Mobile Network Big Data for Development: Demystifying the Uses and Challenges

Sriganesh LOKANATHAN
LIRNEasia, Sri Lanka

Roshanthi Lucas GUNARATNE
Stax Inc., Sri Lanka

Abstract: Given the volumes of data that are now generated by mobile networks due to the almost ubiquitous use of mobile phones by the majority of the population, this data can be considered as big data. Spurred by the exponential growth of mobile connectivity, the attendant large volumes of mobile network big data (MNBD), offer the possibility to obtain rich behavioral insights at a scale that was never possible before. MNBD is also one of the few sources of information on low-income groups. The focus of this paper is to illuminate both academics as well as policy makers on the potential uses of MNBD for public purposes, as well as to articulate the challenges that need to be addressed if such work is to be mainstreamed. This paper also provides a literature review of existing work that has leveraged MNBD to produce insights to inform a host of public policy domains including Transport, Economic Development and Health.

Key words: mobile network, big data, behavioural variables, policy insights, transport, disaster management, regional integration.

■ Why big data and why now?

Whilst the use of the term, "big data" has become ubiquitous, the definitions utilized are varied and sometimes even contradictory. This is understandable given the shared origins of the term in academia, industry and media (WARD & BARKER, 2013). The term is used to describe the characteristics of the data i.e. it involves large volumes of data, but also the process and techniques used to analyze it. The question of volume is itself subjective as what may be considered big today may be small tomorrow, given the exponentially growing deluge of data. One of the more well known and (and one of the earliest) uses of the term was by the consultancy Gartner, which in addition to volume, used additional characteristics such as velocity and variety to describe the data (LANEY, 2001). Velocity refers to

the speed at which data is generated, assessed and analyzed. The term "Variety" encompasses the fact that data can exist as different media (text, audio, video) and come in different formats (structured and unstructured). In the end, the definition will remain anecdotal and despite the significance denoted by the use of the word "big" in the term, will be unquantifiable.

As an amorphous, umbrella term, it is amenable to include data that even predates the use of the term. Included within the scope then is transaction-generated data (TGD) which is sometimes described as "data exhaust". This is data that has been generated as a by-product of doing things (such as providing telephone service, processing payments, and so on), This category was first discussed in 1991, though the term then used was transaction-generated information (McMANUS, 1990). The value of this subset of big data is that it is directly connected to human behavior and its accuracy is generally high. The data generated by mobile networks falls firmly within this TGD category. Given the volumes of data that are now generated by mobile network due to the almost ubiquitous use of mobile phones by the majority of the population, the data can also be considered as big data.

As such mobile network big data (MNBD) may not itself be a completely new phenomena. Both in terms of the underlying data as well as the ways to leverage them, there is a considerable lineage. For example reducing churn has long been of acute interest to mobile network operators. It cannot be denied however that pace of innovation in leveraging MNBD for both private as well as public purposes has accelerated only in recent times. More broadly the use and application of big data has also only recently been "democratized" due to the falling costs of storage and processing power, as well as the rise of open source alternatives to handling and analyzing the data.

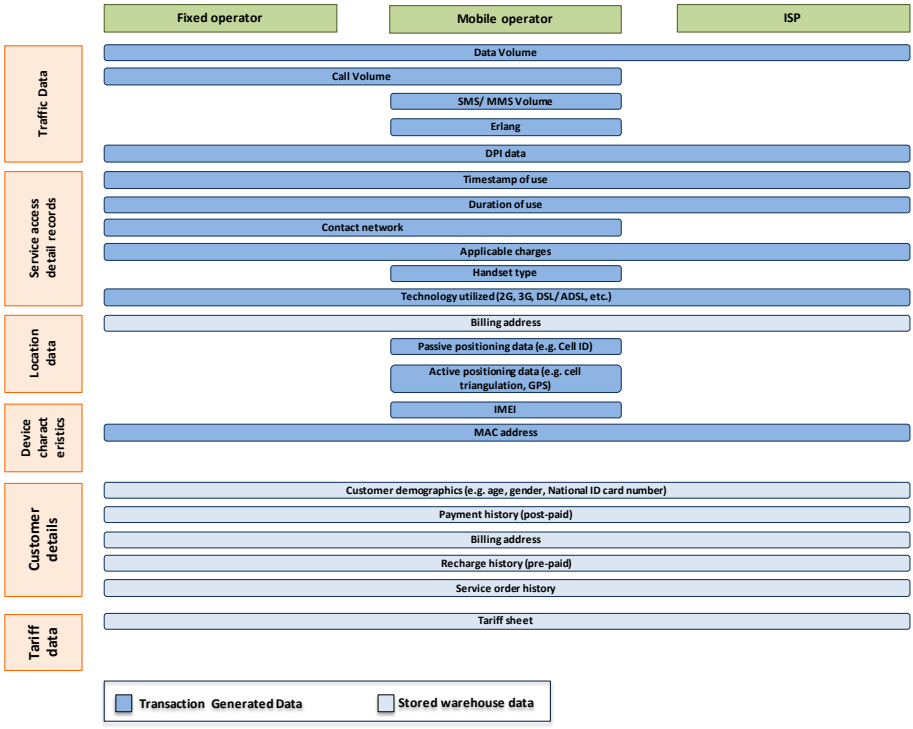
Equally new is the interest in leveraging MNBD for development. With basic mobile subscriptions in the world estimated to be about 96.4% at the end of 2014 (ITU, 2014), MNBD offers the most comprehensive digitized source of data about the poor especially in the developing economies, with generally low levels of "datafication" [1].

[1] MAYER-SCHÖNBERGER & CUKIER (2013) introduced the neologism "datafied" to describe digitized data that has been quantified so as that it may be tabulated and analyzed.

The focus of this paper is to illuminate both academics as well as policy makers on the potential uses of MNBD for public purposes, as well as to articulate the challenges that need to be addressed if such work is to be mainstreamed. The rest of the paper is structured as follows. The 2nd Section provides an overview of the different types of data that are captured by the mobile network and introduces a behavioral classification for the different forms of indicators that may be extracted from the data. The 3rd Section provides a literature review of existing work that has leveraged MNBD to produce insights to inform a host of public policy domains. The 4th Section outlines the challenges and where applicable suggests ways to address them. Finally the 5th section concludes.

■ Mobile network big data

Figure 1 - An overview of telecom network data (*)



(*) Figure adapted from NAEF *et al.* (2014)

The majority of Mobile Network Big Data (MNBD) can be considered as Transaction Generated Data (TGD) [2], captured when customers make or receive a call or SMS, access the internet, use a Value Added Service (VAS), recharge their prepaid accounts, etc.

Types of mobile network big data

The range of data captured by telecom network operators can be categorized based on their main usage by telecom operators: network infrastructure management, marketing and sales, and billing and customer care. The following sub-sections describe the data in greater detail.

Traffic data

Operators use a range of metrics to understand and manage the traffic flowing through their networks. Some of these include:

- Erlang: a dimensionless metric used by operators to understand the offered and utilized load. One Erlang could be equivalent to one person talking for 60 minutes or 2 people talking for 30 minutes each, etc. Erlang data is used to understand the load on a base station at any given time.
- Call, SMS and MMS volumes: These are used for a variety of purposes from billing to customer relationship management as well as for network planning.

Service access detail records

Whenever a user utilizes a telecom service, each access is recorded not just for infrastructure management but also for billing purposes. Irrespective of the type of service accessed, and partly due to the original heavy emphasis on voice-based communication these records can collectively be considered as Call Detail Records (CDRs). They cover not just voice communication, but also SMS/ MMS, internet access, VAS access, etc. Depending on the operator, these records may be stored as separate datasets based on the service or collectively. CDRs will capture the following data artifacts:

[2] The term "meta-data" is also used quite extensively to refer to TGD data from telecommunication operators.

- timestamp of when the service was accessed,
- the duration the service was used for (e.g. duration of a call),
- the numbers of all parties on the communication (For example a CDR would include both the number of originating as well as terminating party),
- applicable charges for the access,
- the type of handset used,
- the technology utilized (2G, 3G, etc.),
- in the case of internet access, the volume of data uploaded/ downloaded as well as the technology utilized.

Figure 2 - A stylized CDR

Calling Party Number	*Called Party Number*	*Caller Cell ID*	*Call Time*	*Call Duration*	*Served IMSI*
76624XXXX	71942XXXX	3134	13-04-2013 17:42:14	00:03:35	138472135843XXXXX

Location data

Mobile networks can capture a range of location variables depending on the sophistication of the network. Passive positioning data is automatically generated by the network and captured in the mobile network's logs for billing purposes and also for network management. CDRs are the main sources for passive positioning data for mobile operators and reside in their data warehouses. Active positioning data is captured after a specially initiated network query to locate a handset using either network or handset based positioning methods. Location data from GPS can also be considered as active positioning data (AHAS, AASA, SILM & TIRU, 2010).

Behavioral variables for development

Given the developmental policy focus of this paper, it is useful to classify the data captured by mobile networks into the following behavioral variable categories:

- *Mobility variables:* The majority of a subscriber's activity in the operator's system has a location attribute, which at the very minimum includes the ID of the antenna that was utilized when making said activity. The antenna in turn has a geo-location.

- *Connectivity variables:* The massive graph of human interactions captured by MNBD, allows for range of social network metrics to be derived.
- *Consumption variables:* From the amount and frequency of airtime top up, to the type of handset, levels of usage of different services, MNBD affords the possibility to understand consumption behavior by mobile users.

■ From variables to policy insights

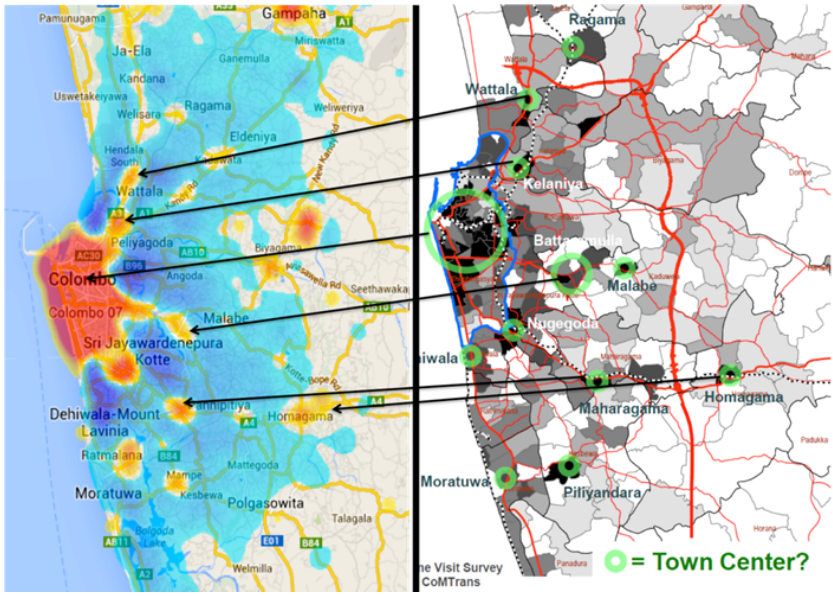
Transport

In developing economies, active and passive positioning data from mobile network big data as well as real-time GPS traces from mobile phones can revolutionize transportation management improving the efficiency and reliability of the overall system. Amongst the three possible mobile network location data, passive positioning data provides the least spatial accuracy (Cell ID). However, even the least sophisticated network produces passive positioning data (e.g. CDRs). The recent literature is tilted towards utilizing these passive location data for producing insights of relevance to transportation policy.

Mobile network big data has been utilized to great effect in transportation, helping measure and model people's movements in both developing as well as developed economies. Trip based Origin-Destination (OD) matrices (traditionally derived infrequently using surveys) have been created using mobile network big data in Korea (YOO, CHON, KANG & KIM, 2005), Spain (CACERES, WIDEBERG & BENITEZ, 2007), Sri Lanka (LOKANATHAN, SILVA, KREINDLER, MIYAUCHI & DHANANJAYA, 2014), United States (CALABRESE, DI LORENZO, LIU & RATTI, 2011) amongst others. Others have used mobile network big data to infer transportation modes (WANG, CALABRESE, DI LORENZO & RATTI, 2010) as well as understand traffic flow (WU *et al*., 2013)

Most recently IBM researchers utilized CDR data from Orange to map out citizens' travel routes and show how data-driven insights could be used to improve planning and management of transportation services in Abidjan, the largest city in Côte d'Ivoire (BERLINGERIO *et al*., 2013). They were able to show how overall travel time could be reduced by 10% by optimizing the network thus offering a partial solution to the city's congestion problems.

Figure 3 - Use of mobile network big data for providing transportation insights (*)



The image on the left depicts relative density of people in Colombo city and the surrounding regions based on mobile operator TGD at 1300 compared to 0000 (midnight the previous day) on Tuesday, 15th January 2013. The yellow to red colors depict areas whose density has increased relative to midnight. The blue color depicts areas whose density has decreased relative to midnight (the darker the blue, the greater the loss in density). The clear areas are those where the overall density has not changed. The image on the right depicts the major transportation transit points identified using a costly survey of 40,000 households to understand mobility patterns, which closely match the main points identified using big data analyses of telecom network big data.

Source: LOKANATHAN et al. *(2014)*

Socio-economic monitoring and planning

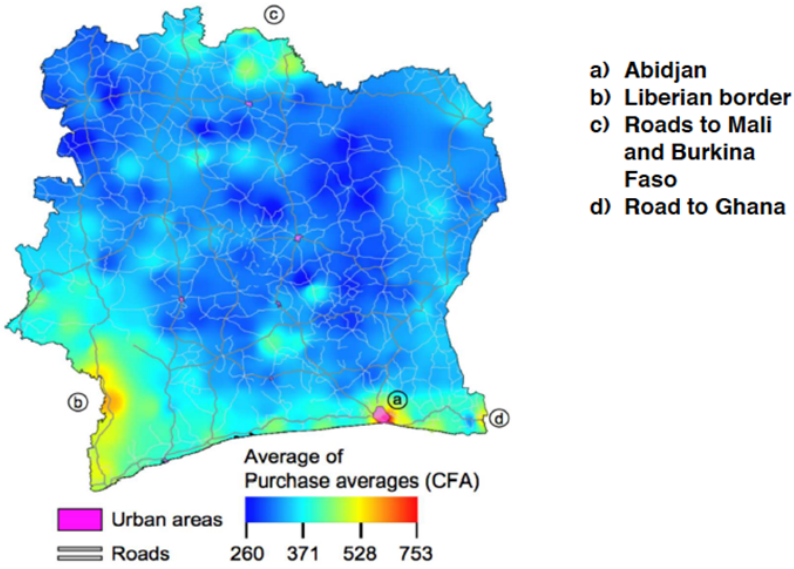
MNBD also affords the possibility for understanding the socio-economic status of regions across the country in near real-time.

FRIAS-MARTINEZ *et al.* (2012) showed correlations between human mobility variables derived from MNBD, and Socio Economic Levels (SELs) from data from the National Statistical Agency in an unnamed country. Her results showed that populations with higher SELs are strongly linked to larger mobility ranges than populations from lower socio-economic statuses. She was able to leverage this to build models that could reverse-engineer the aggregate SELs of small spatial regions, from just MNBD.

Another study by GUTIERREZ, KRINGS & BLONDEL (2013) used MNBD from Côte d'Ivoire (both airtime purchase records as well as communication patterns) to estimate SELs and the diversity and variation in income levels. Their works articulated socio-economic segregation at spatially-fine levels for Côte d'Ivoire. Such research (see Figure 4) can be utilized for poverty mapping.

In fact, mobile network operators often track their revenue from each base station in real time. Access to such data could potentially allow base stations to act as real-time sensors of shocks to the local economy. However given the sensitivity of revenue data for operators, accessing such data by third parties will remain a challenge.

Figure 4 - Poverty mapping in Côte d'Ivoire



Using MNBD (specifically the communication patterns as well as the history of airtime credit purchases) from Orange in Côte d'Ivoire, researchers estimated the relative income of individuals, and the diversity and inequality of income. The above figure shows the poor areas (in blue) in relation to the areas of high economic activity (yellow to red areas).

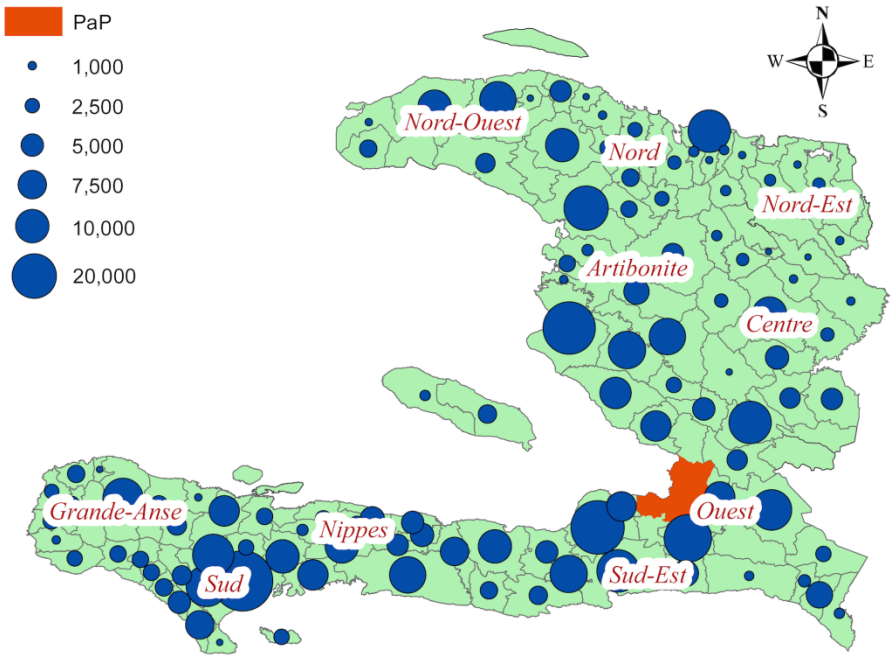
Source: GUTIERREZ et al. *(2013)*

According to the Consultative Group to Assist the Poor (CGAP) and the GSM Association (GSMA), in 2012 nearly 2 billion people were estimated to have a mobile phone but no bank account. A CGAP study (KUMAR & MUHOTA, 2012) hypothesized that MNBD (and specifically the consumption variables) could be leveraged to facilitate financial inclusion by providing

new measures of creditworthiness for the unbanked. For example people who purchased airtime frequently and in a consistent pattern demonstrate predictability in income and planning, which might impact their ability to repay a loan. Conversely, inactive prepaid accounts or ones that consistently run to zero, suggest that their owners may not replay a loan in a timely manner. Such potential has not been lost on the private sector. Cignifi, a big data firm has built a credit scoring model using CDR data and tested it in Tanzania and Brazil against historical lending data from banks and micro-finance institutions (KUMAR & MUHOTA, 2012). Their results suggest that their model is an accurate predictor of default.

Disaster management and syndromic surveillance

Figure 5 - Post-earthquake distribution of Port au Prince (Haiti) population following after 2010 earthquake



Source: BENGTSSON et al. *(2011)*

Mobility data from MNBD can show population displacements after a disaster. Such insights allow first responders and relief agencies to quickly locate affected populations, and improve their targeting of aid and scarce resources. LU, BENGTSSON & HOLME (2012) showed the new locations of

the former residents of Port-au-Prince who were displaced when the 2010 earthquake leveled their city. This retrospective study showcased the value that MNBD can provide in disaster management (see Figure 5).

Mobility data can be similarly leveraged to understand the spread of communicable diseases. WESOLOWSKI *et al*. (2012) utilized passive mobile location data from CDRs and malaria prevalence data to identify the sources and sinks of malaria infections. Fighting and controlling malaria can be enhanced through such new models of disease spread.

Regional/international integration

Traditional analyses of the extent of inter- and intra-regional ties, as well as the level of international ties (economic, connectivity, etc.) have relied on aggregate parameters such as labor market data, commuter and/or travel flows, and other indices of accessibility and socioeconomic status. MNBD can provide new proxy measures of such ties, both at a micro as well as a macro level.

Using CDR data from seven developed and developing countries of varying geographical size as well as population (Belgium, Côte d'Ivoire, France, Italy, Portugal, Saudi Arabia, and United Kingdom), SOBOLEVSKY *et al*. (2013) showed the geospatial dispersion as well as cohesion of societies within a country. Such a large-scale quantitative study of human geography has not been possible before. Their work as well as work by others (e.g. MADHAWA *et al*., 2015 in Sri Lanka) can show how administrative boundaries (which are usually a product of history and geography) relate to actual current community structures.

Using many years of CDR data from Rwanda, BLUMENSTOCK (2011) showed how the volume and direction of international call traffic could be used to understand the nature and strength of inter-personal relationships across international boundaries. The comprehensive coverage afforded by such data can also show the extent of connectivity in times of crises, which cannot be captured when studying only trade documentation.

■ Challenges

Whilst the value of insights from such data is potentially vast, there are still many challenges. Obtaining access to such data (often private) is difficult, even if it is to be leveraged for public purposes. Different data

providers curating and structuring their data according to their own standards, means co-mingling data is often cumbersome. Whilst many agree there are privacy concerns when conducting such analyses, there is little consensus in how to address them. Finally it should be realized that the state of the art is still very much in its embryonic stage, and understanding and addressing the analytical challenges will remain paramount.

The following sub-sections outline these challenges in greater detail and reflect on potential remedies.

Accessing data

With deregulation and liberalization, the majority of the mobile network operators in the world are not state-owned. Therefore obtaining data, even for developmental purposes is difficult. Companies are loath to share information on their clients and the business processes, as it may reveal competitive information or run foul of existing laws and regulations. Furthermore there are few if any incentives for MNOs to share such data for such developmental purposes.

Researchers (mainly from known institutions in developed countries, with some exceptions such as LIRNE*asia*) [3] have recently succeeded in obtaining MNBD but at considerable expenditure of time (in building and leveraging the relationships with operators). Such privileged access mostly comes bound by lengthy legal agreements. They often have to address the parameters related to how the data is to be used, how data is to be anonymized and extracted, time-periods of access, etc.

However, a few operators have recently started to share some of their data (often in aggregate form) publicly, but still through specialized mechanisms. Orange released an aggregated anonymized mobile dataset from Côte d'Ivoir and convened a conference at MIT in early 2013. The latest iteration of their Data for Development challenge, this time using data from Senegal is underway [4]. In 2014, Telecom Italia went further. In addition to MNBD from Milan and the Autonomous Province of Trento, they also curated and provided additional datasets (weather, electricity, public and private transport, social network data, etc.) for the same regions [5].

[3] LIRNEasia is a regional think tank based in Sri Lanka.
See http://lirneasia.net/projects/bd4d/ for more information on the research using MNBD.

[4] See http://www.d4d.orange.com/home

[5] See http://www.telecomitalia.com/tit/en/bigdatachallenge/contest.html

MNBD is not amenable to come under the gambit of open data initiatives that have been gaining popularity. However UN Global Pulse is trying to popularize the concept of 'data philanthropy' that would enable the sharing of data regularly and safely [6]. Others like LIRNE*asia* seek to develop bottom-up, pragmatic, cooperative arrangements with government and private actors. LIRNE*asia* has developed draft guidelines for the ethical use of mobile network big data in collaboration with mobile operators in the South Asian region [7]. Whilst such efforts remain critical, another related question is how to deal with data from multiple sources. This suggests there may be a space for third-party facilitators, who can standardize and curate the data, as well as handle the varying regulatory and privacy burdens, when sharing such data with external researchers and entities. Such a specialized entity can greatly reduce the transaction cost to companies in providing data, whilst facilitating the generation of scientific knowledge. Such an approach was taken by Johnson and Johnson that shared all of its clinical trial data with Yale University's Open Data Access (YODA) Project (KRUMHOLZ, 2014). YODA in turn will facilitate data access to outside researchers, whilst ensuring data protection. In the context of MNBD, it is not yet clear who can play this role: telecom regulators, National Statistical Organization (NSO), or others? The decision however will require the confluence of multiple actors.

Privacy

The privacy issues are taking center stage with the rise of big data, and will be equally important as social scientists look towards private data sources. The conversation on privacy involves not just academics, but also state and private sector, and the general public who often are the primary producers of such data through their activities. But a consensus on how to address the attendant privacy challenges is yet to be reached. Current practices related to data privacy utilize a rights-based approach. The International Telecommunication Union (ITU) for example, defines individual privacy as "the right of individuals to control or influence what information related to them may be disclosed" (ITU, 2006). This approach is best illustrated by the 'inform and consent' policy, whereby companies inform users of what data is being collected about them and how it will be utilized by them and potentially also by the companies' affiliates and partners. But as MAYER-SCHÖNBERGER & CUKIER (2013) rightly point out, the 'inform and consent' model is impractical. Most current user privacy policies are

[6] See http://www.unglobalpulse.org/blog/data-philanthropy-public-private-sector-data-sharing-global-resilience for more information

[7] For an early draft see http://lirneasia.net/wp-content/uploads/2014/08/Draft-guidelines-2.2.pdf

lengthy and written in legalese that makes understanding them difficult for the layperson. They also do not deal effectively with the secondary uses for the data, which often only manifest long after the original data was collected. It becomes impractical then for companies to know in advance all the potential uses or continuously seek permission for each new use.

The lines between personal and non-personal information are further blurred, when data is mashed up. Previously non-personal data, when mixed, could at times reveal insights that can easily be linked to an actual individual (OHM, 2010). A recent study showed how personal attributes such as ethnicity, religious and political views, and even sexual orientation can be inferred from Facebook likes (KOSINSKI, STILLWELL & GRAEPEL, 2013).

Even as computational social scientists utilize anonymization techniques to address privacy concerns, the methods themselves are being called into question (NARAYANAN & SHMATIKOV, 2008). DE MONTJOYE, HIDALGO, VERLEYSEN & BLONDEl (2013) were able to identify 90% of the people using just 4 data points from an anonymized set of 1.5 million CDRs. Even though the data itself does not have any identity information, the authors showed that the real-world identities could be found by cross-referencing their data with other public data. However, till the overall levels of "datafications" rise from its current low levels in developing countries, such concerns may be a bit premature for such countries. Furthermore the large majority of mobile phone connections in the developing world are prepaid accounts, with minimal (if any) registration information associated with them. Even though SIM registration is now mandatory in many countries, often those registered against a SIM are not the actual users.

Even if the means to understand and address privacy concerns are far from clear, there is general agreement that there must be some safeguards. These safeguards may be technological, conceptual, legal or even a combination of all three.

Analytical challenges

"Garbage in, Garbage out" is a computer science axiom that means that bad data will result in invalid results, irrespective of the process. This axiom tends to be overlooked at times in the big data paradigm. When dealing with large quantities of data, often unstructured and often from multiple sources, there is an implicit assumption that the data will be "messy". The belief is that "what we lose in accuracy at the micro-level, we gain in insight at the macro level" (MAYER-SCHÖNBERGER & CUKIER, 2013, p. 36). This is misleading. Data quality and its provenance do matter and the question is important in establishing generalizability of the big data findings. Whilst

establishing data provenance is complex, it is important to understand the processes that may have created the data. For example researchers in Sri Lanka working with MNBD, found strange mobility patterns in some of the data. A subsequent investigation with the operator revealed that their network engineers had in the past reused the same Cell ID when moving antennas. Whilst the practice had been corrected, the older data had remained faulty. Knowing such ground context is important not just in understanding the base raw data, but also when interpreting results. For example Nathan Eagle, a pioneer in using big data for development, upon discovering low population mobility following a flood when analyzing CDR data from Rwanda, theorized the cause to be the outbreak of cholera. However a quick ground survey revealed that the real reason was washed out roads (DAVID, 2013).

Whilst the large data sizes may make questions regarding the sampling rate irrelevant, knowing the representativeness of the data is still important. Even as mobile subscriptions in many developing countries nears 100%, it still doesn't mean that every person in the country owns a mobile phone. Understanding sampling bias will be important. Depending on the research being pursued, questions such as the extent of coverage of the poor, or the levels of gender representation amongst mobile phone users can be very important. Street Bump, a mobile app that notifies Boston City Hall whenever app users drive over a pothole on Boston roads, suffers from a selection bias. This is because the app is biased towards the demographics of the app users, who often hail from affluent areas with greater smartphone ownership (HARFORD, 2014). Theoretically it could be possible that so long as resources are allocated on the basis of insights from such apps, the poorer parts of the city will be become further marginalized. Hence even in the big data paradigm, understanding and accounting for measurement bias, ensuring internal and external validity, and understanding inter-dependencies in the data, all remain important. These are foundational issues not just for “small data” but also for "big data" (BOYD & CRAWFORD, 2012).

The confusion of correlation with causation becomes more pronounced in the big data paradigm. By being observational, big data can measure only correlation and not causality. The techniques of data mining and machine learning, which underpin much of the big data analytics, are primarily about correlation and predictions. When big data started to gain popularity, evangelists were quick to proclaim the end of theory and hypothesis testing, with correlation being all that mattered (see for example ANDERSON, 2008 and even MAYER-SCHÖNBERGER & CUKIER, 2013). While it is true, often correlations can be enough to make decisions, the evanaglists proclamations have not come to bear. The noted behavioral economist Sendhil Mullainathan argues that inductive science (i.e. algorithmically

mining big data sources) will not drown out traditional deductive science (i.e. hypothesis testing) even in a big data paradigm (MULLAINATHAN, 2013). Among the three Vs in the traditional big data definition, volume and variety produce countervailing forces. More volume makes big data induction techniques easier and more effective, while more variety makes them harder and less effective. It is this variety issue that will ensure the need for explaining behavior (i.e. deductive science) rather than just predicting it.

All this is not to say that causal modeling is not possible in the big data paradigm. In fact this is achievable by conducting experiments (VARIAN, 2013). Telecom network operators themselves use such techniques when rolling out new services or, for that matter, for pricing purposes. But third-party researchers will not easily have access to such experimentation possibilities since these are proprietary systems.

Another misconception is that the rise of big data will mean that surveys will go the way of the dodo. Even when leveraging MNBD for development, surveys and supplemental datasets will remain important to sharpen the analyses and especially to verify the underlying assumptions. For example BLUMENSTOCK & EAGLE (2012) ran a basic household survey against a randomized set of phone numbers prior to data anonymization to build a training dataset. This allowed them to understand variations in mobility, social networks and consumption amongst men and women, and between different socio-economic groups that wouldn't have been possible using just the call records.

The broader analytical challenge is that new scientific knowledge from big data, especially when this is using data from private sources, is difficult to verify. Transparency and replicability are critical if the underlying methods and resultant insights are to be honed and improved. This is particularly important given extant embryonic stages computational social science. This underscores how important it will be to open up the private data sources (in a manner that addresses potential privacy concerns) so as to be able to avail of the benefits of proper peer-review.

Skills

Data science is a frontier field and will require broad expertise in a variety of fields. These include a combination of data mining, statistics, domain expertise, and also skills in data preparation, cleaning, and visualization. NSOs may have deep in-house statistical skills, but this is not enough to work with the large volumes of big data which call for computer science and decision analysis skills which are not emphasized in traditional statistical courses (McAFEE & BRYNJOLFSSON, 2012). Currently there is a

mismatch between the supply and demand for talent with the needed broader skill-sets i.e. data scientists. McKinsey predicts that by 2018 the demand for data-savvy managers and analysts in the United States would be 450,000, yet the supply will fall far short at only 160,000 (MANYIKA *et al*., 2011). In the short-term, the work, especially those for public-purposes will have to be done using collaborative teams, which can draw on a variety of skill-sets to collectively analyze and make sense of the data.

■ Conclusion

Spurred by the exponential growth of mobile connectivity, the attendant large volumes of MNBD, offer the possibility to obtain rich behavioral insights at a scale that was never possible before. Against this landscape, academics and policy makers need to have a holistic understanding of the state of the art and the implications of analyzing the only extant source of big data that includes the poor, i.e. MNBD.

It can be seen that mobile TGD can be used to gain useful insights in a variety of domains including Transport, Economic Development, Health, etc. In the Transport sector the use of mobile network big data would be a much less expensive, less intrusive and a less time consuming way of analyzing large-scale mobility than always relying on survey instruments. Of course there are drawbacks with the location resolution being highly dependent on the density of the base stations. There are also a number of studies where the real-time economic development/ socio economic status of regions across the country has been assessed using MNBD. Especially when studying low-income groups this is one of the few sources of information and is a more efficient way for the governments to identify those areas in need of economic and financial aid rather than conducting traditional surveys that are expensive and time consuming. In the health sector, integrating mobility data from mobile networks with epidemiological data, show great potential for tracking the spread of communicable disease.

However it is essential to be cognizant of the fact that the state of the art is still developing and there are analytical and technological challenges to be aware of. Hence understanding issues related to the data provenance, what the data represents, and the ground contexts are important so that incorrect conclusions are not drawn from a blind application of big data techniques to the mobile network operator data. Survey data will remain important not only to bootstrap some of the big data techniques with training data but also to fine-tune models to ground realities. Hence big data will not completely replace traditional surveys but rather complement them. Similarly given the analytical challenges it will be very important that there is transparency and

replicability in the analyses. Being mindful of these analytical challenges does not negate the potential of big data, but rather helps refine and improve the overall process of leveraging mobile operator big data for development. Hence doomsayers and those that hype the potential are both to be taken with a pinch of salt.

While there are many challenges to be overcome not least of which is understanding and addressing the privacy implications of big data, it can be seen that the sharing (subject to appropriate privacy protocols) of privately held data such as mobile phone records can be mutually beneficial to both government as well as the private sector. For example, emerging research in Africa shows how the reduction in airtime top-ups forecast declines in income among the poor. This can allow for targeted and timely policy actions by government to address the underlying problems, which would not be possible with the lagged, and often limited, insights revealed by traditional statistics. Such a collaborative early warning and early action system shows how such sharing could be considered a business risk mitigation strategy for operators in emerging markets. But such cooperation is predicated on opening up the currently privileged access that a few researchers and organizations have been given to mobile operator datasets.

References

AHAS, R., AASA, A., SILM, S. & TIRU, M. (2010): *Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: Case study with mobile positioning data*, Transportation Research Part C: Emerging Technologies, 18(1), 45-54.

ANDERSON, C. (2008): "The end of theory: the data deluge makes the scientific method obsolete", *Wired Magazine*, 16(7).

BERLINGERIO, M., CALABRESE, F., DI LORENZO, G., NAIR, R., PINELLI, F. & SBODIO, M. L. (2013): "AllAboard: a system for exploring urban mobility and optimizing public transport using cellphone data", in *Data for Development: Net Mobi 2013*.

BLUMENSTOCK, J. E. (2011): "Using mobile phone data to measure the ties between nations, in *Proceedings of the 2011 iConference* (pp. 195-202), New York, USA: ACM Press.

BLUMENSTOCK, J. E. & EAGLE, N. (2012): "Divided We Call : Disparities in Access and Use of Mobile Phones in Rwanda", *Information Technologies & International Development*, 8(2), 1-16.

BOYD, D. & CRAWFORD, K. (2012): Critical Questions for Big Data, *Information, Communication & Society*, 15(5), 662-679.

CACERES, N., WIDEBERG, J. P., & BENITEZ, F. G. (2007): "Deriving origin-destination data from a mobile phone network", *IET Intelligent Transport Systems*, 1(1), 15.

CALABRESE, F., DI LORENZO, G., LIU, L. & RATTI, C. (2011): "Estimating Origin-Destination Flows Using Mobile Phone Location Data", *IEEE Pervasive Computing*, 10(4), 36-44.

DAVID, T. (2013): "Big Data from Cheap Phones", *Technology Review*, 116(3), 50-54.

DE MONTJOYE, Y.-A., HIDALGO, C. A, VERLEYSEN, M. & BLONDEL, V. D. (2013): "Unique in the Crowd: The privacy bounds of human mobility", *Scientific Reports*, 3, 1376.

FRIAS-MARTINEZ, V., VIRSEDA-JEREZ, J. & FRIAS-MARTINEZ, E. (2012): "On the relation between socio-economic status and physical mobility", *Information Technology for Development*.

GUTIERREZ, T., KRINGS, G. & BLONDEL, V. D. (2013): "Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets", 1-6. http://arxiv.org/abs/1309.4496

HARFORD, T. (2014, March 28): "Big data: are we making a big mistake?", *Financial Times*, pp. 7-11.
http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html

ITU - International Telecommunication Union:
- (2006): "Security in Telecommunications and Information Technology: An overview of issues and the deployment of existing ITU-T Recommendations for secure telecommunications".
http://www.itu.int/dms_pub/itu-t/opb/hdb/T-HDB-SEC.03-2006-PDF-E.pdf
- (2014): *Measuring the Information Society Report 2014*, Geneva.

KOSINSKI, M., STILLWELL, D. & GRAEPEL, T. (2013): Private traits and attributes are predictable from digital records of human behavior, 2-5.

KRUMHOLZ, H. M. (2014, Feb. 2): "Give the Data to the People", *New York Times*.
http://www.nytimes.com/2014/02/03/opinion/give-the-data-to-the-people.html?hp&rref=opinion&_r=1

KUMAR, K. & MUHOTA, K. (2012): *Can Digital Footprints Lead to Greater Financial Inclusion?*, pp. 1-4, Washington DC.

LOKANATHAN, S., DE SILVA, N., KREINDLER, G., MIYAUCHI, Y. & DHANANJAYA, D. (2014): *Using Mobile Network Big Data for Informing Transportation and Urban Planning in Colombo*, LIRNEasia, August.

LANEY, D. (2001): "3D data management: Controlling data volume, velocity and variety", Gartner.

LU, X., BENGTSSON, L. & HOLME, P. (2012): "Predictability of population displacement after the 2010 Haiti earthquake", *Proceedings of the National Academy of Sciences of the United States of America*, 109(29), 11576-81.

MADHAWA, K., SAMARAJIVA, R., LOKANATHAN, S. & MALDENIYA, D. (2015): "Understanding communities using mobile network big data".

MANYIKA, J., CHUI, M., BROWN, B., BUGHIN, J., DOBBS, R., ROXBURGH, C. & BYERS, A. H. (2011): "Big data: The next frontier for innovation, competition, and productivity".
http://www.mckinsey.com/insights/mgi/research/technologyandinnovation/bigdatathenext_frontier_for_innovation

MAYER-SCHÖNBERGER, V. & CUKIER, K. (2013): *Big Data: A Revolution that Will Transform How we Live, Work, and Think*, Houghton Mifflin Harcourt.

McAFEE, A. & BRYNJOLFSSON, E. (2012): "Big data: the management revolution", *Harvard Business Review*, 90(10), 60-6, 68, 128.
http://www.ncbi.nlm.nih.gov/pubmed/23074865

McMANUS, T. E. (1990): *Telephone transaction generated information: Rights and restrictions*, Cambridge, MA.
http://www.pirp.harvard.edu/pubs_pdf/mcmanus/mcmanus-p90-5.pdf

MULLAINATHAN, S. (2013): *What Big Data Means For Social Science*, "HeadCon'13 Part I". http://edge.org/panel/headcon-13-part-i

NAEF, E., MUELBERT, P., RAZA, S., FREDERICK, R., KENDALL, J. & GUPTA, N. (2014): *Using Mobile Data for Development*, Cartesian, Inc. http://www.cartesian.com/wp_content/upload/Using-Mobile-Data-for-Development.pdf

NARAYANAN, A. & SHMATIKOV, V. (2008): "Robust De-anonymization of Large Sparse Datasets", in *2008 IEEE Symposium on Security and Privacy*, pp. 111-125.

OHM, P. (2010): "Broken promises of privacy: Responding to the surprising failure of anonymization", *UCLA Law Review*, 57(6).

SOBOLEVSKY, S., SZELL, M., CAMPARI, R., COURONNÉ, T., SMOREDA, Z. & RATTI, C. (2013): "Delineating geographical regions with networks of human interactions in an extensive set of countries", *PloS One*, 8(12), e81707.

WARD, J. S. & BARKER, A. (2013): "Undefined By Data: A Survey of Big Data Definitions". http://arxiv.org/abs/1309.5821

WANG, H., CALABRESE, F., DI LORENZO, G. & RATTI, C. (2010): "Transportation mode inference from anonymized and aggregated mobile phone call detail records", In *13th International IEEE Conference on Intelligent Transportation Systems*, pp. 318-323.

WESOLOWSKI, A., EAGLE, N., TATEM, A. J., SMITH, D. L., NOOR, A. M., SNOW, R. W. & BUCKEE, C. O. (2012): "Quantifying the impact of human mobility on malaria", *Science*, New York, N.Y., 338(6104), 267-70.

WU, W., CHEU, E. Y., FENG, Y., LE, D. N., YAP, G. E. & LI, X. (2013): "Studying Intercity Travels and Traffic Using Cellular Network Data", In *Data for Development: Net Mobi 2013*.

YOO, B.-S., CHON, K., KANG, S.-P. & KIM, S.-G. (2005): "Origin-Destination Estimation using Cellular Phone BS Information", *Journal of the Eastern Asia Society for Transportation Studies*, 6, 2574-2588.