europe and the America—to analyze CDRs from their customers in Mexico to study population movements during the 2009 H1N1 epidemic or predict socioeconomic data. Other teams and organizations in academia and the non-profit sector have also actively contributed to expanding the literature and evidence on the potential of and appeal for CDR analytics.

At the same time, challenges and concerns have been raised. For one, access to CDRs is indeed difficult, for technical but also primarily legal and 'institutional' reasons, as Telcos and governments are cautious about privacy and security implications. Several organizations and groups are working on modalities for 'responsible' data sharing[5]; the concept of 'data philanthropy' has emerged as a possible response, although it may be at odds with the notion that CDRs are essentially people's data and not those of telcos. The ethics of data is certainly becoming a major topic, although it remains hard to define. In a post-Snowden world and era where poverty remains pervasive, what is 'ethical': using or not using CDRs, especially in times of crisis—or is it about determining how, where, when, by and for whom?

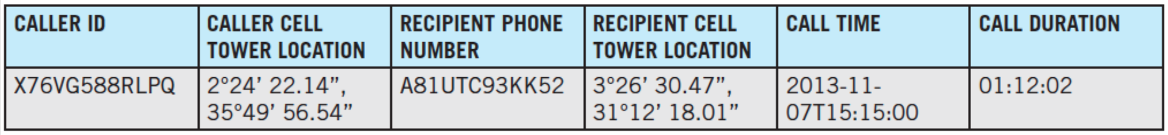
Our central question is: is there a way for the future of CDR analytics as a field of practice and research to learn from history, i.e. to extract societal benefits from CDR analytics while avoiding it become an extractive industry subject to elite capture, bearing in mind the additional risks to personal and group privacy and security? A more concrete question is: what are the key considerations that could help frame or just inform current discussions and attempts at crafting the legal, technical and institutional architecture of the field for the years to come?

This paper aims to contribute to addressing some of these questions by providing contextual elements (Part 1), suggesting political parameters (part 2), proposing ethical principles, centred on the Menlo report[6] (Part 3) and discussing operational options (part 4).

1. Contextual Elements

1.1. Genesis and Excitement

The near ubiquitous spread of cell-phone to almost all areas of the globe has presided over an exponential growth in the volume of cell-phone data produced that is unlikely to abate in the foreseeable future.[7] Global mobile data traffic will multiply by more than 10 between 2012 and 2017 and by 100 since 2009. Cell-phones also increasingly drive Internet penetration in low-income countries.[8] At the same time, progress in data warehousing and management, computing power, including parallel computing, and computer science methods, notably advances in algorithmic analysis, provide unprecedented capacities to extract information from these data.

<table>
<tr><th></th><th>Key terms and concepts</th></tr>
<tr><td><b>Call Detail Records—CDRs</b></td><td>Technical name of cell-phone data. These metadata—data in their own right—are created every day in a rapidly growing amount: By 2017, global mobile data traffic will have multiplied by more than 100 since 2009 (Source: Cisco). This is what a CDR looks like (Source: UN Global Pulse, 2013):

<table>
<tr><th>CALLER ID</th><th>CALLER CELL TOWER LOCATION</th><th>RECIPIENT PHONE NUMBER</th><th>RECIPIENT CELL TOWER LOCATION</th><th>CALL TIME</th><th>CALL DURATION</th></tr>
<tr><td>X76VG588RLPQ</td><td>2°24' 22.14",
35°49' 56.54"</td><td>A81UTC93KK52</td><td>3°26' 30.47",
31°12' 18.01"</td><td>2013-11-07T15:15:00</td><td>01:12:02</td></tr>
</table></td></tr>
<tr><td><b>Big Data for Development and CDR analytics</b></td><td>Field of applied academic and policy research that focus on analyzing and leveraging the "traces of human actions picked up by digital devices" (Letouzé et al, 2013) including "digital breadcrumbs" (Pentland, 2012).) CDR analytics is a subset of "Big Data for Development" (Letouzé, 2012)</td></tr>
</table>

As mentioned above, CDRs are a valuable commercial asset for telcos—large operators may handle over 6 billion CDRs a day.[9]). They are also becoming a much sought-after source of data to yield behavioural 'insights' on human ecosystems. The fast growing space of 'CDR analytics for social good' takes advantage of *"how mobile carriers see the world"*.[10] With CDRs it has become possible to 'follow' and map the movement, actions and interactions of an individual—or, rather, of a phone or SIM card—to look for patterns and trends in the data, especially in conjunction with other datasets, and attempt to model, understand and affect human ecosystems.

A simple way to think about how CDR analytics can be leveraged for development and programing purposes is to distinguish three main functions—using a taxonomy proposed for Big Data more generally[11]:

(i) One is a *descriptive* function—via maps, descriptive statistics etc.

(ii) Another is a *predictive* function, in two senses of the term:
- The first sense refers to predicting as 'proxying', where CDR-based variables are used alone or in combination with others to predict the concomitant level of another variable—poverty for instance;
- The second sense is 'forecasting' where the goal is to assess the likelihood of some event(s) in a near or distant future;

(iii) The third, and least developed to date, is a *prescriptive* function, i.e. the realm of causal inference, where CDR analytics will help unveil causal relationships linking cellphone usage to outcomes, or more generally help prescribe specific interventions.

As is now well known, CDR analytics has had numerous applications, especially falling under functions (i) and (ii); for example to study infectious disease spread, internal migration, spatial dynamics in urban slums, reciprocal giving in the aftermath of a natural disaster, poverty and socioeconomic levels, transportation, notably.[12] Other areas of applicability have also been studied and discussed—including conflict and crime.[13] Recently, much attention has been paid to the work of Flowminder—a Swedish NGO whose researchers pioneered large scale CDR-based mobility analysis in post-earthquake Haiti[14]—on Ebola affected countries (although the initial model is using historical data from Ivory Coast, Senegal, and Rwanda).[15]

As mentioned in the introduction, several Telcos have also been active in the field. Discussing some of their initiatives helps sketch its features and contours more clearly, starting with the one that received the most attention to date, the 2012-13 D4D 'Ivory Coast' Challenge, which was covered extensively in the press.[16]

For that challenge, controlled access was provided to researchers to four datasets derived from anonymized CDRs of phone calls and SMS's between 5 million Orange customers in Côte d'Ivoire between December 1, 2011 and April 28, 2012. The anonymization process relied on the feedback from from two French and British academic "Friendly Test teams" who tried to crack them. Eventually, a three-tier anonymization process was used, with

(i) the generation of random caller IDs for each dataset;

(ii) slight blurring (by 5%) of all 1,200 antenna positions and

(iii) selected exclusion of off network communications, first calls, and data from 'extreme' users who may be too easy to identify.

2012-13 'Ivory Coast' D4D: Datasets made available under controlled access

Datasets			
1. Aggregate data	**2. Fine resolution mobility traces**	**3. Coarse resolution mobility traces**	**4. Communication sub-graphs**
Antenna-to-antenna traffic on an hourly basis for the entire period.	Individual trajectories for 50,000 customers for two-week time windows with antenna location information.	Individual trajectories for 500,000 customers over the entire observation period with sub-prefecture location information.	For 5,000 random customers up to 2 degrees of separation aggregated by two-week time window over five months.

Measures to control the access to and use of the data were two-fold:

(i) access to the data required a project submission to the D4D Challenge signed by a representative of a research institution;

(ii) all teams were asked to sign Terms & Conditions controlling the use of data and the publication of results.

A total of 250 teams submitted proposals and received the data, of which 83 submitted papers. It culminated at the Third NetMob Conference with a 1-day event at MIT in May 2013. The challenge generated papers on a wide range of submissions addressing questions about migration, poverty, public health, urban development and transportation, crisis response, demographic and economic statistics, and more. Four winners were identified (see box), and 30 teams were granted permission to keep the data for further collaborative work

The limitations and lessons from the challenge are discussed further below, as they informed the design of the 2nd D4D challenge and raise questions and concerns of direct relevance to this paper.

Telefónica has also undertaken several research projects and initiatives. One was a 'Datathon' co-organized with the Open Data Institute and MIT organized in September 2013 in London as part of the Campus Party Europe[17]; two papers on crime prediction based on the data provided then were recently published.[18] Other projects were conducted 'in-house—i.e. without the data being released; for example, Telefónica researchers used aggregated anonymized CDRs to analyse which of three public policy interventions was most effective at curbing population movement in Mexico City during the H1N1 epidemic[19]; in another 2011 paper using 2010 data from a "main city in Latin America" (Mexico City), the research team used CDRs in conjunction with survey data on socio-economic levels (SELs) to build a predictive model to 'assign' SELs to various areas on the basis of their digital signatures in CDRs—with a predictive power of 80% [20] The rationale for building such models is to either help estimate changes in SELs over time, or apply them to other locations.

Laureates of the 2012-13 D4D Challenge

Four D4D Challenge prizes were awarded:

First prize

University of Birmingham or the 'best project on scientific and development aspects'—*"Exploiting Cellular Data for Disease Containment and Information Campaigns Strategies in Country-Wide Epidemics"*

Best development insight

IBM Dublin for 'the most "practical" project'—to *"AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data"*

Best Scientific

UC San Diego for a 'project proposing an innovative methodology, a new question addressed and relevant original findings'—*"Analyzing social divisions using cell phone data"*

Best visualization

SynerScope BV, Eindhoven U. and MIT for the 'cleanest and most appealing visual representation'— *"Exploration and Analysis of Massive Mobile Phone Data: A Layered Visual Analytics approach"*.

Another—less successful—attempt was that of The World Bank in Egypt.[21] In 2012, the Wold Bank partnered with Vodafone and IBM Research after the 'Cairo Transport App Challenge'[22] to analyze Cairo's traffic congestion. Under the initial terms of the agreement, Vodaphone was to release historical anonymized CDRs, while the Dublin-based IBM research team would use its AllAboard solution (developed for the D4D Challenge mentioned above, see box) to conduct the analysis.[23] The project was put to a halt and eventually died after the National Telecom Regulatory Authority made requests that the partners were unable or unwilling to meet—including the conditions that CDRs stayed on the Egyptian territory and that only Egyptian researchers should have access to them. IBM Research did not wish to install its AllAboard solution on Egyptian servers and their key research employees were indeed foreign nationals.

There are of course scores of other research papers and initiatives that have used CDR analytics, but this subset of examples does point to the opportunities—on which the above section has focused—and give a sense of the shortcomings of the current state of the field—to which we now turn.

1.2. Limitations and gaps

The promise of CDR analytics to advance our collective understanding of human dynamics is hard to deny. But many uncertainties and challenges remain that too often tend to be obscured and side-lined by at times narrow and short-sighted views.

Let us start by the concrete case of the 1st D4D challenge. For all its success, it also raised controversies and concerns. First, there wasn't a single submission from an Ivorian team. Second, none of the results has since then led to any concrete implementation in Côte d'Ivoire to benefit the people whose data were used. Third, relatedly and most critically for our purpose, the project also raised questions about effective consent, and ethics broadly understood. The second D4D 'Senegal' Challenge has consciously been designed to partially address these criticisms—notably through greater involvement and engagement of Senegalese authorities and of a prize for the best 'ethical' project—although it remains to be seen how many local teams participate, whether and when any findings are turned into actions, and what the criteria for the ethical prize are.

Either way, most of the hard and deep nuts are left to crack, and in general, the practice and use of CDR analytics has been and remains characterized and hampered by biases and gaps—notably a focus on 'getting data and papers out', technical obstacles, institutional fragmentation and the absence of a clear ethical and regulatory framework.

A first basic technical challenge—that is not central to our investigation but is nonetheless related—is differential ownership of cell-phones. The most important word in the phrase 'near ubiquitous' may well be 'near', and we must remain aware of the fact that penetration rates close or above 100% do not ensure representativeness. A few papers have attempted to estimate and correct for sample bias in CDRs; for example, a paper relying on Kenyan data used 'ground truth' survey data to estimate the impact of differential cell-phone ownership on the predictive power of CDR-inferred models of human mobility, finding the CDRs-based models to be surprisingly robust.[24] Other researchers are currently building up on previous work on email and IP data to develop sample bias correction methods.[25] But more research is needed to build solid sample bias correction methods to ensure that CDR analytics does not amplify basic inequities. Differential ownership of analytics *capacities* is another major factor that may contribute to creating and widening a new digital divide between and within countries.

A second set of challenge relates to individual (and group) privacy and security, which have become especially salient since Edward Snowden's revelations on the use of CDRs as part of the US National Security Agency (NSA) surveillance program. Most of the literature is based on carefully 'anonymized' and often aggregated data. But that may not necessarily suffice to alleviate privacy and security concerns. The possibility of 'de-anonymization' of previously anonymized datasets has been known for years[26] (when multiple datasets are combined, one of which contains an ID).

A trickier problem—which is also a blessing for scientists—is the high degree of predictability and conformity of human behaviour. For instance, a paper using D4D Challenge data attempting to derive the 'maximum predictability' in human movement confirmed that human mobility was indeed highly predictable[27], such that only four spatio-temporal points were theoretically sufficient to identify an individual with 95% accuracy in a dataset with no ID.[28]Furthermore, individuals' belonging to specific social groups—in terms of their gender, ethnicity, sexual orientation, etc.—tend to show in Big Data including CDRs, and may be used for targeting purposes—whether or not the individual's identities are known. In other words, 'anonymizing' datasets without aggregation is in effect an uncertain endeavour. This realization has gone on par with, if not spurred, 'privacy-preserving' methodological progress: for instance, researchers have developed a methodology that injects 'noise' in CDRs to make re-identification more difficult.[29] But the fact remains that there is and will remain for the foreseeable future potential risks associated with CDR analytics, or more precisely a trade-off between granularity and security.

Another set of challenges is legal and institutional. Right now, there is simply no coherent and comprehensive set of regulations or guidelines that govern the field of CDR analytics. Responses to growing demands for CDRs from researchers have typically been ad-hoc, granted by Telcos on the basis on personal connections and other arrangements—or for data challenges at their will. Although the concept of 'data philanthropy' has received some traction and may be tactically fruitful, it assumes that CDRs effectively *belong* to Telcos—which is disputable.[30] Current practice and legal and policy arrangements are simply not suited to the opportunities and risks ahead.

Last, and fundamentally, we argue that this is the case because of the absence of clear agreed-upon political parameters and ethical principles in which to ground these discussions. For instance, most discussions contrasts "opportunities" with "challenges" (or "risks"), or the "promise" of CDR analytics with its "perils"—with little explicit recognition of the roles and rights of different actors, of their competing priorities, and the importance of context. Similarly, everybody agrees that CDR analytics must be 'responsible' or 'ethical'—but it is largely unclear what ethical framework and principles ought to be used to inform action. In addition, calls for new ethical standards and norms appear to be made without considering the lessons from decades of research.

This suggests the need to address emerging opportunities and concerns in the field of CDR analytics by identifying political parameters and ethical principles that will help formalize and expand it along clear pragmatic and paradigmatic lines.

"Data philanthropy": benefits and limits

"Data philanthropy" refers to the concept and practice of sharing data held by private corporations for purposes of analysis intended to have positive social impact. Although typically framed as a modern form of corporate social responsibility or charity, it has also been described as being "good for business"—by benefitting consumers and economies.

However, a fundamental problem with data philanthropy is that is seems to assume that the data recorded by private corporations effectively belong to them—and that they may be altruistic or self-interested enough or both to 'give away' some of them. The issue is that there is a much stronger argument to be made that these data *do not* belong to private corporations, but rather to their individual emitters.

Data philanthropy may be a "pragmatic approach' conveying the idea of data being "a public good". But there is a risk that a tactical move may turn into a paradigmatic shift, and that we too swiftly forget the characteristics of public good—and the fact that knowledge, not data, is a public good.

http://www.marketsforgood.org/sharing-data-as-corporate-philanthropy/ http://www.forbes.com/fdc/welcome_mjx.shtml

http://www.unglobalpulse.org/blog/data-philanthropy-public-private-sector-data-sharing-global-resilience

https://hbr.org/2014/11/with-big-data-comes-big-responsibility

2. Political Parameters

2.1. Three sets of considerations

We start by introducing and discussing three types of considerations that implicitly structure all debates about CDR analytics constitute their frame, or boundaries, namely:

(i) *individual* considerations, i.e considerations of and for individual privacy, confidentiality and security—i.e. where the unit of analysis is the individual;
(ii) *commercial* considerations; i.e. considerations of and for Telcos' return on investment and profitability —or any other short-to-medium term financial indicator telecom operators care about;
(iii) *societal* considerations; i.e. considerations of and for the social 'public good' value that can be yielded by analyzing CDRs—such as averting the next cholera outbreak or cutting transportation time.

These suggest three theoretical cases:

(i) A case were no data is collected and therefore no data is shared or analyzed – this is the *extreme individual privacy case*, which reflects an exclusive concern for individuals' rights to privacy.

(ii) A case were all CDRs are collected at all time, but the data are not public and rather remain in the hands of a limited number of actors who use them for commercial purposes, and refrain from sharing them because it could provide valuable information to competitors. This is the *extreme business interests case*.

(iii) A case where all CDRs are collected and made public at all time, reflecting the idea that social 'public good' value can be yielded by opening and analyzing CDRs. This is the *extreme social good case*.

None of the extreme cases are realistic, nor are they desirable: critically, these are *ideal-typical* (or *stereotypical*) categories meant to facilitate the exposition of *kinds,* and not actor-specific, concerns. *Any* agent—an individual, a company, a government—has an interest in insuring a balance between the amount of data collected and the amount of data that is publicly shared.

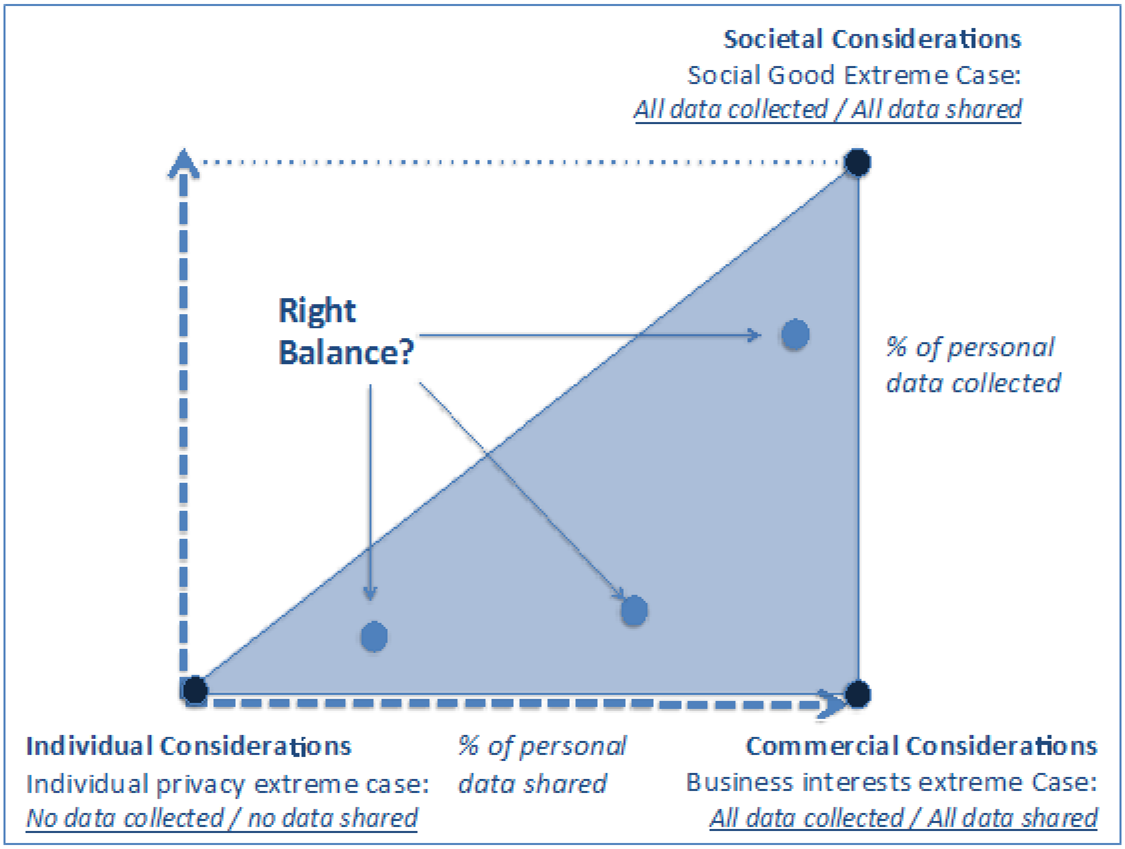
For instance, telecom companies care a great deal about protecting their clients' data for both reputational (thus 'commercial' reasons) but also because they understand how these data may be used against their clients or themselves. Telcos may also wish to contribute to the development of the economies where they operate, reflecting both commercial and societal concerns.

Likewise most governments will not seek to access and analyze all CDRs at all costs and by all means, even for 'the good' of their citizens. Lastly, even (most) uncompromising libertarians may perceive the value of having some of their personal data collected, shared, and analyzed—even as they may insist on strong anonymization and aggregation, and other mechanisms such as 'expiration' dates being put in place—if doing so can help save a life. So the crux or goal is to find the 'Pareto-optimal' equilibrium—or just an acceptable balance—between these considerations.

All of these considerations are subject to threshold effects, such that no corner solution is possible. Indeed there exist levels beyond which no discussion can take place in the eyes of some or all actors. No Telco will accept to disclose publically at all times *all* of the raw data it collects, in light of both 'commercial' and 'individual' considerations—and it will probably find public support for its reluctance. Societies, especially those with vibrant civil society organizations, will unlikely accept that Telcos never share *any* of these data, or that they share them without solid anonymization and aggregation—out of 'societal' and 'individual' considerations.

This static framework does not say where the right balance lies, but it helps assess and discuss the pros and cons of each coordinate in the diagram, all else equal, in a structured and systematic way. It allows greater depth and complexity than when relying on straight axis ranging from 'promises' to 'perils', or by considering individual considerations as a mandatory but essentially secondary part in the dialogue between commercial and societal considerations.

Mapping and reconciling 3 categories of considerations


Societal Considerations
Social Good Extreme Case:
All data collected / All data shared
Right Balance?
% of personal data collected
Individual Considerations
Individual privacy extreme case:
No data collected / no data shared
% of personal data shared
Commercial Considerations
Business interests extreme Case:
All data collected / All data shared

2.2. The importance of contextual factors

However, the 'right' balance cannot be found for all places at all times, but is determined by two kinds of factors—which we will call *systemic* and *idiosyncratic*.

First, systematics factors refer to the effectiveness of frameworks and systems in a given place and time. For instance, there are inherent risks of security breaches through the entire 'data chain'—from acquisition, storage, sharing of data, analysis, and sharing of results. But the problem will be especially salient where and when the operating environment of a company is weak at restricting access or use of CDRs, or where mobile phone companies are faced with oppressive regime who may seek to gain access to sensitive data.

So it may be that a legal framework appropriate in a given country would be ill advised in another, or that the 'right balance' in a given country may change over time with political and technical progress. The 'social value' argument and thus the case for opening up CDRs for analysis will be stronger where and as researchers and policymakers are better at using and relying CDR analytics such that significant additional societal value is created and can be shared—in the form of greater political stability or higher economic growth. Also, how 'commercial' considerations play out and affect the choice of the 'right' balance for a given Telco—which are all faced with these questions—depend on the decision of others: if all participate, then the strength of the argument of a loss of comparative competitive advantage is lessened.

The point is that although inconsistency of the legal or regulatory environment guiding opening and use of CDRs across countries can be problematic it seems implausible and undesirable to settle on global standards and norms.

In addition, idiosyncratic factors—i.e. fast changes in prevailing circumstances—matter. Just as much as the sensitivity of a malfunction detection system designed for an alarm clock need to be enhanced if repurposed to monitor a nuclear plant, the right balance between the three sets of considerations, holding systemic parameters fixed, ought to change if prevailing conditions change dramatically—for instance in the case of an acute public health crisis. This does not mean that individual considerations—by which we mean privacy—are no longer relevant, but their weight must be reassessed against the expected benefits or opportunity costs of opening up the CDRs vs. keeping them locked—in ways that may not be straightforward.

The Ebola crisis offers an interesting case to discuss these points and tensions concretely. Several commentators argued that the crisis made opening up CRDs a near moral imperative, and blamed poor coordination for the absence of effective action in that respect.[31] At the same time, and to play the devil's advocate, one could also argue that these countries' political, economic and historical characteristics, and - for Sierra Leone and Liberia, their decades of ethnic civil wars, raise significant concerns as to the potential misuse of CDR analytics, especially in such volatile times and in their aftermath; it also largely remains to be seen if and how CDR analytics could effectively be used to improve response on the ground.

And so the question becomes: what, in such a case, is or was 'ethical'? Using or not using CDRs? There is no clear-cut answer, and providing one requires relying on some ethical principles absent from the model above.

What we have established so far is that discussions about CDR analytics would benefit from their being framed by the aforementioned political parameters, which can be distilled as follows:

(i) There exist three distinct sets of legitimate considerations that all agents face to varying degrees in different places and at different times;

(ii) The appropriate balance depends on slow changing characteristics (systemic factors) but can and probably should be altered by sudden crisis or events (idiosyncratic factors);

Lastly, it is also clear from the discussion above that the position, modalities and movement of the 'right balance' depend critically on ethical principles that need to be spelled out, to which we now turn.

3. Ethical Principles

3.1. Framework for analysing challenges and opportunities: The Menlo Report

This paper adopts and advocates for the ethical principles for research laid out in the Menlo report as its primary ethical framework.[32]. Other guiding principles and frame of reference are useful to identify and assess the challenges raised by the analysis and management and sharing of personal data such as CDRs. However, we argue that the research ethic frame is the most appropriate to highlight the most consequential and problematic issues as well as opportunities raised by CDRs analytics broadly considered.

The Menlo report, first published in December 2011 and amended in 2012, identified four key ethical principles for computer and information security research and is itself based on landmark guides for ethical research in the biomedical and behavioural sciences, such as the Nuremberg Code, Declaration of Helsinki, and Belmont report.

(i) Beneficience;
(ii) Respect for Persons;
(iii) Justice;
(iv) Respect for Law and Public Interest.

The following table describes these four key ethical principles, which we then unpack and comment in the case of CDR analytics.

The Menlo Report Ethical Principles Guiding ICT Research[33]

Principles	Description[34]
(i) Beneficience	• Do not harm; • Maximize probable benefits and minimize probable harms; • Systematically assess both risk of harm and benefit.
(ii) Respect for Persons	• Participation as a research subject is voluntary, and follows from informed consent; • Treat individuals as autonomous agents and respect their right to determine their own best interests; • Respect individuals who are not targets of research yet are impacted; • Individuals with diminished autonomy, who are incapable of deciding for themselves, are entitled to protection.
(iii) Justice	• Each person deserves equal consideration in how to be treated, and the benefits of research should be fairly distributed according to individual need, effort, societal contribution, and merit; • Selection of subjects should be fair, and burdens should be allocated equitably across impacted subjects.
(iv) Respect for Law and Public Interest	• Engage in legal due diligence; • Be transparent in methods and results; • Be accountable for actions.

3.1. Unpacking the Menlo principles for CDR analytics

1. Beneficience: Understanding risks and benefits

The principle of beneficence refers to *"a moral obligation to act for the others' benefit, helping them to further their important and legitimate interests, often by preventing or removing possible harms"*.[35] Under this principle, researchers must maximize the probability and magnitude of benefits to individual research subjects as well as to society. The recognized benefits are what transform CDRs into valuable assets whose potential should be unlocked.

However, what constitutes a benefit or a risk is not always straightforward or consensual—and, as discussed above, depends to a large degree on the actors considered. CDRs are largely stored and handled by private companies, which are the ones investing in transmission and storage infrastructures. Commercial considerations must therefore be taken into account in framing the risks and benefits of using and sharing CDRs.

The potential benefits and harm of any project making use of CDRs certainly depend on that specific project's objectives..

Furthermore, unlocking the benefits of CDRs will require experimentation and practice that may not have direct value or benefits besides learning – akin to fundamental science which ultimately leads to broader benefits.

2. Respect for Persons: Consent

The issue of consent is gaining attention and is central to the privacy concerns relating to the use and sharing of CDRs. Specifically, users of mobile phone handset rarely grant formal permission for their personal data to be used and shared. If they do it, it is often with little to no choice, since not consenting would limit their access to the technology. Furthermore, the choice given to consumer is typically to either dissent or fully consent regardless of what use of the data can be done several years later, or by a third party should it be accessed. There is little to no way for consumers to exclude specific usage of their data that they do not want, raising major questions around the secondary use of data.

The issue of consent is not purely "informational" – i.e. does a user agree or not with proposed uses of data about themselves. Ultimately it is about potential risks and enabling users to make decisions for themselves. Granting usage of their data may expose individuals to various harms and risks, especially as increased data sharing increase the risk of confidentiality breaches or misuse of the data. The use of their data may also go against their cultural or religious values.

Much of the discussion has been focused on "opt-in / opt-out" which requires the user to either actively consent to terms of use that include data sharing, or to "actively dissent", the default setting being that of consent. More advanced models being discussed include a more flexible process where permissions can be granted in a variety of ways and dependent upon the context of use– either through explicit consent or implicitly through compatible action.

For secondary use, it is generally agreed that uses that are consistent with the original context can carry the permission granted in that context, but that new uses should require new consent. Broad (unlimited) consent remains widely use despite strong opposition on moral and ethical grounds.

An even more advanced model proposes that individuals would permanently "carry" a set of permissions that they grant to algorithm seeking to use their data – no matter what data, enabling them to modify access and permissions at any time.

3. Justice: Bias and inequalities

The principle of justice highlights issues of fairness and equal distribution of risks and benefits. Arguably one of its key aspects is that everyone must have an opportunity to contribute and benefit (e.g. from CDR analysis) even when unequal access to technology exists. Yet, whose data is considered in CDRs analysis is inherently affected by unequal access and use of mobile phones, creating inherent biases and violating the principle of justice.

This creates yet another tension in CDR analysis: It is especially relevant in otherwise data-poor environments, but it is precisely in these environments that access to technology is most unequal, which implies that CDRs are non-representative data. The underlying challenge is that CDRs will typically reflect structural inequalities in any given countries: owning a cellphone is strongly correlated with socio-economic status, and even in countries with high mobile phone penetration, CDRs may be analyzed along criteria that would single out more affluent individuals or areas. These biases hinders the external validity of findings based on CDRs and may potentially reinforce structural inequalities (if, for instance, programs are based ondata from areas with high cellphone usage).

Biases may be unproblematic as long as they are well understood and corrected for. Besides Buckee et al. (2013), correction methodologies for biases in e-mail data have already been proposed (Zagheni and Weber, 2012), although validation is difficult for lack of reliable 'ground-truthing' data. No similar efforts exist to this day with regards to CDRs. Furthermore, it is likely that as cell phone penetration and patterns of use change, there will be a need to constantly adapt methods and algorithms developed to correct biases. This is clearly a challenge and a priority for future research.

Beside the issue of bias in the data, the analysis of CDRs may also lead to unequal targeting of individuals or groups based on their ethnicity, gender, religion, and sexual orientation. The notion of group privacy recoups the rights to groups and their members not to be identified and targeted as such; this concept is likely to gain traction as it is intrinsically related to discriminations, targeting etc. It is indeed possible to predict group-level characteristics—for instance, distinguishing a 40-year old gay male from a 20 year old heterosexual female using various big data streams—credit card transactions, social media data etc, and in all likelihood CDR may also reflect similar characteristics. In such a case, having anonymized, even aggregated, data, may be insufficient to avoid discriminations and negative targeting. These concerns however may be at odds with the increased popularity of the concepts of "hyper-personalization" of marketing, under which individual characteristics are defined so well that they enable corporations to offer highly customized offers and services.

4. Respect for Law and Public Interest

The fourth and last principle framing our discussion highlights the need to engage in legal due diligence; be transparent in methods and results; and be accountable for actions. However, inconsistence of the legal or regulatory environment guiding opening and use of CDRs across countries is problematic where legal protections are insufficient to protect the individual, and where cross-border accountability is difficult to enforce (e.g. if an individual is put at risk because of a foreign organization use of CDRs, what are the recourse for that individual?).

Telcos are especially concerned about their legal exposure if CDRs were to be used to identify, target and/or discriminate against specific individuals or groups. For example, participants in protests can easily be identified through CDRs. Telcos may be confronted to local legal requirements that may be at odd with international law and could potentially be held liable if their data were used in mass atrocities, something not entirely impossible. In repressive environments, Telcos should consider their first priority to protect the sources of information (their customers) and place sensitive data beyond the reach of authorities, even though this may be against their financial and commercial interests. At the same time, Telcos which have access to potentially life-saving information may be morally, if not legally, required to make that information available.

4. Institutional Implications

4.1. Operational requirements

Having used core ethical principles to frame the key challenges emerging in the rapidly growing practice of CDRs analysis, this paper also serves as a call to renew commitment to these principles. Putting these principles into practice requires agreeing on a number of operational requirements without which they will remain dead slogans.

One is to recognize the plurality of actors, the legitimacy of all and the responsibilities of each, which calls for a collegial and coordinated approach to the problem.

Telecom companies contribute—as socio-economic agents—to to enhancing the welfare of societies where they operate. Telecom companies should not mimic the most negative aspects of extractive industries were valuable resources are exported with no or little benefits locally. At the same time, local government, researchers, and organizations are unlikely to have the ability to take advantage of CDRs, including the necessary financial resources and local expertise, without assistance from Telcos. This will require new public-private partnerships that leverage private sector data for public policy. It will also require new collaborations with researchers and investment in research capacities to develop skills and research in cloud and high performance computing, for example through North-South and South-South PhD program development.

Telcos may participate in such partnership if regulators and legislators ensure that investments by telecom operators are fairly rewarded and incentivized. These same legislators must at the same time ensure that the rights of their citizens be fully upheld and will need the appropriate regulatory frameworks to enable (and at times force) access to data for public good.

Researchers may also be stakeholders in the public-private sharing of CDRs, but they too must have defined roles and responsibilities. Researchers should engage in and support efforts to find standardized data sharing tools and protocols. At the same time, the multiple and sometimes competing demands on Telcos to provide data must be coordinated. Data requirements must also be better defined to avoid demands that seeks to capture anything and everything in near real-time, especially when and where historical data may be sufficient for the proposed work. Indeed, the notion that CDRs will help spur 'agile' development in the near future—which would justify getting real-time data for instance—is largely unsubstantiated. Information contained in 'old' CDRs are interesting for research, and their result do not depend too much on the timing of extraction of the data (let's say within the last 2 or 3 years). Some applications may need 'fresh' data, even real time, but these are technically and ethically more difficult. So for these data, we need to be clear about the benefits we expect for individuals and society that justify these efforts/risks. Additionally, aspects of capacity development and participation of researchers from countries whose data are being used should also become standard practice.

These considerations show that the responsible development of CDR analysis will require the involvement, support and good will of all actors involved. Too often the questions raised in this paper are discussed in isolation by a select group of actors, with the individual perspective being the least represented. The recent set up of the UN group is illustrative of the visibility given to corporate and societal perspectives (government) at the expense the individual perspective. Similarly, calls for open data and data philanthropy are largely framed around corporate and societal benefits, with little to no attention paid to individual level considerations.

In addition, it is important to highlight that CDR alone offer only limited insight, and that their richness is unlocked when combined with other data streams. There is therefore a need to create better integration and access across data streams. More traditional forms of data are needed. E.g. tracking poverty or socioeconomic levels using CDRs requires having poverty or socioeconomic data to start with (and a CDR analysis alone would be very sensitive to sample bias). 'Historical data', even aggregated data, are extremely valuable or perhaps in a way even more than 'real-time' raw data because they not just allow but indeed limit/compel 'us' to focus on building methods and tools under greater constraints, above and beyond (i.e. before) attempting to do nowcasting of current populations/variables in any way.

Another key operational principle is to think and act strategically, with a longer term horizon than the next paper or quarterly report. Changing the overall timeframe—thinking and planning for the next five to ten years—does change short-term decisions and priorities. Capacity building and standard and norms settings are absolutely essential ingredients and objectives for the expansion of CDR analysis. This refers to the need to build on and existing models and norms, as well as ongoing work. An example of such an attempt is the WEF's 'personal data initiative' examining among other issues how the process of granting permissions for personal data use and exchange (consent) must be updated for a big data (CDRs) world.

It must also be noted that ethical issues are not exclusive to CDRs: similar concerns are regularly raised, as during the Open Government Partnership Summit sessions on whistle blowing, privacy, and safeguarding civic space—especially in light of the Snowden case—or at the Technology Salon on Participatory Mapping.[36] The fact that similar issues are being discussed from the perspective of a wide range of actors with a wide range of perspectives suggests a high potential for cross-discipline learning.

A last operational principle is context-sensitivity and appropriateness—which we shall illustrate by discussing the value of and case for using non-anonymized data in crisis contexts. The critical use of non-anonymized data offers a good illustration of the need to find a right balance between various interests, but also identify the appropriate mechanisms and principles for the responsible sharing of data.

During a disaster, access to identifiable data from mobile phone operators may be critical to assist with the reunification of families separated by the disaster, or to assist the identification of body remains. Mobile phone data may also be associated with identifiable data for the purpose of tracking services and benefits used by disaster-affected individuals.

In such contexts, the societal value of identifiable CDR data is very high, with the crisis potentially justifying significantly downplaying individual and commercial concerns for some time. Similar uses of CDR data have already taken place, but without guiding principles, these are potentially creating liabilities and risks for affected communities.

4.2. Policy options

It remains to be seen what governance and technical arrangement should dictate the sharing of data and indicators based on CDRs to increase the availability of datasets and the efficiency of data analysis, tool development, knowledge sharing and so on.

A number of pointers can be discussed describing some of the minimum requirements for such arrangements. This could serve as a starting point for mobile phone companies to work with the research community, governments and other civil society actors toward minimum principles or governance structure.

Let us start by considering the implications of the probably contentious proposal of using non-anonymized data in crisis contexts.

One critical path to explore is to learn from advances in the protection of human subjects in research to establish a systematic review process to validate when and where such data should be shared. This could, for example, be done under the supervision of internationally recognized organizations such as the International Federation of the Red Cross.

Specific criteria to judge the benefits and risks should be established under very clear circumstances (sudden onset disaster), and reviews should create a learning process to decrease the risks of inappropriate use of the data. Limits may also be established as to the type of analysis that is permitted (e.g. localization of people reported as missing…). For providers, this may require getting prior informed consent from subscribers with the delicate decision of making this mandatory, or as an opt-in or opt-out decision.

In non-acute crisis contexts, solution should enable mix usage with various levels of privacy setting / concerns / noise or quality degradation in the data depending on the ultimate usage and perhaps actors involved. Access to anonymous CDR's might be granted to a research lab for a specific contract, while only access to aggregated indicators (volumes of calls per day per antenna, etc.) could be accessed by a larger community in a more open fashion. Furthermore, some data may need to be eliminated from public records (e.g. antennas at military sites, data from 'extreme' users.) – specific terms and conditions must be developed to address this "data cleaning" process.

In this context, the level of information loss due to CDRs detail reduction (e.g. How much do we lose by reducing the granularity from Antenna location, to the level up) must result from a systematic and balanced analysis of objectives, risks and benefits. This may require establishing minimal data requirements based on various research use, seeking to answer questions like 'is real time needed?' 'If not, what type of past data?' 'Is there a minimum sample size for a particular analysis?'.

A likely compromise on more systematic sharing of CDRs would enable both individuals and mobile phone companies to maintain and possibly enhance control over CDRs to respect individuals' agency, while Telcos maintain their contractual relation with their customers, the respect of their privacy, and control critical information that may help direct competitors (local marketshare, zone of customer acquisitions). Any solution should also enable CDRs to systematically carry standardized metadata that include any limits on the use of the metadata. In case of aggregation, use of CDRs should be restricted to the most restrictive use granted by any individual whose data are included in the aggregated data. One key aspect of the enhanced control of individuals and metadata that must necessarily accompany CDRs is the ability to maintain ""expiration date" to protect privacy and other individual rights in the long term, echoing ongoing current discussions on the *"erasable future of social media"*.

Whichever approach is chosen should further enable greater participation and capacity development of local actors, while complying with local privacy protection regulations. Local partnership and data processing accreditation are likely to be necessary. A centralized system (real institutional CDR sharing and clearing house for research) is, on the other hand, unlikely and undesirable. Rather, a more distributed model based on principles and standards is more likely to be implemented by various actors, both to enable a better control and develop specific areas of expertise. Due to the many commonalities between analysis (e.g. the use of background maps,) some elements of data/indicators sharing will be effective as well.

Concluding remarks

The risks, constraints and challenges of enabling wider access to CDRs to support social good should not obscure the fact that the combination of exponential growth rates of mobile phone penetration and data production in low- and middle-income countries and intense interest and efforts from social scientists and policy-makers will, in all likelihood, make CDRs analysis, or derived indicators, a standard tool for researchers by the end of the decade.

A broad array of societal opportunities of Big Data in emerging countries are real and there are ways to develop the tools, process and policies to covers both the society needs and the commercial development goals of local companies, often with a combination of the two on the same projects.

Despite the inherent value of CDRs for mobile phone companies, these actors recognize that broad principles of open innovation or open data should apply with limits to guarantee the safety and confidentiality of subscribers.

Selected bibliography and endnotes

Buckee, C.O. et al. 'The impact of biases in mobile phone ownership on estimates of human mobility' in Interface, Journal of the Royal Society, 6 February 2013.

Crawford, K. 'Think Again: Big Data' in Foreign Policy, 9 May 2013.

Frias-Martinez, V. On the Relation between Socio-Economic Status and Physical Mobility. 2012.

Letouzé. E. Big Data for Development: Challenges & Opportunities. New York: UN Global Pulse, 2012.

Letouzé. E., Patrick Vinck and Patrick Meier. Big Data for Conflict Prevention: When the New Oil Meets Old Fires. New York: International Peace Institute, 2013.

NetMob 2013: Third conference on the Analysis of Mobile Phone Datasets, 1–3 May 2013:http://perso.uclouvain.be/vincent.blondel/netmob/2013/

Nurmi, P. Data Analysis from Mobile Networks. Helsinki: University of Helsinki, 2012.

Pentland, A. Reinventing Society in the Wake of Big Data. Edge, 30 December 2012

Talbot, D. African Bus Routes Redrawn Using Cell-Phone Data. MIT Technology Review, 30 April 2013(a).

Talbot, D. How to Mine Cell-Phone Data Without Invading Your Privacy. MIT Technology Review, 13 May 2013(b).

Orange S.A. description of the D4D challenge: http://www.orange.com/fr/D4D/Data-for-Development, and Blondel et al., 2012 - http://arxiv.org/abs/1210.0137

[1] See for example Thomas H Davenport (2012) Enterprise analytics: optimize performance, process, and decisions through big data. Upper Saddle River: FT Press; Andrew McAfee & Erik Brynjolfsson (2012) "Big data: The management revolution" Harvard Business Review (Oct.), 3-9. http://hbr.org/2012/10/big-data-the-management-revolution

[2] http://www.bbc.com/news/business-29617831

[3] http://www.economist.com/news/leaders/21627623-mobile-phone-records-are-invaluable-tool-combat-ebola-they-should-be-made-available and http://www.technologyreview.com/news/530296/cell-phone-data-might-help-predict-ebolas-spread/

[4] Letouzé et al, 2013, Pentland, 2012

[5]http://www.brookings.edu/~/media/research/files/papers/2014/11/12%20enabling%20humanitarian%20mobile%20phone%20data/brookingstechmobilephonedataweb

[6] The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research - http://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/menlo_report_actual_formatted.pdf

[7] Reference – also suggest using a different graphic than the 2012 cisco one, which has been overused and is 2 years old.

[8] Reference GSMA, Others?

[9] Eric Bouillet, Ravi Kothari, Vibhore Kumar, Laurent Mignet, Senthil NathanAnand Ranganathan, Deepak Turaga, Octavian Udrea & Olivier Verscheure (2012) "Processing 6 billion CDRs/day: from research to production (experience report)" pp. 264-67 in Proceedings of the 6th ACM International Conference on Distributed Event-Based Systems: DEBS '12. New York: ACM. DOI: 10.1145/2335484.2335513

[10] UN Glbal Pulse, 2012

[11] Letouzé et al, 2013 and scidev.net for details

[12] (see UN Global Pulse, 2013, Talbot, 2013a, Nurmi, 2012, and Letouzé, 2013

[13] Perry, 2013, Himelfarb, 2014, Letouzé et al, 2013, Ulferlder, 2014, notably, on conflict and Lepri et al, 2014 on crime

[14] http://www.pnas.org/content/109/29/11576.abstract

[15] http://www.technologyreview.com/news/530296/cell-phone-data-might-help-predict-ebolas-spread/ and http://www.worldpop.org.uk/ebola/

[16] Including the MIT Technology Review, Wall Street Journal, Wired Magazine, Le Monde, La Republica, Arte Futurs, and the BBC

[17] http://www.campus-party.eu/2013/Datathon.html

[18] http://arxiv.org/pdf/1409.2983v1.pdf and http://demog.berkeley.edu/announcements/papers/MovesOnTheStreet_BigDataJournalPaperCrime.pdf

[19] http://www.wired.co.uk/news/archive/2013-10/17/nuria-oliver

[20] http://www.vanessafriasmartinez.org/uploads/umap2011.pdf

[21] This section is based on inputs from Isabelle Huynh.

[22] http://cairo.hackathome.com/

[23] The Ministry of transportation was also working on simulated CDR data to model traffic scenario in cooperation with the Japan University in Cairo.

[24] Buckee et al, 2013).

[25] (Zagheni and Weber, fortcoming).

[26] https://www.cs.utexas.edu/~shmat/shmat_oak08netflix.pdf

[27] ((Lu et al, 2013) Barabasi et al, 2010),

[28] (de Montjoye, 2013).

[29] (Martonosi et al, 2013

[30] https://hbr.org/2014/11/with-big-data-comes-big-responsibility

[31] http://www.economist.com/news/science-and-technology/21627557-mobile-phone-records-would-help-combat-ebola-epidemic-getting-look

[32] The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research - http://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/menlo_report_actual_formatted.pdf

[33] The Menlo Report: Ethical Principles Guiding Information and Communication Technology Research - http://www.caida.org/publications/papers/2012/menlo_report_actual_formatted/menlo_report_actual_formatted.pdf

[34] Ibid. See also Pham, Phuong N., and Patrick Vinck. "Technology, conflict early warning systems, public health, and human rights." Health & Human Rights: An International Journal 14.2 (2012).

[35] http://plato.stanford.edu/entries/principle-beneficence/

[36] See more at: http://blog.okfn.org/2013/11/05/ethics-and-risk-in-open-development/#sthash.xMA0K3wi.dpuf