SUMMARY
The goal of this project was to determine how insights from mobile data may be used to maximum effect in support of policy planning and crisis response with minimal risk to privacy. The project aimed to determine the impact that aggregating mobile data to protect privacy has upon the utility of the data for (i) transportation planning and (ii) pandemic control and prevention. Utility of each data set was evaluated by surveying transportation experts and epidemiologists; re-identification risk for each data set was also assessed. Risk of re-identification was subsequently considered together with data utility to determine which level of aggregation is the minimum required to adequately protect individual privacy while preserving its value for policy planning and crisis response. Results of the analysis indicate that the relationship between privacy risk and utility is complex and is highly dependent on the context and purpose of use. Nevertheless, there is a risk of re-identification in applying mobile data for public good, and a better understanding is needed of the critical thresholds for acceptable risk. This project lays the groundwork for development of evidence-based data standards and policy frameworks to ensure proportionality between the risk of harm resulting from misuse of mobile data and the risk of harm resulting from a failure to use it.

BACKGROUND
Mobile network operators, or MNOs, perform real-time analysis of big data, including mobile phone usage records, to gain consumer insights and optimize network performance. In recent years, a few MNOs have begun to explore public good applications of their data by making it available to researchers, and there is now solid evidence that mobile data has high potential to be useful in tackling development and humanitarian challenges (Pastor-Escudero 2014, Decupyer 2014).

However, significant challenges exist in re-using mobile phone data on account of its sensitivity from a privacy standpoint. Media coverage of indiscriminate mass collection of mobile metadata for national security purposes as well as use of that data by MNOs for marketing purposes has created a general sense of unease amongst the public about the visibility and right to determine how their data is used.

There is currently no international consensus on the best approach to anonymising mobile phone data. Furthermore, there is growing evidence that geospatial data, such as mobile phone location data, is more difficult to anonymize than was previously believed (de Montjoye 2012, 2015). Without proper anonymization, re-identification may lead to harms to individuals and vulnerable groups, as well as to mobile network operators. This lack of understanding of how to mitigate the risks of using mobile data, combined with little public awareness of the opportunity cost of not leveraging insights derived from it, has led to pervasive inertia. To date, the wealth of scientific studies has not led to operational examples of real-time insights from mobile data being used for measurable social impact.

Mindful of risks to both subscriber privacy and their core business, those MNOs willing to allow research into their data are increasingly moving away from providing access to disaggregated data sets and are instead electing to aggregate it to such an extent that all of the individual details are blurred or irreversibly removed. While it is generally true that higher degrees of aggregation are more likely to reduce the value of data, what is not known is the degree of aggregation beyond which mobile data is no longer useful for various development and humanitarian purposes.

Given the potential benefits at stake, the global community must act swiftly to develop trusted frameworks that protect fundamental rights and values when using mobile data, while also ensuring that these data may be used in the public interest. The principle of proportionate use – that the use of data should be necessary and in proportion to the benefit of its use (UNHCR, 2014) – applies no less to big data than it does to data traditionally used for development or humanitarian work. Further, it is critical to ensure disclosure of only data with the minimum level of detail required for the purpose at hand.

A prescient example of this dilemma was the 2014 Ebola outbreak in West Africa: the research community proposed that sharing anonymised mobile data, even in an aggregated form, could have supported the response effort by helping to understand disease spread through mobility patterns (Wesolowski 2014) - but it had not happened. This health crisis exemplified a core dilemma in data partnerships: should data privacy requirements be relaxed as the vulnerability of potential beneficiaries increases? If so, to what extent and under which circumstances? Further, if data aggregation is used, how does decreasing spatial or temporal resolution affect data utility and can its effect vary depending on purpose, context and necessity?

1 For example, in some countries an individual or group could be subject to discrimination based on sexual, racial, religious beliefs or medical history. From the perspective of private companies, unfair competition and financial loss could result from if their strategic business information is compromised.

2 These include the right to privacy as recognized by the Universal Declaration on Human Rights and the International Covenant on Civil and Political Rights.
In this context, Global Pulse worked with MIT Connection Science to explore how the utility of Call Detail Records (CDRs) vary as they become more aggregated and so more privacy protecting. As part of this work, within the context of the data minimization, Global Pulse sought to explore to what extent all of the original details in data are necessary so that it may be shared and used minimally.

Measurement of re-identification risk and evaluation of data utility, were viewed through the privacy principle of necessity and proportionality to the risks and benefits of data use in the context of development and humanitarian response.

APPROACH

Data utility (measuring how useful a dataset is) and risk (probability that an individual may be re-identified) were both assessed in two case scenarios. The aim was to understand how data minimization and the principle of proportionate use are best applied for development and humanitarian causes.

This assessment was accomplished by consulting with domain experts who completed a quantitative questionnaire assessing the utility of mobile data, under various levels of aggregation, for their work. This was compared with a measure of re-identification risk in the data under the same aggregation.

Understanding Context and Purpose of data use

In the recent Data for Development (D4D) challenge using mobile data for development purposes in Senegal, several different datasets were

| Re-identification risk was quantified using the metric of unicity (de Montejoye). Assuming that an adversary knows where and when someone made one call, what is the probability that the adversary could use that knowledge to look through the anonymised records of all calls, locate that person’s records in the CDR data set and re-identify them? In order to re-identify someone, that observation must be unique: if there is more than one person who made a call at the same time and place, re-identification is impossible without more observations to look for.

Re-identification becomes harder when data is aggregated; two calls, one made at 9.14AM and a second made at 9.45AM are indistinguishable under hourly aggregation. This risk is reported as a percentage; the chance that a random user could potentially be re-identified using four known observations. Unicity, however, is one of a number of possible measures of the risk of re-identification, which must take context into consideration.

available with varying longitudinality and resolution. The preferred level of resolution was strongly tied to the area of application (transportation, energy, health etc.) (Decordes, 2015).

In this project, two different applications suitable to urban areas were considered: (i) planning of public transport provisions and (ii) control of Dengue fever outbreak. By considering two distinct contexts (transportation – less urgent; pandemics – more urgent), we were able to begin to understand the extent to which the limits of reasonable data usage should be determined by context and purpose.

Measuring Utility

To determine utility, a series of interviews with the domain experts were conducted. A mobile phone dataset in the form of anonymized CDRs (name, phone number, date of birth or other personal identifiers are removed) from a large middle-income country was visualized in a series of maps to reveal both population density and mobility at varying levels of spatial and temporal resolution.

Map visualizations of data sets were produced and used during the interviews with development practitioners and epidemiologists to gather perspectives on how and why mobile phone data becomes more or less useful at different levels of resolution.

Each expert was asked to answer how valuable the data set, aggregated by combinations of 3 spatial and 4 temporal levels of resolution (as shown below) would be, if applied to their corresponding area of expertise. This gave 12 distinct combinations of spatial & temporal aggregation. Their answers were numbers ranging from 0 (no utility) to 10 (high utility). These quantitative questions were complemented by qualitative questions about the potential use of mobile data in order to help interpret their numerical rankings. It is worth mentioning the need to explain the nature of CDR data and its potential to the experts before their judgments.

The images above is one of the maps shown to the experts showing population density estimated with mobile data at ZIP code resolution.

Measuring Re-identification risk using Unicity metric

Our principal objective was to compare the values of risk re-identification at these levels of aggregation with the subjective estimates of utility provided by domain experts.

Re-identification risk at each level of aggregation is displayed below (see call out box on Re-identification for explanation)

<table>
<thead>
<tr>
<th>Zip</th>
<th>1h</th>
<th>6h</th>
<th>12h</th>
<th>1w</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>97%</td>
<td>79%</td>
<td>67%</td>
<td>14.46%</td>
</tr>
<tr>
<td>District</td>
<td>82%</td>
<td>42%</td>
<td>28%</td>
<td>0.17%</td>
</tr>
<tr>
<td>Municipality</td>
<td>55%</td>
<td>14%</td>
<td>6%</td>
<td>&lt;0.17%</td>
</tr>
</tbody>
</table>

3 Due to pathological degenerate behaviour of vastly aggregated records, the calculation of unicity under these extremes becomes infeasibly intensive. Therefore we can only infer its upper limit, however it is likely to be much lower.

www.unglobalpulse.org • info@unglobalpulse.org • 2015
RESULTS & CONCLUSIONS

- **Purpose of data use matters**

There is a qualitative difference between transportation and pandemics scenarios; the two groups of experts had differing models of value in a mobile data set. For the transportation planning group, more detail was always better. However the experts in disease control reported that the value of the data depended on many contextual factors such as the exact disease being studied, its transmission vector and existing data sources that it could be combined with.

- **Utility does not always scale linearly with risk**

We found that from our experts’ perspective not every treatment to make data less identifiable makes the data less useful. Conversely, treating data with less strict privacy preserving techniques doesn’t necessarily confer an increase in operational utility. Therefore, to ensure necessity and proportionality of the purpose, the aggregation treatment chosen must be considered relative to the context, harms and benefits.

In the figures below, for each combination of spatial and temporal aggregation, the y-axis indicates the level of re-identification risk and the x-axis indicates the average level of utility reported by respondents.

The integration of passive data in human decision making and regulatory frameworks is a new frontier, especially within complex scenarios such as development and humanitarian action. For example, the Resolution on Data Protection and Major Natural Disasters adopted by the International Conference of Data Protection and Privacy Commissioners in 2011 suggested that “extraordinary processing” of personal information may be warranted to ensure that in a major national disaster, data is available to respond effectively to the emergency to protect the vital interests of individuals. While the instrument is not legally binding, it recognizes the need for data protection laws to contain sufficient flexibility to ensure that in disaster situations data is available to respond to the emergency to protect the vital interests of individuals. Thus, policy frameworks should recognise and adopt a flexible approach for the context and purpose dependent use of aggregated mobile data to encourage analysis for humanitarian and development causes.

This work is presented as a feasibility study for a longer-term effort and further methodological work examining other scenarios such as disaster response and poverty mapping, other types of big behavioral data, and other complementary definitions of privacy risk.

### TRANSPORTATION SCENARIO

![Graph](image)

### PANDEMSICS SCENARIO

![Graph](image)

Geography, landscape or cultural context may increase the possibility of re-identification. Therefore it is of importance to not only measure the risk for re-identification but to also consider context and influence factors for each specific case. Such an assessment should be followed by further exploration of harms, also dependent on context and surrounding circumstances.

IMPLICATIONS & RECOMMENDATIONS

This project has demonstrated that mobile data aggregation is a viable mechanism to protect individual data privacy, whilst also making use of the potential of valuable digital footprints in development programmes and humanitarian action.

In general, mobile data was found to become less useful with increasing aggregation. This trend, however, varied with the nature of the potential use and is not a simple tradeoff.

It is recommended that further research is conducted by repeating and scaling this project methodology to consider a wider range of scenarios with more specific details and survey more experts, as well as to consider other anonymization methods.

It is recommended that data driven decision-making should consider circumstantial factors such as cultural, religious and social norms as well as the political and legal landscape to properly assess the likelihood of risks and magnitude of harms if the data is re-identified. Such an assessment should be then carefully viewed and analyzed through the principle of necessity of data use.

We recommend the development of a flexible data privacy framework that would provide for a detailed guidance on risk mitigation and aggregation for using mobile datasets and other similar records of human behaviour in particular development and humanitarian contexts.

Implementation of operational decision support tools that incorporate analysis of mobile data should be accompanied by capacity building within UN agencies and other relevant public sector institutions, as well as by broader investments in data literacy.

REFERENCES


14668-English.pdf

AHRC27.37_en.pdf


