**Netcom**
Réseaux, communication et territoires

**30-3/4 | 2016**
**Réseaux sociaux et territoires**

# On the research for big data uses for public good purposes

## Opportunities and challenges

**Adeline Decuyper**

**Publisher**
Netcom Association

**Electronic version**
URL: http://netcom.revues.org/2556
ISSN: 2431-210X

**Printed version**
Date of publication: 16 December 2016
Number of pages: 305-314
ISSN: 0987-6014

**Electronic reference**
Adeline Decuyper, « On the research for big data uses for public good purposes », *Netcom* [Online], 30-3/4 | 2016, Online since 22 March 2017, connection on 29 March 2017. URL : http://netcom.revues.org/2556

*Networks and Communication Studies,*
*NETCOM*, vol. 30 (2016), n° 3-4
pp. 305-314

# *NOTE SCIENTIFIQUE*

## ON THE RESEARCH FOR BIG DATA USES FOR PUBLIC GOOD PURPOSES: OPPORTUNITIES AND CHALLENGES

**DECUYPER ADELINE**[1]

***Résumé*** *– Ces dernières années, beaucoup de nos habitudes ont changé, suite à l'arrivée des big data, et de l'offre de services donnant des informations mises à jour en temps réel, basées sur les données produites par les utilisateurs. Cette révolution a également apporté de nouvelles opportunités pour les compagnies privées, mais aussi pour les gouvernements et chercheurs, les applications potentielles des analyses de ces données étant nombreuses. Dans cet article, nous présentons quelques avancées faites dans ce domaine, et montrons quelques opportunités d'applications pour le bien public, telles que l'aide aux actions humanitaires et la réponse aux situations extrèmes, rendues possibles par ces grandes masses de données, montrant également que dans certains cas, l'analyse de données peut être utilisée pour sauver des vies. Enfin, à côté des nouvelles opportunités, se sont aussi présentés de nouveaux challenges, dont l'évaluation de la robustesse et représentativité d'une base de donnée, ou bien les dangers pour la vie privée, qui seront l'objet d'une dernière section de cet article.*

***Mots-clés*** *- big data, données téléphoniques, aides au développement.*

***Abstract*** *– In the last few years, many of our habits have changed due to the rise of big data and of all the services providing information updated in real time, based on data produced by the crowd of users. This big data revolution has also brought new opportunities for private companies, but also for governments and researchers, as the potential applications are numerous. In this article, we review some of the advances made in this field, and present opportunities offered by big data for public good, in fields*

[1] Postdoctoral researcher at Center for Operations Research and Econometrics, at Université catholique de Louvain, Voie du Roman Pays, 34; B-1348 Louvain-La-Neuve. Tel. +3210479431 – e-mail: adeline.decuyper@uclouvain.be

*such as help for humanitarian action or natural disasters response, indicating that in some cases, data may be used to save lives. Finally, aside from the opportunities, the mass production and use of data has also brought many challenges, such as evaluating the representativity of the data or handling threats to the privacy of users, that we will discuss in the last section of this article.*

***Keywords*** *- big data, data for development, mobile phone data.*

## INTRODUCTION

The last few years have seen the rise of what we now call the "big data revolution", that is rapidly changing the way we move and interact, the way we think and make decisions, or simply the way we live. Further than providing new means of communication between friends or colleagues, smartphone applications now use the massive amount of information collected to put people who have common interests in contact with each other thus creating new relationships between otherwise strangers. Other services that smartphones provide now include advice on which restaurant to chose or which road to take to avoid traffic jams and an endless list of new services providing information updated in real time. It is estimated that all these new applications, together with GPS traces, climate sensors, social media, satellite images and other sources of data produce more than 2.5 billion gigabytes of data every single day (*IBM What is big data?*, 2015).

With this revolution, came new challenges for private companies, governments and researchers alike, brought about by the four V's that are often used to define big data: Velocity, Volume, Veracity and Variety. The first two V's trigger challenges regarding the efficiency of storage and computing power that are needed to handle ever larger amounts of data. The two last V's generally trigger another kind of challenge, namely that of analyzing and using the data to extract sound and significant information from automatically collected data. Indeed, while census information is gathered through surveys, whose data are collected specifically for the purpose of being statistically relevant, big data are collected automatically, and may not have the same statistical qualities. To the four V's that characterize the data itself, two more V's are sometimes associated, brought by the analyses that can be made on those data: Value and Visualization. The data themselves may only be a series of numbers and letters seemingly worthless, but add to that the potential analyses and visualizations that can bring many useful insights, and the data become invaluable.

Automatically collected datasets are produced by many different sources and come in various shapes. Among those, some types of data collected automatically can reveal social interactions, and the networks of relationships between people. These networks can be observed for example through online interactions or phone calls, which have been widely studied for about 15 years (Blondel et al., 2015). Indeed, mobile phone datasets were among the first large datasets available, as the wide adoption of mobile phones already started in the late 90s. Each time a call is made, the provider records

some metadata for billing purposes. The information recorded in Call Detail Records (CDRs) usually contains the hashed ID's of the caller and of the callee, the ID of the cell tower that handled the call, a date and time and sometimes the call duration. These newly available datasets then allowed to observe and model the dynamics of human communication and mobility at a fine grained resolution that was, up to a few years before, impossible to achieve.

In the following paragraphs, we review some of the opportunities and challenges that arose from the use of these large datasets. We first briefly present the advances in the analysis of mobile phone datasets, made in the last fifteen years, mostly with the aim of understanding human and social behavior. Indeed, analyses of mobile phone data have revealed trends of human behavior that were until then impossible to observe through classic census or self-reported data. As the penetration of mobile phones has dramatically increased in the developing world, opportunities for using their data have risen, showing the actual impact that information extracted from those large datasets can have. This will be the topic of the second paragraph, where we present a few opportunities for public good, such as help towards development, or help for a better response after natural disasters. Finally, we discuss a few of the new challenges, that arose from the use of big data, such as representativity of the data, but also the dangers for privacy and the compromises to be found, as these have become major concerns in the last few years.

## MOBILE PHONE DATA REVEAL PATTERNS OF HUMAN BEHAVIOR

The past fifteen years have allowed to use mobile phones as a lens through which social networks and human behavior can be observed. In this framework, the analysis of these data has shown that mobile phone networks are far from random, and that people arrange themselves in communities, or groups of tightly connected people with many strong links inside the group, and fewer to the rest of the population (Ahn et al., 2010, Blondel et al., 2008, Ratti et al., 2010, Tibély et al., 2011). Using additional information on the spatial footprint of mobile phone communications, research has shown that instead of reducing the whole world to a small village, distance still plays a role in the social contacts, as the number and duration of contacts decrease quickly as the distance between two people increases (Lambiotte et al., 2008). Coupling the analysis of communities of people with spatial information has shown where lie the boundaries of social groups, indicating in several cases a strong influence of a language, or administrative border (Blondel et al., 2010, 2011; Bucicovschi et al., 2013).

Using the time component of the information contained in mobile phone call detail records also shed some light on specific trends of human behavior that also present some regularity. Looking at the network with a dynamic point of view has shown that in many cases, the appearance or decay of relationships between people in

the network can be predicted when sufficient information is available on past contacts (Raeder et al., 2011, Miritello, 2013). Furthermore, despite a turnover in links and in the network surrounding individuals, the structure of the network and the distribution of link weights around a person remains very similar through time, representing a type of social signature of a person (Saramäki et al., 2014).

Using both space and time information contained in mobile phone traces allows to observe the mobility traces of users as they use their phone from various places at various times. Human trajectories show a high degree of spatial regularity, placing them very far from random motion models (González et al., 2008). These traces reveal for example which places people visit at different times of the day and of the week, allowing to detect which places are typically work places and which are typically residential or leisure meeting places (Reades et al., 2007). Furthermore, characterizing the mobility habits of users reveal that most people frequently visit only a very small number of preferred locations.

All these studies and use cases were explored using mobile phone data, but could easily be extended to other types of communication data, or mobility traces of individual users. The ultimate goal of this field of study is to understand and model how the populations behave, and how they move or travel. In turn, knowing how the population usually behaves, mobility patterns can be used as a tool to detect anomalous behavior, such as large gatherings, an earthquake or a blackout (Bagrow et al., 2011). Besides, analyzing the mobility habits of a population in an urban environment also allows to better plan urban infrastructures and public transportation networks so as to meet the demand as efficiently as possible. Studies on the geographical extent of social communities, on the other hand, allow to better understand how the population interacts and may help spot segregated groups or places, and determine where better mixing policies may be needed.

## DATA FOR DEVELOPMENT

In the last few years, as the mobile phone penetration increased in developing countries, new opportunities arose for using the mobile phone data, and all the insights gained from previous research, for practical applications to help countries towards development. One effort in this direction was led by Orange in 2013, who shared a dataset from Côte d'Ivoire with more than 150 research teams accross the world for projects ranging from mitigating epidemics of infectious diseases to enhancing the mobility of people within cities (Blondel et al., 2012). While some of the information that can be extracted from mobile phone data is already known in most developed countries, very basic information such as the density of population can be very valuable in developing countries where census data is often unavailable or several years old.

One example of such practical application is in the field of epidemics prediction and prevention. Knowledge on how an epidemic can spread accross a country is

essential for planning an optimal response and mitigation strategy. Estimating and predicting the propagation of infectious diseases has therefore been a widely studied research topic in the last century. Research on characterizing the mobility traces of a population through data analysis helps in that direction, as diseases will spread with the movement of the infected population between different places. Therefore, a large body of research has been focused on using mobility habits and people's movements observed through mobile phone traces and communications, to model the spread of epicemics, and in turn, to evaluate the impact of containment strategies for mitigating epidemics (Kafsi et al., 2013). Among other insights, research has shown that in some cases, launching an information campaign may have a stronger impact on the epidemic than trying to put several cities in quarantine (Lima et al., 2013).

At a more practical level, building up on the insights gained from previous research, the Flowminder Foundation applies advanced mobility models based on mobile phone data to predict the spread of infectious diseases with applications in several countries (Flowminder Foundation). Among many other examples, they provide malaria elimination support in Namibia, where mobile phone data helped build accurate population density maps and population movement patterns, that in turn, helped local decision makers to target the best places for preventive interventions.

Another tangible example of application is the optimization of the network of public transport, based on an estimation of the demand from commuters. Based on the Orange D4D dataset, a research team studied the mobility in Abidjan, and showed that by modifying only a little a few bus routes, they could reduce by 10% the average commuting time of the whole population of the city (Berlingerio et al., 2013). Among other applications, data can also be used to locate places where humanitarian help is needed for example in the case of food crises, or to map poverty and get updated information at a fine grained spatial scale, and therefore adapt the organization to send humanitarian help to places where the situation is most critical (Blumenstock et al., 2015; Decuyper et al., 2014).

In some cases, very practical information gained from the analysis of large datasets may even be used to save lives. With this idea in mind, a project was launched in Japan after the 2011 earthquake and tsunami that killed many, in order to discover what opportunities could be offered by the use of big data in the case of natural disasters, and try to learn lessons, in order to plan a better response for the next earthquakes or tsunamis in the same region. Using several datasets containing Twitter data, mobile phone traces, GPS traces and more, teams of data scientists analyzed the movement and activity of people before and after the earthquake, until the tsunami struck the coast. Among many insights, the analyses revealed that a lot of people died trying to escape from the coast, as they were stuck in traffic jams to get out of coastal cities when the tsunami struck the coast. In some places, the analysis pointed the too narrow bridges as bottlenecks of the traffic jams, suggesting that enlarging some of the bridges may save lives in the future. Using similar approaches, the Flowminder Foundation also provides population displacement information that is useful for

humanitarian agencies to deliver help where and when it is most needed after natural disasters, such as the earthquake that struck Nepal in 2015 or the more recent hurricane Matthew that struck Haïti in 2016.

These very practical examples of insight gained from the analysis of data produced by the activity of the population are examples to follow for future research, to learn from information extracted from data, and allow a better response to events in extreme situations. At a smaller scale, these methodologies can also be applied to other situations, such as detecting the number of people at a large gathering in order to allow a smooth organization. Aside from these, there are many applications of using data for public good, such as the extraction of socio-economic information from the study of social networks, or insights regarding the mobility of commuters. This information may, in turn, be used by policy makers for public infrastructure and transport planning. Be it for transport planning, natural disasters response or epidemics prevention, it has been shown through these various projects that big data can be of use for public good, sometimes up to the point of saving lives, and future research should therefore focus on transforming these first projects into operationalizable insights delivered to the organizations that can put them into practice.

## BIG DATA, BIG CHALLENGES

The use of big datasets triggered a large number of challenges to be handled. Aside from the technical difficulties of handling and storing ever larger amounts of data, one may raise the question of the representativity of the data. When using data for social analysis that is collected automatically, knowing whether the data are representative of the whole population is a difficult question. While the geographical coverage can sometimes be estimated, it is almost impossible to evaluate how representative the sampling of the population is regarding socio-economic status or gender and age in the absence of additional specific information. In short, using a dataset for another purpose than the one it was produced for initially is, most of the time, a difficult task. Furthermore, through the automatic process of collection, the data gathered often contain noise. For example, in a mobile phone database, the error calls or commercial calls will appear, and a Twitter dataset may contain companies as well as individual users, and it is often difficult to design a method for cleaning a given dataset of the noise, while keeping all useful items of information. The failure of Google Flu Trends is one example where the lack of quality of the data and of robustness of the method led to predictions that were very far from what actually happened.

On the other side, the production and availability of new datasets and personal data has outpaced, in the last years, the production of recommendations and laws that regulate the use and share of these data containing a lot of personal and private information.

However, the incentives for sharing potentially sensitive data with a third party are numerous, whether it be for business, concerted research efforts, or for the sake of science. The metadata gathered by mobile phone providers, or by other service providers (smartphone applications, GPS devices, …) are always subject to anonymization and aggregation procedures before being shared with third parties (such as researchers). The analyses made on those datasets are also subject to a non disclosure (or confidentiality) agreement. Yet, previous studies have shown that a dataset that is anonymized only by replacing actual names with user numbers is far from providing a secure situation in which re-identification of a person is impossible. Indeed, it was shown that with as little information as four points in space and time where a specific user was, it may be possible to re-identify them in a mobile phone database (de Montjoye et al., 2013). Of course, the number of points and pieces of information needed can vary depending on the size and precision of the database. In most cases, a dataset could be shared safely, supposing that an attacker wanting to re-identify a person doesn't have any external information. However, this assumption is far from being realistic in practice, and when a dataset is shared, the danger always exists that crossing another database, or other publicly available information can allow the re-identification of specific users. Furthermore, one has to take into account that external information may be rendered public in the future, that could jeopardize the safety of the dataset. Therefore, care has to be taken to share a dataset with trusted third parties, and to robustly anonymize data before sharing them.

## CONCLUSION

In the last few years, we've witnessed big changes in our habits due to the rise of big data. These automatically collected massive datasets have been used for marketing purposes by private businesses as it was their first aim, but other potential uses of these data have also been uncovered through research oriented towards sociology or human behavior modeling. Potential uses have thus been uncovered that could be of use for public good, such as the modeling of mobility habits of inhabitants of a city so as to optimize the public transport system to better correspond to the demand, or uses oriented towards development, for example in African countries, where census information is often scarce. We have seen new opportunities in the analysis of these datasets to serve public good up to the point of saving lives, a goal that could be achieved in the future.

On the other hand, as a consequence of the massive production of data, new challenges have arisen: mass data storage and increasingly high computing power were needed. Furthermore, in the statistical analyses of these data, questions regarding the representativity of the data and statistical significance were also raised, many of which are still under study. Overall, the challenge that got the most media coverage is probably that of privacy protection: sharing those datasets with third parties induce a breach in the privacy of people, regarding the sensitive and personal character of the information contained in these datasets. This last challenge is also a currently ongoing question

ruling the extent to which a dataset may be shared with third parties, researchers or public instances. Keeping in mind that some uses of these datasets may be beneficial for public good, bluntly forbidding any sharing may not be the best scenario to pursue. However, since the information contained in those data are sometimes very sensitive, a compromise has to be found to anonymize a dataset in a robust way, all the while keeping enough valuable information for the analysis to be useful and insightful. Future research in this area should therefore focus on finding new ways of robustly anonymizing a dataset, to allow data sharing for research and for projects for a greater good.

## REFERENCES

AHN Y., BAGROW J., LEHMANN S. (2010), "Link communities reveal multiscale complexity in networks", *Nature*, vol. 466, n° 7307, pp. 761-764.

BAGROW J., WANG D., BARABÁSI A.-L. (2011), "Collective response of human populations to large-scale emergencies", *PLoS One*, vol. 6, v°3, e17680.

BERLINGERIO M., CALABRESE F., DI LORENZO G., NAIR R., PINELLI F., SBODIO M., (2013), "AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data", in: BLOCKEEL H., KERSTING K., NIJSSEN S., ŽELEZNÝ F. (Eds.), *Machine Learning and Knowledge Discovery in Databases*, *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, pp. 663–666.

BLONDEL V.D., DECUYPER A., KRINGS G. (2015), *A survey of results on mobile phone datasets analysis*, EPJ Data Science 4.

BLONDEL V.D., DEVILLE P., MORLOT F., SMOREDA Z., VAN DOOREN P., ZIEMLICKI C. (2011), "Voice on the border: do cellphones redraw the maps?", *Paris Tech Review*.

BLONDEL V.D., ESCH M., CHAN C., CLÉROT F., DEVILLE P., HUENS E., MORLOT F., SMOREDA Z., ZIEMLICKI C. (2012), *Data for development: the D4D challenge on mobile phone data*, arXiv preprint arXiv:1210.0137.

BLONDEL V.D., GUILLAUME J.L., LAMBIOTTE R., MECH E.L.J.S. (2008), *Fast unfolding of communities in large networks*, J. Stat. Mech P10008.

BLONDEL V., KRINGS G., THOMAS I., 2010, "Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone", *Brussels Studies*, 42.

BLUMENSTOCK J., CADAMURO G., ON R. (2015), "Predicting poverty and wealth from mobile phone metadata", *Science*, 350, pp. 1073–1076.

BUCICOVSCHI O., DOUGLASS R.W., MEYER D.A., RAM M., RIDEOUT D., SONG D. (2013), "Analyzing Social Divisions Using Cell Phone Data", in: *D4D Book: Mobile Phone Data for Development. Analysis of Mobile Phone Datasets for the Development of Ivory Coast*.

DECUYPER A., RUTHERFORD A., WADHWA A., BAUER J.-M., KRINGS G., GUTIERREZ T., BLONDEL V.D., LUENGO-OROZ M.A., (2014), "Estimating Food Consumption and Poverty Indices with Mobile Phone Data", *CoRR*, abs/1412.2595.

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.

FLOWMINDER FOUNDATION, http://www.flowminder.org (accessed: 2016/11/28).

GONZÁLEZ M.C., HIDALGO C.A., BARABÁSI A.L. (2008), "Understanding individual human mobility patterns", *Nature*, vol. 453, pp. 779–782.

IBM, *What is big data?*, https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html, (accessed) 2015.

KAFSI M., KAZEMI E., MAYSTRE L., YARTSEVA L., GROSSGLAUSER M., THIRAN P. (2013), "Mitigating Epidemics through Mobile Micro-measures", *arXiv* preprint arXiv:1307.2084.

LAMBIOTTE R., BLONDEL V.D., DE KERCHOVE C., HUENS E., PRIEUR C., SMOREDA Z., VAN DOOREN, P. (2008), "Geographical dispersal of mobile communication networks", *Physica A: Statistical Mechanics and its Applications*, vol. 387, pp. 5317–5325.

LIMA A., DE DOMENICO M., PEJOVIC V., MUSOLESI M. (2013), "Exploiting Cellular Data for Disease Containment and Information Campaigns Strategies in Country-Wide Epidemics", *CoRR*, abs/1306.4534.

MIRITELLO G. (2013), *Temporal patterns of communication in social networks*, Springer Science & Business Media.

RAEDER T., LIZARDO O., HACHEN D., CHAWLA N.V. (2011), "Predictors of short-term decay of cell phone contacts in a large scale communication network", *Social Networks*, vol. 33, 245–257.

RATTI C., SOBOLEVSKY S., CALABRESE F., ANDRIS C., READES J., MARTINO M., CLAXTON R., STROGATZ S. (2010), "Redrawing the map of Great Britain from a network of human interactions", *PLoS One*, vol. 5, n° 12, e14248.

READES J., CALABRESE F., SEVTSUK A., RATTI C. (2007), "Cellular census: Explorations in urban data collection", *IEEE Pervasive Computing*, pp. 30–38.

SARAMÄKI J., LEICHT E.A., LÓPEZ E., ROBERTS S.G.B., REED-TSOCHAS F., DUNBAR R.I.M. (2014), "The persistence of social signatures in human communication", *Proceedings of the National Academy of Sciences*, vol. 111, pp. 942–947.

TIBÉLY G., KOVANEN L., KARSAI M., KASKI K., KERTÉSZ J., SARAMÄKI J. (2011), "Communities and beyond: mesoscopic analysis of a large social network with complementary methods", *Physical Review E*, vol. 83, n° 5, 056125.