Issues in Technology Innovation

November 2014

Enabling Humanitarian Use of Mobile Phone Data

Yves-Alexandre de Montjoye, Jake Kendall, and Cameron F. Kerry

INTRODUCTION

Yves-Alexandre de Montjoye is a Ph.D. Candidate at MIT Media Lab. demontjoye.com

Jake Kendall is a Program Officer in the Financial Services for the Poor Initiative at The Bill & Melinda Gates Foundation. The statements and opinions in this document are the author's alone and in no way reflect the opinions, strategy, or views of the Bill & Melinda Gates Foundation.

Cameron F. Kerry is a Visiting Scholar, MIT Media Lab, Ann R. & Andrew H. Tisch Distinguished Visiting Fellow at The Brookings Institution and Senior Counsel at Sidley Austin LLP.

Mobile phones are now ubiquitous in developing countries, with 89 active subscriptions per 100 inhabitants.[1] Though many types of population data are scarce in developing countries, the metadata generated by millions of mobile phones and recorded by mobile phone operators can enable unprecedented insights about individuals and societies. Used with appropriate restraint, this data has great potential for good, including immediate use in the fight against Ebola.[2]

The metadata generated by millions of mobile phones and recorded by mobile phone operators can enable unprecedented insights about individuals and societies. Used with appropriate restraint, this data has great potential for good, including immediate use in the fight against Ebola.

To operate their networks, mobile phone operators collect call detail records—metadata of who called whom, at what time, and from where. After the removal of names, phone numbers, or other obvious identifiers, this data can be shared with researchers to reconstruct precise country-scale mobility patterns and social graphs. These data have already been used to study importation

1 ITU, (2013) ICT Facts and Figures http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFacts-Figures2013-e.pdf.

2 For a longer piece on the various ways in which mobile data can be used in the development sphere see: Kendall et al. (2014) Using Mobile Data for Development http://www.impatientoptimists.org/Posts/2014/07/Big-Data-and-How-it-Can-Serve-Development.

Our analysis showed (1) the lack of commonly-accepted practices for sharing mobile phone data in privacy-conscientious ways and (2) an uncertain and country-specific regulatory landscape for data-sharing especially for cross-border data sharing.

routes of infectious diseases,[3] migration patterns, or economic transactions.[4] Such data are now being actively sought to inform the fight against Ebola[5] but, despite the promise, this effort appears stalled.[6]

As part of MIT's Big Data initiative, we examined two operational use cases of mobile phone data for development modeled on previous research. The first case, involved the use of location metadata to understand and quantify the spread of infectious diseases (e.g. malaria or Ebola) within and among countries.[6] The second case considered the use of behavioral indicators derived from mobile phone metadata to micro-target outreach or drive uptake of agricultural technologies or health seeking behavior.[7] Here, mobile phone data could be used to define subgroups based on specific traits and behaviors, which would then receive messages or other outreach from the mobile operator.[8] We also considered cases where the data could be used to select individuals to be identified and contacted directly in limited circumstances. These two scenarios are quite distinct from a regulatory and privacy perspective, as we discuss below.

These mobile phone data case studies revealed ways in which, despite the promise, regulatory barriers and privacy challenges are preventing the use of mobile phone metadata from realizing its full potential. More specifically, our analysis showed (1) the lack of commonly-accepted practices for sharing mobile phone data in privacy-conscientious ways and (2) an

3 Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., and Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. Science, 338(6104), 267-270. http://www.sciencemag.org/content/338/6104/267.abstract.

4 Examples include: WorldPop (2014) Ebola http://www.worldpop.org.uk/ebola/; Eagle, N., Macy, M., and Claxton, R. (2010). Network diversity and economic development. *Science, 328* (5981), 1029-1031. http://www.sciencemag.org/content/328/5981/1029; or Eagle, N., de Montjoye, Y., and Bettencourt, L. M. (2009). Community computing: Comparisons between rural and urban societies using mobile phone data. In IEEE *Computational Science and Engineering, 2009*. http://doi.ieeecomputersociety.org/10.1109/CSE.2009.91.

5 Wesolowski, A., Buckee, C. O., Bengtsson, L., Wetter, E., Lu, X., and Tatem, A. J. (2014). Commentary: Containing the Ebola Outbreak-the Potential and Challenge of Mobile Network Data. PLOS Currents Outbreaks.

6 Call for Help & Waiting on Hold, The Economist, Oct. 25, 2014.

7 Sundsøy, P., Bjelland, J., Iqbal, A., Pentland, A., and de Montjoye, Y. A. (2014). Big Data-Driven Marketing: How Machine Learning Outperforms Marketers' Gut-Feeling. In Social Computing, Behavioral-Cultural Modeling and Prediction, 367-374.

8 This is very similar to how some mobile marketing interfaces work where marketers will specify the criteria and identifying characteristics for the people they want to target with specific messages but would not receive actual numbers. Alternatively, anonymized data could be shared with encrypted identifiers which would be passed back to the operator to trigger outreach.

uncertain and country-specific regulatory landscape for data-sharing especially for cross-border data sharing.

While some forward-looking companies have been sharing limited data with researchers in privacy-conscientious ways, these barriers and challenges are making it unnecessarily hard for carriers to share data for humanitarian purposes.[9] [10] We describe these issues further and offer recommendations moving forward.

PROTECTING THE IDENTITY OF SUBJECTS

Mobile phone metadata made available to researchers should never include names, home addresses, phone numbers, or other obvious identifiers. Indeed, many regulations and data sharing agreements rely heavily on protecting anonymity by focusing on a predefined list of personally-identifiable information that should not be shared. In the United States, for example, the privacy rule issued by the Department of Health and Human Services to protect the privacy of patient health records specifies 18 different types of data about patients that must be removed from datasets for them to be considered de-identified.[11]

However, elimination of specific identifiers is not enough to prevent re-identification. The anonymity of such datasets has been compromised before and research[12] shows that, in mobile phone datasets, knowing as few as four data points—approximate places and times where an individual was when they made a call or send a text—is enough to re-identify 95% of people in a given dataset. In general, there will be very few people who are in the same place at the same time on four different occasions, which creates a unique "signature" for the individual making it easy to isolate them as unique in the dataset. The same research also used unicity to shows that simply anonymized mobile phone datasets provide little anonymity even when coarsened or noised.

This means that removing identifying information makes isolating and identifying a specific person in the dataset only slightly more challenging because that person can be identified using available sources of data that link location with a name or another identifier (e.g. geo-tagged posts on social media, travel schedules, etc.). Wholesale re-identification is more difficult, however, because re-identification of a large fraction of the dataset requires access

9 One example is the open Data for Development contest run by Orange, de Montjoye, Y. A., Smoreda, Z., Trinquart, R., Ziemlicki, C., and Blondel, V. D. (2014). D4D-Senegal: The Second Mobile Phone Data for Development Challenge. *arXiv preprint arXiv:1407.4885*. http://arxiv.org/abs/1407.4885.

10 U.N. Global Pulse (2014) Data Philanthropy: Where Are We Now?, http://www.unglobalpulse.org/data-philanthropy-where-are-we-now.

11 45 C.F.R. 164.514.

12 de Montjoye, Y. A., Hidalgo, C. A., Verleysen, M., and Blondel, V. D. (2013). Unique in the Crowd: The privacy bounds of human mobility. *Nature SRep, 3*. http://www.nature.com/srep/2013/130325/srep01376/full/srep01376.html.

to a full list of people and places they have been, which may not be as easy to acquire. Nevertheless, a determined attacker can still re-identify people using such data. Therefore, removing personally identifiable information is only a first step in most instances and more stringent approaches are required unless trust in the recipient of a dataset is high.

Removing personally identifiable information is only a first step in most instances and more stringent approaches are required unless trust in the recipient of a dataset is high.

Recognizing the limits of an approach to anonymity and re-identification that focuses only on identity information like names or ID numbers, governments have sought to expand protection beyond identity to any information that can be used to identify an individual. In 2007, the federal Office of Management and Budget added to its list of identifiers "any other personal information which is linked or linkable to an individual."[13] In Europe, the Directive 95/46/EC cautions that "account should be taken of all the means likely to be used" to identify an individual,[14] and a thorough recent opinion of EU privacy regulators provided technical guidance on the challenges and risks of re-identification.[15]

The challenge of these broad definitions is that they are open-ended. No existing anonymization methods or protocols can guarantee at 100 percent that mobile phone metadata cannot be re-identified unless the data has been greatly modified or aggregated. Hence, open-ended requirements can be unverifiable and, taken to their logical extreme, so strict as to prohibit any sharing of data even when risk of re-identification is very limited.

We believe this places too much emphasis on a limited risk of re-identification and unclear harm without considering the social benefits of using this data such as better managing outbreaks or informing government response after a disaster.[16] Special consideration should be given to cases where the data will be used for significant public good or to avoid serious harm to people. Furthermore, data sharing should allow for greater levels of disclosure to highly trusted data recipients with strong processes, data security, audit, and access control mechanisms in place. For example, trusted third parties at research universities might warrant access to richer, less anonymized data for research purposes and be relied on not

13 Executive Office of The President, Office of Management & Budget, *Safeguarding Against And Responding to The Loss of Personal Information,* Memorandum M-07-16 (May 22, 2007).

14 European Union, Directive 95/46/EC, Recital 26.

15 Article 29 Data Protection Working Party, *Opinion 05/2014 on Anonymisation Techniques*. 0829/14/EN (April 10, 2014).

16 Bengtsson, L., Lu, X., Thorson, A., Garfield, R., and Von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: a post-earthquake geospatial study in Haiti. *PLoS medicine, 8*(8), e1001083.

We contemplate releasing anonymized data to research teams and NGOs in a form that adds technical difficulty to re-identification, limits the amount of data that would be re-identified, and further limiting the risk of re-identification or abuse with a legal agreement that specifies that only specific purposes and other protocols can be applied to the data.

to try to re-identify individuals or to use the data inappropriately.

For both use cases, we defined data-sharing protocols that would allow for the intended analysis, while protecting privacy. We contemplate releasing anonymized data to research teams and NGOs in a form that adds technical difficulty to re-identification, limits the amount of data that would be re-identified, and further limiting the risk of re-identification or abuse with a legal agreement that specifies that only specific purposes and other protocols can be applied to the data. In our analysis, we focused on a middle ground scenario of relatively open sharing of data with multiple research teams and/or NGOs, with some (but limited) accountability and auditability. We did not consider a fully-public release where a very high level of anonymization would be required, nor a release to a highly trusted third party with strong data protection in place that might allow weakly-anonymized data sharing.

For our first use case, we concluded that a 5 percent sampling of the data on a monthly basis, resampled with new identifiers every month for a year and coarsened temporally and spatially into 12-hour periods (7 a.m. to 7 p.m.) and by regions within countries would be the right balance between utility and privacy.[17] It would adequately show individuals' mobility across regions under study and the number of nights spent in infected regions while providing significant—but not absolute—protection of identity and limiting the amount of data that would be re-identified.

For our second use case, we concluded that the behavioral indicators[18] derived from metadata can be shared with the researchers safely, provided outliers have been removed. Researchers could then use this data to segment the population into specific sub-groups based on traits like calling patterns, mobility, number of contacts, etc. People fitting these criteria could then

17 The back-of-the envelope reasoning goes as: We use a spatial resolution of 17 antennas on average (v = 17) and a temporal resolution of 12 hours (h = 12). This means that with 4 points in a given month, we'd have a ~20% chance (\mathcal{E} = .20) at re-identifying an individual in a given month (resp. \mathcal{E} = .55 with 10 points)(see http://www.nature.com/srep/2013/130325/srep01376/fig_tab/srep01376_F4.html). This means that, to have between 20% to 55% chances of re-identifying an individual, we'd need 4 to 10 points every month meaning 48 to 120 points total for a year. Even in this case, as we use a 5% sampling and we resample every month, an individual has only a 45% chance to be in at least one of the sampled month (1 - 0.95 ^ 12 months).

18 Bandicoot, a python toolbox to extract behavioral indicators from metadata http://bandicoot.mit.edu/.

be contacted by the mobile phone operators through text messages or other communications. Their phone numbers would be known only to the mobile phone operators.

Nevertheless, re-identification could be vital in case of emergencies such as an earthquake.

We also considered cases where specific individuals could be contacted based on criteria applied to the data. To do so would require either (a) including in the dataset pseudonymous—but unique—identifiers that make it possible to connect data showing certain traits (such as a likely exposure to disease based on travel patterns) with specific individuals, or (b) including telephone numbers in the dataset so that researchers and/or NGOs can contact the individuals identified directly. Because it enables re-identification, the former would be a departure from good privacy practices unless the data recipient were highly trusted, and the second would be a clear departure because it disclosed unmodified personally identifiable information.

Nevertheless, re-identification could be vital in case of emergencies such as an earthquake.[19] These alternate use cases illustrate further the need to develop mechanisms for trusted third parties to maintain data under strong controls for use, access, security, and accountability.[20]

More generally, promising computational privacy approaches to make the re-identification of mobile phone metadata harder include sampling the data, making the antenna GPS coordinates less precise through voronoi translation for example,[21] or limiting the longitudinality of the data to cover shorter periods of time. These could go as far as to set up systems or collaborations where researchers could pose questions of the data, but where mobile operators would only share with researchers "answers,"[22] such as behavioral indicators or summary statistics.[23] Each of these alternatives could be employed depending on the use the data is put to, the amount and sensitivity of the data that would be uncovered, how and by whom the data will be governed and housed, and the attendant risks of harm.

19 For a discussion of the use of mobile data to direct aid delivery in the 2010 Haiti earthquake see Bengtsson, L., Lu, X., Thorson, A., Garfield, R., and Von Schreeb, J. (2011).

20 We assume here that the mobile operator does not have explicit permission from the data subject to disclose their information. If users were to opt-in to sharing this would then become permissible.

21 https://github.com/yvesalexandre/privacy-tools/.

22 de Montjoye, Y. A., Shmueli, E., Wang, S. S., and Pentland, A. S. (2014). openPDS: Protecting the Privacy of Metadata through SafeAnswers. *PloS One, 9*(7), e98790. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0098790; and de Montjoye, Y. A., Wang, S. S., Pentland, A. (2012). On the Trusted Use of Large-Scale Personal Data. IEEE Data Eng. Bull., 35(4), 5-8.

23 While promising, these solutions are not yet ready for prime-time. Standardized software to process call detail records along with testing and reporting tools are still under development while the use of online systems allowing researchers to ask questions that would be run against the data and only receive answers would imply architectures investments from mobile phone operators.

First, legal uncertainty complicates the design of data-sharing protocols. Indeed, even in countries that have had laws and regulatory agencies in place for some time, the relevant rules have not developed in enough detail to address an issue that is often uncertain even in the most developed legal systems.

ENGAGING GOVERNMENT SUPPORT

The second challenge we identified to humanitarian use of mobile phone metadata is an uncertain and country-specific regulatory landscape for data-sharing. Our study focused on Africa, where data privacy regulation has been evolving along two lines. The Francophone countries—mostly located in West Africa, where current exposure to Ebola is greatest—have tended to adopt privacy frameworks modeled on the 1995 European Privacy Directive and supervised by national data protection authorities. Meanwhile English-speaking countries with common law systems either have not yet adopted comprehensive privacy laws, or have adopted country-specific laws.

This landscape presents a number of barriers to humanitarian use of mobile phone metadata. First, legal uncertainty complicates the design of data-sharing protocols. Indeed, even in countries that have had laws and regulatory agencies in place for some time, the relevant rules have not developed in enough detail to address an issue that is often uncertain even in the most developed legal systems.

Second, as discussed above, questions about the validity of most methods of de-identification persist particularly in countries that use open-ended definitions of anonymization such as the EU one. There exist no widely accepted data-sharing standards to help various actors achieve a rational privacy/utility tradeoff in using mobile phone metadata.

Third, regardless of legal systems, compatible data-sharing protocols—including data de-identification—have to be designed and validated on a country by country basis. For example, data-sharing protocols have to be compatible, which includes having both the phone number and the mobile phone identifier[24] hashed with the same function and salt[25] to allow for mobile phones to be followed across border, even if the user changes SIM cards. These issues make cross-border data sharing or intra-regional tracking of population flows particularly complex and costly. Yet such cross-country sharing is essential in the fight against diseases such as malaria or the current Ebola outbreak.[26]

24 IMEI or International Mobile Station Equipment, a unique number that identifies a mobile phone on the network.

25 One potentially interesting solution here would be to rely on multiple hash functions that can be nested.

26 BBC News, Ebola: Can big data analytics help contain its spread? http://www.bbc.com/news/business-29617831.

Fourth, our second use case contemplated that, in general, only behavioral indicators derived from carriers' metadata would be shared with researchers but that, in specific and limited circumstances where these indicators show an individual would benefit from intervention, the identity could be used to enable remote intervention such as targeted texts sent by the operator, or identification through mechanisms that carefully control the release and use of this information.

In the absence of explicit consent from users to such disclosure and use of data from their mobile phones, these forms of re-identification of data subjects presents obvious privacy challenges and may come into conflict with most privacy legal regimes absent specific exceptions. The EU Privacy Directive provides that data processing must have a lawful basis, but that such a basis may be "to protect the vital interests of the data subject," or "in the public interest, or in the exercise of official authority, and recognizes "public health" as such a public interest."[27] Thus, it will take the support of national governments, their health ministries, and their data protection authorities to enable use of data especially in such exigent situations, but also for a range of humanitarian applications.[28]

There is a clear need for companies, NGOs, researchers, privacy experts, and governments to agree on a set of best practices for new privacy-conscientious metadata sharing models in different development use cases—a wider and higher-level discussion of the kind our MIT working group conducted.

CONCLUSION: ROADMAPS NEEDED

These privacy challenges and regulatory barriers are making humanitarian data-sharing much harder than it should be for mobile phone operators and are significantly limiting greater use of mobile phone metadata in development or aid programs and in research areas like computational social science, development economics, and public health.

To realize the potential of this data for social good, we recommend the following:

27 European Union, Directive 95/46/EC, Article 7 (d), (e). An update to this legislation, the Privacy Regulation proposed by the European Commission in 2012, http://ec.europa.eu/justice/data-protection/document/review2012/com_2012_11_en.pdf, also included an exception from certain requirements for "scientific, historical, statistical, and scientific research purposes,' but this was removed from legislation as passed by the European Parliament. http://www.europarl.europa.eu/meetdocs/2009_2014/documents/libe/pr/922/922387/922387en.pdf.

28 Under the World Health Organization's International Health Regulations, the WHO and member states undertake to conduct "surveillance" for public health purposes and member states are permitted to "disclose and process personal data where essential for purposes of assessing and managing public health risks." WHO, Fifty-eighth World Health Assembly Resolution WHA58.3: Revision of the International Health Regulations, Articles 1 (definition of surveillance), 5.4, and 45 . 2005, http://www.who.int/ipcs/publications/wha/ihr_resolution.pdf.

1. There is a clear need for companies, NGOs, researchers, privacy experts, and governments to agree on a set of best practices for new privacy-conscientious metadata sharing models in different development use cases—a wider and higher-level discussion of the kind our MIT working group conducted. These best practices would help carriers and policymakers strike the right balance between privacy and utility in the use of metadata and could be instantiated by data-protection agencies, institutional review boards, and in data protection laws and policies. This would make it easier and less risky for carriers to support humanitarian and research uses of this data, and for researchers and NGOs to use these metadata appropriately.

Best practices should accept that there are no perfect ways to de-identify data—and probably will never be. There will always be some risk that must be balanced against the public good that can be achieved.

2. Such best practices should accept that there are no perfect ways to de-identify data—and probably will never be.[29] There will always be some risk that must be balanced against the public good that can be achieved. While much more research is needed in computational privacy, widespread adoption of existing techniques as standards could enable this trend of sharing data in a privacy-conscientious way.

3. Standards and practices as well as legal regulation also need to address and incorporate trust mechanisms for humanitarian sharing of data in a more nuanced way. Protection of individual privacy includes not only protection against re-identification, but also data security and protection against unwanted uses of data. Risk of re-identification is not a purely theoretical concept nor is it binary and it should be assessed vis-à-vis the level of trust placed in the data recipient and the strength of their systems and processes. Tracking of migration patterns or analysis of behavior patterns may offer enormous benefits for disease prevention and treatment, but it is possible to envision more malignant uses by actors ranging from disgruntled employees of the data recipient to authoritarian governments. The recognition of trusted third-parties and systems to manage datasets, enable detailed audits, and control the use of data could enable greater sharing of these data among multiple parties while providing a barrier against risks.

29 No silver bullet: De-identification still doesn't work https://freedom-to-tinker.com/blog/randomwalker/no-silver-bullet-de-identification-still-doesnt-work/.

Clear and consistent rules will help but only provided they take a pragmatic and privacy-conscientious approach to anonymization, cross-border transfers, and novel uses that enable public good uses of data and allow for public health emergencies and other valuable research.

There is a need for governments to focus on adopting laws and rules that simplify the collection and use of mobile phone metadata for research and public good purposes. Governments should also seek to harmonize laws on the sharing of metadata with common identifiers across national borders. The African Union took what could be a step in this direction last June, when it approved the African Convention on Cyber Security and Personal Data Protection seeking to advance Africa's digital agenda and harmonize rules among African nations.[30] The treaty, which will not take effect until adopted by 15 member states, commits members to adopting a legal framework that follows the template of the European Privacy Directive. Clear and consistent rules will help but only provided they take a pragmatic and privacy-conscientious approach to anonymization, cross-border transfers, and novel uses that enable public good uses of data and allow for public health emergencies and other valuable research.

Research based on mobile phone data, computational privacy, and data protection rules all may seem secondary when confronted by the challenges of poverty, disease, and basic economic growth. But they are on the critical path to realizing the great potential of information technology to help address these critical problems.

30 Draft African Union Convention On The Establishment Of A Credible Legal Framework For Cyber Security In Africa http://www.au.int/en/cyberlegislation.

Governance Studies
The Brookings Institution
1775 Massachusetts Ave., NW
Washington, DC 20036
Tel: 202.797.6090
Fax: 202.797.6144
brookings.edu/governance.aspx

Editor
Christine Jacobs
Beth Stone

Production & Layout
Beth Stone

EMAIL YOUR COMMENTS TO GSCOMMENTS@BROOKINGS.EDU