Blondel et al. *EPJ Data Science* (2015) 4:10
DOI 10.1140/epjds/s13688-015-0046-0

EPJ Data Science
a SpringerOpen Journal

REVIEW

Open Access

A survey of results on mobile phone datasets analysis

Vincent D Blondel[1]*, Adeline Decuyper[1] and Gautier Krings[1,2]

*Correspondence:
vincent.blondel@uclouvain.be
[1] Department of Applied Mathematics, Université catholique de Louvain, Avenue Georges Lemaitre, 4, Louvain-La-Neuve, 1348, Belgium
Full list of author information is available at the end of the article

Abstract

In this paper, we review some advances made recently in the study of *mobile phone datasets*. This area of research has emerged a decade ago, with the increasing availability of large-scale anonymized datasets, and has grown into a stand-alone topic. We survey the contributions made so far on the *social networks* that can be constructed with such data, the study of *personal mobility*, *geographical partitioning*, *urban planning*, and *help towards development* as well as *security and privacy issues*.

Keywords: mobile phone datasets; big data analysis; data mining; social networks; temporal networks; geographical networks

1 Introduction

As the Internet has been the technological breakthrough of the '90s, mobile phones have changed our communication habits in the first decade of the twenty-first century. In a few years, the world coverage of mobile phone subscriptions has raised from 12% of the world population in 2000 up to 96% in 2014 - 6.8 billion subscribers - corresponding to a penetration of 128% in the developed world and 90% in developing countries [1]. Mobile communication has initiated the decline of landline use - decreasing both in developing and developed world since 2005 - and allows people to be connected even in the most remote places of the world.

In short, mobile phones are *ubiquitous*. In most countries of the developed world, the coverage reaches 100% of the population, and even in remote villages of developing countries, it is not unusual to cross paths with someone in the street talking on a mobile phone. Due to their ubiquity, mobile phones have stimulated the creativity of scientists to use them as millions of potential sensors of their environment. Mobile phones have been used, as distributed seismographs, as motorway traffic sensors, as transmitters of medical imagery or as communication hubs for high-level data such as the reporting of invading species [2] to only cite a few of their many side-uses.

Besides these applications of voluntary reporting, where users install applications on their mobile phones in the aim to serve as sensor, the essence of mobile phones have revealed them to be a source of even much richer data. The call data records (CDRs), needed by the mobile phone operators for billing purposes, contain an enormous amount of information on how, when, and with whom we communicate.

In the past, research on social interactions between individuals were mostly done by surveys, for which the number of participants ranges typically around 1,000 people, and for which the results were biased by the subjectivity of the participants' answers. Mobile phone CDRs, instead, contain the information on communications between millions of people at a time, and contain real observations of communications between them rather than self-reported information.

In addition, CDRs also contain location data and may be coupled to external data on customers such as age or gender. Such a combination of personal data makes of mobile phone CDRs an extremely rich and informative source of data for scientists. The past few years have seen the rise of research based on the analysis of CDRs. First presented as a side-topic in network theory, it has now become a whole field of research in itself, and has been for a few years the leading topic of NetMob, an international conference on the analysis of mobile phone datasets, of which the fourth edition took place in April 2015. Closely related to this conference, a side-topic has now risen, namely the analysis of mobile phone datasets for the purpose of development. The telecom company Orange has, to this end, proposed a challenge named D4D, whose concept is to give access to a large number of research teams throughout the world to the same dataset from an African country. Their purpose is to make suggestions for development, on the basis of the observations extracted from the mobile phone dataset. The first challenge, conducted in 2013 was such a success that other initiatives such as this one have followed [3, 4], and the results of a second D4D challenge were presented at the NetMob conference in April 2015.

Of course, there are restrictions on the availability of some types of data and on the projected applications. First, the content of communications (SMS or phone discussions) is not recorded by the operator, and thus inaccessible to any third party - exception made of cases of phone tapping, which are not part of this subject. Secondly, while mobile phone operators have access to all the information filed by their customers and the CDRs, they may not give the same access to all the information to a third party (such as researchers), depending on their own privacy policies and the laws on protection of privacy that apply in the country of application. For example, names and phone numbers are never transmitted to external parties. In some countries, location data, *i.e.*, the base stations at which each call is made, have to remain confidential - some operators are even not allowed to use their own data for private research.

Finally, when a company transmits data to a third party, it goes along with non-disclosure agreements (NDA's) and contracts that strongly regulate the authorised research directions, in order to protect the users' privacy.

Recently, with the rise of smartphones, other methods of collecting data overcoming those drawbacks have been designed: projects such as Reality Mining [5], OtaSizzle [6], or Sensible DTU [7] consist in distributing smartphones to individuals who volunteered for the study. A previously installed software then records data, and these data are further used for research by the team that distributed the smartphones. This new approach overcomes the privacy problems, as participants are clearly informed and consent to their data being used. On the one hand, these projects gather very rich data, as they usually collect more than just call logs, but also bluetooth proximity data, application usage, etc.... On the other hand, the sample of participants is always much more limited than in the case of CDRs shared by a provider, and the dataset contains information on 1,000 participants at most.

Yet, even the smallest bit of information is enough for triggering bursts of new applications, and day after day researchers discover new purposes one can get from mobile phone data. The first application of a study of phone logs (not mobile, though) appeared in 1949, with the seminal paper by George Zipf modeling the influence of distance on communication [8]. Since then, phone logs have been studied in order to infer relationships between the volume of communication and other parameters (see *e.g.* [9]), but the apparition of mobile phone data in massive quantities, and of computers and methods that are able to handle those data efficiently, has definitely made a breakthrough in that domain. Being personal objects, mobile phones enabled to infer real social networks from their CDRs, while fixed phones are shared by users of one same geographical space (a house, an office). The communications recorded on a mobile phone are thus representative of a part of the social network of one single person, where the records of a fixed phone show a superposition of several social actors. By being mobile, a mobile phone has two additional advantages: first, its owner has almost always the possibility to pick up a call, thus the communications are reflecting the temporal patterns of communications in great detail, and second, the positioning data of a mobile phone allows to track the displacements of its owner.

Given the large amount of research related to mobile phones, we will focus in this paper on contributions related to the analysis of massive CDR datasets. A chapter of the (unpublished) PhD thesis of Gautier Krings [10] gives an overview of the literature on mobile phone datasets analysis. This research area is growing fast and this survey is a significantly expanded version of that chapter, with additional sections and figures and an updated list of references. The paper is organized following the different types of data that may be used in related research. In Section 2 we will survey the contributions studying the topological properties of the static social network constructed from the calls between users. When information on the position of each node is available, such a network becomes a geographical network, and the relationship between distance and the structure of the network can be analyzed. This will be addressed in Section 3. Phone calls are always localized in time, and some of them might represent transient relationships while others rather long-lasting interactions. This has led researchers to study these networks as temporal networks, which will be presented in Section 4. In Section 5, we will focus on the abundant literature that has been produced on human mobility, made possible by the spatio-temporal information contained in CDR data. As mobile phone networks represent in their essence the transmission of information or more recently data between users, we will cover this topic in Section 6, with contributions on information diffusion and the spread of mobile phone viruses. Some contributions combine many of these different approaches to use mobile phone data towards many different applications, which will be the object of Section 7. Finally, in Section 8 we will consider privacy issues raised by the availability and use of personal data.

2 Social networks

In its simplest representation, a dataset of people making phone calls to each other is represented by a network where nodes are people and links are drawn between two nodes who call each other. In the first publications related to telecommunications datasets, the datasets were rather used as an example for demonstration of the potential applications of an algorithm [11] or model [12] rather than for a purpose of analysis. However, it quickly appeared that the so-called *mobile call graphs* (MCG) were structurally different from other complex networks, such as the web and internet, and deserved particular attention, see Figure 1 for an example of snowball sampling of a mobile phone network. We will

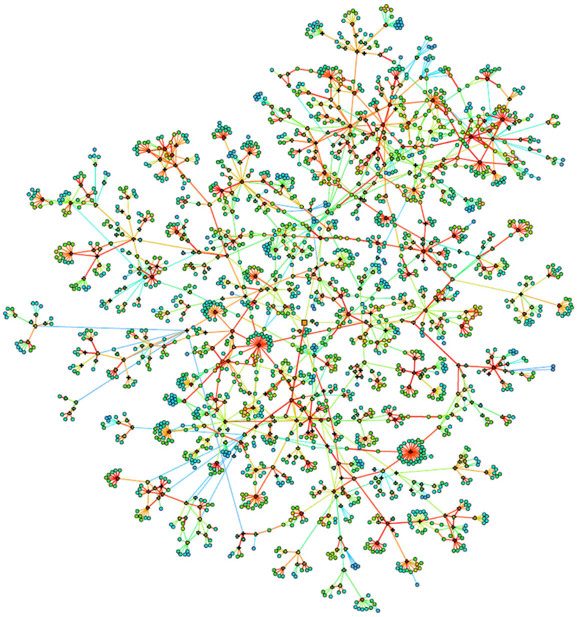


Figure 1 Sample of a mobile phone network, obtained with a snowball sampling. The source node is represented by a square, bulk nodes by a + sign and surface nodes by an empty circle. Figure reproduced from [25].

review here the different contributions on network analysis. We will address the construction of a social network from CDR data, which is not a trivial exercise, simple statistical properties of such networks and models that manage to reproduce them, more complex organizing principles, and community structure, and finally we will discuss the relevance of the analysis of mobile phone networks.

2.1 Construction

While the network construction scheme mentioned above seems relatively simple, there exist many possible interpretations on how to define a link of the network, given a dataset.

The primary aim of social network analysis is to observe social interactions, but not every phone call is made with the same social purpose. Some calls might be for business purposes, some might be accidental calls, some nodes may be call centers that call a large number of people, and all such interactions are present in CDRs. In short, CDRs are noisy datasets. ‘Cleaning’ operations are usually needed to eliminate some of the accidental edges. For example, Lambiotte *et al.* [13] imposed as condition for a link that at least one call is made in both directions (reciprocity) and that at least 6 calls are made in total over 6 months of the dataset. This filtering operation appeared to remove a large fraction of the links of the network, but at the same time, the total weight (the total number of calls passed in by all users) was reduced by only a small fraction. The threshold of 6 calls in 6 months may be questionable, but a stability analysis around this value can comfort that the exact choice of the threshold is not crucial. Similarly, Onnela *et al.* [14] analyzed the differences between the degree distribution of two versions of the same dataset, one containing all calls of the dataset, and the other containing only calls that are reciprocated. Some nodes in the complete network have up to 30,000 different neighbors, while in the reciprocated network, the maximal degree is close to 150. Clearly, in the first case it is hard to imagine that node representing a single person, while the latter is a much more realistic bound. However, even if calls have been reciprocated, the question of setting a meaningful weight on each link is far from easy. Li *et al.* suggest another more statistical approach in [15], and use multiple hypothesis testing to filter out the links that appeared randomly in the network and that are therefore not the mirror of a true social relationship. Further than these considerations of which calls or texts are representative of a true relationship, further corruptions of the data can arise from multiple calls, multiple text messages or calls reaching an answering machine. Depending on the context, it may be preferable to filter out these communications, but they remain, most of the time, difficult to identify in the datasets. Moreover, additional biases can arise from the pricing plans of the operator, putting a preferential price for SMSs or for voice calls, thus influencing the behavior of the users that opted for such a pricing plan [16].

It is sometimes convenient to represent a mobile call network by an undirected network, arguing that communication during a single phone call goes both ways, and set the weight of the link as the sum of the weights from both directions. However, who initiates the call might be important in other contexts than the passing of information, depending on the aim of the research, and Kovanen *et al.* have showed that reciprocal calls are often strongly imbalanced [17]. In the interacting pair, one user is often initiating most of the calls, so how can this be represented in an undirected network by a representative link weight? In a closely related question, most CDRs contain both information on voice calls and text messages, but so far it is not clear how to incorporate both pieces of information into one simple measure. Moreover, there seems to be a generational difference in the use of text messages or preference between texts and voice calls which may introduce a bias in measures that only take one type of communication into account [18].

Besides these considerations on the treatment of noise, the way to represent social ties may vary as well: they may be binary, weighted, symmetric or directed. Different answers to such decisions lead to different network characteristics, and result in diverse possible interpretations of the same dataset. For example, Nanavati *et al.* [19] keep their network as a directed network, in order to obtain information on the strongly connected component of the network, while Onnela *et al.* [14] rather focus on an undirected network, weighted by the sum of calls going in both directions. A few different options for definitions of link weights and measures on nodes are given in Table 1.

It is close to impossible to define unified construction rules functioning for any dataset, given the many sources of variance between two sets of CDRs. Besides the cases mentioned above, let us cite differences in social behaviors between inhabitants of different countries or differences in uses of uncaptured technologies such as email, landline or messaging. The construction of a social network from CDRs should always be of primary importance for the researcher, bearing in mind that there's no 'one size fits all' technique available.

Table 1 Definitions of node and link measures

node strength	degree
	sum of link weights
	local clustering coefficient C_i
link weight	number of calls
	number of SMSs
	total duration (with, *e.g.*, one SMS being counted as 60 seconds)
	overlap O_{ij}

2.2 Topological properties

The simplest information one can get out of CDRs is statistical information on the number of acquaintances of a node, on the local density of the network or on its connectivity. Like social networks, mobile call graphs differ from random networks and lattices by their broad degree distribution [20], their small diameter and their high clustering [21].

The level of clustering of a graph G is measured by its *clustering coefficient*, defined as the proportion of closed triangles among the connected triplets of nodes in G:

$$C(G) = \frac{6 \times \text{number of triangles}}{\text{number of paths of length 2}}. \tag{1}$$

This coefficient takes values in the interval $[0,1]$, and is found to be typically high in social networks and mobile call graphs. An alternative is to take the average of a local measure of the clustering around a given node i, defined as:

$$C(G) = \langle C_i \rangle, \quad \text{where } C_i = \frac{2 \times \text{number of triangles of which } i \text{ is a node}}{k_i(k_i - 1)}, \tag{2}$$

where k_i is node i's degree, and hence $k_i(k_i - 1)$ is the maximum number of possible triangles around node i.

The *diameter* of a graph G measures the greatest distance (in terms of number of edges) between any two vertices, and is typically small in social networks.

As for the degree distribution, while all analyzed datasets present similar general shapes, their range and their fine shape differ due to differences between the datasets, the construction scheme, the size, or the time span of the collection period.

In one of the first studies involving CDR data Aiello *et al.* [12] observed a power law degree distribution, which was well explained by a massive random graph model $P(\alpha, \beta)$ described by its power-law degree distribution $p(d = x) = e^{\alpha} x^{-\beta}$. Power-law distributions have often been observed in empirical datasets, but characterizing their parameters and determining whether the data really corresponds to a power-law distribution is not an easy question, as presented by Clauset *et al.* [22].

Random graph models have often been used in order to model networks, and manage to reproduce some observations from real-world networks, such as the small diameter and the presence of a giant component, such as observed on mobile datasets. However, they fail to uncover more complex features, such as degree-degree correlations. Nanavati *et al.* [19] observed in the study of 4 mobile datasets that besides the power-law tail of the degree distribution, the degree of a node is strongly correlated with the degree of its neighbors.

Characterizing the exact shape of the degree distribution is not an easy task, which has been the focus of a study by Seshradi *et al.* [23]. They observed that the degree distribution of their data can be fitted with a Double Pareto Log Normal (DPLN) distribution, two power-laws joined by a hyperbolic segment - which can be related to a model of social wealth acquisition ruled by a lognormal multiplicative process. Those different degree distributions are depicted on Figure 2. Interestingly, let us note that the time span of the three aforementioned datasets are different, Aiello *et al.* have data over one day, Nanavati *et al.* over one week, and Seshadri *et al.* over one month.

Krings *et al.* dig a bit deeper into this topic, and investigated the effect of placement and size of the aggregation time window [24]. They showed that the size of the time window of aggregation can have a significant influence on the distributions of degrees and weights in

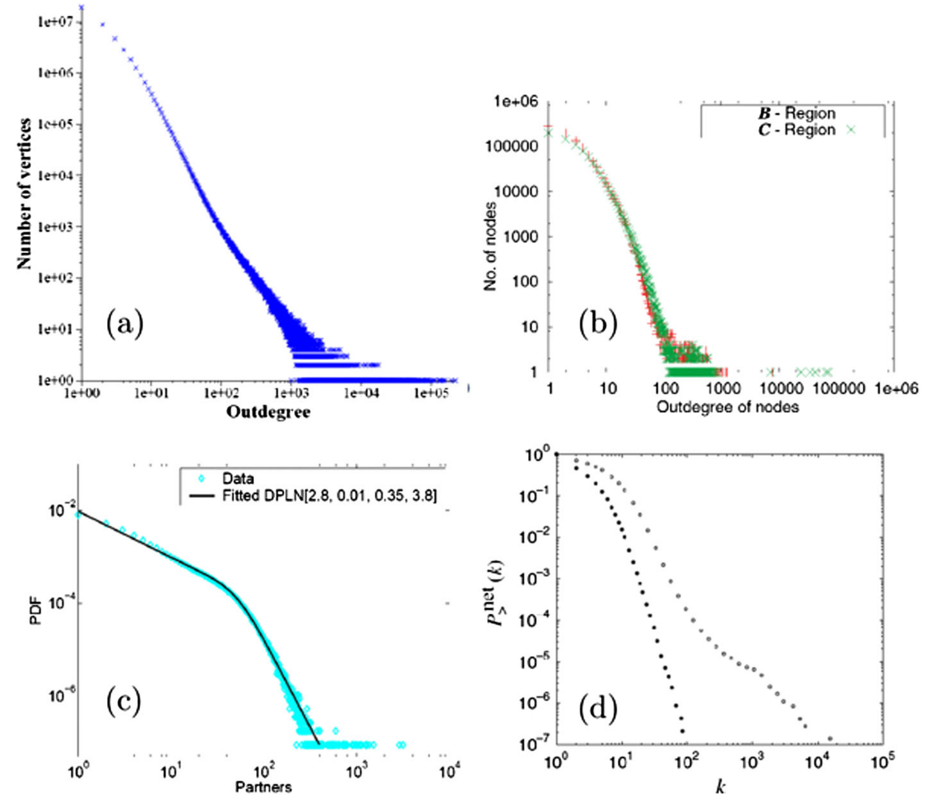


Figure 2 Degree distributions in mobile phone networks. The degree distributions of several datasets have comparable features, but differences in the construction, the time range of the dataset and the size of the system lead to different shapes. Note the bump in (d), when non-reciprocal links are taken into account. **(a)** Aiello, W. *et al.*, 'A random graph model for massive graphs', in: Proceedings of the thirty-second annual ACM symposium on theory of computing, pp 171-180 [12] ©2000 Association for Computing Machinery, Inc. Reprinted by permission. http://doi.acm.org/10.1145/335305.335326. **(b)** Nanavati, A.A. *et al.*, 'On the structural properties of massive telecom call graphs: findings and implications', in: Proceedings of the 15th ACM international conference on information and knowledge management, pp 435-444 [19] ©2006 Association for Computing Machinery, Inc. Reprinted by permission. http://doi.acm.org/10.1145/1183614.1183678. **(c)** Seshadri, M. *et al.*, 'Mobile call graphs: beyond power-law and lognormal distributions.' in: Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining, pp 596-604 [23] ©2008 Association for Computing Machinery, Inc. Reprinted by permission. http://doi.acm.org/10.1145/1401890.1401963. **(d)** Figure reproduced from [25].

the network. The authors also observed that the degree and weight distributions become stationary after a few days and a few weeks respectively. The effect of the placement of the time window has most influence for short time windows, and depends mostly on whether it contains holiday periods or weekends, during which the behavioral patterns have been shown to be significantly different than during normal weekdays.

What information do we get from these distributions? They mostly reflect the heterogeneity of communication behaviors, a common feature for complex networks [20]. The fat tail of the degree distribution is responsible for large statistical fluctuations around the average, indication that there is no particular scale representative of the system. The majority of users have a small number of contacts, while a tiny fraction of nodes are hubs, or super-connectors. However, it is not clear whether these hubs represent true popular users or are artefacts of noise in the data, as was observed by Onnela *et al.* [25] in their comparison of the reciprocated and non-reciprocated network.

The heterogeneity of degrees is also observed on node strengths and link weight, which is also to be expected for social networks. All studies also mention high clustering coefficient, which indicates that the nodes arrange themselves locally in well-organized structures. We will address this topic in more detail further.

However large the datasets studied may be, one may still question the significance of the measures presented above. The data studied is always about only a limited (yet significant) sample of the population, and it is very difficult to examine whether this sample is biased or not, without additional information on the population of the country and the users of the mobile provider considered. We will discuss this topic further in Section 7, but the topological properties presented in this section, such as the degree distributions should be analyzed with care, as one may only expect qualitatively close results for similar datasets on similar populations. If the sample of the population studied is biased, and one only observes people that share specific characteristics (such as age, gender, or a specific profile), then topological properties might be very different from those characterizing the mobile phone networks presented in the above paragraph. We have discussed the impact of the size of the time window of observation on the results. Most studies usually try to infer general trends characterizing the network of acquaintances of a population, based on observation during only a finite time window. The ideal model capturing the network of people would be based on a very long time of observation of interactions between people of the whole population around the globe. This is of course impossible to achieve, and limited time window of observations as well as population sample introduce inherent biases in the results, even though these biases are almost impossible to characterize and quantify.

2.3 Advanced network characteristics

Beyond statistical distributions, more complex analyses provide a better understanding of the structure of our communication networks. The heterogeneity of link weights deserves particular attention. Strong links represent intense relationships, hence the correlation between weight and topology is of primary interest. Recalling that mobile call graphs show high clustering coefficient, and thus are locally dense, one can differentiate links based on their position in the network.

The overlap of a link, introduced in [14] (and illustrated on Figure 3), is an appropriate measure which characterizes the position of a link as the ratio of observed common neighbors n_{ij} over the maximal possible, depending on the degrees k_i and k_j of the nodes and defined as:

$$O_{ij} = \frac{n_{ij}}{(d_i - 1) + (d_j - 1) - n_{ij}}. \tag{3}$$

The authors show that link weight and topology are strongly correlated, the strongest links lying inside dense structures of the network, while weaker links act as connectors between these densely organized groups. This finding has an important consequence on processes such as link percolation or the spread of information on networks, since the weak ties act as bridges between disconnected dense parts of the network, illustrating Granovetter's hypothesis on the strength of weak ties [26].

The structure of the dense subparts of the network provides essential information on the self-organizing principles lying behind communication behaviors. Before moving to the analysis of communities, we will focus on properties of cliques. The structure of cliques

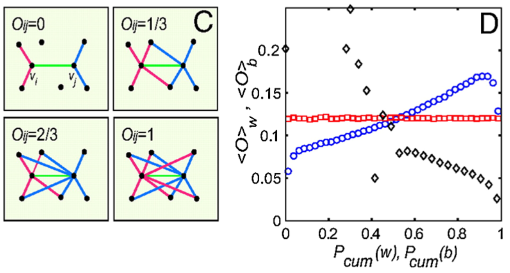


Figure 3 Overlap of a link in a network. (Left) The overlap of a link is defined as the ratio between the common neighbors of both nodes and the maximum possible common neighbors. Here, the overlap is given for the green link. (Right) The average overlap increases with the cumulative weight in the real network (blue circles) and is constant in the random reference where link weights are shuffled (red squares). The overlap also decreases with the cumulative betweenness centrality $P_{cum}(b)$ (black diamonds). Figure reproduced from [14].

is reflected by how weights are distributed among their links. In a group where everyone talks to everyone, is communication balanced? Or are small subgroups observable? A simple measure to analyze the balance of weights is the measure of coherence $q(g)$. This measure was introduced in [27] before its application to mobile phone data in [25], and is calculated as the ratio between the geometric mean of the link weights and the arithmetic mean,

$$q(g) = \frac{(\prod_{ij\in l_g} w_{ij})^{1/|l_g|}}{\frac{\sum_{ij\in l_g} w_{ij}}{|l_g|}}, \qquad (4)$$

where g is a subgraph of the network and l_g is its set of links. This measure takes values in the range $]0,1]$, 1 corresponding to equilibrium. On average, cliques appear to be more coherent than what would be expected in the random case, in particular for triangles, which show high coherence values.

On a related topic, Du *et al.* [28] focused instead on the propensity of nodes to participate to cliques, and in particular on the balance of link weights inside triangles. Their observations differ slightly from Onnela *et al.*: on average, the weights of links in triangles can be expressed as powers of one another. The authors managed to reproduce this singular situation with a utility-driven model, where users try to maximize their return from contacts.

2.4 Communities

The previous analysis of cliques and triangles opens the way for an analysis of more complex structures, such as communities in mobile phone networks. The analysis of communities provides information on how communication networks are organized at large scale. In conjunction with external data, such as age, gender or cultural differences, it provides sociological information on how acquaintances are distributed over the population. From a corporate point of view, the knowledge of well-connected structures is of primary importance for marketing purposes. In this paragraph, we will only address simple results on community analysis, but this topic will be addressed again further in the document, when it relates to geographic dispersal of networks or dynamical networks.

At small scale, traditional clustering techniques may be applied, see [29] and [30] for examples of applications on small datasets. However, on large mobile call graphs involving millions of users, such clustering techniques are outplayed by community detection algorithms.

Uncovering the community structure in a mobile phone network is highly dependent on the used definition of communities and detection method. One could argue that there exist as many plausible analyses as there are community detection methods. Moreover, the particular structure of mobile call graphs induces some issues for traditional community detection methods. Tibely *et al.* [31] show that even though some community detection methods perform well on benchmark networks, they do not produce clear community structures on mobile call graphs. Mobile call graphs contain many small tree-like structures, which are badly handled by most community detection methods. The comparison of three well-known methods: the Louvain method [32], Infomap [33] and the Clique Percolation method [34] produce different results on mobile call graphs. The Louvain method and Infomap both build a partition of the nodes of the network, so that every node belongs to exactly one community. In contrast Clique Percolation only keeps as community dense subparts of the network (see Figure 4).

As observed in Tibely *et al.* the small tree-like structures are often considered as communities, although their structure is sparse. Such a result is counter-intuitive given the intrinsic meaning of communities and raises the question: is community detection hence unusable on mobile call graphs? The results have probably to be considered with caution, but as this is always the case for community detection methods, whatever network is used, this special character of communities in mobile call graphs appears rather as a particularity than a problem. Although they might have singular shapes, communities can provide significant information, when usefully combined with external information. Proof is made by the study of the linguistic distribution of communities in a Belgian mobile call graph [32],

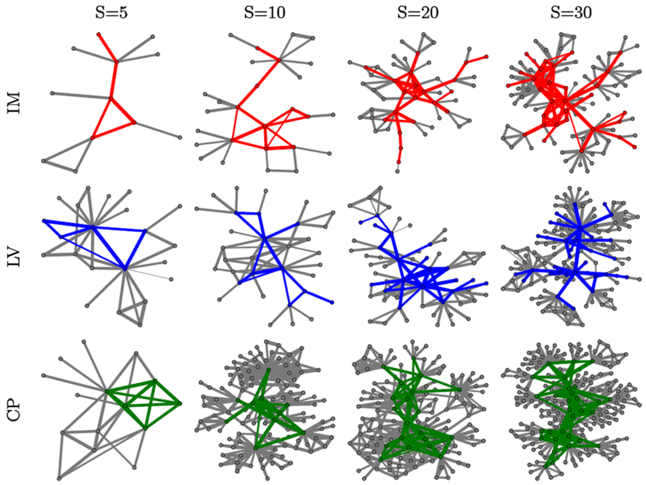


Figure 4 Examples of communities detected with different methods. The different methods are the InfoMap method (IM, red), Louvain method (LV, blue) and Clique percolation method (CP, green). For each method, four examples are shown, with 5, 10, 20 and 30 nodes. The coloured links are part of the community, the grey nodes are the neighbors of the represented community. While IM and LV find almost tree-like structures, CP finds dense communities [31]. Reproduced figure with permission from Tibély, G. *et al.*, Phys Rev E 83(5):056125, 2011. Copyright (2011) by the American Physical Society. http://dx.doi.org/10.1103/PhysRevE.83.056125.

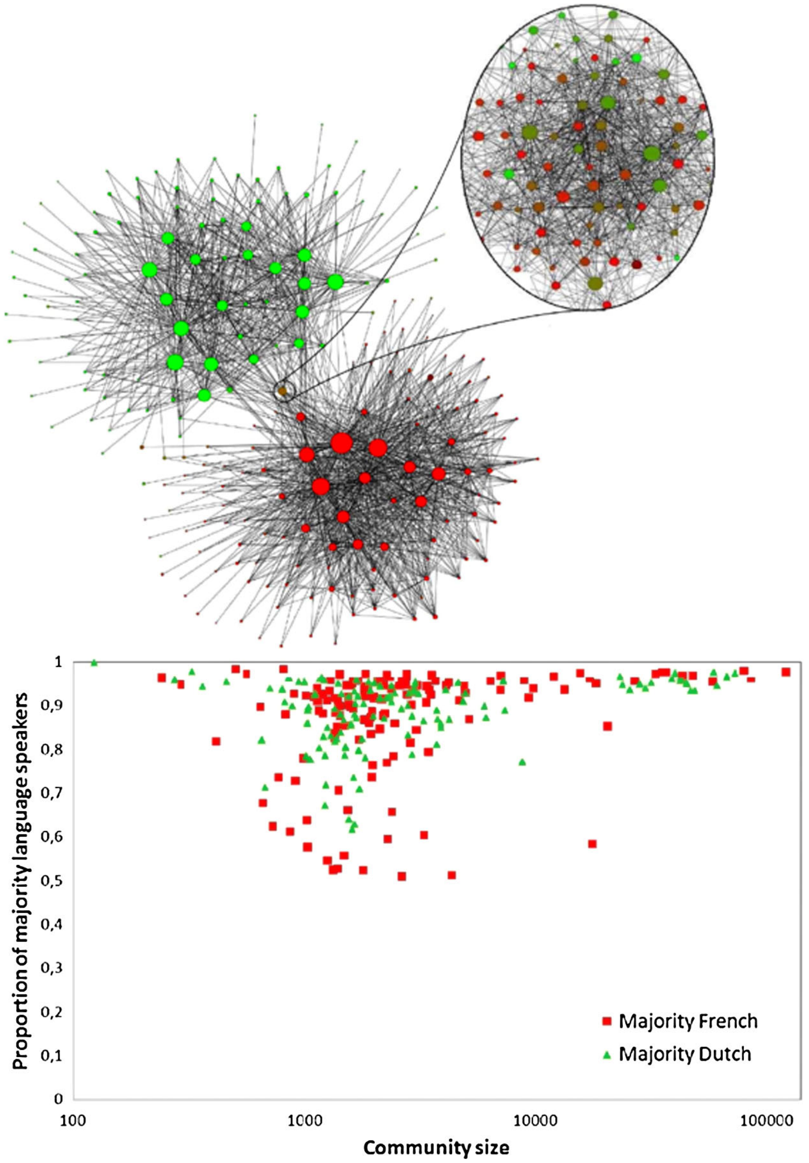


Figure 5 Community detection in Belgium. (Top) The communities of the Belgian network are colored based on their linguistic composition: green for Flemish, red for French. Communities having a mixed composition are colored with a mixed color, based on the proportion of each language. (Bottom) Most communities are almost monolingual. Figures reproduced from [32].

where the communities returned by the Louvain method strikingly show a well-known linguistic split, as illustrated on Figure 5.

The notion of communities in social networks, such as rendered by mobile phone networks, has raised a debate on the exact vision one has of what a community is and what it is not. In particular, several authors have favored the idea of overlapping communities, such that one node may belong to several communities, in opposition with the classical vision that communities are a partition of the nodes of a network. An argument in favor of this vision is that one is most often part of several groups of acquaintances who do not share common interests, such as family, work and sports activities. In [35], Ahn *et al.* show how overlapping communities can be detected by partitioning edges rather than nodes, and illustrated their methods with a mobile phone dataset. For each node, they had additional information about its center of activities, with which they showed that communities were geographically consistent. This discussion of the exact definition of communities and of the best method to detect them can further be influenced by a series of factors. Indeed, one could want to introduce additional information before searching for communities such as, for example, age, gender or specific profiles of people. Moreover, when spatio-temporal information is available, one could want to detect strong communities whose links remain active through time, or detect which geographic areas belong to the same community thus partitioning space, as we will explain further in this paper. Finally, as mentioned in the previous paragraph, if the sample of the population that is available is biased, the structure of the network, and hence also the detection of communities may be influenced. All these additional considerations might significantly change the result of the detection of communities, including their internal topology.

2.5 Social analysis

The use of mobile call data in the purpose of analysis of social relationships raises two questions. First, how faithful is such a dataset of real interactions? Second, can we extract information on the users themselves from their calling behavior?

It has often been claimed that mobile phone data analysis is a significant advance for social sciences, since it allowed scientists to use massive datasets containing the activity of entire populations. The study of mobile phone datasets is part of an emerging field known as computational social science [36]. These massive datasets, it is said, are free from the bias of self-reporting, which is that the answers to a survey are usually biased by the own perception of the subject, who is not objective. Still, the question remains: how much does self-reporting differ from our real behavior, what is the exact added value of having location data? This has been studied by Eagle *et al.* [37] in the well-known Reality Mining project. By studying the behavior of about 100 persons both by recording their movements and encounters using GSM and Bluetooth technology and with the use of surveys, they managed to quantify the difference between self-reported behavior and what could be observed. It appears that observed behavior strongly differs from what has been self-reported, confirming that the subjectivity of the subjects' own perception produces a significant bias in surveys. In contrast, collected data allows to reduce this bias significantly. However, mobile phone data introduce a different bias, namely, that they only contain social contacts that were expressed through phone calls, thus missing all other types of social interactions out [38].

While most studies use external data as validation tool to confirm the validity of results, Blumenstock *et al.* shortly addressed a different question, namely if it was possible to infer information on people's social class based on their communication behavior. Apparently, this task is hard to perform, even if significant differences appear in calling behavior between different classes of the population [39]. While inferring information about users from their calling activity still seems difficult, many studies show strong correlations between calling behavior and other information included in some datasets, such as gender or age. In a study on landline use, Smoreda *et al.* highlight the differences in the use of the domestic telephone based on the genders of both the caller and the callee [40], and show not only that women call more often than men but also that the gender of the callee has more influence than the gender of the caller on the duration of the call. Those same trends have also been observed in later studies of mobile phone datasets [41]. Further than just observing the gender differences in mobile phone use, Frias-Martinez *et al.* propose a method to infer the gender of a user based on several variables extracted from mobile phone activity [42], and achieve a success rate of prediction between 70% and 80% on a dataset of a developing economy. In a later study on data from Rwanda, Blumenstock *et al.* show that differences of social class induce more striking differences in mobile phone use than differences of gender [43].

Further than analyzing the nodes of a network, Chawla *et al.* take a closer look at the links of the network, and introduce a measure of reciprocity to quantify how balanced the relationship between two users is [44]:

$$R_{ij} = \left|\ln(p_{ij}) - \ln(p_{ji})\right|, \tag{5}$$

where p_{ij} is the probability that if i makes a communication, it will be directed towards j. They also test this measure on a mobile communications dataset, and show that there are very large degrees of non-reciprocity, far above what could be expected if only balanced relationships were kept.

Going one step further, instead of inferring information on the nodes of the mobile calling graph, Motahari *et al.* study the difference in calling behavior depending on the relationship between two subscribers, characterizing different types of links. They show that the links within a family generate the highest number of calls, and that the network topology around those links looks significantly different from the topology of a network of utility communications [45].

If we can infer so much information from looking at mobile phone communications, would it be possible to predict existing acquaintances that are unobserved in the dataset available? This question is known as the link prediction problem, and has been addressed by several research teams. As the approach usually takes into account the time component of the dataset, we will address this topic further in Section 4.

3 Adding space - geographical networks

Besides basic CDR data, it happens that geographic information is available about the nodes, such as the home location (available for billing purposes) or the most often used antenna. This allows then to assign each node to one geographic point, and to study the interplay between geography and mobile phone usage. Studies on geographical networks have already been performed on a range of different types of networks [46]. One of the very basic applications is to use mobile phone data to estimate the density of population in the different regions covered by the dataset. Deville *et al.* explored this idea [47], using the number of people who are calling from each antenna, they are able to produce timely estimates of the population density in France and Portugal. In the developing world, census data is often very costly or even impossible to obtain, and existing data is often very old and outdated. Using CDRs can then provide very useful and updated information on the actual density of population in remote parts of the world. Another example is given by Sterly *et al.* who mapped an estimate of the density of population of Ivory Coast using a mobile phone dataset [48], as illustrated on Figure 6.

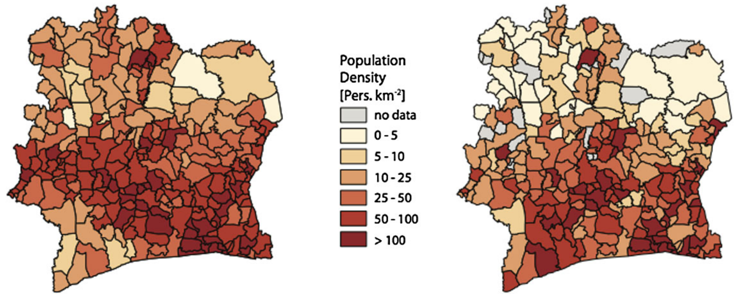
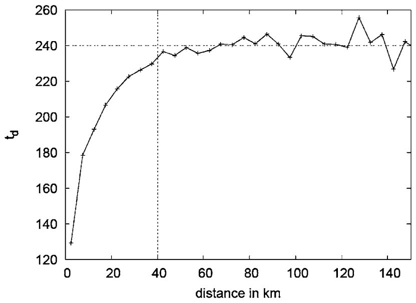


Figure 6 Population density estimates. (Left) Population density estimates from the Afripop project [214]. (Right) Population density estimates from mobile phone data. Figure reproduced from [48].

Figure 7 Average duration of a call depending on the distance between the callers. A saturation point is observed at 40 km. Figure reproduced from [13].



3.1 Relationship space-communication

Lambiotte *et al.* [13], investigated the interplay between geography and communications, and assigned each of the 2.5 million users from a Belgian mobile phone operator to the ZIP code location where they were billed. By approximating the position of the users to the center of each ZIP code area, they showed that the probability of two users to be connected decreases with the distance r separating them, following a power law of exponent –2. The probability of a link to be part of a triangle decreases with distance, until a threshold distance of 40 km, after which the probability is constant. Interestingly, this threshold of 40 km is also a saturation point for the average duration of a call (see Figure 7). A different study on the same dataset also showed that total communication duration between communes in Belgium was well fitted by a gravity law, showing positive linear contribution of the number of users in each commune and negative quadratic influence of distance [49, 50]:

$$l_{ab} = \frac{c_a c_b}{r_{ab}^2}, \tag{6}$$

where l_{ab} represents the total communication between communes a and b, c_a and c_b the number of customers in each commune and r_{ab} the distance that separates them.

While it seems sure that distance has a negative impact on communication, its exact influence is not unique. Onnela *et al.* [51] observed in a different dataset a probability of connection decreasing as $r^{-1.5}$ rather than the gravity model observed by Lambiotte *et al.*, and a later study on Ivory Coast by Bucicovschi *et al.* [52] observe that the total duration of communication between two cities decays with $r^{-1/3}$. However, these differences might be explained by the differences that exist between the studied countries, such as the distribution of the population density. A different study on mobility data from the location-based service Foursquare [53] levelled those variations using a rank-based distance [54], which could also be helpful in this case. Another comparison is presented by Carolan *et al.* [55] who compare two different types of distance, namely the spatial travel distance and the travel time taken to link two cities. Interestingly, it appears that the use of the spatial distance rather than the time taken gives a better fit of the number of communications between two cities with the gravity model. Their observations also show that the gravity model fits the data better when data is collected during the daytime on weekdays than during evenings and weekends.

Instead of studying the communication between cities, Schläpfer *et al.* looked at the relationship between city size and the structure of local networks of people living in those cities [56]. They show that the number of contacts and communication activity both grow with city size, but that the probability of being friends with a friend's friend remains the same independently of the city size. Jo *et al.* propose another approach and study the evolution with age of the distance between a person and the person with whom they have the most contacts [57]. They thus show that young couples tend to live within longer distances than old couples.

Instead of only taking into account the distance between two places to predict the number links between them, Herrera-Yagüe *et al.* make another hypothesis, namely that the probability of someone living in a location i has contacts with a person living in another location j is inversely proportional to the total population within an ellipse [58]. The ellipse is defined as the one whose foci are i and j, and whose surface is the smallest such that both circles of radius r_{ij} centred around i and j are contained in the ellipse. If we name e_{ij} the total population within the ellipse, the number of contacts between locations i and j is thus described by:

$$T_{ij} = K\frac{n_i n_j}{e_{ij}}, \tag{7}$$

where K is a normalisation parameter depending on the total number of relationships to predict, and n_i and n_j are the populations of locations i and j respectively.

Further, Onnela *et al.* also studied the geographic structure of communities, and showed on the one hand that nodes that are topologically central inside a community may not be central from a geographical point of view, and on the other hand that the geographical shape of communities varies with their size. Communities smaller than 30 individuals show a smooth increase of geographical span with size, but bounces suddenly at the size of 30, which could not be clearly explained by the authors, see Figure 8.

3.2 Geographic partitioning

The availability to place customers in higher level entities, such as communes or counties, gave researchers the idea of drawing the 'social borders' inside a country based on the interactions between those entities [59]. Individual call patterns of users are aggregated at a higher level to a network of entities, which can in turn be partitioned into a set of communities based on the intensities of calls between the nodes of this *macroscopic* network. It is important to notice that, in contrast with the *microscopic* network (the network of users), the macroscopic network is not a sparse network at all. Since the nodes represent the aggregated behavior of many users, there is a high chance of having a link between

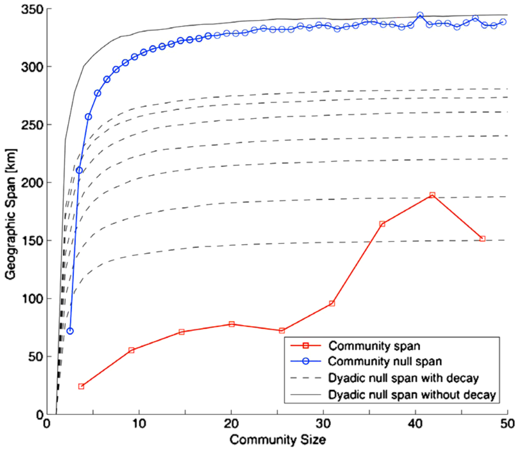


Figure 8 Average geographic span (red) for communities and average geographic span for the null model (blue). A bump is observed for communities of size 30 and more, which could not be reproduced by the different null models. Figure reproduced from [51].

most pairs of communes or counties. Hence, the weights on the links of the macroscopic network are of crucial importance, since they define the complete structure of the network. Such a partition exercise using CDR datasets has been applied, among others, on Belgium, or Ivory Coast [52, 60]. An initial study of the communities in Belgium [61] used the Louvain method optimizing modularity for weighted directed networks to partition the Belgian communes based on two link weights: the frequency of calls between two communes and the average duration of a call. The obtained partitions were geographically connected, with the influence of distance, of influential cities, and the cultural barrier of language being observable in the optimal partitions.

Given that the intensity of communication between two cities can be well-modeled by a gravity law, Expert *et al.* [62] proposed to replace Newman's modularity by a more appropriate null model, given that geographic information was available. The spatial modularity (SPA) compares the intensity between communes to a null model influenced both by the sizes c_a and c_b of the communes and the distance that separates them

$$p_{ab}^{\mathrm{Spa}} = c_a c_b f(r_{ab}). \tag{8}$$

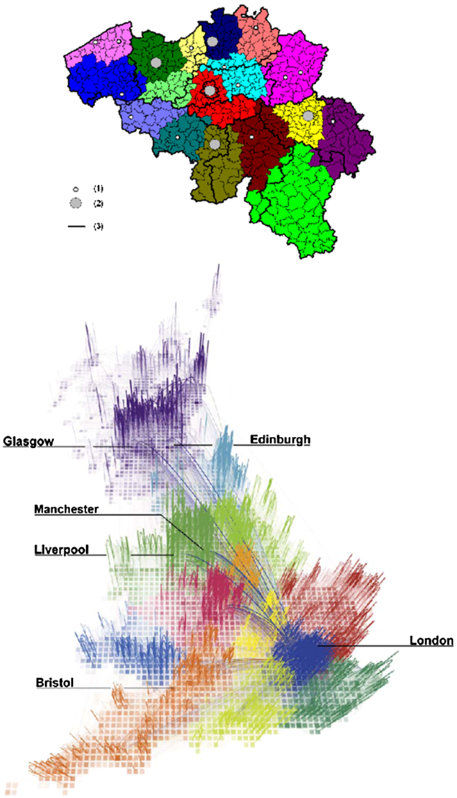
The influence of distance is estimated from the data by a function f, which is calculated for distance bins $[r - \epsilon, r + \epsilon]$ as

$$f(r) = \frac{\sum_{a,b|r_{ab}\in[r-\epsilon,r+\epsilon]} A_{ab}}{\sum_{a,b|r_{ab}\in[r-\epsilon,r+\epsilon]} c_a c_b}. \tag{9}$$

Using their null model, the authors obtained an almost perfect bipartition of the Belgian communes which renders the Belgian linguistic border. Moreover, they showed with a simple example that such a null model allows to remove the influence of geography and obtain communities showing geography-independent features.

On an identical topic, Ratti *et al.* used an algorithm of spectral modularity optimization, to partition the map of Great Britain [63] based on phone calls between geographic locations. Similarly to results obtained on Belgium, they obtained spatially connected commu-

Figure 9 Geographic partitioning of countries. (Top) Communities in Belgium, obtained through modularity optimization. Communities are geographically well-balanced and are centred around important cities (gray dots). Figure reproduced from [61]. (Bottom) Communication network in Great Britain (80% of strongest links). The colors correspond to the communities found by spectral modularity optimization. Figure reproduced from [63].



nities after a fine-grain tuning of their algorithm, which correspond to meaningful areas, such as Scotland or Greater London, see Figure 9. A stability analysis of the obtained partition showed that while some variation appears on the boundary of communities, the obtained communities are geographically centered at the same place. The intersections between several results of the same algorithm showed 11 spatially well-defined 'cores' corresponding to densely populated areas of Great Britain. Interestingly, the map of the cores loosely corresponds to the historical British regions.

A later study using the data of antenna to antenna volumes of communications in Ivory Coast confirmed the very strong influence of language on the formation of communities in a large country. Using the same method as was used by Blondel *et al.* for the Belgian dataset, they show that the borders of the communities formed in Ivory Coast strongly correlate with the language borders, even in the presence of much more than two language groups [52].

Going a bit further, Blumenstock *et al.* introduce a measure of the social and spatial *segregation* that can be observed through mobile phone communication records [64]. They define the *spatial segregation* as the proportion of people from ethnicity t in a region r as:

$$w_{tr} = \frac{N_{tr}}{N_r}, \tag{10}$$

where N_r is the total population of region r. They also define *social segregation* of ethnicity t as the fraction of contacts that individuals of ethnicity t form with the same type of people:

$$H_{tr} = \frac{s_t}{s_t + d_t}, \tag{11}$$

where s_t is the number of contacts that a person of type t has with people from the same ethnicity, and d_t is the number of contacts that people of type t have with people from other ethnicities. With these measures, it is then possible to map the more or less segregated parts of a city, see which ethnicities occupy which regions, and show how strong or weak the links between these ethnicities are.

3.3 Communications reveal regional economy

Lately, with the growth of mobile phone coverage even in the most remote regions of the developing world, a new question has risen, namely: is it possible to use CDR data to evaluate the socio-economic state of the different regions of a country? Being able to estimate and update poverty rates in different regions of a country could help governments make informed political decisions knowing how their country is developing economically.

A first step in that direction was explored by Eagle *et al.* in a study using data from the UK [65]. The authors investigated if some relationship could be found between the structure of a user's social network and the type of environment in which they live. Using both CDRs of fixed landline (99% coverage) and mobile phones (90% coverage), they showed that the social and geographical diversity of nodes' contacts, measured using the entropy of contact frequencies, correlates positively with a socio-economic factor of the neighborhood. Given a node i, calling each of his d_i neighbors j at frequency p_{ij}, and calling each of the A locations a at frequency p_{ia}, his social and spatial diversity are given by

$$D_{\text{social}}(i) = \frac{-\sum_j p_{ij} \log p_{ij}}{\log d}, \qquad D_{\text{spatial}}(i) = \frac{-\sum_a p_{ia} \log p_{ia}}{\log A}, \tag{12}$$

which is 1 if the node has diversified contacts. On Figure 10, the authors compare a composite measure of both diversities with the socio-economic factor of the neighborhood.

In a recent study, this time with data from Africa, Mao *et al.* tried to determine which characteristics of the mobile phone network could best describe the socio-economic status of a developing region [66]. They introduce an indicator named *CallRank*, obtained by running the weighted PageRank algorithm on an aggregated mobile calling graph of Ivory Coast, where nodes are the antennas and the weight of the links are the number

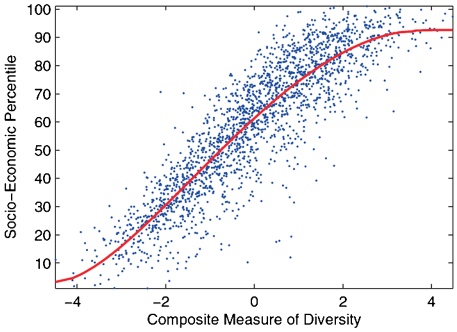
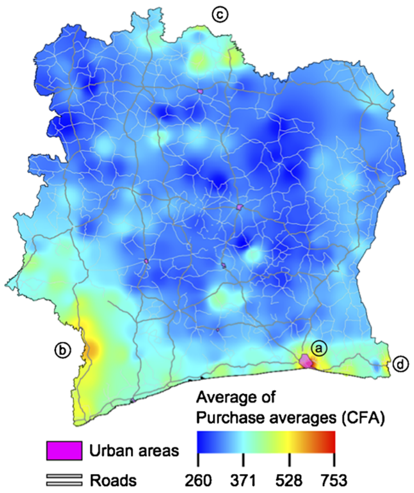


Figure 10 Average social wealth as a function of social and geographic diversity. From Eagle *et al.*, Network diversity and economic development, Science 328(5981):1029 (2010) [65]. Reprinted with permission from AAAS.

of calls between each pair of antenna. They observe that a high CallRank index seems to correspond well to a region that is important for the national economy. However, lacking accurate data to validate the results, they only conclude that this measure is probably a good indicator, without being able to evaluate its accuracy quantitatively. Another analysis of the same dataset was proposed by Smith-Clarke *et al.* who extracted a series of features to see which ones showed the best correlation with poverty levels [67]. The authors show that besides the total volume of calls, poverty levels are also linked to deviations from the expected flow of communications: if the amount of communications is significantly lower than expected from and to a certain area, then higher poverty levels are to be expected in that area. Another indicator of poverty was also explored by Frias-Martinez *et al.* who analyzed the link between the mobility of people and socio-economic levels of a city in Latin-America [68]. The authors propose several measures to quantify the mobility of users, and show that socio-economic levels present a linear correspondence with three indicators of mobility, namely the number of different antennas used, the radius of gyration and diameter of the area of often visited locations, indicating that the more mobile people are, the less poor the area in which they live seems to be. In a further study by the same research group, Frias-Martinez *et al.* go one step further, and propose a method not only to estimate, but also to forecast future socio-economic levels, based on time series of different variables gathered from mobile phone data [69]. They show preliminary evidence that the socio-economic levels could follow a pattern, allowing for prediction with mobile phone data.

Another valuable, and rather new, source of data extracted from mobile phone activity is the history of airtime purchases of each user. Using this data on the network of Ivory Coast, Gutierrez *et al.* propose another approach to infer the socio-economic state of the different regions of a developing country [70]. The authors make the hypothesis that people who make many small purchases are probably less wealthy than those who make fewer larger purchases, supposing that the poorer will not have enough cash flow to buy large amounts at the same time. Figure 11 shows the map of average purchases throughout the country. Here again, lacking external reliable data to validate those results and compare them with socio-economic data, the authors provide an interpretation of the differences observed between the different regions, and show that the hypothesis they make seems plausible.

Figure 11 Average purchases of airtime credit in Ivory Coast. (a) Abidjan, **(b)** Liberian border, **(c)** Roads to Mali and Burkina Faso, **(d)** Road to Ghana. Figure reproduced from [70].



4 Adding time - dynamical networks

A particularity of a mobile call graph is that the links are very precisely located in time. Although each call has a precise time stamp and duration, the previously presented studies consider mobile call graphs as static networks, where edges are aggregated over time. This aggregation leads to a loss of information on the one hand about the dynamics of the links (some may appear or disappear during the collection period) but on the other hand about the dynamics *on* the links. Recently, some authors have attempted to avoid this issue by taking the dynamical component of links into account in the definition of such networks. The topic of dynamical - or temporal - networks has been studied broadly regarding several types of networks [71], but the study of mobile phone graphs as evolving ones is rather recent, and given their inherent dynamical nature, mobile call graphs are excellent sources of information for such studies.

4.1 Dynamics of structural properties

One such question regards the so-called link prediction problem, that is, predict for a future time window, whether a given link will appear or disappear from the network. This problem has already been studied in machine learning and applied to different empirical networks such as e-mails and co-authorship [72] or movie preferences [73]. Some work has already been carried out in this framework using mobile phone data for social network analysis, of which we give a short overview here.

How long does a link last in a network? By analyzing slices of 2 weeks of a mobile phone network, Hidalgo and Rodriguez-Sickert observed that the frequency of presence of links in the different slices, the *persistence*, followed a bimodal distribution [74], as illustrated on Figure 12. The persistence of link (i,j) is defined as:

$$p_{ij} = \frac{\sum_T A_{ij}(T)}{M}, \tag{13}$$

where $A_{ij}(T)$ is 1 if the link (i,j) is in slice T and 0 otherwise, M being the number of slices. Most links in the network are only present in one window, and the probability of a link to be observed in several windows decreases with the number of windows, but there is an unexpectedly large number of links that are present in all windows. These highly recurrent links represent thus strong temporally consistent relationships, in contrast with the large

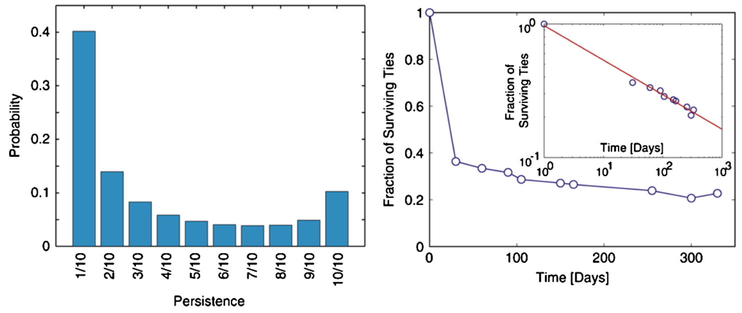


Figure 12 Measures of the strength of links over time. (Left) Distribution of the persistence of links. (Right) The fraction of surviving links as a function of time follows a power-law like decrease [74]. Figures reprinted from Phys A: Stat Mech Appl, 387(12), Hidalgo, C.A. and Rodriguez-Sickert, C. The dynamics of a mobile phone network, pp 3017-3024, Copyright (2008), with permission from Elsevier.

number of volatile connections appearing in only one of the slices. A deeper analysis of correlations between the persistence and static measures further shows that clustering, reciprocity and high topological overlap are usually associated with a strong persistence.

Raeder *et al.* [75] dig a bit deeper into that last topic, by attempting to predict which link will decay and which will persist, based on several local indicators. They quantify the information provided by each indicator with the decrease of entropy on the probability of an edge to persist, and obtain that the most informative indicators are the number of calls passed between both nodes as well as its scaled version. By trying both a decision-tree classifier and a logistic regression classifier, they manage to predict correctly about 70% of the persistent edges and decays.

While most approaches to the link prediction problem use the network structure and similarity between nodes to extract information on future interactions [74–76], Miritello goes a step further and introduces temporal components to build a new link prediction model [77]. The author defines the *temporal stability* of a link as:

$$\Delta_{ij} = t_{ij}^{\max} - t_{ij}^{\min}, \tag{14}$$

where $t_{ij}^{\min}$ and $t_{ij}^{\max}$ are, respectively, the time instants at which the link has first and last been active in the observation period T. High values of temporal stability, close to T, indicate that the link may be active beyond the time window of observation, whereas small values of Δ_{ij} may indicate that the link was only active for a short time. Miritello further shows that introducing temporal components in a link prediction model significantly improves the performance, and presents a threshold model achieving 80% accuracy in predicting whether a link will decay or persist in a future time window.

Addressing a related question, Miritello also observed that it is difficult to distinguish between a link that has decayed and a link that simply has not been observed for a long time, especially when the time window of observation is only a few months long [77].

On a very close topic, Karsai *et al.* studied how the weights of the links in a network vary with time, how strong ties form, and how this process is related to the formation of new ties [78]. They start by measuring the probability $p_k(n)$ that the next communication of an individual that has degree n will occur with the formation of a new $(n + 1)$th tie. This probability depends on the parameter k that corresponds to the final degree of the individual at the end of the observation period. They find that the process of the formation of new ties follows a very consistent pattern, namely

$$p_k(n) = \frac{c(k)}{n + c(k)}, \tag{15}$$

where $c(k)$ is an offset constant that depends on the degree k considered. Using the measured c for each degree class, the authors then show that rescaling the distributions $p_k(n)$ allows to collapse all curves into one (see Figure 13), suggesting that the evolution of the ego-network of each individual is governed by roughly the same mechanism.

The reasons for the decay and persistence of links remain various and unknown. However, Miritello *et al.* addressed a related question, namely how many links can a person maintain active in time [79]? By looking at a large time-window (around 19 months of data), they evaluate how many contacts are new acquaintances, and how many ties are

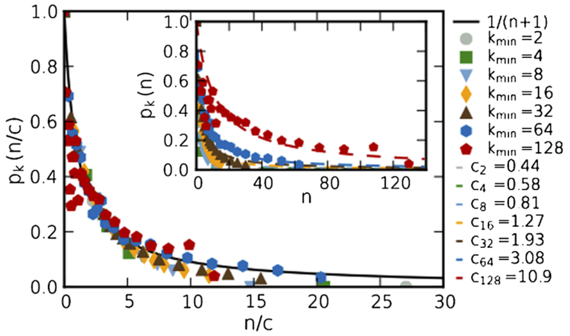


Figure 13 Probability of a new communication to form a new tie. Probability functions $p_k(n)$ calculated for different degree groups. In the inset, symbols show the averaged $p_k(n)$ values for groups of nodes with degrees between the corresponding k_{min} values. Figure reprinted by permission from Macmillan Publishers Ltd: Sci Rep [78], copyright (2014).

de-activated during a smaller time-window. It appears that individuals show a finite communication capacity, limiting the number of ties that they are able to maintain active in time: in the network of a single user, the number of active ties remains approximately constant on the long term. From a social point of view, apart from the balanced social strategy between a user's communication capacity and activity, the authors discern between two kinds of rather extreme behavior that they name *social explorer* and *social keeper*. While the social explorer shows a very high turnover in his social contacts and has a very high activity compared to his capacity, keeping only a very little stable network, the social keeper has a very stable social circle, and only has a very small pace of activating and deactivating ties. The authors further show that the social strategy of an individual can be linked to the topology of its local network. In a related paper, Miritello *et al.* [80] further show that even though people who have a large network tend to spend more time on the phone than those who have few contacts, the total communication time seems to reach a maximum, and the strength of ties starts decaying for people who have more than 40 contacts.

Despite this turnover in links and the fact that links appear and disappear, there seems to be some consistency in a person's network of contacts. In a related study, Saramäki *et al.* showed how a turnover in contacts did not imply a change in the structure of the local network around a person [81]. They study a network of students who, during the time window covered by the dataset, move from high school to college. Despite the very high turnover in a user's contacts, the distribution of the weights on the links around the user, that the authors call the *social signature* of this user, stays very similar through time.

From an evolving network perspective, the question of stability and survival of communities is closely linked to the previous questions. Palla *et al.* studied the temporal stability of a mobile phone network [34], analyzing communities detected on slices of two weeks. They observed that communities have different conditions to survive, depending on their size; small communities require to be stable, while large groups require to be highly dynamic and often change their composition. Recently, several community detection techniques developed for static graphs have been extended to take into account the dynamics of human interactions [82]. One approach of this question is to detect communities in a multislice network, where each slice represents the network at a given point in time, and nodes of a slice are linked to their counterparts in adjacent slices [83]. However, to our knowledge this approach hasn't yet been applied to mobile phone networks. Using another approach on the Reality Mining dataset, Xu *et al.* detect communities using evolving adjacency matrices [84], and show that this approach gives more consistency to the results than detecting communities independently in subsequent snapshots of the network [85].

On a shorter time scale, Kovanen *et al.* identified temporal motifs of sequences of adjacent events involving a small number of nodes (typically 3 or 4) [86]. Events are said to be Δt-adjacent if they have at least one node in common, and the timing between the two events is less than Δt (typically of the order of minutes). The authors analyze the most common motifs present in a mobile phone database and find that the most common temporal motifs of three events involve only two nodes, and motifs that allow a causal hypothesis are more frequent than those that do not.

The availability of timestamps in datasets allows to segment the calls between office hours and home hours. By supposing that calls made during office hours are for a purpose of business, while private calls are made early morning, in the evening or over the weekend, Kirkpatrick *et al.* managed to build two separate networks based on a mobile and landline dataset from the UK [87]. The degree and clustering coefficient distributions of both networks are mostly similar, but a deeper analysis of the network structure shows that some important differences exist between them. By decomposing the network into k-cores and monitoring the speed of information diffusion, they observe that the work network is much more con nected than the leisure network, and that information diffuses almost twice as fast. Addressing a related question, Pielot *et al.* present a method to predict the attentiveness to an instant message, that is, the time it will take to the receiver to attend the message [88]. They use data of the user's interaction with the smartphone to predict with 70% accuracy how fast the receiver will pay attention to the communication.

4.2 Burstiness

The dynamics of many random systems are modeled by a Poisson process, where the average interval between two events is distributed following an exponential, well-characterized by its average. However, it has appeared that human interactions show a different temporal pattern, with many interactions happening in very short times, separated by less frequent long waiting times [89].

The same holds for mobile phone calls. Karsai *et al.* studied the implications of the bursty patterns on the links of a mobile call graph [90]. They observed that indeed, the inter-event time ranges over a multiple orders of magnitude, and in particular, the burstiness of human communication induces long waiting times, which slows down the spreading of information over the network (see Section 6 for more results on spreading processes). In a further paper [91], Karsai *et al.* also analyzed the distribution of numbers of events in bursty cascades, thus better explaining the correlations and heterogeneities in temporal sequences that arise from the effects of memory in the timing of events. In another study, Wu *et al.* find that the distribution of times between two consecutive events is neither a power-law nor exponential, but rather a bimodal distribution represented by a power-law with an exponential tail [92].

It is interesting to note that in the previous papers, the authors observed the inter-event time on links, by sorting links by weight. In [93], Candia *et al.* perform a similar task but for nodes, and measure the inter-event time for nodes, by grouping them based on the number of calls they made. Similarly to Karsai *et al.*'s observations, the inter-event times range over several orders of magnitude, and the distribution is shifted to higher inter-event times for nodes of lower activity. By rescaling with the average of each distribution, the inter-event time distributions collapse into a single curve fitted by a power law with exponent 0.9 followed by an exponential cutoff at 48 days.

$$p(\Delta T) = (\Delta T)^{-\alpha} exp(\Delta T/\tau_c). \tag{16}$$

In a further paper, Karsai *et al.* study bursty trains, and show that the burstiness observed in communication networks is mainly a link property, rather than a node property [94]. They show that bursty trains usually involve the same pair of individuals, rather than one node and several of their neighbors. They further observe that within those bursty trains, there is a strong imbalance within a link with respect to who initiates the communication when voice calls are observed, while trains of SMSs are much more balanced.

The origin of this burstiness in human behavior has been discussed in several papers in the last few years. It is expected, for example, that people will have more activity during the daytime than at night, and that some times of the day will represent peaks of activity [95]. Therefore, could the burstiness of phone calls only be due to the daily patterns present in our lives? Jo *et al.* studied this question and looked at how much of the burstiness of events still remained if they removed the circadian and weekly patterns that appear in a mobile phone dataset [96]. They dilated (contracted) the time of their dataset at times of high (low) activity. They observed that much of the burstiness remained after removing the circadian and weekly patterns, indicating that there is probably another cause of burstiness coming from the mechanisms of correlated patterns of human behavior. Another hypothesis was suggested by Barabási [89] who suggests that the burstiness comes from task prioritizing in human behavior: if an individual decides to always do first the highest priority task on their list, then the high-priority tasks will be executed soon after their arrival on the list, while lower-priority tasks will stay on the list for a much longer time, waiting until all higher priority tasks are executed. This process leads to a fat tailed distribution of waiting times, as was shown in [89].

Mobile phone networks are composed of complex patterns and interactions, but still only little work has been done yet in order to characterize these interactions. The temporal arrival and disappearance of more complex structures than simple edges and the timescales of human communication are only two examples of the wide possible research that still needs to be explored in this matter.

5 Combining space and time - mobility

Given their portability, mobile phones are trusty devices to record mobility traces of users. The availability of spatio-temporal information of mobile phone users has already led to a tremendous number of research projects, and potential applications (see Section 7) which would be too large to review exhaustively here. The increasing number of smartphone applications that offer services based on the geolocation of the user are a proof that this information still has a lot of potential uses that are yet to be discovered. In this section, we concentrate on the contributions that present new observations or methods for analyzing and modeling human mobility, while the contributions that propose new applications or uses of these methods are presented in Section 7.

5.1 Individual mobility is far from random

A mobility trace is represented as a sequence of cell phone towers at which a specific user has been recorded while making a phone call. By studying the traces of 100,000 mobile

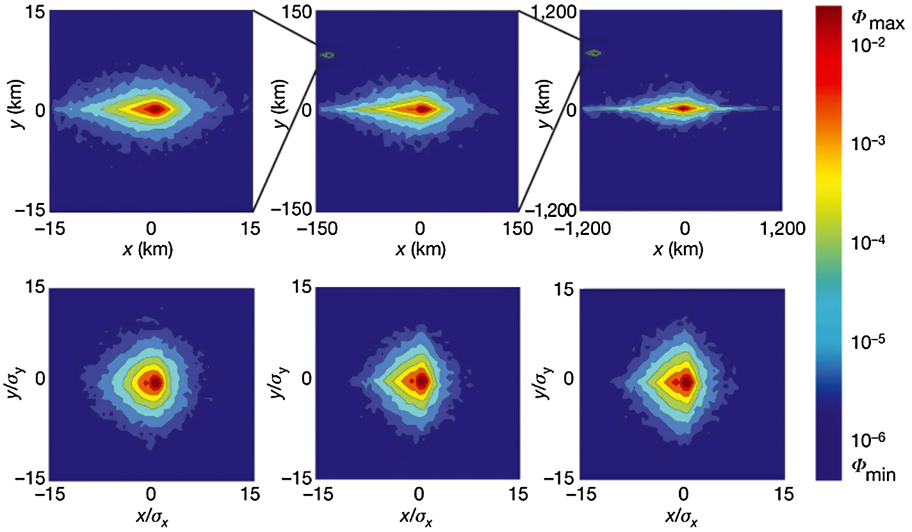


Figure 14 Probability of finding a mobile phone user in a specific location. Probability density function $\Phi(x, y)$ of finding a mobile phone user in location (x, y). The plots, from left to right, were generated for users having a different radius of gyration. After rescaling based on the variance of each distribution, the resulting distribution show approximately the same shape. Figure reprinted by permission from Macmillan Publishers Ltd: Nature [97], copyright (2008)

phone users over 6 months, González *et al.* found that human trajectories show a high degree of temporal and spatial regularity [97], as illustrated on Figure 14. This result contrasts with usual approximations of human motions by random walks or Lévy flights. Their main results showed that all users show very similar patterns of motion, up to a parameter defining their radius of gyration. The regularity is mainly due to the fact that users spend most of their time in a small number of locations. If rescaled and oriented following its principal axis, the mobility of all users can then be described by a single function. These findings are supported by an additional work produced by Song *et al.* [98], who identify significant differences between observational data and two typical models of human displacement: the continuous time random walk and the Lévy flight. Instead, the authors show that a model mixing the propensity of users to return to previously visited locations and a drift for exploration manages to reproduce characteristics present in their data but absent from traditional models. In their model, each time a user decides to change location, they can either choose a new location with a probability that decreases with the number of already visited locations ($p_{\text{new}} \propto S^{-\gamma}$, where S is the number of visited locations, and γ a constant), or they can return to a previously visited location. Despite the simplicity of this model they manage to explain the temporal growth of the number of distinct locations, the shape of the probability distribution of presence in each location, and the slowness of diffusion.

In another approach, Csáji *et al.* show how small the number of frequently visited locations is [99]. They define a frequently visited location of a user as a place where more than 5% of phone calls were initiated. Using a sample of 100,000 users randomly chosen in a dataset of communications of Portugal, the authors find that the average number of frequently visited locations is only 2.14, and that 95% of the users visit frequently less than 4 locations. Instead of making a list of frequently visited locations, Bagrow *et al.* propose another method to group frequently visited locations representing recurrent mobility into one 'habitat' [100]. The primary 'habitats' will therefore capture the typical daily mobility, and subsidiary 'habitats' will represent occasional travel. Interestingly, they show that the mobility within each habitat presents universal scaling patterns and that the radius of gyration of motion within a habitat is usually an order of magnitude smaller than that of the total mobility.

However synchronized and predictable the mobility of most countries presented here seem to be, most of these studies are based on data from developed countries, where the cultural and lingual diversity do not play as big a role as in the developing world. Amini *et al.* analyze and quantify the differences between mobility patterns in Portugal and Ivory Coast, and show that models that perform well for developed countries can be challenged by the cultural and lingual diversity of Ivory Coast, that counts 60 distinct tribes [101]. They show, for example, that commuters in Ivory Coast tend to travel much longer distances than their counterparts in Portugal, and that mobility patterns vary much more across the country in Ivory Coast than in Portugal.

If mobility traces are not random, and if users often return to their previous visited locations, could one state that human mobility could be predicted? Song *et al.* [102] addressed this question and investigated to what extent one could predict the subsequent location of a user based on the sequence of his previous visited locations. This predictability is given by the entropy rate of the sequence of locations at which the user is observed. Importantly, one has to point out that not only the frequency of visits at each location is taken into account, but also the temporal correlations between those visits. Their results show that the temporal correlations of the users' displacements reduces drastically the uncertainty on the presence of a mobile phone user, see Figure 15. Using Fano's inequality, they deduce that an appropriate algorithm could predict up to 93% of a user's location on average. The most surprising finding is that not only users are highly predictable on average, but this predictability remains constant across the whole population, whatever distance users are used to travel. While one would expect that people traveling often and far would be less predictable than those who stay in their neighborhood, Song's results seem to point out that there is no variation in predictability in the population.

While the aim of the previous work was to show how predictable human motion could be, the authors did not provide any prediction algorithm, keeping their contribution on the theoretical side. Calabrese *et al.* went a step further and proposed in [103] a predictive model for the location of people. Their algorithm is both based on the past trajectory of the targeted user and on a general drift of the collectivity, imposed by geographical features

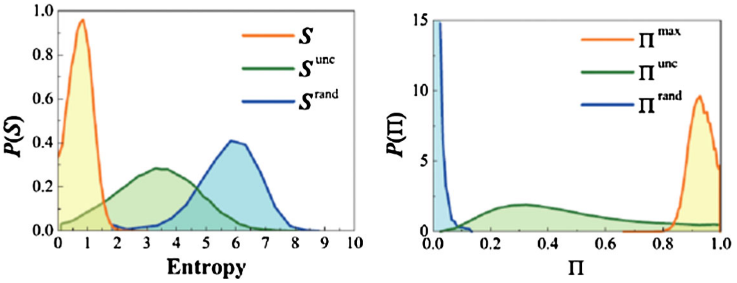


Figure 15 Entropy and predictability of the location of users. (left) Entropy rate of the location of users, for the real, uncorrelated and random data. (right) Maximal predictability of the location of users, for the real, uncorrelated and random data. From Song *et al.*, Limits of predictability in human mobility, Science 327(5968):1018 (2010) [102]. Reprinted with permission from AAAS.

and points of interest. The prediction is then a weighted average between an individual behavior and a collective behavior. The individual behavior is modeled as a first-order approximation of the concept proposed by Song [102], building a Markov chain where states are locations visited by the user and the probability of moving from state i to state j is proportional to the number of times it has been observed in the data. The collective behavior is then modeled as a weighted average between the influence of distance, points of interest and land use. The predictions of their model on a sample of a dataset containing the records of 1 million people on 4 months shows that in 60% of their predictions, they manage to predict correctly the next location of a user.

The Markov chain approach used by Calabrese *et al.* for modeling the individual behavior is also at the base of a study proposed by Park *et al.* [104]. They showed how the temporal evolution of the radius of gyration of a user can be explained by the eigenmode analysis of the transition matrix of the Markov chain. More precisely, the eigenvectors of the transition matrix provide fine-grain information on the traces of individuals.

Instead of looking at the general mobility of people, Simini *et al.* focused on the modeling the commuting fluxes between cities, and introduced the *radiation model* [105], overcoming some of the limitations of the gravity model (recall Section 3). The radiation model is a stochastic model, assigning a person from a county i to a job of another county j with a probability depending on the estimated number of job opportunities close to the county of origin i. The estimated number of job opportunities in a given county is also a stochastic variable proportional to the total population of the county. If we name d_{ij} the distance between counties i and j, the average number of commuters between the two counties depends on the population of both counties (m_i and n_j, respectively), and of s_{ij}, representing the total population in a circle of radius d_{ij}:

$$\langle T_{ij} \rangle = T_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}, \tag{17}$$

where T_i is the total number of commuters from county i. The radiation model, however efficient, still relies on the knowledge of the distribution of the population, which may be difficult to get in some areas such as the developing world. Overcoming this limitation, Palchykov *et al.* suggest a new model using only communication patterns [106]. The *communication model* supposes that the mobility between two places i and j is a function of the distance d_{ij} separating the two locations, and of the intensity of communication between these two locations, c_{ij}:

$$T_{ij} = k \frac{c_{ij}}{d_{ij}^{\beta}}, \tag{18}$$

where k is a normalization constant. The authors find fitting values for the parameter β around 0.98 or 1.08 depending on whether they consider the mobility at intra- or inter-city level, respectively.

As it appears, the massive amount of mobility data, which would on first view be considered as random motion, respects a strict routine. Mathematical models, prediction algorithms and visualization tools (see for example Martino's work [107]) have recently shed light on this routine, allowing to construct better human displacement models which can be used to predict epidemics outbreaks. At individual level, this routine appears to be strictly ruling our daily behavior, as Eagle and Pentland [108] show that six eigenvectors of the mobility patterns of users are sufficient to reconstruct 90% of the variance observed. They also observed that individuals tend to have synchronized behaviors, which will be described in the next paragraph.

5.2 Aggregate mobility reveal synchronized behavior of populations

At a higher level, those datasets allow to consider whole populations from a God-eye point of view. More practically, the availability of such massive data allows us first to observe and quantify the interaction of people with their environment, and second to quantify the synchronicity of those interactions.

Initial projects, such as the Mobile Landscapes [109] project and Real Time Rome [110] have shed light on the potential of such an approach, contributions being essentially visual. However, the next step has been made by Reades *et al.* [111], who used tower signals as a digital signature of the neighborhood. They showed how similar locations presented similar signatures, which implies that a clustering of the urban space is possible, based on the phone usage recorded by its antennas. In particular, the obtained clusters reveal known segmentations of the town, such as residential areas, commercial areas, bars or parks. In short, such a technique may be used as a cheap census method on area usage, which could be of great interest to local authorities. Going a bit further, the same team showed how using an eigendecomposition [112] of the signatures of different locations in town it is possible to extract significant information on differences and similarities in space usage, see Figure 16 for the four principal eigenvectors of the signature of a weekday. With the same goal in mind, Csáji *et al.* [99] used a k-means clustering algorithm on the activity patterns of different areas to detect which places show the same weekly calling patterns,

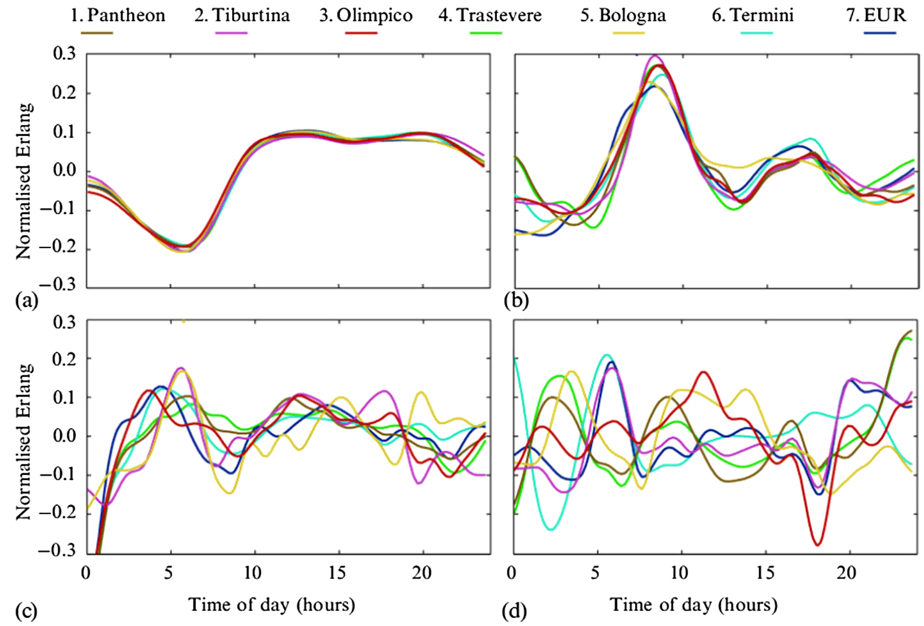


Figure 16 Eigenvectors of the Erlang signature of a weekday. Four principal eigenvectors of the Erlang signature for a weekday of 7 places in Rome. While most of the variance is dominated by the principal eigenvector, representing the normal daily activity, the differences between other eigenvectors indicate differences in space usage. Figure reproduced from [112].

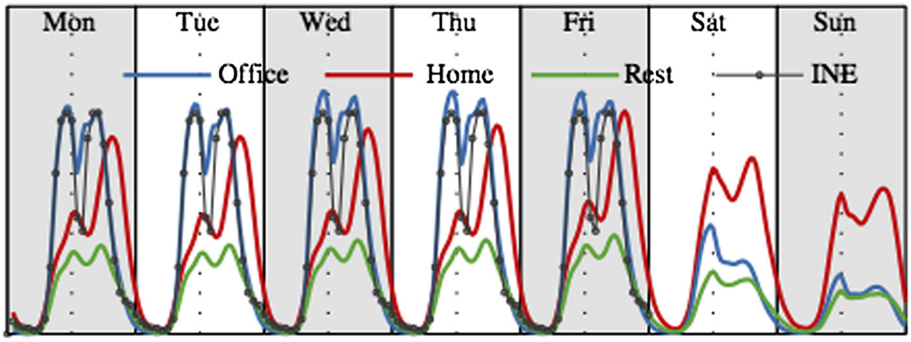


Figure 17 Weekly pattern of clusters. We observe clear differences between calling behavior of work and home locations. Figure reproduced from [99].

and thus identify which places typically correspond to work or home calling patterns (see Figure 17).

Addressing a closely related question, Karikoski and Soikkeli studied data collected from smartphones in the context of the OtaSizzle project at Aalto University, where users agreed to share their data [113]. The authors study whether different contexts trigger different usage patterns of smartphones. From the mobility traces of users, they classify places where the user is observed between: home, work, other meaningful, and elsewhere, the latter representing only pass-by places. They are able to show that depending on the context, users will have different usage patterns. For example, they show that voice calls are longer and more intensively used when people are at home, and that SMSs are more popular in the office context, where the voice calls are the shortest. In a paper studying the same dataset, Jo *et al.* study the contextual and temporal correlations between service usage and thus characterize typical usage patterns of smartphone services [114]. The authors further use k-means clustering to extract typical weekly behavior, and thus classify users between morning-type and evening-type usage patterns. Addressing a very closely related question, Trestian *et al.* show that the mobility and locations of people also influence the choice of smartphone applications they use [115]. Using a similar approach, Naboulsi *et al.* classify call profiles of snapshots of the network, corresponding to an aggregation of the traffic going through the network during a given time window [116]. They measure the similarity between two snapshots, comparing volumes and distribution of the traffic through the network. They further extract typical usage patterns, and propose a method to detect outlying behavior in the network. It is interesting to notice that even though the methods are very similar, this last approach is only based on antenna-to-antenna traffic, and not on individual behavior and mobility patterns, as were the previous studies.

Beyond the analysis of a single city, Isaacman *et al.* explored behavioral differences between inhabitants of different cities [117]. By analyzing the mobility of hundreds of thousands of inhabitants of Los Angeles and New York City, they showed that Angelenos travel on average twice as far as New Yorkers. Finding an explanation for such a significant difference seems possible, if the inhomogeneities of population density and city surfaces are taken into account. See, for example the work of Noulas *et al.* [53], who show using Foursquare location data that using a rank-based distance, the differences between cities are leveled. A rank-based distance measures the distance between two places i and j as the number of potential opportunities (people, places of interest) being closer to i than *j*. Given the geographic distance r_{ij} and the density of opportunities expressed in radial coordinates and centered in *i*, $p_i(r,\theta)$, such a distance reads

$$\operatorname{rank}(i,j) = \int_0^{2\pi} \int_0^{r_{ij}} p_i(r,\theta) r \, dr \, d\theta. \tag{19}$$

In a city of large population density, there will be more opportunities at short geographical distance than in a city with low population density. Hence, users are likely to travel over shorter distances in city of large population density. These distortions of the use of geographical distance are here leveled by the rank-based distance. In a recent study, Louail *et al.* suggest another way to formalize these differences and analyze the spatial structure of cities by detecting hot-spots or points of interest in 31 spanish metropolitan areas [118]. The authors show that the average distance between individuals evolves during the day, highlighting the spatial structure of the hot spots and the differences and similarities between different types of cities. They distinguish between cities that are *monocentric* where the spatial distribution is dependent on land use, and *polycentric* cities where spatial mixing between land uses is more important. In a similar approach, Trasarti *et al.* also analyze the correlations that arise in terms of co-variations of the local density of people, and uncover highly correlated temporal variations of population, at the city level but also at the country level [119].

If the detection of the hot-spots and places of interest in a city is possible, then is it possible to go one step further and infer the type of activity that people engage in, from looking at their mobility patterns ? Jiang *et al.* present a first approach to achieve this in [120], by first extracting and characterizing areas where people will stay or only pass-by, and then infer the type of activity that they engage in depending on the timing of their visit to certain specific locations. In many cases, modeling the mobility of users starts by creating an Origin-Destination matrix that represents how many people will travel between a specific pair of (origin, destination) locations within a given time frame [121–123]. After extracting which places and times of the day correspond to which activities, Alexander *et al.* propose a method to estimate OD-matrices depending on the time of the day and on the purpose of the trip. The authors' results extracted from data in the area of Boston, are surprisingly consistent with several travel survey sources.

5.3 Extreme situation monitoring

If the availability of data containing the time-stamped activity of a large population allows to perform monitoring of routine in population activities, it also enables to observe the population's collective response to emergencies. Many recent papers addressed this interesting question. Candia *et al.*, for first, focused on the temporal activity of users at antennas [93]. They propose a method that is based on the study of the statistical fluctuations of individual users behaviors with respect to their average behavior. As shown on Figure 18, in an anomalous case, users show many high fluctuations from their average, while the overall average is close to that of a normal activity. The variance

$$\sigma(a,t,T) = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}\bigl(n_i(a,t,T) - \langle n(a,t,T)\rangle\bigr)^2} \tag{20}$$

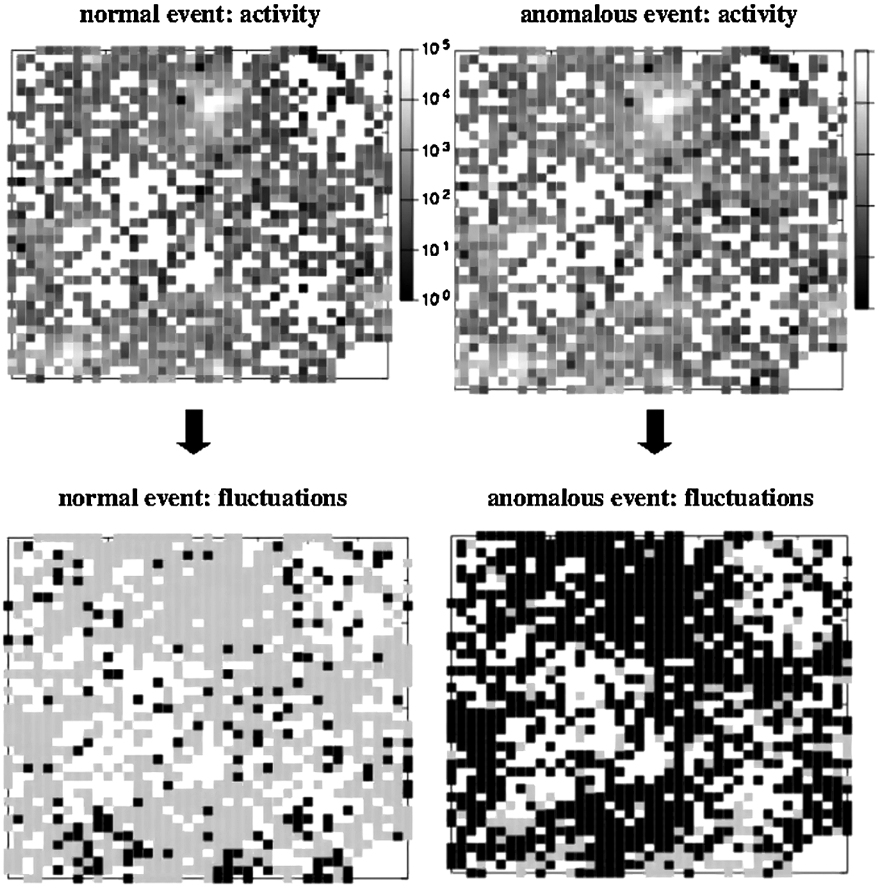
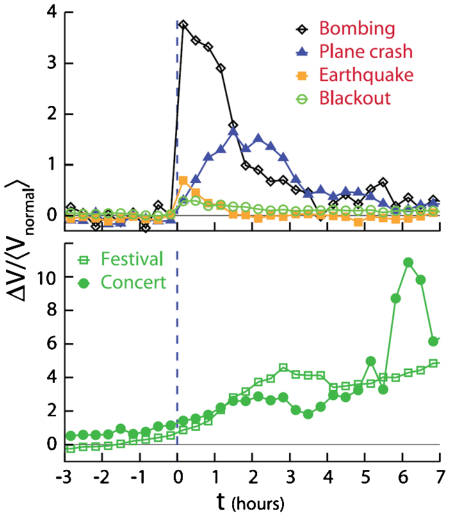


Figure 18 Activity and fluctuations during anomalous events. Activity (top) and fluctuations (bottom) for a normal day (left) and an anomalous event (right). Note that even if no difference is observed on activity, fluctuations are significantly different. Figure reproduced from [93]. ©IOP Publishing. Reproduced by permission of IOP Publishing. All rights reserved.

is computed for each place a, for the time interval $[t, t + T]$ between the different individual behaviors $n_i(a, t, T)$ and the average expected behavior. Comparing this variance with the normally expected variance allows to identify locations where users are acting abnormally, and that such locations are, in case of emergencies, spatially clustered. In cases of extreme emergencies, the response of populations can even be monitored as geographically and temporally located spikes of activity.

In a related paper, Bagrow *et al.* [124] analyzed the reaction of populations to different emergency situations, such as a bombing, a plane crash or an earthquake (Figure 19). They observed such spikes of information when eye witnesses and their neighbors reacted almost directly after the event. The reaction was mostly driven by calls made by nodes who don't usually call at that time, rather than an increase of call rate of usually active nodes. A detailed study of the paths followed by the information during its propagation shows the efficiency of the collective response, with 3 to 4 degrees from eye witnesses being contacted within minutes after the situation. Gao *et al.* further analyzed these dynamics in [125], and observed that the reciprocity of calls, *i.e.*, 'call-back' actions, showed a sharp increase in emergency cases, such as a bombing or plane crash. The same kind of spikes of behavior, though with different characteristics, are also known to appear at

Figure 19 Spikes of activity during emergency situations. The activity has been recorded for users close to the center of activity of several emergency situations, relatively to the normal activity. Figure reproduced from [124].



large-scale events, such as concerts or demonstrations [125, 126]. Altshuler *et al.* have recently also introduced another method they call the *social amplifier* to detect anomalous behavior and thus detect emergencies [127]. Hubs of the network are nodes that have a very high degree, and are thus very well connected to the rest of the network, enabling them to amplify the diffusion of information through the social graph. Using those particular nodes as social amplifiers, the authors show that only analyzing the local behavior of nodes that are close to the hubs of the network can be efficient to detect anomalies of the whole network, and thus detect emergencies. This approach has the advantage that only keeping an eye on a limited fraction of the network is computationally much easier than monitoring and keeping updates on the whole network activity.

Further than *detecting* emergencies, Lu *et al.* studied whether the mobility of populations after a disaster could be predicted, analyzing as case study the mobility of populations before and after the 2010 Haiti earthquake [128]. Interestingly, the predictability of people's trajectories remained high and even increased in the three months following the earthquake. The authors also show that the destinations of people who left the capital were highly correlated with their previous mobility patterns, and thus that, with further research, mobile phone data could be used in the future to monitor extreme situations and predict the movements of populations after natural disasters. These results are very encouraging for many humanitarian organizations who are now trying to use Big Data to save lives. After the earthquake and the following tsunami that struck Japan in 2011, several research teams started a project together combining several big data sources, such as GPS devices, mobile phones, Twitter or Facebook to analyze how the analysis of this data could help save lives in the future, if natural disasters were to strike these regions again. Similar research has been conducted by Kryvasheyeu *et al.* analyzing Twitter data during and after the hurricane Sandy in 2012 measuring the performance of friendship links to raise awareness [129]. This area of research still needs to be explored, especially as so many data sources are now becoming available, combining datasets could prove very useful, and even life-saving for some people.

5.4 Mobility and social ties

The common availability of mobility traces and social interactions in the same dataset allows to address causality questions on the creation of social links. From the work of Calabrese *et al.* it appears that users who call each other have almost always physically met at least once over a one year interval [130]. Users call each other mostly right before or after physical co-location, and interestingly, the frequency of meetings between users is highly correlated with their frequency of calls as well as with the distance separating them.

Going a step further, one may wonder if social ties could be predicted using mobility data. Wang *et al.* [131] showed that indeed, nodes that are not connected in the network, but topologically close, and who show similar mobility patterns are likely to create a link. By combining the mobility similarity and the topological distances in a decision-tree classifier, they manage to improve significantly classical link prediction algorithms, yielding in an average precision of 75% and a recall of 66%. Closely related, Eagle *et al.* showed on 4 years of data how the social network of people changes drastically when moving from one geographical environment to another [132].

On a related topic, Toole *et al.* measure the similarity between the mobility of users to classify social relationships and show how to contextualize social contacts using their mobility patterns [133]. The authors further present a mobility model, based on stochastic decisions to return to a previously visited place or to explore, and to base the choice on social influence or on individual preference. They show that this model achieves good accuracy in reproducing the similarity of mobility traces between social contacts.

6 Dynamics on mobile phone networks

Many networks represent a transport between nodes via their links. In mobile phone networks, the links transport either information (exchanged during phone calls or contained in messages) or non-voice exchanges (SMS, MMS). Information diffusion has opened questions on the speed of the diffusion or on the presence of super-spreaders, with applications in viral marketing or crowd management. The transmission of data has been at the centre of attention only recently, with the rise of new types of computer viruses running on smartphones.

6.1 Information diffusion

A phone call is associated to the transfer of information between caller and callee. However, as paradoxical as it may sound, mobile phone datasets are not appropriate to *observe* real propagations of information. The content of phone calls or text messages is, for evident privacy reasons, unknown. Yet, without having access to the content, it is impossible to decide for sure if an observed pattern of calls reflects the transmission of information or if it happens by chance. One can imagine a network with a number of indistinguishable balls circulating between the nodes. Each time a node receives a ball from one of its neighbors, it decides to keep it for a random time interval and after that to transmit it to one of its neighbors. Suppose now that one decides to track the movement of one specific ball. If the number of balls is small compared to the number of nodes, this can still be doable, as long as each node has maximum one ball in its possession. However, if the number of balls increases to become equivalent to the number of nodes, there is a high probability to confuse the paths of several balls. Add to this that balls might be added, removed or duplicated during the process, and one gets a similar situation as trying to track a piece of information in a mobile phone network.

This artificial example reflects well the issue of tracking information. Peruani and Tabourier addressed this issue and showed that cascades of information, such as observed in mobile call graphs are statistically irrelevant, and correspond thus probably not to real propagations [134]. Tabourier *et al.* show in a further paper [135] that even though large cascades of information spreading don't seem to happen in mobile call graphs, local short chain-like patterns and closed loops seem to be the effects of some causality and could very well be related to information spreading.

In a small number of cases, however, the actual observation of large diffusion of information might be possible. Studying the case of emergencies, such as a plane crash or a bombing, Bagrow *et al.* [124] observed an unusual activity in the geographical neighborhood of the catastrophe. In this case, the knowledge of both the temporal and spatial localization of an unexpected event that is likely to generate a cascade of information allows to assume that the observed sequences of calls are correlated for a specific reason.

If, in most cases, the observation of real propagations seems an unreachable objective, a more complete research has been driven in the simulation of propagation of information on complex networks, which results have been extended to questions related to mobile phone networks. There are several ways of modeling information diffusion on networks. A simple way is used in [14] with an SI or SIR model where at each time step, infectious nodes try to infect their neighbors with a probability proportional to the link weight, which corresponds to a sequence of percolation processes on the network. However, mobile phone networks are known to have very particular dynamics (recall Section 4), which are not taken into account here. Miritello *et al.* [136] used a formalism similar to the one presented by Newman [137] for epidemics, to characterize the *dynamical strength* of a link, which can be used as link weight to map the dynamical process onto a static percolation problem. The dynamical strength, given an SIR model of recovery time T and probability of transmission λ, is given by

$$\mathcal{T}_{ij}[\lambda, T] = \sum_{n=0}^{\infty} P(w_{ij} = n; T)\big[1 - (1 - \lambda)^n\big], \tag{21}$$

which is the expected probability of having n calls between i and j in a time range of T multiplied by the probability of propagation given these n calls, summed over all possible values for n. Using an approximation of this expression, they manage to link the observed outbreaks to classical percolation theory tools.

However, such a formalism still neglects the impact of temporal correlations between calls, which significantly slows down the transmission of information over a network. Social networks often exhibit small-world topologies, characterized by average shortest paths between pairs of nodes being very short compared to the size of the network [138]. However, Karsai *et al.* [90] used different randomization schemes to show that even though social networks have a typical small-world topology, the temporal sequence of events significantly slows down the spreading of information, as illustrated on Figure 20. Kivelä *et al.* [139] analyze this topic further, and introduce a measure they call the *relay time*, specific to each link, that represents the time it takes for a newly infected node to spread the information through that link. By analyzing several computations of this relay time, in randomized and empirical networks, they show that the bursty behavior of

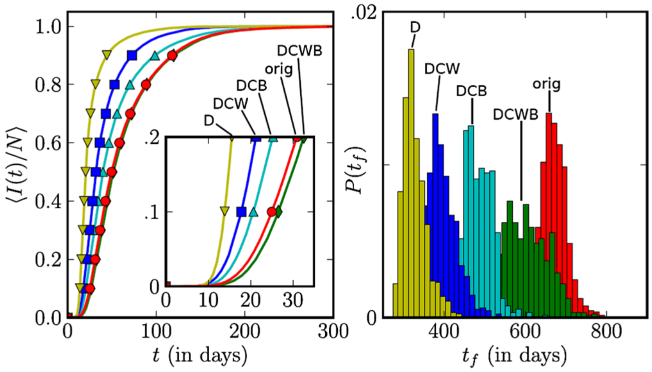


Figure 20 Comparison of the speed of spreading processes using different randomization schemes. (left) Fraction of infected nodes as a function of time for the real (red) data and different randomization schemes. (right) Average prevalence time distribution for nodes. Reprinted figure with permission from Karsai *et al.*, Phys Rev E, 83(2):025102, 2011 [90]. Copyright (2011) by the American Physical Society. http://dx.doi.org/10.1103/PhysRevE.83.025102

links, but also the broad distribution of link weights are the components that slow down the most the spreading dynamics in mobile phone networks. In another study, Karsai *et al.* [78] confirm this influence and show that neglecting the time-varying dynamics by aggregating temporal networks into their static counterparts introduces serious biases of several orders of magnitude in the time-scale and size of a spreading process unfolding on the network.

From a more theoretical point of view, diffusion processes can be seen as particular cases of dynamical systems. Liu *et al.* [140] questioned in this framework the controllability of complex networks. The problem was stated as follows; given a linear dynamical system with time-invariant dynamics

$$\frac{d\mathbf{x}(t)}{dt} = A\mathbf{x}(t) + B\mathbf{u}(t), \tag{22}$$

where $\mathbf{x}(t) = (x_1(t), \dots, x_N(t))^T$ defines the state of the nodes of the network at time t, A is the (possibly weighted) adjacency matrix of the network, and B an input matrix, what is the minimal number of nodes needed for the input such that the state of each node is controllable, *i.e.*, the system is entirely controllable? From control theory, one knows that a sufficient and necessary condition is that the *reachability matrix* $C = (B, AB, A^2B, \dots, A^{N-1}B)$ is of full rank. From previous work, it is known that the minimal number of nodes required is related to the maximal matching in the network, which can be computed with a reasonable complexity. For example, the authors show that in a mobile phone network, one needs to control about 20% of the nodes in order to achieve full controllability of the system. Surprisingly, most nodes needed for controlling the network are low-degree nodes, while hubs, that are commonly used as efficient spreaders, are under-represented in the set of input nodes. While the practical interest of this research still needs to be defined, this first result on controllability of networks might open new ideas in the field of information spreading.

Finally, one may wonder if the patterns of phone usage are efficient in a collaborative scheme. Cebrian *et al.* [141] studied this with a small model, where each node of a mobile phone graph is represented as an agent assorted with a state represented by a binary string. The agents are all given the same function f, that takes their binary string as input and which is hard to optimize, and which computes their personal score. After each communication, the two communicating agents can modify their state in order to increase their personal score. This modification is done with a simple genetic algorithm, which simulates a cross-over of the states of both agents.

Practically, suppose that two agents i and j are respectively in state $\mathbf{x}_i^{(t)}$ and $\mathbf{x}_j^{(t)}$ at time t. These states are both binary strings of length T. The agents choose a random integer c in the interval $[1, T]$ and both update their state as

$$\mathbf{x}_i^{(t+1)} = \arg \max_{x \in \{\mathbf{x}_i^{(t)}, \mathbf{y}_1, \mathbf{y}_2\}} f(x), \tag{23}$$

$$\mathbf{x}_j^{(t+1)} = \arg \max_{x \in \{\mathbf{x}_j^{(t)}, \mathbf{y}_1, \mathbf{y}_2\}} f(x) \tag{24}$$

where $\mathbf{y}_1$ is the vector with the c first entries of $\mathbf{x}_i^{(t)}$ and the $T - c$ last entries of $\mathbf{x}_j^{(t)}$ and $\mathbf{y}_2$ is the vector with the c first entries of $\mathbf{x}_j^{(t)}$ and the $T - c$ last entries of $\mathbf{x}_i^{(t)}$.

The authors observe with this model that the average score on all agents obtained in the real dataset is smaller than for a random topology, which is in line with similar known results from population genetics. Also, perturbation of the time sequence of calls produces a small enhancing of the global fitness.

6.2 Mobile viruses

The study of virus propagations has a long history, may it be biological viruses or more recently computer viruses. Wang *et al.* [142] studied a new kind of virus, which spreads over mobile phone networks. Their work is motivated by the increasing number of smartphones, which have high-level operating systems like computers, which leads to a higher risk of an outbreak. So far, despite the large number of known *mobile viruses*, no real outbreak has been noticed. The reason for this is that mobile viruses function only on the operating system for which they are designed for. An infected phone can hence only transfer the virus to its contacts running on the same operating system. As exposed by Wang *et al.* this situation corresponds to a site percolation procedure on the network of possible contacts. Given the actual market shares of the main operating systems, the authors showed that those were below the percolation transition of the contact network. The study concerns two types of spread available for viruses: the diffusion via Bluetooth and via Multimedia Messaging System (MMS). Both diffusions show major differences in spreading patterns; Bluetooth viruses spread relatively slow and depend on user mobility. In contrast, MMS epidemics spread extremely fast and can potentially reach the whole network in a short time, see Figure 21. However, currently they are contained in small parts of the network, due to the different operating systems. In conclusion, the authors deduce thus that if no outbreak has taken place so far, it is not due to the lack of efficient viruses, but it is rooted in the fragmentation of the call graph. However, the current evolution of the market leads to a situation where some operating systems are gaining a large market share, which could lead to a more risky situation.

In a subsequent study, Wang *et al.* [143] show how the scanning technique, where MMS malware generate random phone numbers to which they try to propagate instead of using the address book of their host, increases the probability of a major outbreak, even

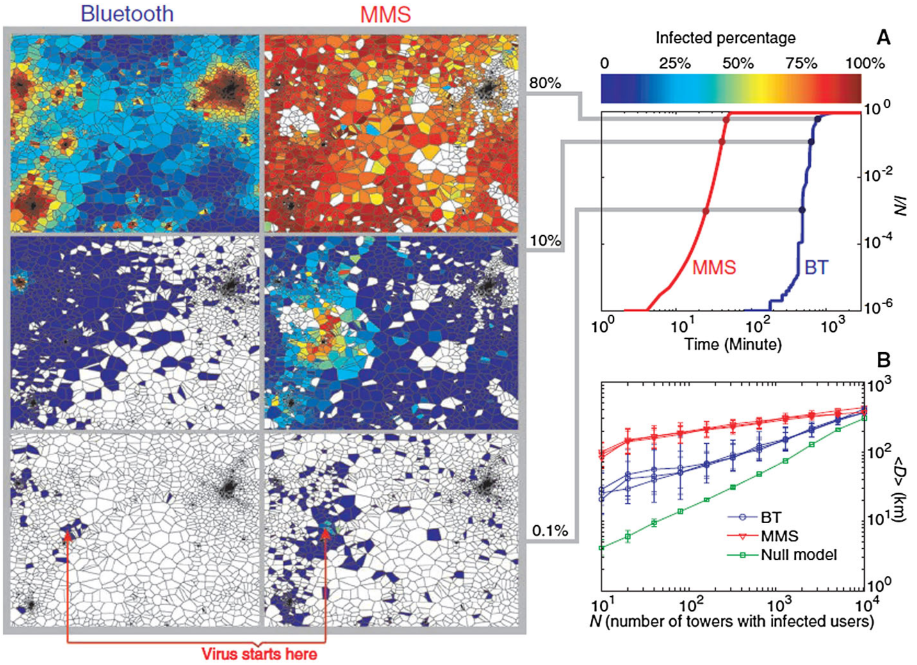


Figure 21 Propagation of a mobile virus, either via MMS or Bluetooth service, over the observed area. From Wang *et al.*, Understanding the spreading patterns of mobile phone viruses, Science 324(5930):1071 (2009) [142]. Reprinted with permission from AAAS.

when the market share of operating systems are too low for having a giant component. Operators can detect such outbreaks by monitoring the MMS traffic of their network and observe suspicious increases of volume. However, given enough time, viruses can infect a large fraction of the network without being detected by operators. Smart anomaly detection schemes may prevent such outbreaks, as well as a reduction of market shares of operating systems. Wang *et al.* also compare the last two strategies in a further paper [144]. They study the effectiveness of topological viruses versus viruses that also use a scanning technique. The authors show that topological viruses, *i.e.*, those that spread through the contact network of infected phones, are the most effective for an operating system that has a large market share, whereas the scanning technique will generate a bigger outbreak in the case of a low market share operating system.

7 Applications in urban sensing, epidemics, development

The last few years have seen the rise of Big Data and of its uses, and in many regards, this is rapidly changing our lives and way of thinking. Further than observing those networks of mobile phone calls, or modeling social behavior, many researchers now engage in finding new ways of using mobile phone data in everyday life.

7.1 Urban sensing

As showed in the previous sections, mobile phone data allows to observe and quantify human behavior as never before. Besides purely sociological questions, this data also opens a number of potential applications, which gives to this data an intrinsic economical value, thinking of geo-localized advertising applications [145]. Recalling that an increasing fraction of the available smartphone applications record the user's geolocation - whether it is necessary for the app to work or not - it is easy to understand that this information is valuable to target the right users when making advertising campaigns, or simply to understand the profile of the application's users. Mobile phones are more and more becoming a way of taking the pulse of a population, or the pulse of a city, and we expect that in the future, more and more cities will make development plans based on information gathered from mobile phone data. In this framework, recent research has shown that mobile phone data could detect where people are [47] and where people travel to [99] including the purpose of their trips [120]. If these findings are applied to a whole city and points of interest are uncovered via mobile phone data (recall Section 5), then the whole organization of urban places can be influenced by the knowledge gained from this data. Urban sensing is only shortly addressed here, but has been a popular topic in the last few years, and we refer the interested reader to a recent survey of contributions in this specific field [146].

We have previously addressed the possibility of using mobile phone signatures as a cheap census technique, Isaacman *et al.* take this analysis a step further and show how one can derive the carbon footprint emissions [147] based on the mobility observed from mobile phone activity.

Many applications of modeling mobility aim towards transport planning and monitoring traffic with evident applications in accident management and traffic jam prevention. Over the last (almost) 20 years, a large number of attempts have been made to enhance prediction using mobile phone data. This topic is only shortly addressed here with a few recent contributions, but for more information on the research in this field, we will refer the interested reader to a review published in 2011 [148]. One example of such an application was proposed by Nanni *et al.*, who create the OD-matrix of Ivory Coast and then assign this matrix to the road network [122] to produce a map (see Figure 22) modeling the traffic of the main roads of the country, showing estimated traffic flows. In a similar approach, Toole *et al.* estimate the flow of residents between each pair of intersections of a city's road map [149]. They show that these estimations, coupled with traffic assignment methods can help estimate congestion and detect local bottlenecks in the city. In a related study, Wang *et al.* examine in more details the usage patterns of road segments, and show

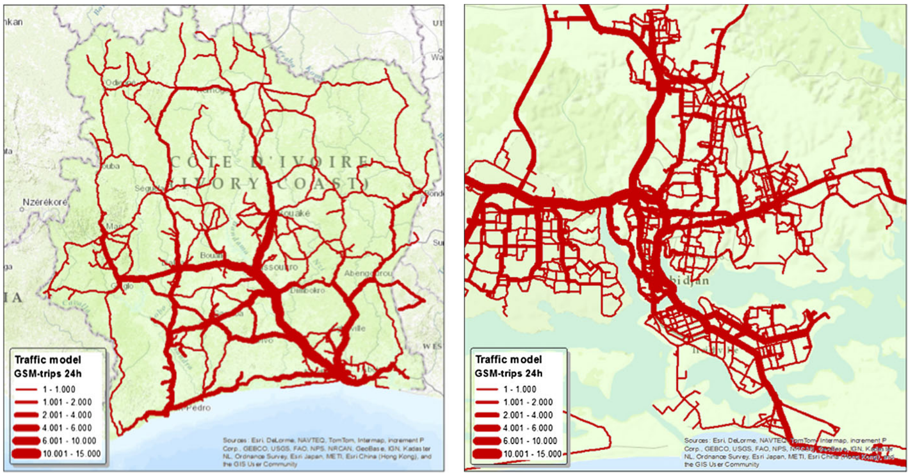

CÔTE D'IVOIRE (IVORY COAST)
Nzérékoré
Abidjan
Traffic model
GSM-trips 24h
1 - 1.000
1.001 - 2.000
2.001 - 4.000
4.001 - 6.000
6.001 - 10.000
10.001 - 15.000
Traffic model
GSM-trips 24h
1 - 1.000
1.001 - 2.000
2.001 - 4.000
4.001 - 6.000
6.001 - 10.000
10.001 - 15.000
Sources: Esri, DeLorme, NAVTEQ, TomTom, Intermap, increment P Corp., GEBCO, USGS, FAO, NPS, NRCAN, GeoBase, IGN, Kadaster NL, Ordnance Survey, Esri Japan, METI, Esri China (Hong Kong), and the GIS User Community

Figure 22 Traffic model for 24 hour period for Ivory Coast (left) and Abidjan Area (right). Figure reproduced from [122].

that a road's usage depends on its topological properties in the road network, and that roads are usually used only by people living a small number of different locations [150]. The authors further show that taking advantage of this observation helps create better strategies for reducing travel time and congestion in the road network of a city.

Going one step further, Berlingerio *et al.* designed an algorithm to detect which means of transport people would choose, including public transportation or private means, to infer how many people used which public transportation routes [121] throughout the day. The authors then proposed a model of the network of local transportation of Abidjan highlighting the routes that are taken most often. Then, they are able to show how specific little changes to the network could improve the average travel time of commuters by 10%. Among other possible uses of information on commuting flows, McInerney *et al.* suggested using the regular mobility of people for physical packages delivery to the most rural areas [151], showing on the one hand, the feasibility of this method, and on the other hand reducing by 83% the total delivery time for rural areas. Other applications of prediction algorithms for the next journey of users include, for example, a recommender system for bush taxis such as suggested by Gambs *et al.* [152], using the predicted next location of users to recommend to pedestrians adapted means of transport that are in their neighborhood.

By monitoring the movements of people towards special planned events, Calabrese *et al.* [153] show that the type of events highly correlates to the neighborhood of origin of the users. Such a cartography of taste can be used by authorities when planning the congestion effects of large events, or for targeted advertising of events (see Quercia *et al.* [154]). In a closely related approach, Cloquet and Blondel use the analysis of anomalous behavior in mobile phone activity to predict the attendance to large-scale events such as demonstrations or concerts. The authors propose, as a first step in that direction, a method to determine the time when no more people will arrive to a certain event [155]. To do this, they propose two methods. The first method uses the mobility of people that are traveling towards the event to model the flux of the arriving or leaving crowd. The second method is based on the recorded interactions between people that are already at the event and other users that are within 20 km. The authors show that using these methods, they are able to predict the time when no more people will join the event up to 43 minutes in advance. Another related application was explored by Xavier *et al.* who analyzed the workload dynamics of a telecommunication operator before and after an event such as a soccer match [156] in order to help the management of mobile phone networks during such events.

Finally, mobility traces can also be used to monitor temporal populations [157], such as tourists. Kuusik *et al.* [158] studied the mobility of roaming numbers in Estonia for 5 consecutive years, showing the potential for authorities to understand and efficiently target visiting tourists.

7.2 Infectious diseases

In recent years, a lot of research has been done in order to use Big Data to help monitor and prevent epidemics of infectious diseases. If one can model information spreading in mobile phone networks (recall Section 6), then the same theory could also be used to model the spreading of real infectious diseases. As mobile phone data can help follow the movements of people (recall Section 5), these movements can also provide information about how a disease could travel and spread across a country. The dynamics at hand usually depend on the type of disease and how it can be transmitted, hence many articles, of

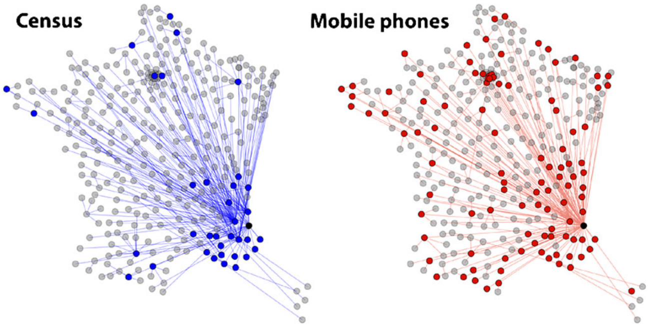


Figure 23 Epidemic invasion trees. Invasion trees observed using the census (left) and the mobile phone network (right), the seed of the simulation is in Barcelonnette (black node). Figure reproduced from [160].

which we will review a few here, propose different models based on the mobility of people to predict the spread of an epidemic.

Using mobile phone traces, Wesolowski *et al.* measure the impact of human mobility on malaria, comparing the mobility of mobile phone users to the prevalence of malaria in different regions of Kenya, and identify the main importation routes that contribute to the spreading of malaria [159]. In another study, Tizzoni *et al.* [160] validate the use of mobile phone data as proxy for modeling epidemics. The authors extract a network of commuters in three European countries by detecting home and work locations for each mobile phone user, and compare this network with the numbers of commuters obtained by census. On these networks of commuters, they trace agent-based simulations of epidemics spreading across the country. They show that the invasion trees and spatio-temporal evolution of epidemics are similar in both census and mobile phone extracted networks of commuters (see Figure 23). Most models assume, lacking additional information, homogenous mixing between people that are physically within the same region or area. Frias-Martinez *et al.* propose another agent-based model of epidemic spreading, using individual mobility *and* social networks of individuals to build a more realistic model [161]. Instead of assuming homogeneous mixing within a given area, an individual will have more probability of meeting an infected agent that is in the same area if they have communicated with each other before. The authors further divide the social network of contacts and the mobility model of an individual between weekday and weekend to achieve better accuracy.

Going a step further, a few contributions to the D4D challenge [162] investigated which would be the best ways to monitor and *influence* an epidemic rather than just predicting its spread. In this framework, Kafsi *et al.* [163] propose a series of measures applicable at the individual level that could help limit the epidemic. They investigate the effect of three different recommendations, namely (1) do not cross community boundaries; (2) stay with your social circle and (3) go/stay home. Considering that either of these three recommendations could be sent via their mobile phone to different users in the network, and that probably only a fraction of the contacted users would participate, the authors evaluate the impact that implementing this system could have on the spreading process. They show that these measures can weaken the epidemic's intensity, delay its peak, and in some regions, even seriously limit the number of infected individuals. Using the same dataset, Lima *et al.* proposed a different approach [164], namely using the connection between people to launch an information campaign about the epidemic, in the hope to reduce the probability of infection if an individual is better informed about the risks. The authors use an SIR model and the observed mobility of mobile phone users to simulate epidemics unfolding on a population, and evaluate the impact of geographic quarantine on the spreading of the disease, as well as the impact of an information campaign reducing the risks of infection for 'aware' individuals. They show that the quarantine measures don't seem to delay the endemic state, even when almost half the population is limited to their own subprefecture, whereas the information campaign, less invasive, seems to limit significantly the final fraction of infected individuals, opening this topic for further research.

This field of research has shown again how valuable mobile phone data could be to save lives, and potentially monitor and limit epidemics of infectious diseases. However, most models and studies are limited by the lack of ground-truth data to compare their results with. Indeed, how would you know who an individual got the disease from, and what was its exact route towards each infected person? Another shortcoming of this area of research comes from the current difficulty of gaining access to those mobile phone datasets, especially to cross-border mobility. If modeling mobility in Africa could be useful to containing the current Ebola outbreak, cross-border mobility would be very valuable data, as discussed in [165]. However, gaining access to these data is more difficult as it involves getting the approval from more than one country for a single dataset. In [166], the authors suggest guidelines to share data for humanitarian use, while preserving the privacy of users.

7.3 Health and stress detection

While infectious diseases are still a major cause of death in developing countries, the attention has slowly shifted, in more developed parts of the world, towards chronic diseases such as cardiovascular diseases or cancer, and their causes. Among one of the studied topics, daily stress in the work environment has become a major problem in the recent years. In this framework, Bogomolov *et al.* have conducted an experiment to find out whether daily stress levels could be predicted from non-invasive sensors, including mobile phone data [167, 168]. Using only one source of data resulted in poor predictive capacity. However, combining mobile phone data with features of personality traits and weather conditions, they produced a predictive model using 32-dimensional feature vectors to classify users between 'stressed' and 'not-stressed', achieving 72% accuracy. Interestingly, among the features extracted from mobile phone data that were selected as useful for the model, many were bluetooth proximity features.

7.4 Viral marketing

In 1970, Katz and Lazarsfeld introduced the breakthrough idea that, more than mass media, the neighborhood of an individual is influencing their decisions [169]. This idea has induced the concept of opinion leaders, that is, persons who have a high influence on their neighborhood, although some debate exists on the exact role played by opinion leaders [170], and introduced the concept of viral marketing. In opposition to direct marketing, the principle of viral marketing is that consumers respond better to information accessed from a friend than to information provided through direct means of communication. Viral marketing searches thus for means of making people communicate about a brand, in order to push friends of an early adopter to adopt the product in their turn. In particular, mobile viral marketing has proved to be an effective means of propagation of such marketing campaigns. The influence of one's neighbors can be observed using CDR data coupled to data on product adoption. In a study of the adoption of 4 mobile services, Szabó and Barabási [171] showed that the adoption of a product by a user was highly correlated to the adoption of their neighbors for some services only, while other services were not showing any viral attribute. A similar study by Hill *et al.* [172] on the adoption of an undisclosed technological service showed again that neighbors of nodes that had adopted the service were 3 to 5 times more likely to adopt the service than the best-practice selection of the company's marketing service. A related result was also obtained in the FunF project by Aharony *et al.* [173], who showed that the number of common installed applications was significantly larger for pairs of users having often physical encounters. Risselada *et al.* [174] further showed that the influence of one's neighbors on the adoption of a product evolved with time, depending on the elapsed time since the introduction of the product on the market.

Even though one could use a simple SI or SIR model to characterize viral marketing, it is more likely in this case, that a user will adopt a product if several of its neighbors have already adopted it and the information comes from several different sources. One of the possible ways to model these dynamics is to use a threshold model: each user is assigned a threshold. A node will adopt a product if the proportion of its neighbors that have adopted the product is above the node's threshold. The model can be either deterministic, and decide a priori a same threshold for all nodes, or stochastic and draw thresholds from a probability distribution. To take into account the timing of contacts between people, one can then add to this model the condition that a node will adopt a product if it has enough contacts with different neighbors that have adopted the product within a given time frame. Backlund *et al.* have studied the effect of timings of call sequences on those models [175]. Here again, they observe that the burstiness of events tends to hinder propagation of adoption of a product, increasing the waiting times between contacts compared to a randomized sequence of contacts.

The identification of 'good' spreaders for a viral marketing campaign is tough work, especially given the usually very large size of the datasets, which makes it hard to extract informational data in a small time frame. With this in mind, the authors of [176] proposed a local definition of *social leaders*, nodes that are expected to play an influential role on their neighborhood. They defined the social degree of a node as the number of triangles in which the node participates, and social leaders as nodes that have a higher social degree than their neighbors. This definition has its use in marketing campaigns, to identify the customers who should be contacted to start the campaign, which proved to be efficient [177]. Moreover, social leaders can also be used to reduce the complexity of a network, by only analyzing the network of social leaders instead of the whole network, with possible uses in visualization and community detection.

7.5 Crime detection

In criminal investigations, the police often request mobile phone records of suspected individuals for inspection, looking for evidence. The analysis of such data can not only reveal behavioral patterns of a single suspected individual, but also uncover potential criminal organizations through mobile relationships. Social networks analysis allows therefore to uncover the structure of criminal networks, but also to quantify the flow of information between its members. In this framework, a research group from the university of Messina propose a toolbox called 'LogAnalysis' to analyze CDRs and the associated social networks, with the aim of detecting criminal organizations [178, 179]. This toolbox allows to measure a series of metrics of the network and of the nodes, such as node centrality or clustering coefficient, and the tool further presents an analysis of the dynamics of the graph. The authors add visualization tools to the analysis, enabling forensic analysts to easily spot nodes that are more central, or visualize clusters and sub-clusters of tightly related individuals.

The approach of this type of research is somewhat different from most studies presented in the above paragraphs on CDR datasets, in that it is not based on studying anonymized datasets and extracting information on the behavior of a population, but rather on studying the network around a specific individual or a specific group of suspects whose identity is clearly known by the forensic analyst carrying out the investigation.

Using a different approach, Bogomolov *et al.* use indicators derived from mobile phone traces to predict whether a certain area will be a crime hot spot in the next month [180]. Using dynamically updated features such as the estimated number of people in each area, or the age, gender and work/home/visitor group splits derived from mobile phone data, their model achieves almost 70% accuracy in predicting whether a given area is at risk of being the scene of a crime in the next month. This type research can therefore be used by the police to achieve a better response time, or direct their attention towards the places that are the most likely to require an intervention.

7.6 Data for development

The last couple of years have seen a spectacular rise of interest for applications of mobile data for the purpose of helping towards development. Many contributions to the 'Data for Development' (or D4D) challenge launched by Orange [162] used different bits of information from the data of mobile phone users to help the development of Ivory Coast. Several of these contributions have already been reviewed in the previous paragraphs, for the full set of research projects, see [181].

While in the developed world, much information of what can be inferred from mobile phone data is already known (population density, some of the mobility traces, ...), this information can be very valuable in the developing world where census data is often unavailable or several years old. Modeling the mobility of people in developing countries can provide very useful information for local governments when making decisions regarding changes in local transportation networks, or urban planning. Indeed, in rural areas of low income countries where the most recent technologies are not always available, up to date information on how many people commute from one place to another can be very useful and help policy makers to decide on the next steps towards development. Sometimes, very basic information such as drawing the road network can be difficult in remote places. Salnikov *et al.* used the D4D challenge dataset to detect high traffic roads by selecting displacements only within a certain range of velocities [182]. They were able to redraw the main road structure of the country and even identified unknown roads, which they validated a posteriori. Between techniques for cheap census, mobility planning and fighting infectious diseases applications, we expect that in the next few years, the developing world will profit from the availability of such rich databases, and research will provide useful insights into how to better help towards development.

7.7 Data representativity

Finally, one may raise the question of the significance of the data: given that only a fraction of a country's population is reached by one operator, to which extent may the results on a dataset be generalized to larger populations? Clearly, quantitative results obtained in these studies, such as the degrees of nodes, cannot be taken for granted, but one may expect that as long as the population sample is not biased, qualitative observations such as the broadness of degree distribution or the organization of nodes in communities are significant information on the structure of communication networks. However, the question of knowing whether the sample is biased or not is almost impossible, especially given the lack of information about the users in CDR databases.

Frias-Martinez *et al.* raised this question in [183], regarding *e.g.* the socio-economic level that could be biased among mobile phone users compared to the whole population. They validate their results by performing a series of statistical tests to compare the population in their sample to the overall population using census data, and show that no significant difference was observed. However, in the general case, data about users in CDR databases is often missing, and census data may not always be available for comparison. Regarding mobility models, one could argue that active mobile phone users are more likely to be on the move than the rest of the population. A mobility model based on mobile phone users is therefore likely to overestimate the number of people within a population that are traveling. Buckee *et al.* raised this question regarding those models, further arguing that bias in models of mobility could, in turn, influence the spreading of modeled epidemics [184]. Onnela *et al.* also address this problem studying how paths differ depending how much of the network is observed [185]. They show that, counterintuitively, paths in partially observed networks may appear shorter than they actually are in the underlying full network.

Ranjan *et al.* studied a related question regarding the mobility of users [186]: given that one only sees data points where and when a user has made a phone call, to which extent are these points representative of a user's mobility? They found that sampling only voice calls of an individual will most of the time do well to uncover locations such as home and work, but will also, in some cases, incur biases in the spatio-temporal behavior of the user. In a recent study, Stopczyncki *et al.* widen their coverage by coupling databases from many sources on the same set of users in the context of the Sensible DTU project [7]. While this approach clearly captures more than just studying mobile phone records, its coverage is limited (1,000 subjects) as the users had to give their explicit consent to share their data: facebook interactions, face-to-face encounters, and answers to a survey. The authors are therefore able to analyze a bigger picture than other studies based on only mobile phone data and show that only studying mobile phone data may not be enough to capture a user's comprehensive profile. Learning from these studies, one should therefore be cautious when drawing conclusions from such analyses, and keep in mind that observing the traces left by mobile phones is only observing selected parts of the whole picture.

8 Privacy issues

The collection and availability of personal behavioral data such as phone calls or mobility patterns raises evident questions on the security of users'privacy. The content of phone calls or text messages is not recorded, but even the simple knowledge of communication patterns between individuals or their mobility traces contains highly personal information that one typically does not want to be disclosed. During the past decade, a fairly high amount of personal data was made available to researchers via, among others, CDR datasets. The companies sharing their data do not always know how much personal information can be inferred from the analysis of such large datasets, and this has led, so far in other cases than mobile phone data, to a few scandals in the recent years [187, 188]. In turn, these incidents led, in 2012, to a procedure of adaptation of legal measures in Europe [189]: the previous european law on the protection of privacy and data sharing dated back from 1995 [190], long before the era of what is now called 'Big Data'.

Moreover, the use of data has become global, and an organization based in a specific country uses data generated by its users from all over the world, hence the need for similar regulations in different countries. So far, this has not yet been achieved, as legislations in different parts of the world are very different from one another. In the USA, for example, there is no specific law regulating the data protection and privacy, but instead laws are specific on a sector-by-sector basis. Data protection in the finance or health sectors are therefore regulated by separate authorities [191]. In Europe, on the contrary, the new directive is designed to apply everywhere in Europe, to the people and organizations who collect and manage personal data [192].

The procedure often used when a company shares private data with a third party such as a research group is the following: the company keeps on secured machines the exact private information such as names, addresses or phone numbers on their customers, as well as the CDRs, which contain the phone number of the caller, the callee, the time stamp of the call, the tower at which the caller was connected, idem for the callee, and additional information such as special service usage and so on. The anonymization procedure consists then in replacing each phone number by a randomly generated number, such that each user has a unique random ID, from which it is impossible to retrieve the original phone number by reverse engineering procedures. The CDRs are then modified such that phone numbers are replaced by the corresponding ID. After this procedure, the CDRs are anonymized, and can be transferred to a third party. The standard procedure then implies that the third party signs a non-disclosure agreement, stipulating that they cannot make the CDR data available, and the agreement usually also restricts the range of potential research questions to be explored with the data. The safety of users privacy is then guaranteed both by the removal of information allowing to identify users and by the assumption that the third party doesn't make use of the data for any malicious intent.

8.1 De-anonymization attacks

Some research has been produced on mobile phone datasets to challenge this apparent feeling of security, however, recent results are opening new ways of considering the privacy problem. Using CDR data containing mobility traces, Zang and Bolot [193] show how it is possible to uniquely identify a large fraction of users with a small number of preferred locations. Their methodology goes as follows: for each user, it is possible to list the top N locations at which calls have been recorded. The authors show then that depending on the granularity of the locations, a non-negligible fraction of users may be uniquely identified by only 2 locations. For example, if locations are taken at cell level, up to 35% of the users of a 25 million communication network can be uniquely identified with 2 locations, which will be likely to correspond to home and work. Thus, while the anonymization procedure is intended to impeach any linkage between the dataset and individuals, using this procedure allows to potentially retrieve the mobility and calling pattern of targeted users given the access to as little information as home and work addresses. If additional data, such as year of birth or gender of users would be available - which is common in most datasets - it would be possible to identify very large fractions of the network. However, in this attack scheme, one has to know quite well the profile of the user for them to be found in the database. Using a different approach, de Montjoye *et al.* [194] show that knowing only four points in space and time where a user was allows to uniquely re-identify the user with 95% probability. Using only very little information that could be available easily to an attacker, the authors thus show how unique each user's trajectory is. They further show that blurring the resolution of space or time does not reduce much the information needed to re-identify a user in the database, thus keeping the database very vulnerable if faced with this type of de-anonymization attack.

Other possible attacks have also been considered on anonymized online social networks. Although those attacks are not likely to be applied in the case of mobile phone data, we quickly mention some of them, as it is likely that breaches found in different applications might be similar to potential breaches in mobile phone datasets.

For example, Backstrom *et al.* [195] describe a family of local attacks, which enable to retrieve the position of some targets in the network, and hence to uncover the connections between those patterns. The authors showed that on a network of 4.4 million nodes, by controlling the links of 7 dummy nodes they manage to uncover the presence or absence of 2,400 links between 70 target nodes, without being detected by the database manager. On a wider scale, Narayannan and Shmatikov [196] show that it is possible to retrieve the identity of a large part of a social network by combining it with an auxiliary network. Such a situation happens when users are present in two separate datasets. The authors show then that even if this overlap is available for only a fraction of the users, it is still possible to retrieve the information for a large part of the network. Add to this that other types of databases may be available (for example Twitter, or Facebook in addition to CDRs), and the possibility of de-anonymization is even greater. Such situations have already led to problematic situations, where specific people were re-identified in supposedly anonymized medical records, or movie preferences databases [187, 188, 197]. Indeed, two separate databases coming from different sources may be anonymized and safe to be released separately, but can still present a great danger for privacy if an attacker combines and crosses the information contained in both databases.

Against these possible threats of privacy breach, one may wonder if solutions are proposed to counter such attacks. If research on mobile datasets only considers average behaviors, rather than exact patterns, a simple countermeasure is to perform small modifications of the dataset, that would not alter the general aspect of it but that would have dramatic consequences on the algorithms used by attackers, who search for exact matchings between statistics on the network and a priori known properties of the targets.

Another protection against such attacks, and particularly when mobility data is involved, is to produce new random identifiers for each user at regular time intervals. By regenerating random identifiers, it makes it impossible to use longitudinal information in order to assess the preferred locations of a user. As shown by Zang and Bolot [193], by changing every day the ID of each user, only 3% of the nodes can still be identified using their top 2 locations. While this method seems efficient to protect the privacy of users, it reduces substantially the possible information to retrieve from such a dataset for research purposes. Using a similar approach also proved useful against the attack scheme considered by de Montjoye *et al.*, as Song *et al.* show in [198] that changing the ID of each user every six hours reduces substantially the fraction of unique trajectories in the dataset. A compromise between preserving the anonymity and keeping enough information in the dataset is difficult to achieve. In collaboration with the Université catholique de Louvain, the provider Orange tried to achieve this for their first D4D challenge before releasing a dataset to a wide community of researchers (more than 150 research teams participated). Through releasing four different datasets anonymized differently [162] and containing information of different spatio-temporal resolutions, they could guarantee the preservation of the anonymity of users. Yet, the loss of information was not too dramatic, as many studies showed very good results using the provided aggregated information. The challenge was a success and a second one followed in 2014-2015, using a wider dataset from Senegal [3]. Other similar initiatives include the Telecom Italia Big Data Challenges 2014 and 2015 [4], whose goal is to show the variety of applications that can be derived from the use of Big Data, including mobile phone data, but also weather, Twitter, public transport, energy.... These data are aggregated and anonymized and are therefore made openly available online to all who wish to analyze them [199].

Another question that is closely linked to this research is how to quantify the anonymity of a database. Latanya Sweeney proposed a measure that is *k-anonymity* [200], defining that a database achieves k-anonymity if for any tuples of previously defined entries of the database, there are at least k users corresponding to it, making it impossible to re-identify a single user with only information on these entries of the database. Of course, the larger k is, the most difficult it becomes to achieve this, especially in a CDR database containing spatio-temporal information about each call. Moreover, when the attacker is looking for a particular person in the database, enabling him to reduce the number of potential corresponding users to a small number is sometimes already a lot of information, and too big a risk to release the database publicly. Another potential solution to preserving privacy was suggested by Isaacman *et al.* [201] who suggest using synthetic data to model the mobility of people. They used mobile phone data from two american cities to validate their model, showing that their model, based on only aggregated data and probability distributions, could reproduce many of the features of mobility of users, without any of them corresponding to a real person. Mir *et al.* further proposed an evolved version called DP-WHERE [202] of the previous model, adding controlled noise to the set of empirical probability distributions. This noise then guarantees that the model achieves *differential privacy*, that is, that the analyses will not be significantly different whether or not a single individual is in the database from which the model is derived, even if this individual has an unusual behavior. However, on may wonder if these synthetic data could be used to carry out analyses that were not previously tested on the real database, as no guarantee exists on the outcome of analyses that were not foreseen by the researchers that tested the model for compatibility with empirical data.

8.2 Personal data: ownership, usage, privacy

Phone companies collect data about their users, about their habits, their mobility, their acquaintances. Still, the legislation up to 2013 was fuzzy [203], chilling companies to share such data for research and making customers feel that George Orwell's predictions are coming true, especially after the scandal in 2013 revealing how much personal information the NSA was collecting from many sources [204].

Such data represents an enormous added value, both to companies, for marketing purposes and client screening, and to authorities for traffic management or epidemic outbreak prevention. It is often forgotten, but the use of mobile phone datasets also has a huge positive potential in the developing world, as many of the proposed project to the Data for Development challenge showed [181], may it be for supervising the health status of populations, generating census data or optimizing public transport. Ironically, even though the research community has shown the potential that such data have to save lives and that using these data is technically possible, it is still often difficult to access the data because of privacy concerns, even when the data is aggregated and non-disclosure agreements are signed.

Such opportunities, both for corporates and authorities need to develop standardized procedures for the acquisition, conservation and usage of personal data, which is not yet the case. The communication about these procedures to customers hasn't been clear, as are the possibilities for a user to 'opt-out' if they don't want to have their personal data released.

With this intent, several voices have recently been raised in order to urge authorities to develop a 'New Deal' [205] on data ownership, in which users would own their personal data as well as the decisions to provide it - in exchange of payment - to companies interested in their usage. A transparent system armed with the necessary protocols and regulation for a transparent use of personal data would also facilitate the access to data for researchers [206], and could so benefit to the entire society.

9 Conclusion and research questions

The first analyses of mobile phone datasets appeared in the late 90's, and the result of this decade of research contains a large number of surprises and several promising directions for the future. In this paper, we have reviewed the most prominent results obtained so far, in particular in the analysis of the structure of our social networks, and human mobility. We decided not to cover some closely related questions, such as churn prediction (see [207–210]) or dynamic pricing [16, 211], which are rather business-related topics, and for which a vast literature is available.

The recent availability of mobile phone datasets have led to many discoveries on human behavior. We are not all similar in our ways of communicating, and differences between users can range to several orders of magnitudes. Our networks are clustered in well-structured groups, which are spatially well-located. With the raise of communication technology, some have predicted that the barrier of distance would fall, shrinking the world into a small village. However, mobile phone data suggests instead that distance still plays a role, but that its impact is nuanced by the varying population density. Regarding our mobility behavior, individuals appear to have highly predictable movements [212], while as populations we act and react in a remarkable synchronized way. In this context, the availability of mobile phone data has for the first time allowed to observe populations from a God-eye point of view, monitoring the pace of daily life or the response to catastrophes.

The ubiquity of mobile phones - there are nowadays more mobile phones than personal computers in use - which allows us to obtain such precise results raises also the thread of viral outbreaks, from which mobile phones have been safe until now. Mobile viruses could be a potential risk for users' privacy, as it is also the case that the anonymized datasets provided by operators to third parties for research could potentially be de-anonymized too.

The availability of such enormous datasets creates a huge potential that could benefit to society, up to the point of saving lives. The research that has been conducted so far only represents the tip of the iceberg of what could potentially be done, when adequately exploited. However, it is the necessity of authorities to ensure that such datasets could not be misused.

9.1 Further research

The number of possible research questions on mobile phone datasets is gigantic. In this last part, we will present one research direction that we believe to be highly important and still not addressed in its most general form.

A large number of research has been conducted on the analysis of social networks, based on CDRs. As it appears from the different publications on this topic, there exist some common features but also many differences in the structure of the constructed network. Recall as simplest example the degree distributions, which show different functional forms for most datasets.

These differences may, of course, be linked to cultural differences between the different countries of interest, but there are probably other, quantifiable, reasons. The datasets differ greatly in the market shares of the operators, in the time span of the data collection period, in the size of the network and in the geographical span of the considered country. The method of network construction is also always different and has a tangible impact on the network structure. The use of directed or undirected links, weights and thresholds for removing low-intensity or non-mutual links all greatly impact the structure and hence the statistical features of the obtained network.

Hence, we believe that a serious analysis, both on theoretical and on empirical side of the influence of these factors on the general structure of mobile phone networks may lead to a general framework, allowing to interpret differences between results obtained on several datasets with the knowledge of potential side-effects.

This question is closely related to the even more general question of the significance of information provided by CDR data. Recalling what was said in Section 2, CDR datasets are noisy data, some links appear there by chance, while other have not been captured in the dataset. It would thus be interesting to question the stability of the obtained results, provided that the real network is different from what has been observed in the data. This links with the work of Gourab [213], who analyzed the stability of PageRank under random noise on the network structure. Again, in this framework, no real theoretical result has yet been achieved, allowing to characterize which results are significant, and which are not.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

VB conceived and supervised the study. AD and GK reviewed the papers and drafted the manuscript. All authors edited and approved the final version of the paper.

Author details

[1]Department of Applied Mathematics, Université catholique de Louvain, Avenue Georges Lemaitre, 4, Louvain-La-Neuve, 1348, Belgium. [2]Real Impact Analytics, Place Flagey, 7, Brussels, 1050, Belgium.

Acknowledgements

We would like to thank Franscesco Calabrese, Yves-Alexandre de Montjoye, Vanessa Frias-Martinez, Marta González, Jukka-Pekka Onnela, Jari Saramäki and Zbigniew Smoreda for their valuable comments and advice in finalizing this survey. AD is a research fellow with the Fonds de la Recherche Scientifique - FNRS.

Received: 9 February 2015 Accepted: 23 June 2015 Published online: 05 August 2015

References

1. The world in 2014: ICT facts and figures. International Telecommunication Union. http://www.itu.int/
2. Kwok R (2009) Personal technology: phoning in data. Nature 458(7241):959-961
3. de Montjoye YA, Smoreda Z, Trinquart R, Ziemlicki C, Blondel VD (2014) D4D-Senegal: the second mobile phone data for development challenge. ArXiv preprint arXiv:1407.4885
4. Telecom Italia big data challenge. http://www.telecomitalia.com/tit/en/bigdatachallenge/contest.html
5. Eagle N, Pentland A (2006) Reality mining: sensing complex social systems. Pers Ubiquitous Comput 10(4):255-268
6. Karikoski J, Nelimarkka M (2010) Measuring social relations: case otasizzle. In: IEEE second international conference on social computing (SocialCom). IEEE Press, New York, pp 257-263
7. Stopczynski A, Sekara V, Sapiezynski P, Cuttone A, Madsen MM, Larsen JE, Lehmann S (2014) Measuring large-scale social networks with high resolution. PLoS ONE 9(4):e95978
8. Zipf GK (1949) Human behavior and the principle of least effort: an introduction to human ecology. Addison-Wesley Press, Reading
9. Cortes C, Pregibon D, Volinsky C (2001) Communities of interest. In: Hoffman F et al (eds) Advances in intelligent data analysis. LNCS, vol 2189. Springer, Berlin, pp 105-114
10. Krings G (2012) Extraction of information from large networks. PhD thesis, Université catholique de Louvain
11. Abello J, Pardalos PM, Resende MGC (1999) On maximum clique problems in very large graphs. In: External memory algorithms. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, vol 50. Am Math Soc, Providence, pp 119-130
12. Aiello W, Chung F, Lu L (2000) A random graph model for massive graphs. In: Proceedings of the thirty-second annual ACM symposium on theory of computing. ACM, New York, pp 171-180
13. Lambiotte R, Blondel VD, de Kerchove C, Huens E, Prieur C, Smoreda X, Van Dooren P (2008) Geographical dispersal of mobile communication networks. Physica A 387(21):5317-5325
14. Onnela JP, Saramaki J, Hyvonen J, Szabo G, Lazer D, Kaski K, Kertesz J, Barabasi AL (2007) Structure and tie strengths in mobile communication networks. Proc Natl Acad Sci USA 104(18):7332-7336
15. Li M-X, Palchykov V, Jiang Z-Q, Kaski K, Kertész J, Miccichè S, Tumminello M, Zhou W-X, Mantegna RN (2014) Statistically validated mobile communication networks: evolution of motifs in European and Chinese data. New J Phys 16:083038. doi:10.1088/1367-2630/16/8/083038
16. Kim Y, Telang R, Vogt WB, Krishnan R (2010) An empirical analysis of mobile voice service and SMS: a structural model. Manag Sci 56(2):234-252
17. Kovanen L, Saram J, Kaski K (2011) Reciprocity of mobile phone calls. JDySES 2(2):138-151
18. Ling R, Bertel TF, Sundsøy PR (2012) The socio-demographics of texting: an analysis of traffic data. New Media Soc 14(2):281-298. doi:10.1177/1461444811412711
19. Nanavati AA, Gurumurthy S, Das G, Chakraborty D, Dasgupta K, Mukherjea S, Joshi A (2006) On the structural properties of massive telecom call graphs: findings and implications. In: Proceedings of the 15th ACM international conference on information and knowledge management. ACM, New York, pp 435-444
20. Barabási AL (2009) Scale-free networks: a decade and beyond. Science 325(5939):412-413
21. Watts DJ, Strogatz SH (1998) Collective dynamics of small-world networks. Nature 393(6684):440-442
22. Clauset A, Shalizi CR, Newman ME (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661-703
23. Seshadri M, Machiraju S, Sridharan A, Bolot J, Faloutsos C, Leskovec J (2008) Mobile call graphs: beyond power-law and lognormal distributions. In: Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 596-604
24. Krings G, Karsai M, Bernhardsson S, Blondel VD, Saramäki J (2012) Effects of time window size and placement on the structure of an aggregated communication network. EPJ Data Sci 1:4
25. Onnela JP, Saramaki J, Hyvonen J, Szabo G, de Menezes MA, Kaski K, Barabasi AL, Kertesz J (2007) Analysis of a large-scale weighted network of one-to-one human communication. New J Phys 9:179
26. Granovetter MS (1973) The strength of weak ties. Am J Sociol 78:1360-1380
27. Onnela J-P, Saramäki J, Kertész J, Kaski K (2005) Intensity and coherence of motifs in weighted complex networks. Phys Rev E 71(6):065103
28. Du N, Faloutsos C, Wang B, Akoglu L (2009) Large human communication networks: patterns and a utility-driven generator. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. ACM, New York, pp 269-278
29. Kianmehr K, Alhajj R (2009) Calling communities analysis and identification using machine learning techniques. Expert Syst Appl 36(3):6218-6226
30. Zhang H, Dantu R (2008) Discovery of social groups using call detail records. In: On the move to meaningful Internet systems: OTM 2008 workshops. Springer, Berlin, pp 489-498
31. 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. Phys Rev E 83(5):056125
32. Blondel VD, Guillaume JL, Lambiotte R, Mech ELJS (2008) Fast unfolding of communities in large networks. J Stat Mech 2008:P10008
33. Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci USA 105(4):1118-1123
34. Palla G, Barabási AL, Vicsek T (2007) Quantifying social group evolution. Nature 446(7136):664-667
35. Ahn YY, Bagrow JP, Lehmann S (2010) Link communities reveal multiscale complexity in networks. Nature 466(7307):761-764
36. Lazer D, Pentland A, Adamic L, Aral S, Barabási A-L, Brewer D, Christakis N, Contractor N, Fowler J, Gutmann M, Jebara T, King G, Macy M, Roy D, Van Alstyne M (2009) Computational social science. Science 323(5915):721-723. http://www.sciencemag.org/content/323/5915/721.full.pdf. doi:10.1126/science.1167742
37. Eagle N, Pentland AS, Lazer D (2009) Inferring friendship network structure by using mobile phone data. Proc Natl Acad Sci USA 106(36):15274-15278
38. Wiese J, Min J-K, Hong JI, Zimmerman J (2015) 'You never call, you never write': call and SMS logs do not always indicate tie strength. In: Proceedings of the 2015 conference on computer supported cooperative work - CSCW'15. ACM, New York, pp 765-774

39. Blumenstock JE, Gillick D, Eagle N (2010) Who's calling? Demographics of mobile phone use in Rwanda. Transportation 32:2-5
40. Smoreda Z, Licoppe C (2000) Gender-specific use of the domestic telephone. Soc Psychol Q 63(3):238-252
41. Kovanen L, Kaski K, Kertész J, Saramäki J (2013) Temporal motifs reveal homophily, gender-specific patterns, and group talk in call sequences. Proc Natl Acad Sci USA 110(45):18070-18075
42. Frias-Martinez V, Frias-Martinez E, Oliver N (2010) A gender-centric analysis of calling behavior in a developing economy using call detail records. In: AAAI spring symposium: artificial intelligence for development
43. Blumenstock JE, Eagle N (2012) Divided we call: disparities in access and use of mobile phones in Rwanda. Inf Technol Int Dev 8(2):1-16
44. Chawla NV, Hachen D, Lizardo O, Toroczkai Z, Strathman A, Wang C (2011) Weighted reciprocity in human communication networks. Technical report arXiv:1108.2822
45. Motahari S, Mengshoel OJ, Reuther P, Appala S, Zoia L, Shah J (2012) The impact of social affinity on phone calling patterns: categorizing social ties from call data records. In: The 6th SNA-KDD workshop '12
46. Barthélemy M (2011) Spatial networks. Phys Rep 499(1):1-101
47. Deville P, Linard C, Martin S, Gilbert M, Stevens FR, Gaughan AE, Blondel VD, Tatem AJ (2014) Dynamic population mapping using mobile phone data. Proc Natl Acad Sci USA 111(45):15888-15893
48. Sterly H, Hennig B, Dongo K (2013) 'Calling abidjan' - improving population estimations with mobile communication data. In: Mobile phone data for development - analysis of mobile phone datasets for the development of Ivory Coast. Orange D4D challenge, pp 108-114
49. Krings G, Calabrese F, Ratti C, Blondel VD (2009) Urban gravity: a model for inter-city telecommunication flows. J Stat Mech Theory Exp 2009:07003
50. Krings G, Calabrese F, Ratti C, Blondel VD (2009) Scaling behaviors in the communication network between cities. In: International conference on computational science and engineering. IEEE Press, New York, pp 936-939
51. Onnela JP, Arbesman S, González MC, Barabási AL, Christakis NA (2011) Geographic constraints on social network groups. PLoS ONE 6(4):16939
52. Bucicovschi O, Douglass RW, Meyer DA, 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, pp 42-54
53. Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: universal patterns in human urban mobility. PLoS ONE 7(5):37027
54. Liben-Nowell D, Novak J, Kumar R, Raghavan P, Tomkins A (2005) Geographic routing in social networks. Proc Natl Acad Sci USA 102(33):11623-11628
55. Carolan E, McLoone SC, McLoone SF, Farrell R (2012) Analysing Ireland's interurban communication network using call data records. In: IET Irish signals and systems conference (ISSC 2012). IET, Stevenage, pp 1-6
56. Schläpfer M, Bettencourt L, Grauwin S, Raschke M, Claxton R, Smoreda Z, West GB, Ratti C (2014) The scaling of human interactions with city size. J R Soc Interface 11:20130789
57. Jo H-H, Saramäki J, Dunbar RI, Kaski K (2014) Spatial patterns of close relationships across the lifespan. Sci Rep 4:6988
58. Herrera-Yagüe C, Schneider CM, Smoreda Z, Couronné T, Zufiria PJ, González MC (2014) The elliptic model for communication fluxes. J Stat Mech Theory Exp 2014(4):04022
59. Grady D, Brune R, Thiemann C, Theis F, Brockmann D (2012) Modularity maximization and tree clustering: novel ways to determine effective geographic borders. In: Handbook of optimization in complex networks. Springer, Berlin, pp 169-208
60. Blondel VD, Deville P, Morlot F, Smoreda Z, Van Dooren P, Ziemlicki C (2011) Voice on the border: do cellphones redraw the maps? ParisTech Review
61. Blondel V, Krings G, Thomas I (2010) Regions and borders of mobile telephony in Belgium and in the Brussels metropolitan zone. Bruss Stud 42(4):1-12
62. Expert P, Evans TS, Blondel VD, Lambiotte R (2011) Uncovering space-independent communities in spatial networks. Proc Natl Acad Sci USA 108(19):7663-7668
63. Ratti C, Sobolevsky S, Calabrese F, Andris C, Reades J, Martino M, Claxton R, Strogatz SH (2010) Redrawing the map of Great Britain from a network of human interactions. PLoS ONE 5(12):14248
64. Blumenstock JE, Fratamico L (2013) Social and spatial ethnic segregation: a framework for analyzing segregation with large-scale spatial network data. In: Proceedings of the 4th annual symposium on computing for development. ACM DEV-4 '13. ACM, New York, article no 11
65. Eagle N, Macy M, Claxton R (2010) Network diversity and economic development. Science 328(5981):1029-1031
66. Mao H, Shuai X, Ahn YY, Bollen J (2013) Mobile communications reveal the regional economy in Côte d'Ivoire. In: Mobile phone data for development - analysis of mobile phone datasets for the development of Ivory Coast. Orange D4D challenge, pp 17-34
67. Smith-Clarke C, Mashhadi A, Capra L (2014) Poverty on the cheap: estimating poverty maps using aggregated mobile communication networks. In: Proceedings of the 32nd annual ACM conference on human factors in computing systems. ACM, New York, pp 511-520
68. Frias-Martinez V, Virseda J, Frias-Martinez E (2010) Socio-economic levels and human mobility. In: Qual meets quant workshop - QMQ
69. Frias-Martinez V, Soguero-Ruiz C, Frias-Martinez E, Josephidou M, (2013) Forecasting socioeconomic trends with cell phone records. In: Proceedings of the 3rd ACM symposium on computing for development. ACM, New York, article no 15
70. Gutierrez T, Krings G, Blondel VD (2013) Evaluating socio-economic state of a country analyzing airtime credit and mobile phone datasets. ArXiv preprint arXiv:1309.4496
71. Holme P, Saramäki J (2012) Temporal networks. Phys Rep 519(3):97-125
72. Huang Z, Lin DKJ (2009) The time-series link prediction problem with applications in communication surveillance. INFORMS J Comput 21(2):286-303. doi:10.1287/ijoc.1080.0292
73. Yu K, Chu W, Yu S, Tresp V, Xu Z (2006) Stochastic relational models for discriminative link prediction. In: Advances in neural information processing systems, pp 1553-1560

74. Hidalgo CA, Rodriguez-Sickert C (2008) The dynamics of a mobile phone network. Physica A 387(12):3017-3024
75. Raeder T, Lizardo O, Hachen D, Chawla NV (2011) Predictors of short-term decay of cell phone contacts in a large scale communication network. Soc Netw 33(4):245-257
76. Kossinets G, Watts DJ (2006) Empirical analysis of an evolving social network. Science 311(5757):88-90
77. Miritello G (2013) Temporal patterns of communication in social networks. Springer, Berlin
78. Karsai M, Perra N, Vespignani A (2014) Time varying networks and the weakness of strong ties. Sci Rep 4:4001
79. Miritello G, Rubén L, Cebrian M, Moro E (2013) Limited communication capacity unveils strategies for human interaction. Sci Rep 3:1950
80. Miritello G, Moro E, Lara R, Martínez-López R, Belchamber J, Roberts SGB, Dunbar RIM (2013) Time as a limited resource: communication strategy in mobile phone networks. Soc Netw 35(1):89-95
81. Saramäki J, Leicht EA, López E, Roberts SGB, Reed-Tsochas F, Dunbar RIM (2014) The persistence of social signatures in human communication. Proc Natl Acad Sci USA 111(3):942-947
82. Aynaud T, Fleury E, Guillaume J-L, Wang Q (2013) Communities in evolving networks: definitions, detection, and analysis techniques. In: Dynamics on and of complex networks, vol 2. Springer, Berlin, pp 159-200
83. Mucha PJ, Richardson T, Macon K, Porter MA, Onnela J-P (2010) Community structure in time-dependent, multiscale, and multiplex networks. Science 328(5980):876-878. http://www.sciencemag.org/content/328/5980/876.full.pdf. doi:10.1126/science.1184819
84. Xu KS, Kliger M, Hero AO (2010) Evolutionary spectral clustering with adaptive forgetting factor. In: IEEE international conference on acoustics speech and signal processing (ICASSP). IEEE Press, New York, pp 2174-2177
85. Xu KS, Kliger M, Hero AO, III (2011) Tracking communities in dynamic social networks. In: Social computing, behavioral-cultural modeling and prediction. Springer, Berlin, pp 219-226
86. Kovanen L, Karsai M, Kaski K, Kertész J, Saramäki J (2011) Temporal motifs in time-dependent networks. J Stat Mech Theory Exp 2011(11):11005
87. Kirkpatrick S, Kulakovsky A, Cebrian M, Pentland A (2012) Social networks and spin glasses. Philos Mag 92(1-3):362-377. doi:10.1080/14786435.2011.634858
88. Pielot M, de Oliveira R, Kwak H, Oliver N (2014) Didn't you see my message?: predicting attentiveness to mobile instant messages. In: Proceedings of the 32nd annual ACM conference on human factors in computing systems. ACM, New York, pp 3319-3328
89. Barabási AL (2005) The origin of bursts and heavy tails in human activity. Nature 435:207-211
90. Karsai M, Kivelä M, Pan RK, Kaski K, Kertész J, Barabási AL, Saramäki J (2011) Small but slow world: how network topology and burstiness slow down spreading. Phys Rev E 83(2):025102
91. Karsai M, Kaski K, Barabási AL, Kertész J (2012) Universal features of correlated bursty behaviour. Sci Rep 2:397
92. Wu Y, Zhou C, Xiao J, Kurths J, Schellnhuber HJ (2010) Evidence for a bimodal distribution in human communication. Proc Natl Acad Sci USA 107(44):18803-18808
93. Candia J, González MC, Wang P, Schoenharl T, Madey G, Barabási AL (2008) Uncovering individual and collective human dynamics from mobile phone records. J Phys A, Math Theor 41:224015
94. Karsai M (2012) Correlated dynamics in egocentric communication networks. PLoS ONE 7(7):40612. doi:10.1371/journal.pone.0040612
95. Malmgren RD, Stouffer DB, Motter AE, Amaral LA (2008) A Poissonian explanation for heavy tails in e-mail communication. Proc Natl Acad Sci USA 105(47):18153-18158
96. Jo H-H, Karsai M, Kertész J, Kaski K (2012) Circadian pattern and burstiness in mobile phone communication. New J Phys 14(1):013055
97. González MC, Hidalgo CA, Barabási AL (2008) Understanding individual human mobility patterns. Nature 453(7196):779-782
98. Song C, Koren T, Wang P, Barabási AL (2010) Modelling the scaling properties of human mobility. Nat Phys 6:818-823
99. Csáji B, Browet A, Traag VA, Delvenne J-C, Huens E, Van Dooren P, Smoreda Z, Blondel VD (2013) Exploring the mobility of mobile phone users. Physica A 392(6):1459-1473
100. Bagrow JP, Lin Y-R (2012) Mesoscopic structure and social aspects of human mobility. PLoS ONE 7(5):37676
101. Amini A, Kung K, Kang C, Sobolevsky S, Ratti C (2013) The differing tribal and infrastructural influences on mobility in developing and industrialized regions. In: Mobile phone data for development - analysis of mobile phone datasets for the development of Ivory Coast. Orange D4D challenge, pp 330-339
102. Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. Science 327(5968):1018-1021
103. Calabrese F, Di Lorenzo G, Ratti C (2010) Human mobility prediction based on individual and collective geographical preferences. In: 13th international IEEE conference on Intelligent transportation systems (ITSC). IEEE Press, New York, pp 312-317
104. Park J, Lee DS, González MC (2010) The eigenmode analysis of human motion. J Stat Mech Theory Exp 2010:11021
105. Simini F, Gonzalez MC, Maritan A, Barabasi A-L (2012) A universal model for mobility and migration patterns. Nature 484(7392):96-100. doi:10.1038/nature10856
106. Palchykov V, Mitrovic M, Jo H-H, Saramaki J, Pan RK (2014) Inferring human mobility using communication patterns. Sci Rep 4:6174
107. Martino M, Calabrese F, Di Lorenzo G, Andris C, Liang L, Ratti C (2010) Ocean of information: fusing aggregate & individual dynamics for metropolitan analysis. In: Proceedings of the 15th international conference on intelligent user interfaces. ACM, New York, pp 357-360
108. Eagle N, Pentland AS (2009) Eigenbehaviors: identifying structure in routine. Behav Ecol Sociobiol 63(7):1057-1066
109. Ratti C, Williams S, Frenchman D, Pulselli RM (2006) Mobile landscapes: using location data from cell phones for urban analysis. Environ Plan B, Plan Des 33(5):727-748
110. Calabrese F, Colonna M, Lovisolo P, Parata D, Ratti C (2011) Real-time urban monitoring using cell phones: a case study in Rome. IEEE Trans Intell Transp Syst 12(1):141-151
111. Reades J, Calabrese F, Sevtsuk A, Ratti C (2007) Cellular census: explorations in urban data collection. IEEE Pervasive Comput 6:30-38
112. Reades J, Calabrese F, Ratti C (2009) Eigenplaces: analysing cities using the space-time structure of the mobile phone network. Environ Plan B, Plan Des 36(5):824-836

113. Karikoski J, Soikkeli T (2011) Contextual usage patterns in smartphone communication services. Pers Ubiquitous Comput 17(3):491-502
114. Jo H-H, Karsai M, Karikoski J, Kaski K (2012) Spatiotemporal correlations of handset-based service usages. EPJ Data Sci 1:10
115. Trestian I, Ranjan S, Kuzmanovic A, Nucci A (2009) Measuring serendipity: connecting people, locations and interests in a mobile 3G network. In: Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference. ACM, New York, pp 267-279
116. Naboulsi D, Stanica R, Fiore M (2014) Classifying call profiles in large-scale mobile traffic datasets. In: Conference on computer communications - INFOCOM 2014. IEEE Press, New York, pp 1806-1814
117. Isaacman S, Becker R, Cáceres R, Kobourov S, Rowland J, Varshavsky A (2010) A tale of two cities. In: Proceedings of the eleventh workshop on mobile computing systems and applications. ACM, New York, pp 19-24
118. Louail T, Lenormand M, Cantú OG, Picornell M, Herranz R, Frias-Martinez E, Ramasco JJ, Barthelemy M (2014) From mobile phone data to the spatial structure of cities. Sci Rep 4:5276
119. Trasarti R, Olteanu-Raimond A-M, Nanni M, Couronné T, Furletti B, Giannotti F, Smoreda Z, Ziemlicki C (2014) Discovering urban and country dynamics from mobile phone data with spatial correlation patterns. Telecommun Policy 39:347-362
120. Jiang S, Fiore GA, Yang Y, Ferreira J, Frazzoli E, González MC (2013) A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: Proceedings of the 2nd ACM SIGKDD international workshop on urban computing. ACM, New York, article no 2
121. 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. LNCS, vol 8190. Springer, Berlin, pp 663-666
122. Nanni M, Trasarti R, Furletti B, Gabrielli L, Van Der Mede P, De Bruijn J, De Romph E, Bruil G (2013) MP4-A project: mobility planning for Africa. In: Mobile phone data for development - analysis of mobile phone datasets for the development of Ivory Coast. Orange D4D challenge, pp 423-446
123. Angelakis V, Gundlegård D, Rajna B, Rydergren C, Vrotsou K, Carlsson R, Forgeat J, Hu TH, Liu EL, Moritz S, Zhao S, Zheng Y (2013) Mobility modeling for transport efficiency - analysis of travel characteristics based on mobile phone data. In: Mobile phone data for development - analysis of mobile phone datasets for the development of Ivory Coast. Orange D4D challenge, pp 412-422
124. Bagrow JP, Wang D, Barabási AL (2011) Collective response of human populations to large-scale emergencies. PLoS ONE 6(3):17680
125. Gao L, Song C, Gao Z, Barabási AL, Bagrow JP, Wang D (2014) Quantifying information flow during emergencies. Sci Rep 4:3997
126. Xavier FHZ, Silveira LM, Almeida JM, Malab CHS, Ziviani A, Marques-Neto HT (2013) Understanding human mobility due to large-scale events. In: NetMob 2013 - third international conference on the analysis of mobile phone datasets
127. Altshuler Y, Fire M, Shmueli E, Elovici Y, Bruckstein A, Pentland AS, Lazer D (2013) The social amplifier - reaction of human communities to emergencies. J Stat Phys 152(3):399-418
128. Lu X, Bengtsson L, Holme P (2012) Predictability of population displacement after the 2010 Haiti earthquake. Proc Natl Acad Sci USA 109(29):11576-11581
129. Kryvasheyeu Y, Chen H, Moro E, Van Hentenryck P, Cebrian M (2015) Performance of social network sensors during hurricane Sandy. PLoS ONE 10(2):e0117288. doi:10.1371/journal.pone.0117288
130. Calabrese F, Smoreda Z, Blondel VD, Ratti C (2011) Interplay between telecommunications and face-to-face interactions: a study using mobile phone data. PLoS ONE 6(7):20814
131. Wang D, Pedreschi D, Song C, Giannotti F, Barabási AL (2011) Human mobility, social ties, and link prediction. In: 17th ACM SIGKDD conference on knowledge discovery and data mining (KDD'11). ACM, New York, pp 1100-1108
132. Eagle N, de Montjoye YA, Bettencourt LMA (2009) Community computing: comparisons between rural and urban societies using mobile phone data. In: 2009 international conference on computational science and engineering. IEEE Press, New York, pp 144-150
133. Toole JL, Herrera-Yaqüe C, Schneider CM, González MC (2015) Coupling human mobility and social ties. J R Soc Interface 12(105):20141128
134. Peruani F, Tabourier L (2011) Directedness of information flow in mobile phone communication networks. PLoS ONE 6(12):28860
135. Tabourier L, Stoica A, Peruani F (2012) How to detect causality effects on large dynamical communication networks: a case study. In: Fourth international conference on communication systems and networks (COMSNETS). IEEE Press, New York, pp 1-7
136. Miritello G, Moro E, Lara R (2011) Dynamical strength of social ties in information spreading. Phys Rev E 83(4):045102
137. Newman MEJ (2002) Spread of epidemic disease on networks. Phys Rev E 66:016128. doi:10.1103/PhysRevE.66.016128
138. Newman MEJ, Barabasi AL, Watts DJ (2006) The structure and dynamics of networks. Princeton University Press, Princeton
139. Kivelä M, Pan RK, Kaski K, Kertész J, Saramäki J, Karsai M (2012) Multiscale analysis of spreading in a large communication network. J Stat Mech Theory Exp 2012(3):03005
140. Liu YY, Slotine JJ, Barabási AL (2011) Controllability of complex networks. Nature 473(7346):167-173
141. Cebrian M, Lahiri M, Oliver N, Pentland A (2010) Measuring the collective potential of populations from dynamic social interaction data. IEEE J Sel Top Signal Process 4(4):677-686
142. Wang P, González MC, Hidalgo CA, Barabási AL (2009) Understanding the spreading patterns of mobile phone viruses. Science 324(5930):1071-1074
143. Wang P, González MC, Menezes R, Barabási AL (2010) New generation of mobile phone viruses and corresponding countermeasures. Arxiv preprint arXiv:1012.3156
144. Wang P, González MC, Menezes R, Barabási AL (2013) Understanding the spread of malicious mobile-phone programs and their damage potential. Int J Inf Secur 12(5):383-392

145. Baccelli F, Bolot J (2011) Modeling the economic value of location and preference data of mobile users. In: 2011 Proceedings IEEE INFOCOM. IEEE Press, New York, pp 1467-1475
146. Calabrese F, Ferrari L, Blondel VD (2014) Urban sensing using mobile phone network data: a survey of research. ACM Comput Surv 47(2):25
147. Isaacman S, Becker R, Cáceres R, Kobourov S, Martonosi M, Rowland J, Varshavsky A (2011) Identifying important places in people's lives from cellular network data. In: Pervasive computing. LNCS, vol 6696. Springer, Berlin, pp 133-151
148. Steenbruggen J, Borzacchiello MT, Nijkamp P, Scholten H (2011) Mobile phone data from GSM networks for traffic parameter and urban spatial pattern assessment: a review of applications and opportunities. GeoJournal 78:1-21
149. Toole JL, Colak S, Alhasoun F, Evsukoff A, Gonzalez MC (2014) The path most travelled: mining road usage patterns from massive call data. ArXiv preprint arXiv:1403.0636
150. Wang P, Hunter T, Bayen AM, Schechtner K, González MC (2012) Understanding road usage patterns in urban areas. Sci Rep 2:1001
151. McInerney J, Roger A, Jennings NR (2013) Crowdsourcing physical package delivery using the existing routine mobility of a local population. In: Mobile phone data for development - analysis of mobile phone datasets for the development of Ivory Coast. Orange D4D challenge, pp 447-456
152. Gambs S, Killijian M-O (2013) Towards a recomender system for bush taxis. In: Mobile phone data for development - analysis of mobile phone datasets for the development of Ivory Coast. Orange D4D challenge, pp 457-466
153. Calabrese F, Pereira F, Di Lorenzo G, Liu L, Ratti C (2010) The geography of taste: analyzing cell-phone mobility and social events. In: Pervasive computing. LNCS, vol 6030. Springer, Berlin, pp 22-37
154. Quercia D, Lathia N, Calabrese F, Di Lorenzo G, Crowcroft J (2010) Recommending social events from mobile phone location data. In: IEEE international conference on data mining. IEEE Press, New York, pp 971-976
155. Cloquet C, Blondel VD (2014) Forecasting event attendance with anonymized mobile phone data. Big Data Res (in press)
156. Xavier FHZ, Silveira LM, Almeida JM, Ziviani A, Malab CHS, Marques-Neto HT (2012) Analyzing the workload dynamics of a mobile phone network in large scale events. In: Proceedings of the first workshop on urban networking. ACM, New York, pp 37-42
157. Manfredini F, Tagliolato P, Di Rosa C (2011) Monitoring temporary populations through cellular core network data. In: Murgante B et al (eds) Computational science and its applications - ICCSA 2011, Part II. LNCS, vol 6783. Springer, Berlin, pp 151-161
158. Kuusik A, Ahas R, Tiru M (2009) Analysing repeat visitation on country level with passive mobile positioning method: an Estonian case study. In: XVII scientific conference on economic policy, pp 1-3
159. Wesolowski A, Eagle N, Tatem AJ, Smith DL, Noor AM, Snow RW, Buckee CO (2012) Quantifying the impact of human mobility on malaria. Science 338(6104):267-270
160. Tizzoni M, Bajardi P, Decuyper A, Kon Kam King G, Schneider CM, Blondel VD, Smoreda Z, González MC, Colizza V (2014) On the use of human mobility proxies for modeling epidemics. PLoS Comput Biol 10(7):1003716
161. Frias-Martinez E, Williamson G, Frias-Martinez V (2011) An agent-based model of epidemic spread using human mobility and social network information. In: Privacy, security, risk and trust (PASSAT) and IEEE third inernational conference on social computing (SocialCom). IEEE Press, New York, pp 57-64. doi:10.1109/PASSAT/SocialCom.2011.142
162. Blondel VD, 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
163. Kafsi M, Kazemi E, Maystre L, Yartseva L, Grossglauser M, Thiran P (2013) Mitigating epidemics through mobile micro-measures. ArXiv preprint arXiv:1307.2084
164. Lima A, De Domenico M, Pejovic V, Musolesi M (2013) Exploiting cellular data for disease containment and information campaigns strategies in country-wide epidemics. ArXiv preprint arXiv:1306.4534
165. Wesolowski A, Buckee CO, Bengtsson L, Wetter E, Lu X, Tatem AJ (2014) Commentary: containing the Ebola outbreak - the potential and challenge of mobile network data. PLoS Curr Outbreaks. doi:10.1371/currents.outbreaks.0177e7fcf52217b8b634376e2f3efc5e
166. de Montjoye YA, Kendall J, Kerry CF (2014) Enabling humanitarian use of mobile phone data. Issues in Technology Innovation 26
167. Bogomolov A, Lepri B, Ferron M, Pianesi F, Pentland AS (2014) Pervasive stress recognition for sustainable living. In: IEEE international conference on pervasive computing and communications workshops (PERCOM workshops). IEEE Press, New York, pp 345-350
168. Bogomolov A, Lepri B, Ferron M, Pianesi F, Pentland AS (2014) Daily stress recognition from mobile phone data, weather conditions and individual traits. In: Proceedings of the ACM international conference on multimedia. ACM, New York, pp 477-486
169. Katz E, Lazarsfeld PF (1970) Personal influence, the part played by people in the flow of mass communications. Transaction Publishers, New Brunswick
170. Watts DJ, Dodds PS (2007) Influentials, networks, and public opinion formation. J Consum Res 34(4):441-458
171. Szabó G, Barabási AL (2006) Network effects in service usage. Arxiv preprint arXiv:physics/0611177
172. Hill S, Provost F, Volinsky C (2006) Network-based marketing: identifying likely adopters via consumer networks. Stat Sci 21(2):256-276
173. Aharony N, Pan W, Ip C, Pentland A (2010) Tracing mobile phone app installations in the 'friends and family' study. In: Proceedings of the 2010 workshop on information in networks (WIN'10)
174. Risselada H, Verhoef PC, Bijmolt THA (2014) Dynamic effects of social influence and direct marketing on the adoption of high-technology products. J Mark 78(2):52-68
175. Backlund V-P, Saramäki J, Pan RK (2014) Effects of temporal correlations on cascades: threshold models on temporal networks. Phys Rev E 89:062815. doi:10.1103/PhysRevE.89.062815
176. Blondel V, de Kerchove C, Huens E, Van Dooren P (2006) Social leaders in graphs. Lecture notes in control and information sciences, vol 341, pp 231-237
177. de Kerchove d'Exaerde C (2009) Ranking large networks: leadership. optimization and distrust. PhD thesis

178. Ferrara E, De Meo P, Catanese S, Fiumara G (2014) Detecting criminal organizations in mobile phone networks. Expert Syst Appl 41(13):5733-5750
179. Catanese S, Ferrara E, Fiumara G (2013) Forensic analysis of phone call networks. Soc Netw Anal Min 3(1):15-33
180. Bogomolov A, Lepri B, Staiano J, Oliver N, Pianesi F, Pentland A (2014) Once upon a crime: towards crime prediction from demographics and mobile data. In: Proceedings of the 16th international conference on multimodal interaction. ACM, New York, pp 427-434
181. Blondel VD, de Cordes N, Decuyper A, Deville P, Raguenez J, Smoreda Z (eds) (2013) Mobile phone data for development - analysis of mobile phone datasets for the development of Ivory Coast. Orange D4D challenge
182. Salnikov V, Schien D, Youn H, Lambiotte R, Gastner MT (2014) The geography and carbon footprint of mobile phone use in Côte d'Ivoire. EPJ Data Sci 3:3
183. Frias-Martinez V, Virseda J (2012) On the relationship between socio-economic factors and cell phone usage. In: Proceedings of the fifth international conference on information and communication technologies and development. ACM, New York, pp 76-84
184. Buckee CO, Wesolowski A, Eagle NN, Hansen E, Snow RW (2013) Mobile phones and malaria: modeling human and parasite travel. Travel Med Infect Dis 11(1):15-22
185. Onnela JP, Christakis NA (2012) Spreading paths in partially observed social networks. Phys Rev E 85(3):036106
186. Ranjan G, Zang H, Zhang ZL, Bolot J (2012) Are call detail records biased for sampling human mobility? ACM SIGMOBILE Mob Comput Commun Rev 16(3):33-44
187. Singel R (2009) Netflix spilled your Brokeback Mountain secret, lawsuit claims. Threat Level (blog), Wired
188. Barth-Jones DC (2012) The 're-identification' of governor William Weld's medical information: a critical re-examination of health data Identification risks and privacy protections, then and now
189. E Commission (2012) Commission proposes a comprehensive reform of data protection rules to increase users' control of their data and to cut costs for businesses. Reference IP/12/46
190. EU (1995) Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Off J EC 38:31-50
191. Jay RP (ed) (2014) Data protection and privacy in 26 juristictions worldwide. Gideon Roberton
192. European Commission (2012) Commission proposes a comprehensive reform of data protection rules to increase users' control of their data and to cut costs for businesses. http://europa.eu/rapid/press-release_IP-12-46_en.htm?locale=en
193. Zang H, Bolot J (2011) Anonymization of location data does not work: a large-scale measurement study. In: ACM MobiCom. ACM, New York, pp 145-156
194. de Montjoye YA, Hidalgo CA, Verleysen M, Blondel VD (2013) Unique in the Crowd: the privacy bounds of human mobility. Sci Rep 3:1376
195. Backstrom L, Dwork C, Kleinberg J (2007) Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In: Proceedings of the 16th international conference on World Wide Web. ACM, New York, pp 181-190
196. Narayanan A, Shmatikov V (2009) De-anonymizing social networks. In: 30th IEEE symposium on security and privacy. IEEE Press, New York, pp 173-187
197. Sweeney L (2000) Simple demographics often identify people uniquely. Health (San Francisco) 671:1-34
198. Song Y, Dahlmeier D, Bressan S (2014) Not so unique in the crowd: a simple and effective algorithm for anonymizing location data. In: Proceeding of the 1st international workshop on privacy-preserving IR: when information retrieval meets privacy and security (PIR 2014), pp 19-24
199. Telecom Italia big data challenge - open data. http://theodi.fbk.eu/openbigdata/
200. Sweeney L (2002) *k*-Anonymity: a model for protecting privacy. Int J Uncertain Fuzziness Knowl-Based Syst 10(5):557-570
201. Isaacman S, Becker R, Cáceres R, Martonosi M, Rowland J, Varshavsky A, Willinger W (2012) Human mobility modeling at metropolitan scales. In: Proceedings of the 10th international conference on mobile systems, applications, and services. ACM, New York, pp 239-252
202. Mir DJ, Isaacman S, Cáceres R, Martonosi M, Wright RN (2013) DP-WHERE: differentially private modeling of human mobility. In: IEEE international conference on big data. IEEE Press, New York, pp 580-588
203. Madan A, Waber BN, Ding M, Kominers P, Pentland AS (2009). Reality mining and personal privacy
204. Landau S (2013) Making sense from Snowden. IEEE Secur Priv 11(4):54-63
205. Pentland A (2009) Reality mining of mobile communications: toward a new deal on data. The global information technology report 2008-2009, 1981
206. Eagle N (2009) Engineering a common good: fair use of aggregated, anonymized behavioral data. In: First international forum on the application and management of personal electronic information
207. Hung SY, Yen DC, Wang HY (2006) Applying data mining to telecom churn management. Expert Syst Appl 31(3):515-524
208. Dasgupta K, Singh R, Viswanathan B, Chakraborty D, Mukherjea S, Nanavati AA, Joshi A (2008) Social ties and their relevance to churn in mobile telecom networks. In: Proceedings of the 11th international conference on extending database technology: advances in database technology. ACM, New York, pp 668-677
209. Richter Y, Yom-Tov E, Slonim N (2010) Predicting customer churn in mobile networks through analysis of social groups. In: Proceedings of the 2010 SIAM international conference on data mining (SDM 2010). SIAM, Philadelphia, pp 732-741
210. Dierkes T, Bichler M, Krishnan R (2011) Estimating the effect of word of mouth on churn and cross-buying in the mobile phone market with Markov logic networks. Decision Support Systems 51(3):361-371
211. Fitkov-Norris ED, Khanifar A (2001) Dynamic pricing in cellular networks, a mobility model with a provider-oriented approach. In: Second international conference on 3G mobile communication technologies. IET, London, pp 63-67
212. Barabási AL (2010) You're so predictable. Phys World 23:22-26
213. Ghoshal G, Barabási AL (2011) Ranking stability and super-stable nodes in complex networks. Nat Commun 2:394
214. Afripop Project. http://www.worldpop.org.uk