Mobile Traffic Analysis: a Survey

Diala Naboulsi, Marco Fiore, Stephane Ribot, Razvan Stanica

Mobile Traffic Analysis: a Survey

Diala Naboulsi, Marco Fiore , Stephane Ribot , Razvan Stanica

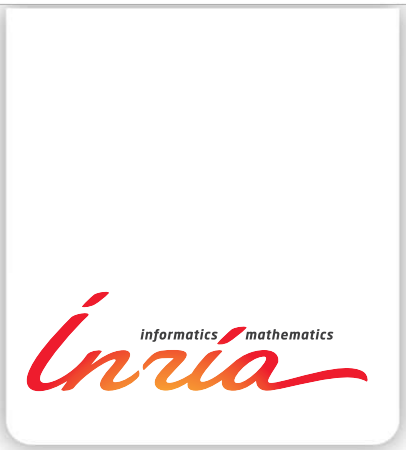
RESEARCH REPORT

N°

March 2015

Project-Teams

ISSN 0249-6399 ISRN INRIA/RR—FR+ENG



Mobile Traffic Analysis: a Survey

Diala Naboulsi*, Marco Fiore† *, Stephane Ribot ‡, Razvan Stanica *

Project-Teams

Research Report n° — March 2015 — 56 pages

Abstract: This report surveys the literature on analyses of mobile traffic collected by operators within their network infrastructure. This is a recently emerged research field, and, apart a few outliers, relevant works cover the period from 2005 to date, with a sensible densification over the last three years. We provide a thorough review of the multidisciplinary activities that rely on mobile traffic datasets, identifying major categories and sub-categories in the literature, so as to outline a hierarchical classification of research lines. When detailing the works pertaining to each class, we balance a comprehensive view of state-of-the-art results with punctual focuses on the methodological aspects. Our approach provides a complete introductory guide to the research based on mobile traffic analysis. It allows summarizing the main findings of the current state-of-the-art, as well as pinpointing important open research directions.

Key-words: Mobile traffic, data analysis, cellular networks

* Université de Lyon, INRIA, INSA-Lyon, CITI-INRIA, F-69621, Villeurbanne, France, diala.naboulsi@insa-lyon.fr
† CNR – IEIIT, Corso Duca degli Abruzzi 24, 10129 Torino, Italy, marco.fiore@ieiit.cnr.it
‡ Université de Lyon, Magellan Research Center, Lyon, France, stephane.ribot3@univ-lyon3.fr

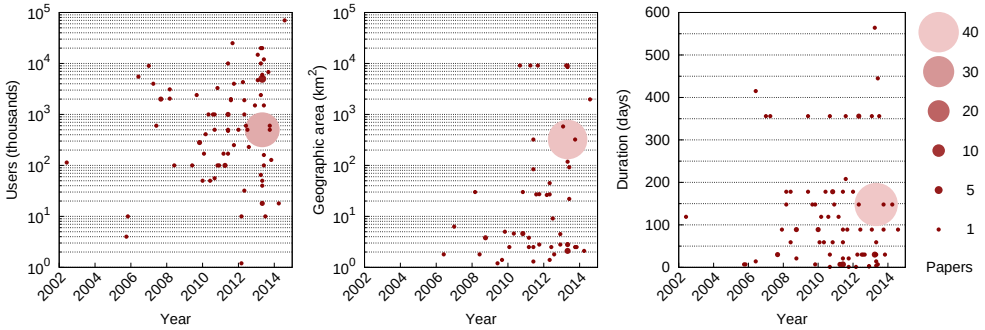
RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

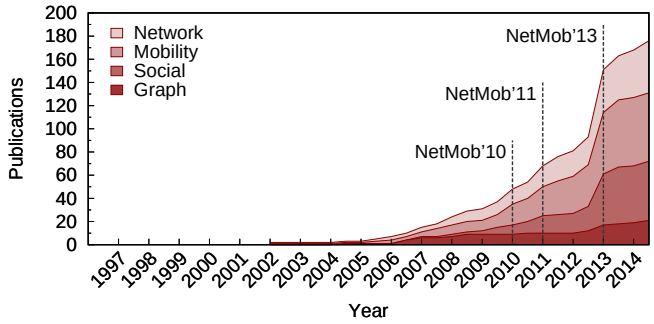
Une synthèse des analyses de trafic mobile

Résumé : Ce rapport résume les études portant sur l'analyse de trafic mobile collecté par les opérateurs de réseaux cellulaires sur leurs infrastructures. Ce domaine de recherche est assez récent, la majorité des travaux date d'après 2005, avec une concentration assez forte sur les trois dernières années. Notre article couvre des études pluridisciplinaires, que nous classifions en des catégories majeures ainsi que des sous-catégories significatives. En plus, notre synthèse ne se limite pas seulement à la discussion des résultats des différents travaux, mais englobe aussi les méthodologies appliquées. Ainsi, nous fournissons aux lecteurs un guide d'introduction assez complet au domaine de recherche de l'analyse de trafic mobile, récapitulant les principaux résultats déjà trouvés aussi bien que les pistes de recherche futures possibles.

Mots-clés : Trafic mobile, analyse de donnees, réseaux cellulaires



(a) Dataset users (left), area (center) and duration (right) over time



(b) Publication volume over time

Figure 1: Evolution of mobile traffic literature. (a) Main features of mobile traffic datasets. Circle color and size denote the number of papers with identical properties. (b) Cumulated publications based on mobile traffic analysis. Different color shades map to the four main categories we identify in our survey. Vertical lines pinpoint major dedicated events, as per the labels.

1 Introduction

Mobile traffic analysis is a rapidly emerging research field that encompasses a wide range of disciplines. We summarize its scope as *the study of massive traffic datasets collected by mobile network operators to improve the understanding of natural or technological phenomena occurring at large scales, and to design solutions to issues they may yield.* This definition is necessarily generic, as it has to accommodate works that exploit mobile traffic of different type, in diverse ways, and for many and varied purposes. Yet, our definition traces a clear boundary on the mobile traffic sources we consider in this survey, which solely concerns datasets collected at the operator's side of the mobile communication system. Therefore, works dealing with data gathered on the subscribers' side, i.e., via dedicated monitors running at the user equipment, are out of the scope of our review.

Even then, the diversity of the literature is staggering. Mobile traffic analyses can build on datasets whose nature varies depending on the precise collection point within the operator network, on the amount of subscribers concerned, and on the duration and timing of the measurement campaign. Different works can rely on information that describes user position with spatial granularities that range from cell sectors to whole cities, and with temporal granularities that span from milliseconds to hours. Some datasets may contain no or minimal notion of the actual service provided to each subscriber (e.g., voice, texting, data), whereas others may detail the protocols, applications, and URLs involved in each network transaction. Differences also appear in terms of customer base, geographical and temporal coverage. Fig. 1a portrays

such heterogeneity, as scatterplots of the main features of datasets employed by papers that appeared over the past decade. It is clear that: (i) the number of subscribers, the geographical surfaces and the timespan covered by mobile traffic datasets can differ by several orders of magnitude; (ii) there is no clear trend over time, and the growing number of points, i.e., works, just leads to more diversity; (iii) with one notable exception[1] there is a tendency for each paper to use its own mobile traffic dataset.

Despite these differences, what makes mobile operator data unique is that they typically contain information about hundreds of thousands of customers for whole weeks, as demonstrated by Fig. 1a. It is precisely the availability of terabytes of data related to large numbers of individuals over long time periods that makes mobile traffic so appealing to many research communities. The likes of sociologists, epidemiologists, physicists, transportation or telecommunication experts see in mobile network operators' datasets a clear opportunity to bring their analyses to an unprecedented scale while retaining a sufficient level of detail. No traditional data collection technique can offer a comparable perspective on human activities, and this fact alone is sufficient to explain the dramatic surge in mobile traffic studies over the last few years. As shown in Fig. 1b, the number of papers carrying out mobile traffic analyses was nearly at zero in 2005. Since then, it has been swelling at a 90% compound annual growth rate.

One of the main causes behind the success of mobile traffic analyses is the increasing availability of datasets. Mobile operators have been always monitoring mobile traffic in their network, for troubleshooting, efficiency, and billing purposes. However, they have been traditionally very cautious about sharing the collected data. This trend is now changing, also thanks to seminal works that proved how mobile traffic data can be an extremely valuable asset for fundamental research with a return for the operators themselves.

Not only the latter appear today more prone to open their data to the wider research community, but, in some cases, they are even fostering fundamental and applied research on mobile traffic through targeted challenges. Significant examples are the Data for Development (D4D) Challenge by Orange[2], whose second edition is ongoing at this time, and the Telecom Italia Big Data Challenge[3]. In these initiatives, mobile operators publicly disclose datasets of mobile traffic, and ask the community to carry out analyses that can answer specific societal challenges.

The impact of operators' challenges is notable. In Fig. 1b, we mark the dates of the main international venue dedicated to mobile traffic analysis, i.e., NetMob[4]. The jump in the number of publications observed in early 2013 corresponds to the 2013 edition of that conference, where the results of the first D4D Challenge were presented. This gives a rather clear idea of how similar initiatives can prompt research activities in the field. An event like NetMob is also interesting in that it captures the heterogeneity of applications of mobile traffic analysis. Sessions span over many domains, from transportation systems to graph theory, from health to privacy, from social structures to network management.

The aim of this manuscript is to provide an introductory guide to the state of the art in mobile traffic analysis. To the best of our knowledge, there exists only two, very recent, previous effort in that direction. Shang *et al.* [1] provide an overview of several works that collect and employ cellular phone data for studies on social networks, mobility, monitoring and estimation, or business applications. Blondel *et al.* [2] compile a significantly more extensive review of results on the analysis of mobile phone datasets, considering research on social networks, mobility, geography, urban planning, help towards development, and security. We believe that our survey extends both these works, introducing more comprehensive classification and discussion. On the one hand, we include in our study the vast literature on networking analyses that is neglected in previous reviews, and which is of capital interest to technology-oriented (e.g., computer science, telecommunications, engineering) research communities. On the other hand, we

[1] The larger circle in Fig. 1a maps to the fourty-some papers using the Data for Development (D4D) Challenge dataset, presented later in this section.

[2] `http://www.d4d.orange.com`

[3] `http://www.telecomitalia.com/tit/en/bigdatachallenge/`

[4] `http://www.netmob.org/`

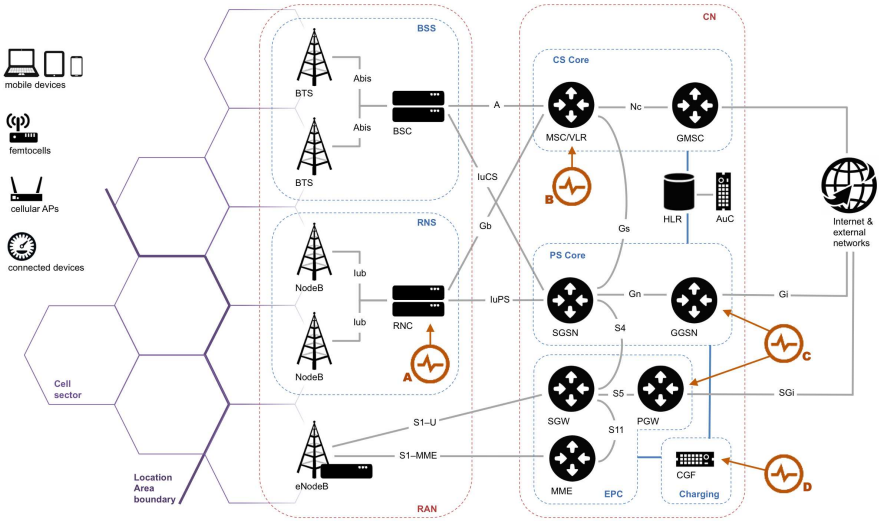


Figure 2: Simplified architecture of the cellular network encompassing different 2G, 3G and LTE technologies, and positions of probes for passive monitoring.

provide a compact treatise, focusing on major findings and methodologies rather than discussing sample results; in the same perspective, we also include per-category tables that provide an immediate guidance through the many and varied works on mobile traffic analysis, by summarizing the nature of datasets they employ, and the popularity and intertwining of research topics they address.

The document is structured as follows. We start by introducing, in Sec. 2, some basic notions about mobile traffic datasets, discussing current solutions for their collection and anonymization. Sec. 3 presents an overview of our proposed classification of mobile traffic analyses. Specifically, we separate the literature into four main categories, which are thoroughly surveyed in Sec. 4–6. A general discussion, including outtakes and pointers to main open issues, is then provided in Sec. 7. Finally, we draw conclusions in Sec. 8.

2 Mobile traffic data collection

The scope of this survey encompasses works dealing with data collected by probes that record traffic at different locations within the cellular network infrastructure[5], whose architecture is outlined in Fig. 2. Such a network grants access to telecommunication services and to the Internet by a wide range of devices: not only portable devices carried by mobile users, such as smartphones or tablets, but also meters or other types of machine-to-machine (M2M) communicating devices, as well as femtocells and cellular-connected Wi-Fi access points that bring local connectivity without the need for cabling.

2.1 Cellular network architecture: an overview

The network is composed of two main parts: a Radio Access Network (RAN), which provides wireless access to the individual devices, and a Core Network (CN), which manages all operations needed to

[5]According to the definition of Smoreda *et al.* [3], this corresponds to *passive* monitoring of mobile traffic, in contrast to *active* collection performed by operator-side platforms that periodically query end devices, typically for positioning information intended to enable location-based services.

transfer voice and data among different portions of the RAN as well as to and from external networks, including the Internet. The RAN is composed of base stations, each in charge of one or multiple cell sectors that jointly cover the geographical surface the network serves. End devices connect to the base station overseeing the cell section they are currently located in. Mobile devices may trespass the cell sector boundaries while exchanging data with the RAN, which generates a handover (HO) event to the new serving base station. Moreover, cell sectors are clustered into Location Areas (LA)[6] that represent the spatial granularity at which the device position is known at all times by cellular network, and it is thus used for paging. As a consequence, devices moving to a different LA are required to inform the network via a location update (LU) event, even if they do not have any ongoing communication at that time.

From a more technical perspective, base stations are referred to as Base Station Subsystem (BSS) and Radio Network Subsystem (RNS) in 2G (GSM, GPRS, and EDGE) and 3G (UMTS and HSPA) architectures, respectively. In both cases, base stations are composed of separated antennas (Base Transceiver Station, i.e., BTS, or NodeB) and controlling hardware (Base Station Controller, i.e., BSC, or Radio Network Controller, i.e., RNC). In the LTE architecture, the eNodeB gathers all base station functionalities.

At the CN, and considering 2G and 3G architectures, voice and texting services are managed via the Circuit Switched (CS) Core, whereas data (i.e., IP-based) services are handled by the Packet Switched (PS) Core. The main entities of the CS Core are the Mobile Switching Center (MSC) and the Gateway MSC (GMSC), which enable voice/text switching within the mobile network and with networks of different operators, respectively. In the PS Core, Serving Gateway Support Nodes (SGSN) and Gateway GPRS Support Node (GGSN) are the interfaces towards the devices and the Internet, respectively, and take care of packet-switched data transfers. In LTE, new entities are introduced to form the Evolved Packet Core (EPC). These manage the device control (Mobility Management Entity, or MME) and data (Serving Gateway, or SGW) planes, and interface them with other IP-based networks (Packet Data Network Gateway, or PGW).

2.2 Mobile traffic probes

Monitoring probes can be deployed at different locations within the architecture described above.
RNC probes, marked as *A* in Fig. 2, can be used to capture signaling events concerning any Radio Resource Control (RRC) operation. This allows to record fine-grained state changes of each device, and thus to detect device network attach and detach operations, start and conclusion of sessions, HO and LU events, related to any call, texting, or data transfer activity. Moreover, it allows collecting performance indicators on data transmission, such as the uplink and downlink throughput experienced by the device.
MSC probes, marked as *B* in Fig. 2, are similar to RNC probes, in that they can collect similar statistics. However, as MSCs are located in the CS Core, these probes can only track signalling related to voice and texting (and not to data traffic). Moreover MSCs control multiple base stations and thus events that are managed locally by a BSC or RNC (e.g., intra-base station handovers occurring among cell sectors under control of a same BSC or RNC) are transparent to the probe.
GGSN/PGW probes, marked as *C* in Fig. 2, tap at links in proximity of data gateways on the PS Core or EPC[7]. They record Packet Data Protocol (PDP) Context information on each data traffic session of every single end device, which necessarily transits by the GGSN/PGW in order to reach external IP networks (the Internet, typically). PDP Context includes session start and end time, device and user identifiers, traffic volume, type of service (i.e., transport- and application-layer protocols, class of service – such as web, email, streaming audio/video – and name of the application in some cases). In addition, GGSN/PGW probes can associate location information to PDP Context sessions. To that end, the probes

[6]The notion of Location Area, introduced originally in 2G, evolved with the development of new generations of mobile networks. Similar concepts, such as Routing and Tracking Area are described in 3G/4G. However, in this paper, we use Location Area as a generic term, denoting all these different technical definitions

[7]Many operators have co-located GGSN and PGW, which allows gathering information on 3G and LTE traffic at once [4].

can, e.g., monitor the authentication, authorization and accounting (AAA) procedures triggered by each PDP Context establishment or update. The messages generated during these procedures are exchanged over the G_i interface with the Remote Authentication Dial in User Service (RADIUS) Accounting server. They allow to map the IP address assigned to a device during a session to its IMSI and, more importantly, to its actual cell. In current network configurations, no information concerning voice or texting activities can be collected by GGSN/PGW probes.

CGF probes, marked as *D* in Fig. 2, retrieve data from the Charging Gateway Function. The latter is responsible of providing Call Detail Records (CDR) information to the billing domain of the mobile operator, where fees to be charged to the owners of the end devices are determined. It is precisely CDR that are collected by CGF: these contain start timestamp, duration, and originating cell sector of each voice, texting and data traffic activity of every device. Less frequently, CDR include additional information on the last cell sector of the activity and on HO events occurred during the activity.

The probes listed above all have strengths and weaknesses. As a general rule, probes located closer to the end devices (i.e., following the alphabetical order in Fig. 2) provide a more detailed view of the mobile traffic, but are more difficult to deploy and often less dependable in terms of uptime.

As an example, RNC probes deployed at all RNS allow observing all significant events occurring in the network, and thus provide accurate information about which cell sector each device is at all times[8]. This represents the ideal data for any study of user mobility or mobile traffic consumption. However, not all RNC equipment is designed to support probes, which, in any case, induce non-negligible computational and storage overhead on the RNC hardware. Moreover, RNCs are geographically distributed, which forces (i) the deployment and maintenance of a large number of probes[9] to cover a significant geographical area, and (ii) significant additional long-haul capacity to transfer all events to a central server.

On the contrary, a small number of GGSN/PGW probes deployed at the few data gateways necessary to cover a whole country allows to monitor mobile traffic much more efficiently. In addition, the information provided by such probes provides a rather detailed description of the IP traffic generated by each device, largely sufficient for studies on mobile traffic consumption. On the downside, no voice or texting data is currently recorded by GGSN/PGW probes. More critically, these probes only yield very approximated positioning information, updated only at the establishment of the PDP Context by an end device[10], or when the device moves across different SGSN or 2G/3G/LTE coverage areas. The latter events are quite rare, whereas cell sector changes that trigger HO or even LU events – instead very frequent in cellular networks – are not reported up to GGSN or PGW and thus go unnoticed. As a result, GGSN/PGW probes often have stale views of device locations.

The tradeoff is shifted in the case of CGF probes. On the one hand, the CDR they collect are readily available to mobile operators, typically at a single server for the whole network, and contain clean, well formatted information on millions of devices. This made such kind of mobile traffic source extremely popular in research. In addition, the mobility information yielded by CDR is more accurate than that provided by GGSN/PGW probes: despite the fact that CDR only include the starting cell sector of each activity, they track voice and texting sessions in addition to data ones, which leads to a higher sampling frequency of device position. Clearly, this also implies that voice and texting behaviors can be studied using CDR, which is instead not possible with PDP Context data. On the other hand, however, CDR do not provide any insight on the type of data traffic generated by the devices: the rich information on protocol- and service-level operations granted by GGSN/PGW probes is lost at CGF probes, which only

[8] We recall that cell sectors represent the finest spatial granularity achievable by passive monitoring in cellular networks, at least unless complex triangulations based mechanisms, using transmit power or timing advance information, are performed by the operator.

[9] This number can be two orders of magnitude larger than that of, e.g., CGF probes, at comparable geographical coverage.

[10] This maps to the time at which the device opens a data connection to the network. We remark that, once the connection established, a device may keep it open even if it switches to an idle state, and thus does not actually transfer data. The device can then become active again, and generate traffic over the same connection that was never closed. This leads to PDP Contexts that are not updated for hours even if the devices change location.

observe traffic volumes.

2.3 Mobile traffic anonymization

Independently of collection location, mobile traffic data contain information on many aspects of subscribers' life, including their activities, interests, schedules, movement, and preferences. It is precisely the possibility of accessing to such information at unprecedented scales that proves of critical importance for studies in many and varied research fields.

However, accessing such a rich source also raises concerns about potential infringements of the privacy rights of mobile customers: among others, individuals can be identified, their movements can be tracked, and their mobile traffic can be monitored. As a result, regulators have been working on laws intended to protect the privacy of mobile users. As an example, the European Data Protection Directive 95/46/EC mandates that all mobile traffic datasets be anonymized so that no individual is identifiable, before any cross-processing can be run on the data. Moreover, Directive 2002/58/EC states that anonymized data shall be analyzed only for the time necessary to provide the intended value-added service.

However, directives such as those above do not indicate any precise anonymization technique or privacy preservation model to be adopted during or after data collection. The reason is that there is still a high degree of uncertainty on this subject. On the one hand, there are many different notions of privacy that are not necessarily subset of each other, such as k-anonymity [5], l-diversity [6], t-closeness [7], and differential privacy [8], just to cite a well-known few. Which definition should be adopted, and under which conditions, is open to discussion. On the other hand, current anonymization algorithms aimed at guaranteeing the different privacy notions above are thought for standard tabular databases of static attributes, which are quite different in nature from mobile traffic datasets of subscribers' spatio-temporal activity. In fact, even the debate on whether user re-identification represents an actual threat to subscribers or not is still on-going [9, 10].

Overall, no definitive solution exists today to protect mobile users from privacy breaches that represent a certain risk – in the first place because the latter are not yet clearly defined. The result is that, so far, operators have considered naive techniques to preserve the privacy of customers. In most of the previous works, subscribers are anonymized by replacing their unique identifiers[11] with random sequences that allow to pinpoint a single user but hide his/her actual identity. Several works have focused on the issues of such an approach, and proposed solutions based on generalization and suppression of data. We refer the reader to Sec. 6.2.3 for a technical discussion of the topic.

3 Survey organization

The literature on mobile traffic analysis is very heterogeneous – a consequence of the large number of disciplines for which datasets collected by mobile network operators represent an important asset. Structuring the relevant works in a comprehensive way is not trivial: one needs to harmonize research originating from domains such as physics, sociology, epidemiology, transportation, and, obviously, networking. At the same time, forcing a neat separation among results obtained in each of such domains is a limiting approach, which would lose the significant overlaps and reciprocal references existing across disciplines.

Our survey is thus organized around research subjects, each of which features multidisciplinary contributions. The global outline of the proposed classification is shown in Fig. 3. At the top layer, we identify three macro-subjects of research at the interface of multiple domains: they deal with the analysis of social, mobility, and network properties, respectively. Then, a hierarchy of topics is developed within each macro-subject. Below, we provide an overview of the themes addressed across the classification.

[11] Typically, IMSI, IMEI, or the phone number.

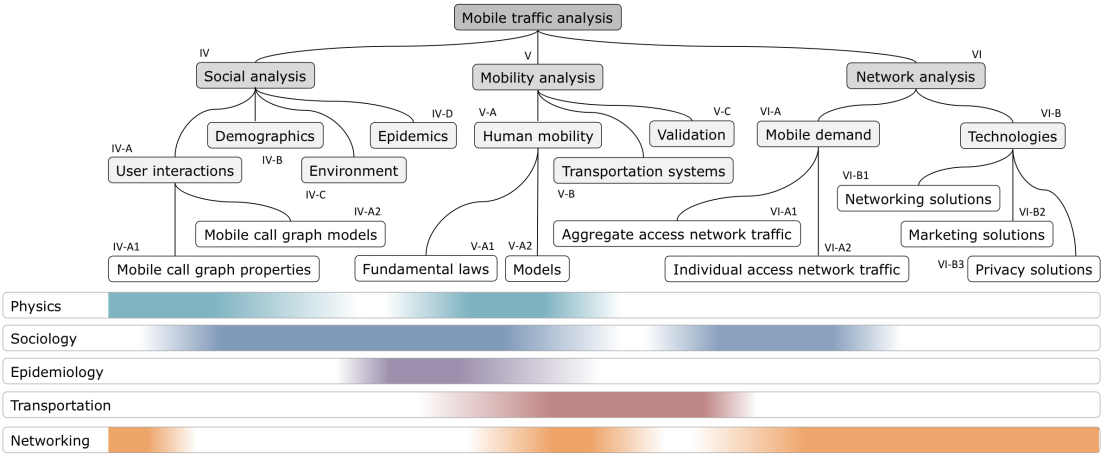


Figure 3: Proposed classification of the mobile traffic analysis literature, including the spectrum of disciplines related to topics.

Social analyses (Sec. 4) investigate the relationships between mobile traffic and a wide set of social features. The major research focus is on the characterization of the social structure of mobile users' interactions (Sec. 4.1), and on the study of how demographic, economical, or environmental factors influence the way users consume mobile services (Sec. 4.2 and Sec. 4.3). We also consider in this category works that leverage social features, inferred from mobile traffic, for the characterization and mitigation of disease epidemics (Sec. 4.4).

Mobility analyses (Sec. 5) deal with the extraction of mobility information from mobile traffic. Mobility is intended here in its broadest acceptation, and includes generic human movements at both individual or aggregate levels (Sec. 5.1), as well as specialized patterns that concern specific users, e.g., traveling on transportation systems (Sec. 5.2). We also review in this section the quite extensive literature on the dependability of mobile traffic data as a source of mobility information (Sec. 5.3).

Network analyses (Sec. 6) take a more technical perspective, as they are interested in understanding the dynamics of the mobile traffic demand, and how to evolve the mobile network infrastructure to better accommodate it. Works in this category thus focus on either the characterization of mobile service usages (Sec. 6.1) or on the exploitation of such knowledge to devise improved technological solutions (Sec. 6.2).

The vast majority of the categories outlined above are interdisciplinary by their own nature. In the lower portion of Fig. 3, we provide a representation of the relevance of five major research domains to the different topics of mobile traffic analyses. Relationships are necessarily not sharp, but we can remark that mobility studies are those attracting the highest variety of contributions. Most categories are significant to two or three disciplines. The only non-multidisciplinary subjects concern the development of novel solutions for mobile networks: being quite specific and very technical topics, it is understandable that they attract contributions solely from the networking community.

Throughout our discussion of the classification hierarchy in Fig. 3, we try to balance two aspects: (i) the comprehensive overview of the main results achieved by mobile traffic analyses in the considered theme, across disciplines; (ii) the introduction to significant details of the methodology adopted to obtain such results. The former represent the primary output of the research activities, and are presented in the main text. Methodological aspects that go into some technical depth are instead introduced only when required and in footnote, so as not to break the flow of the text. We thus suggest that readers interested in grasping fundamental outcomes of state-of-the-art research in mobile traffic analysis go through the main text, skipping technical footnotes. Readers willing to dig into some detail on a specific subject may

instead refer to the technical footnotes associated to that topic.

As a final remark, we report in Fig. 3 the number of the (sub-)section where each classification subject is addressed, for the reader's ease of reference.

4 Social analysis

The scale and granularity of social studies has been historically limited by the considerable costs of collecting meaningful data. Extensive, statistically reliable population surveys require significant economic and organization efforts, may take a long time, and cannot be guaranteed to be free of biases introduced by the sample selection or survey methodology.

From this perspective, the availability of datasets describing the dynamics of millions, such as those collected by mobile operators, is a definite game changer. Still, social studies often require information that is not present in mobile traffic data: the latter is thus complemented with traditional surveys, including national and regional demographics and statistics, or supplementary personal notions, including users' age, gender, employment, or revenue.

We identify four main research directions where social studies have enjoyed particular benefit from mobile traffic analysis. The first is the investigation of the structure of interactions among mobile subscribers, typically represented as a so-called mobile call graph. Results on properties and models of such particular graphs are surveyed in Sec. 4.1. The second subject is the exploration of the interactions among demographic factors and mobile communications, by means of cross-correlation of mobile traffic and personal subscriber information databases. The main results on this subject are presented in Sec. 4.2. The third topic concerns the relationships between the environment, in terms of both geographical and temporal features, and the communication structure. We review the related works in Sec. 4.3. The fourth research direction relates to epidemiology, since mobile traffic provides massive information on human movements and interactions that are critical to better understanding how viral diseases propagate. We discuss these latter works in Sec. 4.4.

To ease the reader's access to the studies on his/her topic of interest, we give an overview of the works discussed in this section in Tab. 1. The table also provides a quick access regarding the size and geographical coverage of the used datasets, as well as information regarding the supplementary data used in these studies.

4.1 User interactions

Understanding the complex structure of mobile user interactions is a challenging task that has implications in physics, sociology and also networking, since this knowledge can be used to, e.g., understand service adoption or anticipate evolutions in the customer base (see also Sec. 6.2.2 on this subject).

The vast majority of studies on mobile data characterization employ graph representations that allow adopting well-known analysis techniques issued from graph theory. We present the main results of these studies in Sec. 4.1.1. Another significant research line aims at understanding the reasons behind the structure of such graph representations, and develop so-called graph generative models. Such models can create synthetic graphs of mobile data whose features mimic those of graphs extracted from real-world datasets. Sec. 4.1.2 is dedicated to works in the field of graph generative models.

4.1.1 *Mobile call graph properties*

Mobile traffic datasets are very often represented as *mobile call graphs*. A mobile call graph $\mathbb{G} = (\mathbb{V}, \mathbb{E})$ is a mathematical structure describing a set of mobile users, which map to the set of vertices $\mathbb{V}$, and their interactions (i.e., exchanged voice calls or text messages), which correspond to the set of edges $\mathbb{E}$

	Analysis		Dataset							Focus											
	Name	Date	Operator	Area	Time	Users	V	T	ED	FT	Di	GS	GM	AG	EL	EF	Ge	UL	SE	EC	PC
Mobile call graph	Nanavati [11]	11/06	–	4 Indian regions	1 month	2.7 M	✓		–	From	✓	✓	✓								
	Doran [12]	12/12	–	unknown country	3 weeks (2011)	3 M	✓		–	From	✓	✓									
	Onnela [13]	02/07	–	European country	18 weeks	7.2 M	✓		–	From	✓	✓									
	Lambiotte [14]	09/08	Mobistar	Belgium	16 months	2.5 M	✓	✓	–	From	✓	✓	✓								
	Seshadri [15]	08/08	Sprint	4 USA regions	2 months	2 M	✓		–	From	✓		✓								
	Karsai [17]	02/14	–	European country	18 weeks	6.2 M	✓		–	From	✓	✓									
	Onnela [18]	05/07	–	European country	18 weeks	7.2 M	✓		–	From	✓	✓									
	Hidalgo [19]	05/08	–	unknown country	1 year (2004/05)	2 M	✓		–	From		✓									
	Miritello [20]	04/13	Telefonica	Spain	19 months	20 M	✓		–	From		✓		✓							
	Palla [21]	04/07	–	–	1 year	4 M	✓		–	From		✓									
Demographics	Yang [24]	06/09	–	Chinese city	6 months	300 K	✓		Demographic	To				✓							
	Sarraute [25]	08/14	–	Mexico	3 months	500 K	✓	✓	Demographic	To				✓							
	Stoica [26]	11/10	Mobistar	Belgium	6 months (2006/07)	3.3 M	✓	✓	–	To				✓							
	Mehrotra [27]	03/12	–	Rwanda	4 years (2005/09)	1.2 K	✓	✓	Demographic	To				✓							
	Wang [28]	05/13	Sprint	USA	1 month (2010)	20 M	✓		Demographic	From				✓							
	Brea [29]	08/14	–	Mexico	3 months	70 M	✓	✓	Demographic	From				✓							
	Blondel [30]	03/08	Mobistar	Belgium	6 months	2.04 M	✓	✓	Demographic	From		✓			✓						
	Toomet [31]	05/12	–	Tallin (Estonia)	1 year (2009)	32 K	✓	✓		From					✓		✓				
	Morales [32]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Demographic	From					✓						
	Bucicovschi [33]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Demographic	From					✓						
	Soto [34]	06/11	–	City in Latin America	6 months (2010)	500 K	✓	✓	Demographic	From						✓					
	Smith [35]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Economic indicators	From						✓	✓				
	Mao [36]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Economic indicators	From					✓	✓					
	Wakita [37]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	–	From						✓	✓				
	Fajebe [38]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Commodity prices	From						✓					
	Lim [39]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Economic indicators	From						✓					
	Frias-Martinez [40]	01/12	Telefonica	City in Latin America	6 months	500 K	✓	✓	Demographic	From						✓					
	Krings [41]	05/13	–	Brazil	2 months	6 M	✓	✓	Employment details	From					✓	✓					
Environment	Onnela [42]	04/11	–	European country	1 month	3.4 M	✓	✓	–	To						✓	✓				
	Krings [43]	07/09	Mobistar	Belgium	6 months (2006)	2.5 M	✓		User billing address	To							✓				
	Schmitt [44]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	–	From								✓			
	Eagle [45]	08/09	–	African country	4 years (2005/08)	1.4 M	✓		Regional census	From						✓	✓	✓			
	Almeida [46]	09/99	Telecel	Lisbon	3 days (1997)	–	✓		–	From								✓			
	Soto [47]	06/11	Telefonica	Madrid and Barcelona	1 month (2009)	3 M	✓	✓	Land usage	From								✓			
	Trestian [48]	11/09	–	5000 km^2	1 week	281 K	✓		Data traffic	From								✓	✓		
	Vieira [49]	08/10	Telefonica	2 metropolis	4 months	1 M	✓		–	From								✓			
	Pulselli [50]	06/08	Telecom Italia	Milan, Italy	2 months (2004)	–	✓		–	From							✓	✓			
	Naboulsi [51]	04/14	Orange	Abidjan, Ivory Coast	5 months (2011/12)	18 K	✓	✓	–	From								✓	✓		
	Girardin [52]	10/08	Telecom Italia	Rome, Italy	3 months (2006)	–	✓	✓	–	From							✓	✓			
	Candia [53]	07/08	–	230400 km^2	–	–	✓		–	From									✓		
	Calabrese [54]	11/10	AirSage	Boston	6 weeks (2009)	1 M	✓	✓	Event list	From									✓		
	Dixon [55]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	–	From									✓		
	Gowan [56]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Event list	From									✓		
	Bagrow [57]	03/11	–	European country	3 years	10 M	✓	✓	Event list	From				✓					✓		
	Linardi [58]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Event list	From									✓		
Epidemics	Wesolowski [59]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Health surveys	From										✓	
	Enns [60]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Demographic	From										✓	
	Gavric [61]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Health surveys	From										✓	
	Baldo [62]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Demographic	From										✓	
	Ndie [63]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Health surveys	From										✓	
	Chunara [64]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Health surveys	From										✓	
	Azman [65]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Health and meteo surveys	From										✓	✓
	Tizzoni [66]	09/11	Orange	3 countries	–	6.8 M	✓		Demographic	From				✓						✓	
	Frias-Martinez [67]	05/12	Telefonica	Mexico	6 months (2009)	1 M	✓		Health surveys	From										✓	
	Frias-Martinez [68]	09/11	Telefonica	Mexican city	6 months (2009)	2.4 M	✓		Health surveys	From				✓						✓	
	Saravanan [69]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Health surveys	From				✓						✓	✓
	Leidig [70]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Health surveys	From										✓	✓
	Kafsi [71]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	Demographic	From				✓						✓	✓
	Lima [72]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓	–	From										✓	✓

Table 1: Main features of works that leverage mobile traffic data for social analysis. In the analysis columns, date is in MM/YY format. In the dataset columns, V is voice, T is texting, and ED is complementary external data. In the focus columns, FT indicates if the paper studies social properties appearing in the dataset (From) or the impact of societal issues on mobile phone data (To), Di is mobile call graph distributions, GS is graph structure, GM is graph generative models, AG is age and gender, EL is ethnicity and language, EF is economic factors, Ge is geography, UL is urbanization and land use, SE is special events, EC is epidemics characterization, PC is epidemics prevention and control.

connecting pairs of vertices. This generic definition can accommodate a number of variations, depending on whether, e.g., edges are directed or undirected, weighted or unweighted, or subject to filtering rules.

In fact, there is no unique definition of a mobile call graph, and a variety of alternatives is found in the literature, as presented next. However, independently of the graph construction methodology, there exists a limited set of metrics that yield most of the significant information about the mobile call graph structure. We employ these metrics to classify relevant works in the following.

Degree distribution. The vertex degree distribution is the statistical distribution of the number of vertices connected by edges to a single other vertex. It conveys information about the basic structure of communications among mobile users.

In a seminal work, Nanavati *et al.* [11] construct an unweighted directed graph, which preserves the caller-callee relationship (as edges point to the latter), but loses any information on the number or duration of interactions between pairs of users (as edges do not depend on the intensity of the interactions). The authors observe that the in- and out-degree of vertices[12] both follow power law distributions[13]. The parametrization of the power law is however different for the in- and out-degree, with an exponent taking values between 2.7 and 2.9 for the in-degree, and between 1.5 and 2 for the out-degree. Nevertheless, the correlation between the two metrics at a same node is strong, implying that mobile users that call more people also tend to be called by a larger set of individuals. However, vertices with a very high in-degree (e.g., customer service numbers) or out-degree (e.g., salesmen) lose that correlation. Similar conclusions are drawn by Doran *et al.* [12], although with slightly different power law parameters (an exponent of 3.41 for the in-degree and 2.63 for the out-degree).

The node degree power law distribution seems to be consistent over different modeling choices, as shown in [13], where the authors consider a *mutual* mobile call graph, with an undirected edge connecting two vertices if at least one reciprocated pair of calls was exchanged between the corresponding users. In this graph, the notions of in- and out-degree coincide, and the node degree is characterized by a power law with a much faster decay, an exponent of 8.4, implying that the number of high-degree vertices is much lower than that measured when including one-way interactions.

In yet another different approach, Lambiotte *et al.* [14] consider a *constrained* mobile call graph, where an undirected edge connects two vertices if a minimum number of reciprocated calls exists between the corresponding users during a given time period. Specifically, the authors consider that at least 6 reciprocated calls must be present in a 6-month dataset for the relative edge to be present. The vertex degree distribution follows a power law in this case as well, with an exponent of 5.0. Yet, the results indicate that the power law models accurately only the tail of the empirical distribution, but not its head. A similar conclusion is drawn by Seshadri *et al.* [15] on multiple versions of an undirected mobile call graph. The authors consider both unweighted and weighted versions of the graph, with two types of edge weights: the total call duration between the pair of users, and the total number of calls they exchanged. In all cases, power laws are found to fit the tail of the degree distributions, but not the head. Instead, a Double Pareto Log Normal (DPLN) distribution[14] yields a good fit for the full vertex degree range.

Other power law distributions. Power laws characterize not only the tail of the vertex degree distribution, but other features of mobile call graphs as well. A first example is that of edge weights, as shown by Karsai *et al.* [17] in undirected weighted graphs, with the edge weight representing the number of calls between pairs of users. However, Onnela *et al.* [13] find that a different weight definition, the total call duration between two users, can introduce a cutoff in the distribution, leading to an exponentially-truncated power law[15].

[12] The in-degree of a vertex, d_{in}, is the number of directed edges that end at the vertex. Equivalently, out-degree of a vertex, d_{out}, is the number of directed edges that originate at the vertex.

[13] Denoting as d the in- or out-degree, then $P(d) \sim d^{-\gamma}$, where the exponent γ is inversely proportional to the presence of highly connected vertices (also referred to as *hubs*) in the graph.

[14] The DPLN distribution is a mixture of lognormal distributions. Its complete formulation is rather complex, and, for the sake of brevity, we do not provide it here. A detailed discussion is provided in [16].

[15] Denoting as w the edge weight, then $P(w) \sim w^{-\gamma} e^{-w/k}$, where k is the weight at which the exponential cutoff occurs, i.e.,

A second case is that of spatio-temporal properties. Karsai *et al.* [17] disaggregate the mobile call graph over time, and study the users' activity rate, i.e., the probability of a vertex to be involved in an interaction at each unit time. They find the distribution of the activity rate to be heavy-tailed, with an exponent of 2.8. On the spatial side, Lambiotte *et al.* [14] associate geographical information from billing ZIP codes to the vertices of the graph, and find that a power-law gravity model[16] well approximates the probability that two mobile users living at a given distance are connected in the graph, i.e., call each other.

Assortativity. A graph is assortative if its vertices tend to connect to other vertices with similar degree. This property, also known as assortative mixing, is typical of social networks. On the contrary, in a disassortative network high-degree nodes tend to connect to low-degree ones and vice versa.

In the case of directed mobile call graphs, Nanavati *et al.* [11] show that assortative mixing is only present for the in-degree, whereas the out-degree graph is even weakly disassortative. Undirected graphs appear instead to be always assortative, as shown by Onnela *et al.* [13].

In [13], the authors extend the assortativity analysis to edge weights, comparing the average weight of a vertex's edges to that of its neighbors. The outcome is dependent on the definition of edge weight: the graph is weight-assortative if edges are associated with the number of calls exchanged by mobile user pairs, but it is not in case total call durations are used as edge weights.

Structural role of vertices and edges. Several studies have focused on the identification of vertices and edges that are especially important within the structure of the mobile call graph, so as to pinpoint mobile users and calling interactions that play key roles in the communication network.

The PageRank[17] algorithm is used by Nanavati *et al.* [11] to assess the importance of vertices in the mobile call graph. The results show that the rank, i.e., importance, of a user is tightly correlated to the in-degree of its vertex, or, in other words, to the volume of calls it receives.

Onnela *et al.* [13] focus on the importance of edges, rather than vertices. Specifically, they map edge significance to the role that an edge plays in maintaining the mobile graph structure robust, i.e., well connected. They find that several measures allow to rank edges according to their importance for the graph robustness: removing edges with the lowest weight, the lowest overlap[18], or the highest betweenness centrality[19] results in a rapid disintegration of the graph. In a follow-up, Onnela *et al.* [18] delve deeper into the relevance of the edge weight, which they name the strength of the tie between a pair of users. Interestingly, they find that the weight is correlated to the logical positioning of the edge within the mobile graph structure. High-weight edges, i.e., strong ties, connect members of a same community, whereas weak ties tend to build links among communities. This explains why weak ties are critical to the graph connectivity. The result is confirmed in a recent work by Karsai *et al.* [17].

Instead, Doran *et al.* [12] are only in partial agreement with the conclusion above. They rank edges according to their outlying behavior, i.e., how significantly the edge weight and overlap[18] deviate from the mean value in the graph, either positively or negatively. Their results suggest that the mobile call graph is composed of well-connected communities featuring non-outlying edges. These communities are kept together by a backbone of outlying edges.

---

for which it becomes very unlikely to find edges. In [13], $\gamma = 1.9$ whereas k is equal to $3.4 \cdot 10^5$ s, implying that reciprocated calls lasting more than 30 minutes/week are rare.

[16] The gravity model commends that a measure decreases as a power of distance. Denoting as d_{ij} the geographical distance between mobile users i and j, their probability to be connected is $P(d_{ij}) \sim d_{ij}^{-\gamma}$. In [14], $\gamma = 2$.

[17] PageRank is a random-walk-based algorithm used to rank webpages in the Google search engine. PageRank computes the rank $r(i)$ of a vertex i as $r(i) = q/N + (1-q)\sum_{j:j\to i} r(j)/d_{out}(j)$, where N is the total number of vertices in the graph, $j \to i$ indicates an edge from j to i, $d_{out}(j)$ is the out-degree of vertex j, and $1-q$ is the *damping factor*, i.e., the probability to stop the random walk and start it again at a random graph vertex – the latter being modeled by the term q/N.

[18] The overlap of an edge connecting two vertices i and j is defined as $o_{ij} = n_{ij}/[(d(i)-1)+(d(j)-1)-n_{ij}]$, where $d(i)$ is the degree of vertex i, and n_{ij} is the number of neighbors common to i and j.

[19] The betweenness centrality of an edge connecting vertices i and j is defined as $b_{ij} = \sum_{v\in\mathcal{V}}\sum_{w\in\mathcal{V}/v}\sigma_{vw}(i,j)/\sigma_{vw}$, where $\sigma_{vw}(i,j)$ is the number of shortest paths between vertices v and w that traverse the (i,j) edge, and σ_{vw} is the total number of shortest paths between v and w.

Finally, Hidalgo and Rodriguez-Sickert [19] identify a relationship among the importance of vertices and edges in the mobile call graph structure. They measure the former as the vertex degree and the latter as the frequency with which an edge appears in the graph, i.e. edge persistence, and find that low-degree vertices tend to create more persistent edges.

Cluster structure. Real-world networks typically have non random organizations that are the outcome of, e.g., social interactions, or spatio-temporal constraints. As a result, their vertices and edges build precise internal structures that are not found in random graphs.

A classical feature of real-world networks is the presence of clusters, i.e. groups of vertices that are more tightly connected with each other than with other vertices in the graph. A typical metric for the level of clustering in a graph is the clustering coefficient[20]. Nanavati *et al.* [11] and Onnela *et al.* [13] measure the average clustering coefficient of either directed or undirected mobile call graphs, and find it to be similar to that of many other empirical networks that have non-random structures. Lambiotte *et al.* [14] add a geographical dimension to the analysis, by studying the distribution distances between ZIP areas of mobile users whose vertices form a triangle in the mobile call graph. They find that triangles are typically characterized by shorter geographical distances and, incidentally, call durations.

Another measure of the cluster structure within a graph is the presence of cliques, i.e. complete subgraphs in which all the nodes are connected to each other. Onnela *et al.* [13] confirm that this feature holds in mobile call graphs as well, as the number of cliques they observe is much more important than what expected in a random graph.

Temporal dynamics. A few studies have considered the evolution of mobile call graphs over time. Miritello *et al.* [20] perform a massive study on the evolution of an individual's mobile call graph, using a 19-month dataset covering more than 20 million mobile customers in Spain. The authors show that subscribers tend to renew their social circle slowly, as more than 75% of the ties remain active over the full timespan of the dataset. Moreover, a conservation principle of the social network size is observed, with a very similar number of activated and deactivated ties per individual. The results are also related to user demographics, since male users display larger social circles than women, and younger users have more contacts than older ones. Palla *et al.* [21] complement these results, by showing that large groups persist in time even in the presence of important membership turnovers, while small groups have a significant lifetime only if their composition remains unchanged.

4.1.2 *Mobile call graph models*

The natural step beyond the characterization of a complex network is the definition of models that capture its most significant properties. This has been the case with, e.g., classical models of the Internet or World Wide Web, such as Jellyfish [22] and Bow-Tie [23] models. A correct model of the mobile call graph has a number of applications, including: (i) explaining the generative process behind the formation of mobile call graph structures; (ii) creating call interaction networks from synthetic populations of mobile users; (iii) anticipating the evolution of the mobile demand.

Treasure-Hunt model. Nanavati *et al.* [11] were the first to propose a model of directed mobile call graphs. Their Treasure-Hunt model divides graph vertices into three groups, depending on whether they belong to the graph strongly connected component (SCC), are able to reach such a component (IN), or are reached by it (OUT). It then tells apart edges that connect IN-IN (entry), IN-SCC (in-tunnel), SCC-SCC (maze), SCC-OUT (out-tunnel), OUT-OUT (treasure), or IN-OUT (shortcut) pairs. The Treasure-Hunt model is shown to fit the directed mobile call graphs from mobile traffic datasets collected in four different regions.

[20]The clustering coefficient of a vertex i is defined as $c_i = 2t_i/d_i(d_i - 1)$, where d_i is the degree of i and t_i is the number of triangles to which i belongs. The average clustering coefficient of a graph is the average of all c_i's.

Lognormal multiplicative process. Seshrandi *et al.* [15] propose a method to build a synthetic mobile call graph, by studying the generative process of such a graph. To that end, the authors leverage datasets from two different time periods, and study the evolution of the user population and interactions. They conclude that the temporal growth of the graph follows a lognormal multiplicative process, already successfully used to model income distributions. Lognormal multiplicative processes result in the DPLN distributions that the authors found to characterize the vertex degree distributions, as discussed in Sec. 4.1.1.

Migration model. Lambiotte *et al.* [14] argue that classical models neglect the geographical distances associated to edges present in mobile traffic datasets. They thus propose a generative model where vertices are represented by agents, which can migrate from one region to another. Upon migration, an agent can either maintain its previous edges, or create new ones with vertices in the new region it moved to. The authors show that the migration model captures the geographical diversity of triangles in the graph, which are mostly composed of short-distance edges, yet at time include long-distance edges, as mentioned in Sec. 4.1.1.

4.2 Demographics

The most direct usage of mobile traffic for sociology purposes is probably the study of how communication and mobile device usage patterns relate to demographics. A number of such factors can be expected to shape the behavior of mobile users, including, e.g., their age, gender, and interpersonal ties. Below, we review the main studies that focus on such issues.

Age and gender. Age and gender are among the primary features from demography that play a major role in defining the behavior of a user. This was first indicated by Yang *et al.* [24], in an early study where they unveil a strong correlation between social and demographic elements. Using a six-month mobile traffic dataset covering a large Chinese city, and mixing it with subscribers' age and gender information, the authors find outthat people in the same age group communicate among them more often and for a much longer time, a result that holds throughout all age classes. Gender also plays a significant role, the results showing that calls between female users have a much longer duration than calls between male users.

Sarraute *et al.* [25] confirm the age homophily at a country scale, considering 500,000 users over the entire Mexico. However, cultural differences seem to play an important role on gender-related patterns, as men make more and longer calls than women in Mexico, i.e., the opposite of what happens in PRC. Gender impact on mobile communications has been further investigated in a number of other countries. Stoica *et al.* [26] study an even larger dataset of 3 million subscribers in Belgium, and show once again differences between genders, with average call duration longer for women. Mehrotra *et al.* [27] outline that gender also affects intra-day and inter-day calling dynamics in Rwanda. Specifically, they prove women to call much more than men at nighttime, whereas the trend is reversed during daytime. Gender differences also emerge with respect to special events, with women increasing their activity in proximity of, e.g., Valentine's Day or political elections, and men doing the same during Year's End holidays.

The significant impact of demographic factors on phone usage implicitly invites to develop techniques to automatically infer personal data of mobile subscribers from their calling profiles. Wang *et al.* [28] identify social characteristics like the age group, income level, and residential region of 20 million individuals, by leveraging homophily properties of the mobile call graph in combination with ground-truth data on a small user subset. The accuracy is in the 70-80% range in all cases. A similar approach is adopted by Brea *et al.* [29], who focus on age prediction of 74 million Mexican citizens. By using the correlation between demographic properties of users that are connected in the mobile call graph, the authors successfully classify up to 72% of the population into four age categories.

Ethnicity and language. In addition to genetic characteristics, also social features characterizing large groups of individuals have attracted significant attention in terms of mobile traffic analyses. In this per-

spective, most works have addressed the problem of recognizing ethnic groups from the network data, and understanding their properties and dynamics.

In a seminal work, Blondel *et al.* [30] analyse mobile traffic of 2 million users in Belgium, and show that the two main ethnic groups in the country, i.e., Walloons and Flemish, can be clearly inferred from the mobile call graph. To that end, they extract communities, i.e., sets of subscribers with strong communication ties between each other and with weaker connections to individuals outside the set. The problem of community detection, computationally expensive to solve in large graphs, is addressed by proposing an original technique, called *Louvain method*[21], which has hence risen to become the standard approach for community detection in all types of large datasets, not necessarily limited to mobile traffic.

Toomet *et al.* [31] study a mobile traffic dataset of Tallinn, Estonia, and identify two separate ethnic groups in the city. In addition, they investigate the spatial segregation between the two communities, and find that, while segregation exists in residential and work neighborhoods, the rest of the activities, e.g., shopping or entertainment, take place in a virtually non-segregated environment.

Morales *et al.* [32] separate[22] ethnic communities in Ivory Coast. Linguistic identity plays, rather unsurprisingly, a fundamental role in the ethnical separation. Also, mobile communication is shown to occur by preference within ethnic groups. An equivalent analysis is carried out by Bucicovschi *et al.* [33] in the same country, using a spatial approach[23]. The results confirm those above, as they show a general geographical correlation between the presence of mobile communication groups and the distribution of languages.

Economic factors. The socio-economic status of subscribers is characterized by three main factors: income, education and occupation. If measured at an individual level, these measures can indicate the role the person plays in the society. If averaged over a certain population, they are an important instrument to measure the development of a country or a region.

Soto *et al.* [34] define a comprehensive list of 279 mobile user features, and use machine learning methods to show that the economic levels of a customer can be predicted with an accuracy higher than 80% with only 38 such features[24]. As the result is obtained by still combining a quite large number of features, Smith *et al.* [35] argue that such a micro-measurement approach is too complicated and may lack transparency in the end. Therefore, the latter authors also use machine learning techniques, but target regions rather than individuals, and limit their analysis to four properties only: the sum of communication flows between the regions, the gravity residuals[25], the diversity[26], and the introversion[27]. They show that a limited training sample, as low as 10% of the total mobile traffic data, allows determining the poverty index of Ivory Coast regions, although the spatial granularity can be improved significantly with more complete training. In fact, correlations between poverty and mobile traffic on a per-region basis can be also found using simpler metrics, e.g., the volume of outgoing calls. Indeed, Mao *et al.* [36] find a negative relationship of the latter with economic indicators such as the poverty rate and annual income

[21]The Louvain method is a scalable heuristic based on modularity, i.e., a benefit function designed to measure the strength of a possible partition of a network into components. The Louvain method efficiently detects communities through an iterative two-steps process, repeated until the maximum modularity is achieved: the first step aims at optimizing the modularity locally, while the second step aggregates the nodes in the same community to create a new network.

[22]In [32], individual trajectories and language maps are employed to draw ethnical links among users. Then, a K-means clustering is run on the resulting graph, so as to identify the groups of users sharing strong interactions of ethnical nature. K-means is a partitional clustering algorithm that allows separating a set of items into K disjoint categories.

[23]In [33], a combination of gravity and Potts models is employed. A q-state Potts model is a spin model used to represent multi-body systems in statistical mechanics, and has important applications in segmentation problems.

[24]Key features include the number of weekly calls, the reciprocity of communication, the median of total number of calls, the individual area of influence, the radius of gyration, the total number of towers used, and the traveled distance.

[25]The gravity residuals are the errors between the real and estimated flows among each pair of regions u and v. The latter is $F_{uv} = g m_u m_v / d_{uv}^2$. There, g is a constant, m_u is the population of region u, and d_{uv} is the euclidean distance between the centroids of regions u and v.

[26]Considering v_{ij} to be the fraction of region i flow that goes to region j, the diversity of i is $\Delta(i) = -\sum_j v_{ij} log(v_{ij}) \log(v_{ij}) / \log(k_i)$, where k_i represents the number of regions to which region i is connected.

[27]The introversion of a region i is $I(i) = f_{ii} / \sum_{j \neq i} f_{ij}$, where f_{ij} is the flow between regions i and j.

of 19 regions in Ivory Coast. The authors explain this result by the fact that the communication fee is generally paid by the initiator of the call, and people in richer regions have greater means to start a call. Also, by exploring communities in the mobile call graph of each region, they show that rich areas have a tendency to split in many small communities, whereas poor areas display less heterogeneity and segregation in the communication patterns.

Wakita *et al.* [37] use mobile traffic to determine the industrialization level and the economic status of different regions in Ivory Coast. The authors first identify large cities as hubs of antennas with high social tie strength[28]. Then, they use time series of the average daily human activity to tell apart residential, working, and mixed zones in urban and non-urban areas. Their results show that the economy of cities in Ivory Coast is still largely dependent on agriculture, as urban areas do not show a clear separation of residential and working zones, except for the capital city, Abidjan. Further proofs are provided by Fajebe *et al.* [38], who find positive correlations between the mobile communication volume and the availability of commodities such as coffee, cocoa or palm oil in different regions of the same country.

Original metrics and tools have also been introduced in the attempt to fill the gap between mobile communications and economic development. Lim *et al.* [39] propose the concept of *social capital*, i.e., a series of social attributes with an economic impact. Using classical clustering approaches on the mobile call graph, the authors show that communities of mobile users with similar social capital can be found in the Ivory Coast population. Similarly, Frias-Martinez *et al.* [40] propose a tool named *CenCell* that infers the socio-economic level of mobile subscribers from the behavioral patterns obtained from their call records. CenCell attains 50% to 70% accuracy, depending on the classification type. On a related note, Krings *et al.* [41] leverage community detection techniques[29] in mobile traffic datasets so as to identify business leaders in the Brazilian economic system. The authors analyse the mobile communications of 6 million business subscribers working in 334,000 companies in Brazil, and individuate companies and their leaders with a 70% accuracy.

4.3 Environment

Not only the demographics aspects, but also the geographical and social environment where users reside affects their mobile communication patterns. Below, we summarize the main results concerning prominent environmental features that have an impact on mobile traffic.

Geographical distance. Geographical locations can induce important biases on many human habits, and telecommunication patterns are no exception. In a seminal work, Onnela *et al.* [42] focus on the most basic geographical property, i.e., physical distance. Using a one-month, country-wide dataset, they assign to each of the 3.4 million subscribers a geographical coordinate, corresponding to the base station they use the most. By studying the mobile call graph at the light of the distance of each user pair, they find that the probability of a tie, i.e., mobile contact, between two users follows a power law with respect to their distance[30]. Interestingly, the tie strength, the call volume between the two users, is shown not to vary with distance.

Analyzing the communities in the mobile call graph also allows Onnela *et al.* [42] to unveil the geographical properties of groups of individuals who maintain an important communication activity among themselves. The geographical span[31] of a community is found to depend on the size of the community:

[28]The strength of a social tie between two antennas i and j is computed as $w_{ij} = c_{ij}/(p_i p_j)$, where c_{ij} represents the number of calls made between the antennas, and p_i is the estimated population covered by antenna i.

[29]In [41], the authors employ the Louvain method to tell apart companies and sub-groups in each company. They then use an original metric of leadership, suggesting that leaders are not necessarily the users who communicate the most, but those who have ties with entities in all sub-companies, and that are also tightly linked to each other.

[30]Denoting as l the distance between a user pair, then the probability of a tie between the two users is $P(l) \sim l^{-\gamma}$, where γ = 1.5 in [42]. This means that a vast majority of communications are geographically bounded, yet there exists a heavy tail of long-distance ties.

[31]The geographical span of a community C is an indicator of how spread out are the n members of the community, and is defined

it is almost constant at around 50 km for communities with less than 30 users, then it sharply increases over 100 km for larger communities.

Krings *et al.* [43] group mobile customers by their billing address, and obtain a communication network between 571 cities in Belgium. By studying this graph, the authors show that inter-city communication follows a gravity model[32]. This result thus corroborates that mobile communication distance tends to be heavy tailed. On a related aspect, Schmitt *et al.* [44] also suggest that the average call duration increases as the inter-subscriber distance increases.

Urbanization and land use. Living in an urban or rural environment yields sociological differences that reflect on mobile traffic. Eagle *et al.* [45] use four years of mobile traffic data collected over a whole country to study the differences emerging between urban and rural users. The authors find that subscribers in urban areas communicate 50% more and with more people than those in rural areas, although the latter have, on average, longer conversations with their interlocutors. Schmitt *et al.* [44] complement these results, showing that some segregation exists between urban and rural regions, as users in rural zones tend to communicate more among them than with individuals living in cities. These trends do not change when considering migrations among the two types of areas: Eagle *et al.* [45] show that the call volume of individuals moving in urban areas increases, while the call volume towards the rural region of origin decreases.

In the urban context, several studies found a significant relationship between land use, i.e., the type of activity a geographical area is destined to, and mobile traffic in the region. In an early work, Almeida *et al.* [46] group base stations in Lisbon according to the land use of the area where each base station is located. They then study mobile traffic within the different groups, and find its temporal evolution to be similar in residential and suburban areas. Areas including major transport arteries yield instead a diverse temporal profile. Soto *et al.* [47] confirm that the nature of mobile traffic depends on the local land use, by adopting a reverse approach: they cluster[33] base stations on their traffic volume, and find the resulting groups to be associated to work, residential, hybrid, nightlife, and leisure regions – which are thus characterized by unique traffic profiles. As a result, also mobile traffic hotspots, i.e., high-activity locations, depend on land use. Trestian *et al.* [48] identify day, noon, evening and night hotspots in a metropolitan region, and find them to be correlated with the nature of the geographical area they reside in. Similarly, Vieira *et al.* [49] show how base stations in downtown undergo heavy loads during mornings of weekdays, whereas base stations in commercial and business areas become hotspots during the rest of the weekdays. In the weekend, hotspots appear around commercial and business centers in the morning and afternoon, and at commercial and night life areas in the evening and at night.

The difference in the spatial distribution of mobile traffic between working days and weekends is recorded by other works as well. Pulselli *et al.* [50] employ geographical plots of the aggregate daily demand in Milan, Italy, and note activity to be concentrated in the city center during weekdays, and in peripheral residential areas during weekends. Similar behaviors are found in considerably different environments, such as Abidjian, Ivory Coast, as discussed by Naboulsi *et al.* [51]. Again, land use appears to be a main explanation: as an example, Girardin *et al.* [52] detect a high level of activity close to the train station in Rome, Italy, during weekdays, whereas significant mobile traffic is generated during the weekends around the Colosseum, a major tourist attraction of the city.

Special events. Human-inhabited environments often feature special events that induce unusual mobile communication patterns. Events such as political happenings (e.g., elections or manifestations), entertainment occasions (e.g., concerts, sports games), and accidents (e.g., power outages or exception road

as $D = \frac{1}{n}\sum_{i \in C}\sqrt{(\tilde{x}-x_i)^2+(\tilde{x}-y_i)^2}$, where $(\tilde{x},\tilde{y})$ are the coordinates of the community geographical center, and (x_i, y_i) are the geographical coordinates of a user i belonging to the community.

[32]The gravity model defines the communication intensity c_{ij} between two cities i and j as $c_{ij} = p_i^{\alpha} p_j^{\beta} / d_{ij}^{\gamma}$, where p_i is the population of city i, and d_{ij} is the geographical distance between i and j. In [43], $\gamma = 2$.

[33]The authors apply K-means, with K chosen by a stopping rule maximizing the ratio of the inter-cluster to intra-cluster distances.

congestion) can produce anomalies in the cellular access network load, which can be detected by, e.g., clustering the spatiotemporal dynamics of mobile traffic. An early attempt is that by Candia *et al.* [53], who propose to detect anomalous events by measuring the gap between the current and mean number of calls occurring within groups of closely-located base stations. They find the methodology to be highly sensible to the gap threshold. Similar approaches have been taken, more recently, by Calabrese *et al.* [54] and Dixon *et al.* [55]: both leverage large variations in mobile traffic volumes to identify large-scale social events, national holidays, or power network outages. The former authors can even track back the origin location of crowds participating to events taking place in Boston, MA, USA.

More complex techniques for special event detection have also been proposed. Gowan *et al.* [56] use a hierarchical clustering technique to isolate the special communication patterns emerging during soccer games. Naboulsi *et al.* [51] introduce a dedicated framework to detect general outlying behaviors, based on the hourly geographical variations of mobile traffic. The authors can detect a number of special events, including national holidays, political happenings, and sport events.

Attention has also been paid to events that are not the result of social behaviors, but of natural or human-caused disaster situations. In an extensive study, Bagrow *et al.* [57] focus on emergency situations. Using a dataset covering 10 million users for two years, the authors select four such events occurring in the target region: a bombing, a plane crash, a mild earthquake, and a power outage. The mobile traffic activity following these events is compared with that of regular days, as well as with that recorded in presence of special planned events, such as concerts and festivals. While all the special events, both emergency and non-emergency, result in increased call volumes over the typical patterns, the mobile activity growth is immediate for actual emergencies, and more gradual for planned events. Moreover, the magnitude of the increase is correlated with the severity of the event: the bombing results in the highest number of calls, followed by the plane crash, the earthquake and the blackout. Diversity emerges also from a geographical perspective: in all cases, the activity change is the highest in proximity of the event epicenter, and exponentially decays with distance. When communication hops in the mobile call graph are considered, major emergencies propagate farther away from the epicenter: the activity following the bombing and plane crash events is shown to quickly reach three-hop neighbors of the eyewitness population. A similar study is carried out by Linardi *et al.* [58] on violent incidents occurring in Ivory Coast between 2011 and 2012. The authors show that such events are not preceded by any unusual calling activity, but are followed by an increased mobile traffic volume. Moreover, they also highlight an important medium-term effect, with a significant increase in the call volume enduring for several days after each violent episode.

As a final remark, we stress that correlations between special events and mobile traffic are also very relevant to network studies. While the works reviewed in this section concern the problem of detecting special events from the analysis of mobile traffic, networking research has mainly focused on the dual problem, i.e., the characterization of the impact of social events on the mobile demand. Indeed, the latter is critical to the design of networking solutions that can better accommodate any exceptional dynamics generated by unusual situations. Thus, we refer the interested reader to the relevant, although more networking-oriented, works in Sec. 6.1.1, *special dynamics* tag.

4.4 Epidemics

Mobile traffic encloses data about the movement of large masses of individuals. This kind of information, other than interesting per se, as thoroughly discussed in Sec. 5, is paramount to a better understanding of the spreading dynamics of infectious diseases. Indeed, by cross-correlating mobile traffic datasets with statistics on the propagation of contagious pathologies, it is possible to draw original models and propose containment solutions that effectively operate in very large-scale scenarios.

Epidemics characterization. Many works have investigated whether patterns present in mobile traffic can be correlated with the diffusion of contagious diseases. Indeed, identifying such relationships would

pave the way to very effective but extremely cheap techniques to anticipate and control outbreaks.

In a seminal work, Wesolowski *et al.* [59] study networks of mobile user movements and maps of malaria prevalence in Kenya, so as to identify relationships among common trajectories of human mobility and parasite infection. The authors are able to pinpoint several importation routes that foster the diffusion of malaria among different regions of Kenya. A similar approach is adopted by Enns *et al.* [60], and Gavrić *et al.* [61], who compare mobility and communication networks derived from mobile traffic to maps of malaria and HIV prevalence, respectively. The former authors find that the regions of Ivory Coast showing the strongest connections, in terms of both movements and mobile communication, are also those where the malaria parasite is the most present. The latter authors draw regression models based on mobile communication features that attain very strong correlations with HIV prevalence in the country. At the light of these results, both works suggest to account for mobility information when designing infectious diseases control strategies; this appears especially important for movements among regions of varying prevalence, so as to avoid malaria being carried from areas of high infection to areas of low infection.

Simpler analyses do not appear to yield equally significant information. For instance, Baldo *et al.* [62] explore spatial correlations among influenza cases and calls occuring in proximity of main hospitals in Ivory Coast, but their results show that there is no correlation between the two metrics. Ndie *et al.* [63] explore instead correlations between call exchange rates and HIV prevalence rates among different regions of Ivory Coast, but do not find significant correlations.

Mobile traffic also encloses data about the movement of a vast amount of individuals during contagion outbreaks, which allows refining traditional epidemics representations, such as the Susceptible-Infected-Recovered (SIR) model and its variants. The SIR model builds on a macroscopic approach, and divides the population into groups of people who are (i) susceptible to catch the disease, (ii) infected by the disease and capable to transmit it, and (iii) recovered from – and thus immune to – the disease. Each individual can then transit from the first to the third phase above. The standard SIR model can be augmented with geographical mobility information derived from fine-grained mobile traffic, as done by Chunara *et al.* [64]. They developed an extended SIR model by including an additional model stage, where so-called carrier individuals (hence the new model name, SCIR) diffuse meningitis through their physical movements. An alternative approach is proposed by Azman *et al.* [65], who parametrize a SIR model with transition rates that depend on mobility curves fitted on mobile traffic as well as on meteorological data. However, Tizzoni *et al.* [66] question the validity of SIR model variations based on mobile traffic. They consider three different European countries, namely France, Spain and Portugal, and they compare the results of a SIR model run on mobile subscriber commuting movements against those obtained when the same model is applied to reliable census data. The authors show that mobile traffic leads to overestimate the actual commuting flows, which in turn introduces some bias in the infection process. Still, the network data allows inferring somehow meaningful arrival times of the disease at different regions of a country, with an error of 2-3 weeks.

While the analyses above provide a macroscopic view of the epidemics, other works have focused on a fine-grained microscopic-level characterization. Frias-Martinez *et al.* [67, 68] have been using an agent-based model to capture social patterns that can explain the spreading of infectious diseases. Operating on a per-individual basis, their model can take personal features into account, unlike what happens with, e.g., aggregated SIR models. The results obtained using such a detailed model indicate that different countermeasures adopted by the Mexican government in occasion of the 2009 H1N1 flu outbreak have retarded the infection peak, and decremented its impact by 10%. However, the decisions did not impact the spatial evolution of the virus. Saravanan *et al.* [69] enriched the microscopic approach based on mobile agents above with the notion of importance of individuals. The latter information is extracted from the mobile call graph[34], and reveals especially useful for designing epidemics control policies targeted on

[34]In [69], influential members are identified by means of the Shapley value, which assigns a high score to subscribers who maintain a high number of connections with users who are instead scarcely connected. Formally, the Shapely value of a user i is

individuals.

Epidemics prevention and containment. Mobile traffic can be used not only to understand disease spreading, but also as an instrument of control. This has led to the proposal of solutions that mitigate the spreading of diseases and involve, to different extents, mobile communications.

Leidig *et al.* [70] propose to reduce the diffusion of the infection by rapidly spreading awareness of the danger. To that end, ego networks[35] are leveraged to identify[36] a limited set of key individuals who can propagate information about the disease in a rapid and reliable manner. Kafsi *et al.* [71] present three other strategies that aim to the same goal. The first strategy extracts trajectories from mobile traffic so as to detect geographically localized communities of users: then, a policy is enforced that forbids inter-community movements during outbreaks. The second strategy leverages the mobile call graph, and identifies social communities within it: then, inter-community contacts, deemed to foster the infectious process, are prohibited. The third strategy is adaptive with respect to the disease spreading status, as it avoids trips of mobile subscribers from regions of high prevalence to areas of low prevalence.

A more comprehensive study is provided by Lima *et al.* [72], who evaluate the contagion via a legacy SIR model, when (i) no countermeasure is adopted, (ii) geographic quarantine is enforced, and (iii) an information campaign is run among the population. The authors leverage country-wide mobile traffic data of Ivory Coast to model individual mobility, and to outline the mobile call graph on which informative communication occurs in the last case before. Results show that a geographic quarantine, despite being invasive, expensive and hard to enforce, only reduces the endemic size, but does not slow down the disease spreading. Instead, a collaborative information campaign attains a significantly lower fraction of infected individuals, even for low participation rates of the subscriber population.

5 Mobility analysis

Mobile data is an excellent source of knowledge on the movement of individuals. It can provide information about the mobility dynamics of populations of millions, impossible to obtain otherwise. Moreover, it allows doing so at virtually no operating cost. It is thus unsuprising that mobile data has rapidly established as a key new source in the field of mobility modeling, complementing and replacing traditional approaches based on, e.g., surveys or traffic counters. The novel mobility models obtained from mobile data are expected to affect a number of fields, including urban planning, road traffic engineering, human sociology, epidemiology of infectious diseases, or telecommunication networking.

In the following, we review the body of most relevant works that leverage mobile data to study human mobility. We distinguish three major subcategories. Sec. 5.1 discusses research on the characterization of human mobility, whose goal is to better understand and model how people travel at different spatial and temporal scales. Sec. 5.2 surveys the exploitation of mobile data for transportation research, where the aim is characterizing the usage of road and public transport infrastructures. Sec. 5.3 presents results on the reliability of mobile data for studies on mobility.

Tab. 2 summarizes the works reviewed in these sections, and provides an overview of the features of datasets they employ. It also outlines which works deal with each research aspect of mobility-oriented analyses of mobile traffic: it thus represents a useful quick reference for the reader.

$s_i^h = \sum_{j \in V_i^h} 1/(1+d_j)$, where V_i^h is the set of users within h hops from i in the mobile call graph, and d_j is the degree of user j.

[35] An ego network is a subset of the mobile call graph (see Sec. 4.1.1), pruned so as to form a tree structure rooted at a specific individual. It thus represents mobile user interactions from the perspective of that individual.

[36] In [70], the authors employ a dedicated measure of an individual's importance, which approximates the number of communities formed by his/her neighbors in the ego network.

		Analysis		Dataset							Focus														
		Name	Date	Operator	Area	Time	Users	V	T	D	S	Lo	Tr	RM	PM	SM	IM	AM	RT	OD	Mu	TP	Co	VP	VA	VI
Human mobility	Fundamental laws	Halepovic [80]	10/05	3G	–	1 week (2004)	4 K			✓		✓				✓	✓									
		Paul [81]	04/11	–	Nationwide	1 week (2007)	∼100 K			✓		✓		✓												
		Frias-Martinez [90]	12/10	–	city in emerging country	5 months	100 K	✓				✓														
		Isaacman [85]	06/11	AT&T	NY, LA	3 months (2009/10)	170 K	✓	✓			✓														
		Trestian [48]	11/09	–	5000 km^2	1 week	281 K			✓		✓		✓												✓
		Sridharan [93]	04/13	–	USA	1 month	10 M	✓				✓														
		González [97]	06/08	–	–	6 months	100 K	✓	✓				✓	✓												✓
		Calabrese [77]	10/11	AirSage	MA, USA	3 weeks	250 K	✓	✓	✓		✓	✓	✓						✓				✓	✓	
		Mitrovic [98]	05/13	Orange	Ivory Coast	5 months (2012)	50 K	✓	✓				✓			✓										
		Zang [100]	09/07	3G	3 cities	1 month (2006)	2 M	✓	✓	✓				✓												
		Schneider [88]	05/13	Orange	Paris, France	5 months (2012)	40 K	✓	✓		✓	✓		✓											✓	
		Song [101]	01/10	–	–	3 months	50 K	✓	✓						✓											
		Lu [102]	10/13	Orange	Ivory Coast	5 months (2012)	500 K	✓	✓						✓			✓								
		Lu [103]	05/12	Digicel	Haiti	1 year (2010)	1.9 M	✓							✓											
		Isaacman [104]	02/10	AT&T	NY and LA, USA	2 months)	170 K	✓								✓										
		Isaacman [105]	03/11	AT&T	NY and LA, USA	4 months (2009/10)	170 K	✓	✓							✓								✓		✓
		Rubio [94]	03/10	Telefonica	–	4 months	410 K	✓	✓				✓			✓										
		Dixon [55]	05/13	Orange	Ivory Coast	5 months (2012)	500 K	✓	✓							✓										
		Calabrese [54]	05/10	AirSage	Boston, USA	2 months	1 M	✓	✓	✓		✓	✓			✓										
		Girardin [52]	10/08	Telecom Italia	Rome, Italy	3 months (2006)	–	✓	✓							✓										
		Bengtsson [107]	08/11	Digicel	Haiti	7 months (2010)	1.9 M	✓								✓									✓	
		Linardi [58]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓							✓										
		Csáji [91]	06/13	Orange	Portugal	–	100 K	✓				✓				✓		✓						✓		
	Models	Song [92]	09/10	–	–	1 year	1 M	✓	✓			✓	✓	✓			✓									
		Cho [83]	08/11	–	one country	1 year	2 M	✓	✓			✓	✓	✓	✓		✓									
		Simini [108]	04/12	–	–	4 weeks	4.3 M	✓										✓								
		Isaacman [109]	06/12	AT&T	NY and LA, USA	3 months (2011)	600 K	✓	✓									✓								
		Mir [110]	10/13	AT&T	NY and LA, USA	3 months (2011)	600 K	✓	✓									✓								
		Yang [111]	05/13	France Telecom	Portugal and Rwanda	–	–	✓										✓		✓			✓		✓	
Transportation systems		Qiu [113]	01/07	–	Shanghai, PRC	1 year	9 M				HO								✓		✓					✓
		Bar-Gera [116]	12/07	Orange/ITIS	Ayalon freeway, Israel	3 months	–				HO								✓							✓
		Schlaich [73]	01/10	T-Mobile	Sout-West Germany	3 months	–				LU	✓	✓						✓							✓
		Janecek [117]	09/12	–	one highway, Austria	1 month (2011)	–				✓								✓		✓				✓	
		Calabrese [119]	03/11	Telecom Italia	Rome, Italy	Time	–	✓	✓		✓								✓		✓					✓
		Ma [79]	11/13	AirSage	Sacramento, USA	1 month (2010)	128 K	✓	✓	✓		✓	✓							✓					✓	
		Bekhor [75]	06/13	Orange/ITIS	Israel	1 week	160 K	✓			LU	✓	✓							✓				✓		
		Nanni [89]	05/13	Orange	Ivory Coast	5 months (2012)	500 K	✓	✓			✓								✓						
		Mamei [86]	05/13	Orange	Ivory Coast	5 months (2012)	500 K	✓	✓			✓								✓				✓		
		Wang [76]	09/10	AirSage	Middlesex, USA	1 month	56 K	✓	✓	✓		✓							✓		✓					
		Doyle [122]	06/11	Meteor	Ireland	1 week (2009)	1 M	✓	✓	✓											✓					
		Zilske [123]	05/13	Orange	Abidjan, Ivory Coast	5 months (2012)	18 K	✓	✓													✓				
		Berlingerio [78]	05/13	Orange	Abidjan, Ivory Coast	5 months (2012)	18 K	✓	✓			✓	✓									✓				
		Cici [87]	02/13	Telefonica	Madrid, Spain	3 months	4.7 M	✓	✓			✓										✓				
		Zhang [95]	09/14	–	Shenzhen	1 year (2014)	10 M	✓	✓				✓							✓		✓				✓
		Furletti [124]	08/12	–	Pisa, Italy	1 month (2012)	230 K	✓															✓			
		Scepanovic [82]	05/13	Orange	Ivory Coast	5 months (2012)	500 K	✓	✓			✓	✓			✓							✓			
		Liu [74]	05/13	Orange	Ivory Coast	5 months (2012)	500 K	✓	✓			✓											✓		✓	
Validation		Tizzoni [66]	09/13	Orange	3 cities	–	6.8 M					✓								✓			✓	✓	✓	
		Wesolowski [106]	02/13	–	Kenya	1 year (2008/09	14.8 M	✓	✓							✓									✓	
		Iovan [125]	05/13	Orange	Paris, France	1 day (2009	4 M	✓	✓	✓	✓															✓
		Ranjan [84]	07/12	3G	S.Francisco, USA	1 month (2011)	500 K	✓	✓	✓		✓														✓

Table 2: Main features of works that leverage mobile traffic data towards mobility analysis. In the analysis columns, date is in MM/YY format. In the dataset columns, V is voice, T is texting, D is data, and S is signalization (either HO, i.e., handover, LU, i.e., location updates, or ✓, i.e., both). In the focus columns, Lo is properties of visited locations, Tr is properties of travels between locations, RM is regular mobility patterns, PM is prediction of mobility patterns, SM is special mobility patterns, IM is individual mobility models, AM is aggregate mobility models, RT is road traffic conditions, OD is origin-destination flows, Mu is multimodal mobility, TP is transport planning, Co is commuter mobility, VP is validation of population distributions, VA is validation of aggregate mobility, and VI is validation of individual mobility.

5.1 Human mobility

The characterization of generic human mobility from mobile data aims at two different objectives: (i) the investigation of the fundamental laws that govern movement patterns, or (ii) the proposal of mathematical or simulative models capable of reproducing such patterns. For a more coherent discussion, we separate works relating to these two objectives in the following.

5.1.1 *Fundamental laws*

Laws derived from mobile data analysis can relate to multiple facets of human mobility, which we use below to structure the relevant literature.

Visited locations. How individuals visit geographical locations[37] represents the very first subject addressed by studies that employ mobile data to infer human mobility laws. The seminal work by Halepovic and Williamson [80] uses a relatively small dataset of 4,156 users, and describes the mobility of users in terms of the number of cells they visit. The authors find mobility to be generally low, as 55% of users only appeared at one location; yet, the distribution is heavy-tailed, i.e., there exist users who visit hundreds of cells in one week. The imbalance in user mobility is later confirmed by Paul *et al.* [81] in a much larger, nationwide dataset. They show that 60% of the customers are static, but 1% travel through 50 cells or more in a day, on average. Subsequent works have confirmed the heavy tail of the visited location distribution, e.g., that by Scepanovic *et al.* [82].

Halepovic and Williamson also outline the presence of one clear preferred location for every user, which they refer to as the *home* location[38]. Later, Isaacman *et al.* [85] prove the definition to be correct, as mobile data analysis can reveal the important locations of a user, including home and work locations, with a typical accuracy of 1 mile. Trestian *et al.* [48] confirm that trend, finding mobile subscribers to spend between 55% and 90% of their time, depending on their level of mobility, at the same three locations. Schneider *et al.* [88] study the distribution of the number of different locations visited daily by mobile subscribers. They find the distribution to be log-normal[39], with a small average value around three, thus confirming the low mobility of most users.

The temporal features of visits to locations are more thoroughly explored by Song *et al.* [92], on a larger dataset of 50,000 users. They study the time spent by an individual at a given location, finding that it follows a truncated power-law distribution[40]. They also investigate the number of distinct locations visited by a user over time, showing that humans have a decreasing tendency to visit new locations over time[41].

[37] Identifying the locations where a user stops from mobile data is non-trivial, as the latter provide an irregular sampling over time with low geographical accuracy. Stop locations are typically mapped to pauses in the movement longer than a time threshold [73–75], possibly allowing the user to dwell at multiple antennas within a space threshold [54, 76–79].

[38] The extraction of home (and work) locations from mobile data is again non trivial, and different authors used diverse approaches. As in the case of Halepovic and Williamson, several works tag as home the most popular location for each user [75, 83] Other authors adopt the same technique, but limit the study to night time, when users are more probably at home [48, 54, 77, 78, 84], Likewise, work locations are typically identified as the most frequent location during working hours [74, 84–86]. All such mechanisms can be complemented with antenna clustering, so that locations map to a group of nearby antennas rather than to a single one [66, 85, 87–89]. Notably different approaches have been proposed by Frias-Martinez *et al.* [90], who use ground-truth data to train a genetic algorithm, and by Csáji *et al.* [91], who classify antennas depending on their weekly time series and unveil three classes of frequent locations, easily mapped to work, home, and *other* locations.

[39] Denoting as n the number of visited locations, then $P(n) \sim exp\left[-(ln(n)-\mu^2)/(2\sigma^2)\right]/(\sigma n\sqrt{(2\pi)})$, with μ = 1 and σ = 0.5.

[40] Denoting as t the time spent at a location, then $P(t) = t^{-\gamma}exp(-t/k)$, where γ is the tail weight, and k is duration at which the exponential cutoff occurs. In [92], γ = 0.8 and k = 17 hours – the latter value matching the typical daily activity period of an individual.

[41] Denoting as $n(t)$ the number of visited locations at time t, then $n(t) = t^{\mu}$, with μ = 0.6. For random walks μ = 0.8 and for Lévy flights μ = 1, implying that these random models yield a much stronger tendency to visit new locations over time than found in real-world mobile data.

Recently, Sridharan and Bolot [93] have shown that a single distribution can describe the scaling properties of multiple features related to the locations visited by a user. Specifically, the Double Pareto LogNormal (DPLN) distribution[42] describes well the area of the minimum rectangle bounding all locations visited by a user, or the distance between groups of popular locations. Interestingly, such a property is invariant of the locale (i.e., the city considered) or geographical span (i.e., city- or country-wide) of the analysis.

Travel distance. The distribution of distances[43] between subsequent locations has also attracted significant attention. In a seminal work, González *et al.* [97] employ voice and text mobile data from 100,000 users to show that such travel distances follow again a truncated power-law distribution[44]. This result is in agreement with those of Halepovic and Williamson [80], as both imply that a large portion of the population is characterized by limited mobility, but there exists a non-negligible number of highly mobile individuals who travel over long distances.

The travel distance law above refers to the case where the displacements of all users are aggregated into a single distribution. Interestingly, González *et al.* [97] find that also distances traveled by *each* user follow truncated power-law distributions, with different cutoff values that map to the user's radius of gyration[45]. Since the low spatial and temporal granularity of voice and text mobile data used by González *et al.* might have biased the analysis (see also Sec. 5.3), Song *et al.* [92] carry out a similar study using data from 1,000 users whose location was recorded every hour, thanks to a location-based service they subscribed to. Yet, their results confirm[46] the truncated power-law nature of distances.

In fact, the truncated power law scaling of travel distance appears to be a global property, invariant of countries or continents. While the analyses above were performed on mobile data collected in European countries, similar results have been obtained by Calabrese *et al.* [77] and Mitrovic *et al.* [98] from mobile data collected in Massachusetts, USA, and in Ivory Coast, respectively[47]. The same distribution is retrieved also when computing all travel distances with respect to a user's home location[48], according to Cho *et al.* [83]. The same authors unveil an interesting twist, by relating mobile data to social georeferenced networks – namely, Gowalla and Brightkite. Cross-referencing the datasets allowed them to conclude that short-distance travels (below 100 km) are better explained by routinary behaviors, such as home-workplace patterns, while long-distance travels are much more influenced by social ties, such as the presence of friends.

As a final remark, we stress that the scaling properties of travel distances inferred from mobile data appear to hold over large geographical scales (i.e., in the case of country-wide and inter-urban movements) only. Recent works based on finer-grained sources, such as GPS-based tracking, public transport usage, or individual surveys, have shown that human travels within cities follow a different, exponential scaling [99].

Spatiotemporal regularity. One of the most talked-about results of mobile data analysis is that individ-

[42] See footnote 14 for more details on DPLN.

[43] Travel distances are computed over the trips or trajectories of each user. The extraction of the latter from mobile data is typically performed by mapping trips to sequences of geographical points (i.e., the positions where the user carries out some mobile traffic activity) between each two successive stop locations [54, 73, 77, 78, 94]. Some works also add a second phase where trips that are too short are aggregated [75], or trips that form a small-distance loop within a brief time interval are discarded [79]. Recently, state-of-the-art techniques used in GPS trajectory recontsruction have been also adapted to the case of mobile traffic data [95]. Database-inspired approaches have also been explored by Vieira *et al.* [96], who develop a query system to retrieve user trajectores from call detail record databases, under complex geographical and temporal conditions.

[44] Denoting as d the travel distance between two subsequent locations, then $P(d) = (d + d_0)^{-\alpha} exp(-d/k)$. According to González *et al.* [97], α = 1.75 and the exponential cutoff k = 400 km.

[45] The radius of gyration r_g is a unidimensional measure of the distance traveled by a user, which also keeps into account the direction of movement. It is computed as $r_g = \sqrt{1/n \sum_i (r_i - 1/n \sum_i r_i)^2}$, where $r_i, i \in [1, n]$ is a bi-dimensional vector describing the i-th location of the user.

[46] In [92], α = 1.55 and k = 100 km: the latter is limited by the 1-hour periodicity of sampling.

[47] Calabrese *et al.* found α = 0.78 and k = 60 km, while Mitrovic *et al.* found a cutoff at around 100 km.

[48] In the study by Cho *et al.*, α = 1.7 and k = 100 km: the latter is limited by the small geographical coverage of the dataset employed.

uals tend to have strong regularity in their movement patterns. That is true in both spatial and temporal dimensions, as first claimed by González *et al.* [97]. These authors show that: (i) the popularity of locations visited by a user follows a Zipf's law[49], thus individuals tend to have a few preferred locations, and a long tail of seldom visited ones; (ii) there is a strong habit by users to return to previously visited locations within 24 hours, which highlights the temporal periodicity of movements. Here again, the results by González *et al.* confirm those by Halepovic and Williamson [80], in that most individuals spend the vast majority of their time at a limited number of frequently visited locations.

Song *et al.* [92] employ their finer-grained mobile data, where the location of 1,000 users is monitored on an hourly basis, to validate both the Zipf's distribution of location popularity[50], and the 24-hour periodicity of movement patterns. Daily periodicity is also detected by Paul *et al.* [81] in a very large-scale dataset of millions of subscribers, and by Trestian *et al.* [48], who show that more than 70% of the mobile users revisit at least one same location on every single day. Cho *et al.* [83] confirm the strong geographic and temporal regularity of human mobility, observing that users tend to return to the same places and travel at similar times of the day. In addition, mobile traffic yields a strong periodicity in user movements not only at a daily scale, but also at a weekly scale, as proven by Calabrese *et al.* [77] and Zang and Bolot [100].

The latter authors also identify strong regularity in the precise sequences of cells visited by mobile users over time. A similar level of detail on the geographical regularity of users' movements is considered by Schneider *et al.* [88], who employ *motifs*, i.e., closed sequences of transitions among activity locations, and unveil that each customer's daily pattern in a 40,000-user dataset can be described through one of just 17 motifs. The limited number of motifs, which include no more than 6 locations each, further proves how human mobility dynamics are simpler than one could expect.

Predictability. The strong regularity of human mobility raises the question of how easy to predict are individuals' movements. In a seminal work, Song *et al.* [101] tries to answer that question, by investigating the theoretical maximum predictability of individual mobility patterns in a 50,000-user mobile traffic dataset. To that end, they define a measure of entropy that captures spatiotemporal ordering of the visited locations[51]: when computed over all users, the measure shows that users' movements yield very low randomness[52]: on average, 93% of individual movements are potentially predictable [53]. The authors also prove that such a result is due not only to the limited number of favorite locations frequently visited by each user, but also to the strong spatiotemporal correlation in such visits.

Also, Song *et al.* [101] show that movement predictability stays constant throughout very heterogeneous sets of users (e.g., for different gender, age, geographical attachment). Lu *et al.* [102, 103] confirm that such a high predictability of human movements holds also in the case of developing countries, and even after major events like natural disasters. Other factors may instead have an impact on the predictability. As an example, by using a similar analysis on a one-year-long mobile traffic dataset of 2 million users, Cho *et al.* [83] show that the entropy in the visited locations is lower (and thus users' locations are more predictable) at night hours, when people are at home, and much higher during weekends, when travel destinations are more varied.

Factors affecting mobility. A large number of factors can affect the diverse human mobility laws identified above. A typical example is that of movement patterns in areas with diverse topological features

[49] Given the rank l of a location, its level of popularity is described by $P(l) = l^{-\beta}$, with $\beta = 1$.

[50] In [92], $\beta = 1.2$.

[51] Given the complete mobility $\mathbb{l}_i$ of a user i, expressed as a sequence of locations $\mathbb{l}_i = \{l_1, l_2, \ldots, l_N\}$, the entropy in his/her mobility is expressed as $S = -\sum_{\mathbb{l}'_i \subset \mathbb{l}_i} P(\mathbb{l}'_i) log_2[P(\mathbb{l}'_i)]$. There, $\mathbb{l}'_i$ is one of all possible subsequences found in $\mathbb{l}_i$, and $P(\mathbb{l}'_i)$ denotes the probability of finding that precise subsequence in $\mathbb{l}_i$.

[52] The entropy distribution has a peak at a value corresponding to an uncertainty in the user's whereabouts of 1.74, i.e., less than two locations. For comparison, the entropy of random mobility implies an uncertainty of 64 locations on a similarly sized dataset.

[53] The maximum predictability is obtained from the entropy measure by applying Fano's inequality, which states that if a user with entropy S moves among N locations, then his/her predictability is bounded by a maximum value Π_{max} that depends solely on S and N, through $S = -\Pi_{max} log_2(\Pi_{max}) - (1 - \Pi_{max}) log_2(1 - \Pi_{max}) + (1 - \Pi_{max}) log_2(N - 1)$.

or development levels. In their studies, Isaacman *et al.* [104, 105] show that mobile data collected in two cities in the USA, New York and Los Angeles, yield very different mobility features. Similarly, Rubio *et al.* [94] demonstrate that mobile data can be used to show differences in the way people move in developed and emerging economies. In the latter context, and namely in developing countries such as Ivory Coast and Kenya, different levels of mobility, measured as the radius of gyration of subscribers, are found to be related to geographical regions and income levels by Scepanovic *et al.* [82] and Wesolowski *et al.* [106].

Seasonality, as shown by Isaacman *et al.* [104, 105], holidays, as shown by Dixon *et al.* [55], and public events, as shown by Calabrese *et al.* [54], are other examples of phenomena that can all affect in a significant manner the movement of individuals, by changing their standard attraction locations as well as the sheer volume of human mobility.

Even when dealing with non-typical mobility, mobile data analysis can reveal extremely useful. In an early work, Girardin *et al.* [52] show that mobile data can, e.g., help understanding the mobility of tourists. Within a different scope, Bengtsson *et al.* [107] demonstrates that mobile data analysis yields very accurate estimations of the mobility of people after natural disasters or large-scale epidemics. The same holds for violent episodes, as those that occurred in Ivory Coast between 2011 and 2012 and investigated by Linardi *et al.* [58]. In particular, the authors show a strong impact of such episodes on human mobility, with a reduction of inward mobility and an increase in the number of users leaving the region where the violence happened. An interesting result is that the impact on mobility is even observed beforehand, suggesting that these violent events are predictable, due to societal tensions prior to the outbreak.

From an engineering standpoint, correlations have also been found between the level of mobility of users and their demand in terms of mobile data traffic. Early results in that direction can be found in the works by Halepovic and Williamson [80], Mitrovic *et al.* [98], and Dixon *et al.* [55]. We point however the interested reader to more thorough discussions of traffic-mobility correlations carried out by networking papers that are reviewed in Sec. 6.1.2.

As a closing remark on this discussion, a relevant question is that of which factors are the most important in order to fully characterize users' mobility. An interesting study by Csáji *et al.* [91] proves[54] that the only relevant features are the average position of the user and the location of the two cells the user is most frequently attached to. This suggests that basic geographical information is already largely sufficient for a comprehensive analysis of subscriber movement patterns.

5.1.2 *Models*

Models of human mobility can either describe the movement of individual users, or aggregate dynamics of whole populations. Next, we classify models derived from mobile data analysis according to their granularity.

Individual mobility models. Individual mobility models represent the movement of each user independently. A first mobile traffic-inspired model of individual mobility was proposed by Halepovic and Williamson [80]. Their stochastic approach builds on (1) the empirical distribution of the number of cells visited by a user, and (2) the empirical distribution of cell changes by a user. The model generates the movements of a given user by extracting realizations of the theoretical functions fitting such two distributions.

Song *et al.* [92] propose a more refined model, based on analysis of a dataset containing the activity of one million users for one year. The model relies on two complementary phases. The *preferential return* phase, occurring with probability $(1-\rho)N^{-\delta}$, lets the user return to one of the previously visited

[54] Csáji *et al.* [91] use Principal Component Analysis (PCA) on 50 features that can be used to represent mobile users, and include, e.g., the number of visited locations, their geographical dispersal, the quantity and duration of calls. They find that that 95% of the information is yield by just 5 features.

N locations, chosen proportionally to the number of past visits. The *exploration* phase, occurring with probability $\rho N^{-\delta}$, lets the user choose a new location never visited before, thus incrementing N by one unit[55]. The residence time at a location (in both phases) and the distance of a new location (in the exploration phase) are drawn from the heavy-tail probability distributions identified to characterize individual human mobility as discussed in Sec. 5.1.1. The model is demonstrated to correctly reproduce travel distances and residence times at locations, and it respects the location ranking as well as the number of new visited locations over time, as seen in Sec. 5.1.1. However, the model only captures long-term scaling features, and neglects the temporal periodicity (e.g., regular returns at every 24 hours) and the sequential patterns (e.g., home-work-home) in the visited locations.

A third relevant model is devised by Cho *et al.* [83] to predict the location of each individual at different hours of the week. The model considers the N most popular locations for each user[56] and creates a spatial probability distribution that shifts over time among gaussian-shaped distributions centered at such location. A social component is also added to the model, in a way that a portion of the movements becomes driven by the previous locations visited by friends of the user. The model anticipates the exact user location 42% of the time, although a simpler model that assumes the user to be at his/her top location on a hourly basis attains 40% accuracy.

Aggregated mobility models. Aggregated mobility models describe the mass movement of a large number of users with low spatial granularity, e.g., among municipalities. Simini *et al.* [108] first found that such mobility is well described by the *radiation model*[57]. Such a model is found to match the distribution of traveled distances computed from mobile data of 4.3 million users over 4 weeks. Also, it significantly improves the well-known gravity model[58], although the latter has been shown to be highly representative of specific scenarios, e.g., commuting distances in Portugal studied by Csáji *et al.* [91]. A simpler approach, involving Markovian modeling is proposed by Lu *et al.* [102], who show how a first-order model is already sufficient to correctly predict 90% of daily human mobility in Ivory Coast.

Radiation and Markovian models are intended to capture mobility at low spatial granularity (i.e., large geographical regions), and are shown not to hold in the case of intra-urban mobility by Liang *et al.* [99]. Thus, when considering movements within a single urban area, different models are needed. Isaacman *et al.* [109] propose WHERE, a framework that extracts probability distributions (of home/work locations, commuting distances, and calls) from mobile and US Census data, and mixes such distributions so as to generate a synthetic model of mobility and calling behaviors. The framework yields daily traveled distances similar to those extracted from real-world mobile data. An extension to WHERE is proposed by Mir *et al.* [110], by including *differential privacy*[59].

More recently, an improved model, based on a combination of the gravity and radiation models was proposed by Yang *et al.* [111]. Evaluation against mobile traffic datasets shows that such a mixed model is effective at different scales and in scenarios from diversely developed countries worldwide.

[55] Calibration on mobile data yields $\delta = 0.21$, while ρ is specific to each user and can be extracted from a normal distribution with mean $\bar{\rho} = 0.6$.

[56] In [83], $N = 2$, as adding more locations yields minor improvements.

[57] The radiation model determines the mobility flux m_{ij} between two regions i and j as $m_{ij} = p_i K_c (p_i p_j / (p_i + p_{ij})(p_i + p_j + p_{ij}))$, where p_i and p_j are the populations in region i and j, K_c is the fraction of population that commutes to work, and p_{ij} is the the population in the circle centered at i and of radius equal to the distance between i and j.

[58] The gravity model has been long considered as the reference model for long-range mobility of people, animals and goods. It defines the mobility flux m_{ij} between two regions i and j as $m_{ij} = p_i^{\alpha} p_j^{\beta} / d_{ij}^{\gamma}$, where d_{ij} is the geographical distance between i and j.

[59] Differential privacy formalizes in a mathematically rigorous way the principle that results of a data analysis should not be significantly affected by the presence/absence of a single individual in the database, for any individual. Rigorously, an algorithm $\mathcal{A}$ is considered as ϵ-differentially private, i.e. providing a level of privacy equal to ϵ, if it fulfills the condition $e^{-\epsilon} P[\mathcal{A}(D_2) = O] \leq P[\mathcal{A}(D_1) = O] \leq e^{\epsilon} P[\mathcal{A}(D_2) = O]$, where D_1 and D_2 represent any couple of datasets that differ in one element and O is any output of the algorithm.

5.2 Transportation systems

Despite early criticisms, such as those expressed by Rose [112], the evaluation and enhancement of transportation systems has been among the first practical applications of mobile traffic analysis. Usage of mobile network data for intelligent transportation system (ITS) was first envisioned in the late '90s, and comprehensive literature reviews have been compiled by Qiu *et al.* [113], and Caceres *et al.* [114]. In the following, we summarize the main findings, classifying the different works according to the topic they address.

Travel time and traffic state. Wunnava *et al.* [115] authored an early survey on the efforts by private companies (typically contracted by telecom operators) to extract travel time and traffic state information from mobile network data[60]. They find mobile traffic-based technologies to be mature and to provide correct travel time estimates in presence of *free flow* road traffic conditions. However, the authors conclude that mobile traffic does not appear sufficient to accurately estimate congested road traffic conditions. Since then, a number of academic efforts, discussed below, has been carried out in that direction, as outlined below.

Qiu *et al.* [113] show that processing of handover information can lead to an average error in estimated travel times which is within 5–15% of those computed from traditional induction loop detectors. A similar conclusion is reached by Bar-Gera *et al.* [116], who compare speed and travel time measurements from mobile call data and handovers with the equivalent data from dual magnetic loop detectors: the error they record is within 10% of the actual value. The same authors also validate the results obtained from mobile data and loops against actual GPS recordings from sample vehicles, and find that both techniques yield an acceptable 5–20% error to the ground-truth GPS data. Schlaich *et al.* [73] further confirm the appropriateness of mobile traffic analysis for travel time estimation, with results close to those reported by transportation authorities. Also, they are able to correctly separate fast (e.g., private cars) and slow (e.g., trucks) traffic, as well as to identify special traffic patterns (e.g., congestion due to an accident), by using just mobile call information and location updates.

More recently, Janecek *et al.* [117] combined handover and location update data in their study. Specifically, they propose to use coarse-grained location updates, available from all switched-on mobile terminals, to estimate travel times and detect congestion. If congestion is observed, fine-grained handovers from terminals engaged in calls (around 1/20 of the total switched-on terminals in their scenario) is employed to localize and possibly classify the congestion event in a more accurate manner. The authors show that their location update-based technique can identify traffic anomalies faster than traditional systems (i.e., roadside sensors, toll data from trucks, GPS data from taxis, and FM radio broadcasts based on drivers' indications). Also, using handover data allows to identify the precise type of congestion, e.g., wide moving jams or milder synchronized flows.

A different perspective is taken by Caceres *et al.* [118], who study the problem of traffic volume estimation, rather than travel times. To that end, they map handovers to highway road traffic crossing cell boundaries. The authors develop a mapping function based on a wide range of physical properties, which is found to capture road traffic with a 20% relative error w.r.t. real-world traffic counts from detectors.

The good results obtained with highway traffic are not easily reproduced in the more complex and heterogeneous urban environment. The only work dealing with travel time estimation in city scenarios is that by Calabrese *et al.* [119], who develop a framework, named LocHNESs, allowing real-time localization and tracking of vehicles from mobile traffic. They obtain a 10–18% error with respect to ground-truth GPS data.

Origin-destination matrices. Origin-destination (O-D) matrices describe the number of trips performed, during a given time period, between each pair of locations within a geographical area. They are a standard way to represent the *travel demand* in transportation engineering. Using mobile traffic to infer O-D

[60] In the context of transportation research, handover and location updates are the kind of data typically used to estimate travel times [73, 113, 114, 116, 117], with rare exceptions [76].

matrices of human mobility in urban regions[61] was first envisioned by Bolla *et al.* [120]. A small scale evaluation on a minimal subset of mobile traffic collected during one morning was then performed by White *et al.* [121]. The first tests on mobile traffic datasets of significant scale were performed by Calabrese *et al.* [77] in Massachusetts, USA. They find a good agreement between O-D matrices obtained from mobile data and county-to-county trips extracted from a US Census survey. Results do not scale well when considering the more precise mobility among sub-county areas, although the authors state that it is not necessarily a problem of mobile traffic data; rather, it is an issue of the survey itself.

A similar result is obtained by Ma *et al.* [79], who can recreate a faithful O-D matrix that models the travel demand around the system interchange of two major highways in California, USA. The result is validated by the good match against an inter-city travel survey, as well as against flow directions at one highway ramp computed via automated plate number recognition. Other works successfully employed mobile traffic data to derive country-wide O-D matrices, such as those of Israel, by Bekhor *et al.* [75], or Ivory Coast, by Nanni *et al.* [89] and by Mamei *et al.* [86]. In particular, the aggregate mobility model mixing gravity and radiation approaches proposed by Yang *et al.* [111] can be effectively employed to generate O-D matrices of commuting patterns at very different scales, from individual cities to whole countries.

It is to be said that, despite the sucess stories above, some works question the capability of deriving accurate nationwide O-D matrices from mobile traffic only. As an example, by using census data as a reference, Tizzoni *et al.* [66] underline the poor capacity of the mobile traffic to properly account for the actual attractiveness of different destinations for commuters of a given location. A solution to this issue was recently proposed by Zhang *et al.* [95], who combine mobile traffic information with public transport data, i.e., taxi/bus GPS logs and subway transits, in order to derive O-D matrices. The proposed framework, named mPat, is capable of building dynamic O-D matrices, termed mobility graphs, from data collected on-the-fly, with an accuracy of 75%.

Multimodality. Multimodal transportation derives from the combined utilization of different means of transport (e.g., private or public, motorized or not, mass or individual). The analysis of mobile traffic has been employed to identify multimodal aspects of transportation systems, and namely to quantify populations using different means of transport[62].

Wang *et al.* [76] show that mobile traffic can be coupled with travel time information (from the Google Maps service in their case), so as to successfully infer the type of transportation (car, public mean or pedestrian) employed by an individual. The percentages of utilization of different transport modes obtained from mining mobile traffic are found to be close to those recorded in surveys. Calabrese *et al.* [119] employ their LocHNESs framework to separate users onboard cars from those moving on foot with an error in the mode estimation of 3–19%. Doyle *et al.* [122] can correctly identify the transportation mode, between road and rail, of around 80% of the mobile users that traveled between Dublin and Cork, in Ireland, using a one-week mobile traffic dataset.

Planning of transportation systems. Recent works have targeted the simulation and improvement of city-wide transportation systems from mobile traffic analysis. As far as simulation is concerned, Zilske and Nagel [123] use mobile traffic data to parameterize the MATSim road traffic generator in the scenario of Abidjan, Ivory Coast. They find that it is possible to directly inject mobile traffic-based trips into the

[61] O-D matrices are easily obtained by aggregating individual trips over a discretized space [86]. Since mobile users only represent a portion of the whole population, a scaling factor is needed for a comprehensive representation of mobility. That can be achieved using reference data on, e.g., population distribution [77], road traffic counts [79], mobile operator customer information [75], or surveys [75]. Recently, Nanni *et al.* [89] also propose filtering the aggregated trips, so as to only include important mobility flows in the O-D matrices. To that end, the importance of the flow between each pair of locations l_1 and l_2 is measured via a *lift* measure computed as $P(l_1, l_2)/(P(l_1) \cdot P(l_2))$, where P is the probability of occurrence of a location (or an ordered sequence of locations) in the mobile traffic dataset. Only flows whose lift measure is higher than a threshold are accounted for.

[62] Different techniques have been devised to tell apart the transportation mode of users from their mobile traffic. Many of them leverage diverse measures computed on speed estimates [113, 117, 119]. Others rely on comparison against real-world travel times [76], or training on sets of trips whose transportation mode is known [122].

road network without intermediate interpretative steps, and still obtain plausible results.

Concerning the enhancement of transportation systems, Berlingerio *et al.* [78] identify, from mobile calls, thirty common mobility patterns[63] in the city of Abidjan. Such patterns, mostly mapping to home-work commuting flows, are used to plan improvements to the existing public bus transit network: the authors show that, by adding 4 new routes, the overall travel times could be reduced by 10%. Cici *et al.* [87] use instead mobile traffic to study the potential for car sharing in Madrid, Spain. Their results indicate that a reduction in the number of cars of up to 67% can be attained when drivers share their cars and agree to take detours of 600 m at most in their routes. Finally, Zhang *et al.* [95] identify underserviced routes in Shenzhen, PRC, by comparing the trajectories inferred from mobile traffic to public transport flows. The authors propose a system of new bus lines that can reduce travel times of commuters along such routes of around 25% in typical days.

Commuting patterns. Mobile data has recently emerged as an interesting source of information for the characterization of the mobility patterns of commuters[64]. Furletti *et al.* [124] can successfully tell apart commuters from other user categories, such as residents and tourists in Pisa, Italy. Scepanovic *et al.* [82] rank[65] regions in Ivory Coast according to their importance in the country-wide commuting process, whereas Liu *et al.* [74] find commuting activity sequences to be the dominating cause of mobility in the same area. Finally, the mobility modeling methodologies developed by Yang *et al.* [111] and Tizzoni *et al.* [66] explicitly target the representation of commuting patterns.

5.3 Validation

As discussed above, the analysis of mobile traffic can lead to important insights on human mobility, from a number of different perspectives. However, the limited granularity of the datasets (see the discussion about mobile traffic sources in Sec. 2.2) may question the validity of the results. Significant effort has then been put in assessing the reliability of these data as a source for studies concerning mobility.

Specifically, the focus has been on the dependability of CDR, as they are by far the most common type of mobile traffic data employed by the works in Sec. 5.1 and Sec. 5.2, but, at the same time, they yield rather inaccurate positioning information. In fact, Smoreda *et al.* [3] speculate that the popularity of CDR is a consequence of their wide availability and ease of collection, which makes them preferred over more precise mobile traffic sources, such as, e.g., signaling events, handover or location update records.

Overall, the conclusion of the studies on the reliability of CDR is that the latter do introduce a certain bias in the study of mobility. However, this bias can affect the final results at different extents – or even not affect them at all. As that mainly depends on the type of analysis, in the following we separate reliability evaluations on this aspect.

Geographical distributions of populations. Multiple independent studies have proven that the density of mobile users' home locations extracted from CDR provides a very good approximation of the actual population distribution. Isaacman *et al.* [105] unveil the match between national census data and the density of mobile users registered at different ZIP-code areas in New York and Los Angeles, USA. Excellent agreements between the spatial distribution of mobile phone locations at night time and that of the population are also observed by Calabrese *et al.* [77] in eastern Massachusetts, and Bekhor *et al.* [75] over the whole Israel country. High correlations[66] between the geographical distributions of CDR-based and

[63] Mobility is described in terms of paths followed by flows of users, rather than just origin and destination pairs. The precise paths are identified using the widely adopted Prefixspan algorithm to mine sequential patterns in the sequences of stop locations of all users.

[64] Different techniques have been adopted to identify commuters in the mobile user population. In [124], commuters are extracted by clustering together users with similar temporal profiles via Self Organizing Maps, a class of neural network based on unsupervised learning. In [82], commuting patterns are mapped to round-trips returning to the origin within a same day. In [74], visited locations are classified as home, work, or other, and commuting is mapped to home-work-home and home-other-home sequences.

[65] Performed by running the PageRank algorithm on the commuting graph.

[66] R^2 values of 0.75 and 0.92 are obtained in [86] and [91], respectively.

census populations have been likewise calculated by Mamei *et al.* [86] and Csáji *et al.* [91].

Aggregated mobility flows. Results change when considering whether the aggregated movement of large flows can be reliably inferred from CDR. The controversy concerns both routinary and exceptional mobility situations.

On the one hand, Schneider *et al.* [88] show that regular mobility motifs[67] inferred from CDR map well to those obtained through reliable population surveys. In addition, Bengtsson *et al.* [107] use CDR to estimate the distribution of people that left Port-au-Prince, Haiti, in the months following the 2010 earthquake, and their results match with those of a large retrospective survey carried out by the United Nations. Finally, a number of studies, including those by Calabrese *et al.* [77], Ma *et al.* [79], Yang *et al.* [111], and Liu *et al.* [74] demonstrate how mobile traffic can be leveraged to generate O-D matrices or commuting patterns that are equivalent to those obtained from census data or population surveys.

On the other hand, other studies found CDR to lead to an overestimation of large-scale mobility flows. Such is the conclusion of Tizzoni *et al.* [66], who identify significant statistical differences between commuting flows inferred from national census data and those observed from mobile phone data in three European countries, namely Portugal, Spain and France. Wesolowski *et al.* [106] show that inhabitants of Kenya that use more often their mobile phones also tend to travel farther and more frequently. As they generate a large number of entries in the CDR dataset, such high-end users risk to bias the average level of mobility of the population towards unrealistically high values. The authors cross-validate this result using a survey of 33,000 individuals regarding mobile phone ownership, cellular phone expenses, income, and a variety of other social and economical parameters. Additionally, they prove that the actual bias is dependent on the geographical region considered: they thus propose a methodology based on mobile phone ownership and usage information, which allows compensating for the bias and producing statistics that are representative of the entire population of a given district.

Individual mobility features. When it comes to the analysis of individual mobility, two approaches have been adopted in order to assess the reliability of CDR data. We refer to those as intra-CDR and CDR-to-ground-truth, respectively.

As far as the intra-CDR approach is concerned, several works have evaluated the quality of the results provided by voice and texting CDR, using as a benchmark similar information extracted from high-frequency data traffic CDR. As a matter of fact, as shown by Iovan *et al.* [125], there is a strong positive correlation among the length and span of mobile user movements and their access frequency to the cellular networks; however, above a given activity level[68], the correlation disappears, which implies that high-frequency mobile traffic data becomes at that point reliable proxy for the actual movement of users. A confirmation, although with slightly different numbers comes from Trestian *et al.* [48], who compare the daily travel distance of users computed from mobile traffic data featuring diverse levels of granularity. They find that trajectories extracted from CDR sampled at every hour or at every 20 minutes are comparable, whereas lower sampling frequencies yield a loss of information.

In this context, Ranjan *et al.* [84] find that voice and texting CDR are sufficient to infer important locations[69] of each mobile user. Similarly, González *et al.* [97] prove that mobility flows among the important locations of a user are well modeled by voice and texting CDR. However, data traffic CDR provide a much more complete view of individual mobility, exceeding important locations. Ranjan *et al.* [84] find data traffic CDR to allow a better inference of (i) transient locations along trajectories, (ii) radius of gyration, and (iii) geographical spread of activities[70].

[67] See Sec. 5.1 for a definition of motif.

[68] In [125], the authors identify fifty events per day, i.e., an average sampling rate of around 30 minutes, as the access frequency threshold needed to fully capture user mobility.

[69] The notion of *significant locations* is used in [84], which maps to the subset of all visited locations that account for over 90% of a user's activity.

[70] The difference in the activity spread is measured through the Jensen-Shannon divergence, a method to assess the similarity between two distributions, popular because it is symmetric and bounded in the [0,1] interval. Considering two distributions P_S and P_O, the Jensen-Shannon divergence is defined as $JSD(P_S||P_O) = \frac{1}{2}(D(P_S||P_M) + D(P_O||P_M))$, where $P_M = (P_S +$

The second approach consists instead in comparing mobility extracted from CDR with some ground-truth information obtained from a different, reliable source. In the context of generic user mobility, such reliable source is typically a small subset of individuals participating in the experiment: e.g., Isaacman *et al.* [105] employ ground-truth reference provided by five volunteer who periodically logged their position, and measured a typical error in the order of 0.5-1 mile in the locations retrieved from voice and texting CDR. This approach is also very common in transportation studies of travel times, where ground-truth data can be obtained from, e.g., induction loops or GPS probes: that is the case of Qiu *et al.* [113], Bar-Gera *et al.* [116], and Schlaich *et al.* [73], who all find good agreement between reference and mobile traffic-inferred data. Another example is that of Calabrese *et al.* [119], who leverage ground-truth data from heterogeneous sources to show that CDR processed via their LocHNESs framework could attain individual positioning precision of around 100, 200 or 700 m, in urban, subsurban and extra-urban environments, respectively.

Relevant to the last approach is also the work by Zhang *et al.* [95], who compare individual trajectories inferred from CDR with those extracted from public transport data, i.e., taxi/bus GPS logs and subway transits. The authors find that mobile traffic can in fact capture trips that are more varied in nature than those depicted by transportation data, both in terms of lengths and geographical coverage. Moreover, CDR can track a larger number of users if collected over time intervals of six days or more.

6 Networking analysis

Large-scale mobile traffic data clearly yield enormous potential when it comes to understanding and improving cellular network systems. On the one hand, mobile traffic information is paramount in drawing a clear picture of how the access network resources are consumed by mobile users. We thus present works aiming at characterizing traffic dynamics in cellular networks, at both aggregate and per-user levels, in Sec. 6.1. On the other hand, the characterization of mobile traffic is a first step towards the design and evaluation of solutions concerning not only technological aspects of cellular systems, but also privacy and marketing ones, as reviewed in Sec. 6.2.

We summarize the works covered in this section in Tab. 3, and present the main characteristics of the studied datasets therein. We also highlight the networking-oriented research aspects discussed in each paper, providing the reader with a quick guide through the articles.

6.1 Mobile demand

The characterization of access network traffic has been addressed from two diverse perspectives: (i) a mobile operator viewpoint, where the traffic is considered from the network perspective and aggregated over many users within coverage of a same base station or within a same geographical area; (ii) a mobile user viewpoint, which focuses on the behavior of individuals in terms of their cellular network access. Next, we separately review works taking the two approaches.

6.1.1 *Aggregate access network traffic*

Studies on mobile traffic from an operated cellular network viewpoint aim at understanding the spatiotemporal dynamics of the global user demand. The focus is not only on the typical variability due to the routinary mobility and activities of the network customers, but also on anomalous behaviors induced by particular social events or technological issues.

Temporal dynamics. There is general agreement on the fact that mobile traffic tends to follow regular temporal patterns.

$P_O)/2$, and $D(P_S||P_M) = \ln(P_S/P_M)P_S$ is the Kullback-Leibler divergence between P_S and P_M.

		Analysis		Dataset						Focus															
		Name	Date	Operator	Area	Time	Users	V	T	D	TD	SD	Sp	VT	AD	UC	TM	DT	LS	PT	AI	E	D	CT	SA	PS
Mobile demand	Aggregate	Williamson [126]	11/05	–	100 cells	1 week (2004)	10 K			✓	✓				✓		✓									
		Paul [81]	04/11	–	One country	1 week (2007)	100 K			✓	✓	✓			✓		✓									
		Keralapura [127]	09/10	–	USA	1 day (2008)	500 K			✓	✓					✓										
		Shafiq [128]	06/11	–	One state	1 week (2010)	$\sim$ M			✓	✓				✓			✓								
		Zhang [129]	08/12	–	–	1 week	50 K			✓	✓							✓								
		Mucelli [130]	09/14	–	Mexico city, Mexico	1 week (2013)	2.8 M			✓	✓				✓	✓										
		Naboulsi [51]	04/14	Orange	Abidjan, Ivory Coast	5 months (2012)	18 K	✓	✓		✓	✓														
		Wang [131]	04/13	–	2 cities	Months (2007/11)	2.4 M	✓	✓		✓															
		Girardin [132]	06/09	AT&T	NY, USA	1 year (2007/08)	–	✓	✓		✓	✓														
		Hohwald [133]	06/10	–	Metropolis	6 months	50 K	✓	✓		✓															
		Cardona [134]	12/14	–	European country	7 months (2011/12)	40 K			✓	✓														✓	
		Shafiq [135]	12/13	–	USA	1 week (2010)	–			✓		✓						✓								
		Ratti [136]	11/06	–	Milan	2 weeks (2004)	–	✓	✓			✓														
		Willkomm [137]	10/08	–	NC, USA	3 weeks	–	✓				✓			✓											
		Csáji [91]	06/13	Orange	Portugal	–	100 K	✓				✓														
		Cerinsek [138]	05/13	Orange	Ivory Coast	5 months (2012)	5 M	✓	✓			✓				✓										
		Hoteit [139]	12/12	Orange	Paris	2 days (2012)	$>$1.5M			✓		✓	✓													
		Shafiq [140]	03/12	–	Metropolis	32 hours (2010)	$\sim$ 10 K			✓		✓														
		Trestian [48]	11/09	–	5000 km^2	1 week	281 K			✓		✓					✓									
		Trasarti [141]	05/13	–	Paris, France	–	–	✓	✓			✓														
		Zong [142]	05/13	Orange	Ivory Coast	5 months (2012)	5 M	✓	✓			✓														
		Xavier [144]	12/12	Oi Telecom	Rio de Janero, Brazil	3 days (2011)	–	✓	✓				✓													
		Gowan [56]	05/13	Orange	Ivory Coast	5 months	5 M	✓	✓				✓													
		Shafiq [145]	06/13	-	2 metropolis	Several days (2012)	100 K	✓		✓			✓							✓				✓		
		Xavier [146]	05/13	–	4 cities (Brazil)	4 days (2011/12)	–	✓	✓				✓													
		Paraskevopoulos [147]	05/13	Orange	Ivory Coast	5 months (2012)	5 M	✓	✓				✓													
		Pastor-Escuredo [148]	05/13	Orange	Ivory Coast	5 months (2012)	5 M	✓	✓				✓													
		Elzen [149]	05/13	Orange	Ivory Coast	5 months (2012)	–	✓	✓				✓	✓												
		Bodlaj [150]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓					✓												
		Rodriguez [151]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓					✓												
		Smith [35]	05/13	Orange	Ivory Coast	5 months (2011/12)	500 K	✓	✓					✓												
	Individual	Dasgupta [154]	03/08	–	–	5 months (2007)	3.1 M	✓							✓									✓		
		Ben Abdesslem [153]	03/14	–	European country	8 weeks (2011/12)	3 M			✓					✓											
		Candia [53]	07/08	–	230400 km^2	–	–	✓							✓		✓									
		Lin [155]	10/07	–	Northern PRC	–	600 K	✓	✓							✓									✓	
		Becker [156]	06/11	–	Morristown, USA	2 months (2009/10)	475 K	✓	✓							✓										
		Couronné [157]	10/11	Orange	Paris, France	1 day	4 M	✓	✓								✓									
Technologies	Networking	Zang [100]	09/07	3G	3 cities	1 month (2006)	2 M	✓	✓	✓									✓							
		Xu [158]	06/11	–	LA, USA	1 week (2010)	–			✓									✓							
		Shafiq [161]	06/14	–	USA	1 month (2012)	500 K			✓										✓					✓	
		Balachandran [160]	09/14	–	Metropolis, USA	1 month (2012)	1 M			✓										✓						
		Gerber [162]	03/11	–	USA	2 days (2010)	$\sim$ M			✓											✓					
		Finamore [163]	12/13	–	European metropolis	1 day (2012)	$>$ 200 K			✓											✓					
		Yu [164]	04/13	–	Metropolis, PRC	1 month (2011)	65 K			✓												✓				
		Wang [165]	05/09	–	–	6 months	100 K	✓	✓														✓			
		Agarwal [166]	05/13	Orange	Ivory Coast	5 months (2012)	500 K	✓	✓														✓			
		Zhu [167]	04/09	–	US	2 weeks (2008)	2 M		✓	✓													✓			
		Zhu [168]	05/13	Orange	Ivory Coast	2 weeks (2011)	500 K	✓	✓														✓			
	Marketing	Wei [169]	08/02	–	Southern Taiwan	4 months (2001)	114 K	✓	✓																✓	
		Belo [170]	07/13	–	European country	1 year (2008/09)	10 K	✓	✓																	✓
		Szabo [171]	11/06	–	–	14 months (2004/05)	5.5 M	✓	✓	✓																✓
	Privacy	Zang [172]	09/11	–	50 states, USA	3 months (2010)	25 M	✓																		✓
		Montjoye [173]	03/13	–	Western country	15 months (2006/07)	1.5 M	✓	✓																	✓
		Song [174]	07/14	–	–	1 week	630 K	✓																		✓
		Acs [175]	08/14	Orange	Paris, France	1 week (2007)	2 M	✓	✓																	✓

Table 3: Main features of works analyzing mobile traffic data towards understanding resource consumptions and designing technological solutions. In the analysis columns, date is in MM/YY format. In the dataset columns, V is voice, T is texting, D is data. In the focus columns, TD is traffic temporal dynamics, SD is traffic spatial dynamics, Sp is special dynamics, VT is visualization techniques, AD is activity distributions, UC is users categories, TM is traffic-mobility correlations, DT is device and traffic types, LS is localization solutions, PT is network parameter tuning, AI is architecture improvements, E is energy efficiency, D is device to device, CT is churning and traffic plans, SA is service adoption, PS is privacy solutions.

Williamson *et al.* [126] observe that traffic at 100 base stations presents a repetitive daily pattern over different weekdays, with a characteristic binary profile of low demand at night and high demand during the day. This first result has been largely confirmed by later works, over much larger scales. As an example, Paul *et al.* [81] confirm that this regularity exists at a nation-wide scale. They study the distribution of daily traffic over the whole US, and find it to evolve over time in very similar ways on different weekdays. The same diurnality of aggregate mobile traffic is remarked by Keralapura *et al.* [127], Shafiq *et al.* [128], Zhang *et al.* [129], and Mucelli *et al.* [130]. The phenomenon is not affected by seasonality, rather it remains stable over different months, as shown by Naboulsi *et al.* [51]. Specifically, the latter authors develop a dedicated clustering strategy[71] and group hourly usage profiles that yield similar load distributions. This technique improves the simple visual inspection of time series or probability distributions employed by previous works, and allows observing how night and day hours form the two categories with the most different mobile usage behaviors.

Although the most significant load difference is between night and day, some variability in the mobile traffic can also be noted among different daytime hours. In this case as well, several works agree that fluctuations tend to follow a common pattern over all weekdays. However, the precise behavior of such fluctuations seems to depend on the dataset considered: Williamson *et al.* [126] detect several daytime peaks, the largest of which appearing late in the afternoon; Wang *et al.* [131] identify two daytime peaks in data collected in PRC, but just one in data from San Francisco, USA; Naboulsi *et al.* [51] observe the most significant diversity to occur between hours falling in the interval from 8 am to 4 pm and other times of the day.

Weekends also yield aggregate mobile traffic demands that are quite unlike those measured during weekdays. Specifically, Williamson *et al.* [126] remark that weekends are characterized by loads that are remarkably lower than those recorded during weekdays, a conclusion later supported by the works of Girardin *et al.* [132], Zhang *et al.* [129], Wang *et al.* [131], Naboulsi *et al.* [51], and Hohwald *et al.* [133]. However, the latter authors also underline that weekend calls last longer on average: it is thus a dramatic drop in the number of calls that leads to the lower demand on those days.

Regular, although less intense, variations are also observed over time scales longer than a week. Cardona *et al.* [134] detect seasonal variations in users' consumptions: namely, they find a 20% increase in monthly data usages towards the end of the year with respect to the summer period.

In all cases, the temporal regularity of aggregate mobile traffic is especially useful when it comes to predicting the future network load. Shafiq *et al.* [128] show that a simple Markovian model is capable of accurately anticipating the temporal evolution of the demand based on its past history.

Spatiotemporal dynamics. Temporal dynamics are aggregated over the whole access network, and thus hide the geographical variability of mobile traffic. When separating the demand of individual base stations or topographical regions, different spatiotemporal profiles of mobile traffic emerge.

In a seminal work, Girardin *et al.* [132] consider the evolution of mobile traffic over different areas of interest in New York, NJ, USA. They observe that these regions exhibit similar average mobile demands during working days, but a neat variability during weekends. On a finer daily temporal scale, the same authors detect heterogeneity in the geographical distribution of mobile traffic during the evening hours, while the consumption stays quite similar among different regions over the rest of the day. The spatial heterogeneity of the radio access load has been later confirmed by Paul *et al.* [81], Shafiq *et al.* [135], and Naboulsi *et al.* [51], among others.

Building on these observations, several works aim at rigorously categorizing geographical regions, according to their mobile usage profiles. In an early work, Ratti *et al.* [136] show that some base stations in Milan, Italy, are characterized by a high level of activity during the evening while others present

[71] The authors combine two clustering techniques: the Unweighted Pair Group Method with Arithmetic Mean algorithm (UPGMA) and the K-means algorithm. UPGMA is a hierarchical clustering algorithm that starts from one-item clusters, and then merges at each iteration the two clusters at minimum distance. K-means is a partitional clustering algorithm that allows separating a set of items into K disjoint categories.

high demands at office hours. Interestingly, geographically locating base stations with the two behaviors above allows the authors to observe a neat movement of activity from the suburbs towards the city center between 9 am and 1 pm. A more comprehensive approach is taken by Willkomm *et al.* [137], who group[72] base stations in NC, USA, based on the time series of their mobile traffic load, and find three representative temporal patterns at base station level: those with permanent low traffic, those with low traffic during weekday nights only, and those with low traffic during weekdays nights as well as during weekends. Three classes of base stations are also identified by Csáji *et al.* [91], using weekly time series; the authors map such categories to base stations in home, work and *other* locations. Finally, Cerinsek *et al.* [138] find[73] five classes of base stations with similar daily and weekly traffic profiles. The authors also show that three of the base station clusters present geographical correlation, as they are located in close proximity.

The geographical heterogeneity of mobile traffic becomes even more evident when separating the load on a per-application basis. Hoteit *et al.* [139] notice that TCP- and UDP-based services increase the diversity among base stations in Paris. Shafiq *et al.* [140] observe that usage of popular applications is not spatially uniform, but stongly depends on location. The authors group[74] base stations according to the type of application traffic they receive: this leads to the identification of four classes of base stations that mainly manage web browsing, email, audio and mixed traffic, respectively. In addition, base stations showing similar usage are often located nearby, which allows associating different applications to specific geographical regions. These results are aligned with those by Trestian *et al.* [48], who show that services are consumed differently at home and work locations.

An original twist to the analysis of spatiotemporal dynamics of mobile traffic is proposed by Trasarti *et al.* [141], who investigate correlations between the mobile load observed in different geographical areas at successive time instants. By applying this approach to data from Paris, France, they detect that, e.g., an increase in activity at the local international airport is followed with high probability by an augmented mobile demand at a major train station of the city. A quite unique approach is also adopted by Zong *et al.* [142], who build a graph describing cell-to-cell interactions[75], and study its dynamics over several months. The authors show that traditional network growth models, such as the preferential attachment model[76] do not apply to cell-to-cell mobile traffic graphs, and propose a better-fitting generative model, named latent node radius[77].

Also relevant to this section are studies that show how the spatio-temporal dynamics of the mobile demand are affected by land-use. We refer the reader to Section 4.3 for a detailed discussion of such analyses.

Special dynamics. Special events of, e.g., natural, social, political, economical or technical origin, can affect human activity routines, which influence in turn cellular network usage. The networking literature is mainly concerned with the investigation of how special events impact cellular network usage. For the dual analysis of how mobile traffic can be leveraged to infer social events, we refer the reader to Sec. 4.2, *special events* tag.

[72] The authors apply the K-means algorithm with $K = 10$.

[73] The authors adopt a clustering strategy over vectors that represent the mobile call activity at each base station. Their proposed methodology first reduces the set of samples to analyze using a generalization of the K-means method, named *leaders method*. Then, it runs Ward's hierarchical clustering algorithm to unveil relations among the selected samples.

[74] The authors use the K-means algorithm, where K is chosen according to a gap statistic that relies on comparing the intra-cluster distance for the studied data to the one resulting from a reference null distribution.

[75] Graph vertices map to cells, and unweighted edges connect cells among which the mobile traffic volume is larger than a minimum threshold.

[76] The preferential attachment model is known to describe, e.g., the growth of the World Wide Web graph [143]. The model commends that the likelihood of connecting new vertices to existing ones is directly proportional to the degree of the latter. The resulting graph yields a power-law degree distribution.

[77] In the latent node radius model, a new vertex i is assigned a latent radius $r(i)$, and the probability that i connects with an existing vertex j of degree $d(j)$ is $P(i,j) = a[r(j) - d_{ij}] - b \cdot d(j)$, where d_{ij} is the spatial distance between vertices i and j, and a and b are model parameters.

A wide range of large-scale social events are found to induce notably higher mobile demands. Significant examples are provided by Girardin *et al.* [132] during the New York waterfall exhibition, Thanksgiving, Christmas, New Year's Eve, Easter, and July 4th, and by Hoteit *et al.* [139] during and after the final match of the European soccer cup, in 2012. At times, however, special events can result in a localized decrease of the mobile activity, as observed by Xavier *et al.* [144] in the area around the soccer stadium in Rio de Janeiro during a match. Similarly, Gowan *et al.* [56] detect peaks in the call duration before soccer games in Ivory Coast, followed by an important drop as games start. Shafiq *et al.* [145] illustrate how crowded sports and conference events can increase the access workload and result in significant voice and data performance degradation, including two orders of magnitude more probable connection failures. Clearly, the nature of the event determines whether, where and in which way mobile traffic is varied, and different events can result in opposite dynamics. Several meaningful examples are provided by Xavier *et al.* [146].

Paraskevopoulos *et al.* [147] delve deeper in the analysis of the localization of the effects of special events, by proposing a strategy to cluster[78] base stations based on their traffic profiles during special events. The authors find that base station load can be affected in antithetical ways by a same event, depending on the base station geographical location within a city. The same spatial heterogeneity holds nationwide, as proven in different contexts by several studies, as follows. Gowan *et al.* [56] cluster[79] base stations using their call duration profile before, during and after soccer matches: the authors find matches to affect the mobile traffic in the outskirts of large cities much more evidently than in all other areas of Ivory Coast. Pastor-Escuredo *et al.* [148] discover that natural hazards such as wildfires lead to a growth of morning calls on the aftermath of the event in rural areas and small cities of Ivory Coast; mobile traffic is instead reduced in large urban areas of the country during the same period. Elzen *et al.* [149] show that confrontations between political and ethnical factions in developing countries can lead to an increase or decrease in the mobile traffic activity, depending on the location respective to the region where clashes occur.

Visualization techniques. Finally, relevant to this section are also several works that target the effective and scalable visualization of aggregate access network traffic. Their goal is enabling the rapid identification of the important properties of mobile traffic by flexible visual inspection. Bodlaj *et al.* [150] employ colored lines, dispersed lines and star rays to portray the number of calls and the call duration between base stations deployed within a given geographical region. They show how different levels of information are captured by the diverse representations. The popular Data-Driven Documents (D3) visualization language is employed by Rodriguez *et al.* [151] and Smith *et al.* [152] to create and control dynamic and interactive geographical graphics of individual base station statistics, call volumes, user movement, or lack thereof. Finally, Elzen *et al.* [149] adopt a layered visual analytics approach, which facilitates the investigation of the properties of massive mobile traffic datasets by allowing interactive analysis, zooming and filtering of the data. The authors show that visual analytics is effective in identifying major trends as well as special events in the dataset.

6.1.2 *Individual access network traffic*

Characterizing mobile traffic on a per-user basis primarily aims at understanding how individual customers consume mobile services. Studies on the subject analyze the heterogeneity in the demand generated by single users, its variability over time and space, and how it is affected by mobility and consumed services.

[78] The authors group base stations using the UPGMA (see footnote 71) algorithm on vectors describing traffic at each base station at key time instants.

[79] The authors employ Ward's minimum variance linkage. The algorithm starts from the set of individual base stations, and merges at each iteration the pair of clusters that yields the minimum joint intra-cluster variance.

Activity distributions. The behavior of mobile users is defined in the first place by when and where they access the cellular network. In a seminal work, Williamson *et al.* [126] study the calling behavior of 4,156 mobile users and find that they use the cellular network in a very heterogeneous way. The distribution of per-user activity follows a power law[80], which implies that a vast majority of users performs a few calls per week, yet there exists a non-negligible amount of high-activity customers generating hundreds of calls per week. The imbalance among users in terms of mobile access has been later confirmed by Dasgupta *et al.* [154] at a much larger scale, considering 3 million users. The same authors also show that the skewness does not affect calls only, but mobile data traffic as well. This latter observation is corroborated by Paul *et al.* [81] at an even larger, nationwide scale, by showing that high-end users can generate 100,000 times the median data traffic of all customers. The result is that 10% of the users consume 60% of the access network bandwidth. Shafiq *et al.* [128] provide further confirmation, as they remark that mobile traffic is dominated by a small fraction of users, 5% of which being responsible for 90% of the total demand. Even when focusing on specific types of traffic, mobile customers can be quite heterogeneous in their access: e.g., Ben Abdesslem *et al.* [153] show that 20% of the users are responsible for 78% of the total number of YouTube requests from mobile devices.

Coherently with the heterogeneous load they induce, users also tend to have very diverse temporal patterns in accessing the cellular network. Candia *et al.* [53] show that the inter-call time also follows a truncated power-law distribution[81]. However, the result may vary across datasets, as Willkomm *et al.* [137], observe that call inter-arrivals follow instead an exponential distribution. Mobile users appear less diverse when it comes to call durations: Willkomm *et al.* [137], and Dasgupta *et al.* [154] observe a clear tendency of calls to be short, with a peak at around 1 minute. Also when considering mobile data traffic usage, subscribers do not show very different activity durations: Mucelli *et al.* [130] find that 80% of users in Mexico City are active for at most 4 hours per day, while less than 5% consume services for more than 10 hours per day.

Mobile user categories. An interesting problem is that of distinguishing categories of mobile users, so as to make a limited number of typical user profiles emerge. Clustering strategies are typically applied to address this challenge.

Lin *et al.* [155] proved the feasibility of the approach by separating[82] the calling behavior of 600,000 mobile customers in PRC. A number of per-user features is considered to that end, including the duration of different types of calls, the duration of idle periods, and the volume of data traffic generated. The authors identify three classes of users with especially interesting features that tell them apart from the average customer conduct: (i) frequent long-distance callers, (ii) frequent local callers, who also make large use of texting, and (iii) users who seldom access the cellular network. Becker *et al.* [156] adopt a similar approach[83]. on a simpler description of each user's behavior, represented by the calling and texting load generated on an hourly basis over a week. They find seven typical user profiles, two of which are especially interesting, as they can be mapped to commuters with a high level of mobility, and to students, respectively. Cerinsek *et al.* [138] group[73] users according to their daily and weekly activities, and detect two major behaviors that map to morning and late evening users. Recently, Mucelli *et al.* [130] have separated[84] user profiles according to their total data volume and number of sessions over a period of 2 weeks. Their strategy leads to the identification of six subscriber classes.

A more complex co-clustering solution is devised by Keralapura *et al.* [127], with a specific focus on grouping mobile users based on their web browsing activity. The authors build a dedicated framework,

[80] Denoting as c the call frequency, then $P(c) \sim c^{-\gamma}$, where γ is inversely proportional to the tail weight and thus to the presence of users performing a very high number of calls in the dataset. In [126], $\gamma = 1.021$, which implies a very heavy-tailed distribution.

[81] Denoting as t the inter-call time, then $P(t) \sim t^{-\gamma} e^{-t/k}$, where k is the inter-call time at which the exponential cutoff occurs, i.e., it becomes very unlikely to find users whose calls are separated by such long intervals. In [53], $\gamma = 0.9$, implying again a very heavy tail. Instead, k is around 48 days, i.e., calls by a same users occurring at more than 2 months of distance are rare.

[82] The authors employ a K-means algorithm, with K=15.

[83] The authors employ a K-means algorithm, with K=7.

[84] The authors combine the UPGMA (see footnote 71) and the K-means algorithm, similarly to [51].

named Phantom[85], and run it on a one-day 500,000-user dataset. They find that just ten clusters can capture all possible browsing behaviors of mobile users, and that such profiles are scarcely affected by time. Interestingly, the heterogeneity in mobile access already observed in terms of traffic demand also exists for the browsing activity: many users have a limited set of browsing interest, but a non-negligible number of users is present, who have very diverse browsing interests.

Traffic-mobility correlations. Several works have investigated how the mobility of a user affects the mobile traffic he/she generates. An early study by Williamson *et al.* [126] observes no significant correlation between the level of network activity of a user and the number of cells he visits. However, the conclusion is based on a small dataset of 4,156 users. More recent works based on larger populations proved the opposite. As an example, Couronné *et al.* [157] show that a strong correlation exists between the number of locations visited by a user and the number of communication events he generates. Similarly, Paul *et al.* [81] observe that the median traffic of high-mobility users is twice that of subscribers with a low mobility level.

Trestian *et al.* [48] provide a more in-depth analysis, by considering the impact of mobility not only on the aggregate traffic but also on the actual applications consumed by customers. To that end, they compute correlations among the mobility of a user during a mobile data transfer session and the kind of service accessed on that session. Significant differences emerge. Streaming music is mostly listened by users while stationary and it rapidly disappears as user mobility increases. On the contrary, email shows a strong positive correlation with mobility, i.e., it is accessed more and more frequently as subscribers become increasingly mobile. Other applications, such as social networking, show instead maximum access probability in presence of moderate mobility.

Finally, an interesting observation is made by Candia *et al.* [53], who indicate the fraction of users who call and travel at the same time remains stable over time, notwithstanding the large spatiotemporal variations of the aggregate network activity discussed in Sec. 6.1.1.

Device and traffic types. Traffic consumption also depends on the types of device used to access the cellular network, and on the kind of applications such devices run. Indeed, different families of mobile (smart)phones have heterogeneous computational and storage capabilities; moreover, they only represent a portion of the devices accessing the mobile network, which is also used by, e.g., femtocell routers, vehicles uploading so-called floating car data, or metering devices that represent the first instances of the emerging machine-to-machine (M2M) networking paradigm. Clearly, all these devices tend to generate diverse types of traffic.

Shafiq *et al.* [128] propose a first analysis of the load induced by different types of mobile devices. They consider two families of smartphones and a class of wireless modems providing cellular connectivity to laptops and netbooks, finding that devices belonging to each class tend to generate very dissimilar traffic. Morever, diversity emerges even among devices of a same class, favored by varied user behaviors.

In a subsequent work, Shafiq *et al.* [135] compare the traffic generated by smartphones and M2M devices. They observe that M2M devices induce a lower aggregate demand, which is however strongly biased towards the uplink direction, unlike that of smartphones. Moreover, the authors point out how M2M devices are not all the same, as different temporal dynamics emerge in the traffic generated by diverse types of devices. In order to shed light on the heterogeneity of M2M device traffic, a time series clustering strategy[86] is proposed, which results into two major classes of M2M traffic. The first class shows a diurnal behavior that maps to the working and non-working hours, whereas the second class

[85] Phantom adopts original operations to make co-clustering scalable to very large datasets comprising hundreds of thousands users and URLs. Specifically, it first groups browsed URLs into website categories based on their subject. It then runs a co-clustering algorithm, based on the recursive, divisive hierarchical partitioning of data and automatic identification of stopping conditions. Once co-clusters have been found, website categories are expanded back to browsed URLs, and the same co-clustering algorithm is run within each co-cluster output by the previous step, so as to obtain the final co-clusters of users and browsed URLs.

[86] The authors apply Daubechies-1 wavelet transforms to decompose traffic time serie into sines and cosines, adopting Coifman and Wickerhauser's method to detect the optimal decomposition level. After that, they cluster the decomposed time series via Ward's method, with l^2-norm distance metric and Davies-Bouldin index to determine the optimal number of clusters.

reflects a flat consumption shape over the whole day.

Also relevant to the mobile traffic type is the nature of the applications that generate it. Zhang *et al.* [129] analyze mobile data traffic and find that applications providing similar services can in fact yield quite heterogeneous packet inter-arrivals. They thus identify[87] sub-categories of social, news, and video applications that show comparable packet, flow and session-related metrics.

6.2 Technologies

The information extracted from mobile traffic has been leveraged to devise and evaluate technological solutions that relate to cellular systems. We separate works that employ findings from mobile traffic analysis to propose novel approaches to (i) networking algorithms, protocols, and architectures, (ii) marketing strategies, and (iii) mobile user privacy.

6.2.1 *Networking solutions*

Original solutions that target the improvement of the cellular network operation represent a natural outcome of mobile traffic analysis. As a result, various aspects of that subject have been studied in the literature.

User localization. Accurate identification of the location of mobile users within a cellular network is a first important task that can be improved by inference of information from mobile traffic. The result can enable more efficient monitoring and management of the radio resources, as demonstrated by Zang and Bolot [100]. The authors first profile individual user movements and retrieve their popular locations from mobile traffic. Such information is then used to restrain paging, typically performed over large location areas that include hundreds of cells, to frequently visited cells only. The authors show that paging cost can be reduced by 90% in different urban scenarios, at the cost of a 10% increase in the paging delay due to misses.

A more accurate localization strategy, named AccuLoc, is proposed by Xu *et al.* [158]. AccuLoc compensates for the traditional limitations of user tracking from mobile traffic data collected at the cellular network core. Once trained with ground truth information retrieved, e.g., from fine-grained signalling events collected at the access portion of the network (see Sec. 2), AccuLoc allows to locate mobile users within four cell sectors with an accuracy of 70%, by just employing standard call detail records. The precise positioning of mobile users paves the road to the rigorous characterization of access network usages.

Network parameter tuning. A proper characterization of mobile traffic can unveil problems that may be mitigated by dynamically tuning controllable settings at the access network. Shafiq *et al.* [145] study how the cellular network becomes locally overloaded during especially crowded events, and find that legacy mobile device state transitions[88] lead to inefficient radio resource utilization. They show how simply tuning state transition timings[89] can completely avoid performance degradation in presence of special events.

Balachandran *et al.* [160] employ fine-grained mobile traffic data on web browsing sessions[90] and

[87] The authors cluster applications using the K-means algorithm and apply Principal Component Analysis to understand the impact of each metric.

[88] User equipments typically cycle through three states: IDLE, i.e., inactive; FACH, where a link is established over physical radio channels that are shared among multiple terminals; DCH, where a link is established over a dedicated radio channel. Mobile operators implement proprietary state machines for transitions with fixed timeouts for state demotions, some of which have been reverse-engineered [159].

[89] The authors study the impact of a single parameter, i.e., the DCH-to-FACH demotion timeout, and demonstrate that slightly decreasing its value by 1-2 seconds during event days is sufficient to reduce access delays and energy consumption.

[90] Namely, HTTP records.

radio-level signalization[91] to understand how technical network factors (including handovers, failures, power levels, throughput, competing users) impact mobile user browsing experience (measured in terms of incomplete downloads, abandoned sessions, and session length). They show that a limited set of parameters fully characterizes – and can be used to anticipate[92] – the Quality of Experience (QoE) of subscribers. Notably, such a set does not include factors that are often considered important by network operators, which are instead enabled to monitor QoE metrics through radio network information only.

A very similar approach is adopted by Shafiq *et al.* [161], who focus on video streaming to mobile users rather than web browsing. By using equivalent mobile traffic data[93] they assess how a vast range of technical network parameters affect video abandonment. Their results provide guidance to network operators on how to improve user QoE when it comes to video streaming – for example, a 1-dB higher signal-to-interference ratio reduces the video abandonment probability by 2%. Then, the authors propose a model that relates scalable network statistics[94] to video abandonment. The model can be leveraged to predict complete download of a video by a mobile user with 87% accuracy by observing only the initial 10 seconds of a session.

Network architecture improvements. Gerber *et al.* [162] go beyond parameter tuning, and analyze detailed mobile traffic data[95] with the goal of exploring the advantage brought by significant modifications to the cellular network operations. In particular, they focus on content caching at different levels of the cellular network architecture. They find that cache hit ratios between 27% and 33% can be achieved when content is stored within the cellular network core[96]. Finamore *et al.* [163] focus on content caching as well, but they consider a "push" strategy, according to which the content in the cellular network is pre-staged to the mobile device cache before it is demanded. The authors evaluate three different caching strategies that leverage content popularity, volume, and both, respectively. They observe that the content popularity-based strategy can lead to a reduction of up to 20% of the downlink traffic for a smartphone cache size of 100 MB, in case any popular content can be cached.

Energy efficiency. Considering the energy aspect, Yu *et al.* [164] analyze mobile traffic with the aim of evaluating the energy consumption due to the establishment and release of a radio link between user equipments and base stations. The analysis confirms that a significant amount of power is wasted during inactivity times when the interface of a user equipment switches between different states[88]. The authors investigate the temporal correlations[97] of mobile traffic workloads and propose a prediction model for future data transmissions, which allows cutting down unnecessary waiting times. Their proposed scheme saves 56% of energy on average.

Device-to-device communication. The analysis of mobile traffic allows to explore original networking paradigms that go beyond the traditional user equipment-to-base station communication. Particular attention has been paid to device-to-device (D2D) communication, by considering that users who are nearby can exchange data without resorting to the cellular infrastructure, but via technologies such as Bluetooth, or, more recently, Wi-Fi Direct and LTE Direct.

In a seminal work, Wang *et al.* [165] investigate how a combination of near-distance D2D communication and long-distance texting would affect the spread of self-propagating malware among the mobile

[91] Namely, Radio Resource Control (RRC) measurement reports.

[92] Linear regression and decision trees models prove to be simple yet efficient techniques to predict web browsing QoE from soft and inter-radio access handovers, energy per chip of the pilot channel, received signal strength indicator, and number of users.

[93] Namely, HTTP records from which URL, host and requested content information can be extracted, and RRC measurement reports.

[94] The model leverages pruned decision and regression trees to determine whether a user will complete a streaming video session. It is based on radio network statistics and information collected from TCP/IP headers: as such, it does not require deep packet inspection of, e.g., HTTP headers, which would imply much larger data collection and significantly reduce scalability.

[95] Namely, HTTP records.

[96] The maximum hit ratio of 33% is obtained by assuming infinite caching at the GGSN level. A 27% cache hit ratio is obtained when balancing positive effects of caching with its cost: in that case, caching is found to be best implemented at SGSNs.

[97] The temporal correlation is calculated from the the entropy of individual time series of packet arrivals.

terminal population. By assuming a Susceptible-Infected (SI) model[98], the authors remark that D2D communication allows the malware to reach all susceptible devices, but at low speed. The spread via texting is much faster, but limited by the presence of communities in the mobile call graph[99], and by the market share of different operating systems. Agarwal *et al.* [166] adopt a similar approach to assess the effectiveness of D2D communication in disseminating information at nation-wide scales without resorting to the cellular infrastructure. The authors show that one single device can propagate the information of 90% of a 5,000-user population spread over the whole Ivory Coast. Zhu *et al.* [167] focus on viruses reproducing via texting only, and leverage the structure of the mobile call graph in order to restrain their spread among subscribers. Specifically, they identify graph partitions[100], and inject security patches to selected users that link the different partitions.

The efficacy of D2D communication can be also leveraged to offload the access network from part of its load. Zhu *et al.* [168] explore several opportunistic routing methods to that end, evaluating their performance on mobile traffic data. They conclude that all methods perform well in densely populated and geographically constrained areas, where D2D communication can be a promising solution for the delivery of delay-tolerant contents to mobile users. Shafiq *et al.* [145] focus on radio access rather than traffic load. They consider that multiple devices can leverage D2D communication to share a single connection to the cellular network. During especially crowded events, this approach can reduce failed connections up to 95%.

6.2.2 *Marketing solutions*

Mobile traffic data constitute a valuable source of information to devise marketing strategies. Their analysis allows an operator to understand the behavior of customers, their calling patterns and habits, and thus to formulate adequate and targeted offers.

Churning and traffic plans. Mobile users tend to change their operator over time, which leads to so-called churning. Collecting mobile traffic data over long time periods can help understand and predict the churning phenomenon. In a seminal work, Wei *et al.* [169] propose to predict future churners by studying the volume and frequency of calls by each user.[101]. The performance evaluation over a dataset of 114,000 customers indicates results in a correct prediction of 70% of the churners, with a 20% false positive ratio. Dasgupta *et al.* [154] also focus on the prediction of churners by considering the impact of the social relationships among customers on the churning behavior. Given a set of initial churners, the authors employ a diffusion model[102] over the mobile call graph[99] to successfully predict 60% of future churners.

In order to attract churners, Lin *et al.* [155] group 600,000 customers into different categories according to their calling and texting behaviors. The authors then tailor new traffic plans adapted to each group, which is claimed to positively affect new customer subscriptions to the considered operator. An opposite perspective is adopted by Cardona *et al.* [134], who aim at avoiding churning. More precisely, the authors investigate the cost savings that customers could achieve, combining different data traffic pricing plans and consumption schemes. They observe that collaborative pricing plans can be very beneficial for

[98] The SI model defines the infection rapidity as $dI/dt = \beta S \cdot I/N$ where β is the effective infection rate, S the number of susceptible terminals, I is the number of infected terminals and N is the size of the terminal population.

[99] See Sec. 4.1.

[100] Two graph partitioning strategies are considered. Balanced graph partitioning aims at forming even partitions in terms of node degree. Clustered graph partitioning separates the graph into partitions with minimum cut weight.

[101] The authors employ a multi-classifier class-combiner technique, especially designed to cope with the fact that the percentage of churners in a dataset is typically low. The proposed solution starts by generating from the mobile traffic training set a number of training subsets, each characterized by a known distribution of churners and non-churners. Then, a classification model is generated for each subset with a base classifier. Finally a meta-classifier combines the predictions made by each individual base classifier.

[102] The model is based on Spreading Activation (SPA) techniques applied in cognitive psychology. It allows to predict potential churners by exploring their social connections with current churners.

customers, and savings of up to 45% can be attained with group plans, i.e. plans that allow users in a pre-defined group to share a certain amount of allowed capacity. Even higher savings, reaching up to 70% of the baseline cost, are granted by open sharing plans, i.e., plans providing each customer with an individual traffic volume, and allowing him to sell his excess, unused capacity. User-driven collective consumption schemes through tethering are also shown to properly complement such strategies, inducing additional gains, especially in dense urban areas.

Service adoption. Belo and Ferreira [170] study the impact of the mobile call graph[99] on the diffusion of telecommunication-related products among customers. They identify different adoption incentives for several types of products. They also observe that social influence among mobile users can have a positive or a negative effect on the diffusion of the product, depending on the characteristics of the latter. A similar perspective is taken by Szabo and Barabasi [171], who evaluate the impact of the social relationships on the adoption of services. The authors observe a strong correlation between social networking services adopted by a mobile subscriber and those utilized by his/her contacts. The correlation is instead not present in the case of technical-oriented services for, e.g., browsing or emailing.

On a related point, mobile traffic can be also analyzed so as to understand the details behind the current level of adoption of a given service. As an example, Shafiq *et al.* [161] investigate the market of video players for mobile devices, by mining a fine-grained 500,000-user dataset. The authors point out that usage distributions among different video players is limited, as 80% of the relevant traffic load is generated by the five top players only. Ben Abdesslem *et al.* [153] focus on one specific video streaming service, i.e., YouTube, and unveil interesting features of the mobile traffic it generates. Namely, the authors find a significant tendency to replay, with 37% of the users requesting at least 10 different streams over a month who replayed more than 20% of their videos. Also, they find video popularity to follow a Zipf's distribution[103], and propose a classification of the different processes that bring videos to become viral.

6.2.3 *Privacy solutions*

As discussed in Sec. 2.3, mobile traffic data contains sensible information on individual subscribers, whose privacy needs to be properly protected. Unfortunately, the common practice of replacing mobile users' identifiers with so called *pseudo-identifiers*, is not sufficient to that end.

The unfitness of the de-facto standard approach to mobile traffic data anonymization has been the focus of several works. Zang and Bolot [172] compare the most popular locations visited by mobile users in order to break the anonymity granted by random identifiers[104]. When applying their approach to data collected at typical spatial (i.e., cell or sector) and temporal (i.e., at each event generated by a user) granularity, the authors find that considering only the single top location of each user does not pose privacy issues. However, if the two most frequently visited locations (presumably, home and work locations, see Sec. 5.1) are considered, 10% to 50% of the users are uniquely identifiable. This percentage grows to more than 50% when looking at the three top locations.

The authors also study how reducing the spatial granularity of the original data, by aggregating mobile traffic over geographical ares of different size, improves anonymity. They find that only city-wide aggregation results in a reliable anonymization. Temporal domain approaches, implemented by periodically changing the random identifiers, are instead found to ensure privacy if the updating procedure is repeated at most every 24 hours, i.e., a same user cannot be tracked over two subsequent days.

[103] Given the rank l of a video, its level of popularity is described by $P(l) = l^{-\beta}$, with β = 1.07.

[104] The authors introduce the concept of k-anonymity. They extract the top N locations of each user, and group all users sharing the same set of locations. Then, the mobile traffic dataset is said to grant k-anonymity if each user group contains at least k individuals, who thus share the same preferred locations. This makes a user indistinguishable from at least other k-1 subscribers in the dataset. If k = 1 for some N, anonymity is breached for that N, as the user can be unequivocally identified within the dataset by just looking at his/her N preferred locations.

The low level of anonymity granted by random identifiers, and the relevance of choosing adequate spatial and temporal resolutions for the preservation of mobile user privacy are confirmed by de Montjoye *et al.* [173]. Their analysis does not assume knowledge of popular locations, but only of randomly sampled pairs of user position and time[105]. The authors show that two randomly chosen spatiotemporal points are enough to uniquely characterize 50% of the users, i.e., half of the time there exists only one user in the dataset who visits those two locations at those two times. Increasing the knowledge to four randomly chosen points allows to uniquely characterize more than 95% of the users. These results refer to fine-grained mobile traffic data, where the user position is identified at cell or sector level every hour. The authors thus explore the impact of a reduced spatiotemporal granularity, but find that even coarse resolutions may provide little anonymity[106].

An attempt at finding a solution is made by Song *et al.* [174]. First, the authors confirm the previous obervations on the uniqueness of pseudo-anonymized trajectories: more than 60% of users their dataset is uniquely traced with only 2 random points, and the percentage grows to 95% also with a set of 4 random points. Then, to preserve user privacy, they periodically change each user's pseudo-identifier. An updating interval of 6 hours reduces the trajectory uniqueness to 40% in the case of two randomly selected points, however the benefit is less important for a higher number of points.

Finally, Acs *et al.* [175] focus instead on preserving user anonymity in the specific case where subscriber trajectories are aggregated into spatiotemporal density information. Their main concern is that, even under such aggregation, areas visited by a low number of users may reveal individual mobility patterns. To handle that, they introduce a data-driven differentially private scheme[107] that combines sampling, clustering and filtering processes of per-cell information to generate aggregate density information for areas grouping several cells. Their strategy is shown to provide high privacy guarantees: removing any single user does not lead to any sensible variation in the outcome of the analysis, which basically means that all subscribers are well hidden in the dataset. At the same time, the solution also preserves a good data utility level.

7 Outlook

The results presented in the previous sections are many and varied, and span across a wide range of subjects and disciplines. However, as anticipated at the beginning of this manuscript, analyses of mobile traffic have become a popular tool only during the last few years, and a large number of questions remain open. In that perspective, our comprehensive overview of the literature puts us in a unique position to comment on the open problems and future research directions for research in the field of mobile traffic analysis. Below, we propose a discussion of such issues, organized along the lines of our literature classification.

7.1 Social analysis

In Sec. 4.1, we saw that the studies on the structure of mobile user communications have unveiled most of the major properties of social interactions occurring via mobile devices. As a result, we have today a rather clear understanding of the shape of mobile call graphs, for which some models have also been

[105] Given a set of p random spatio-temporal points, named I_p, the authors compute the number $S(I_p)$ of individual movement traces that include the set of points I_p. An individual is considered uniquely identifiable if $|S(I_p)|$=1, i.e., he is the only one whose movement trace includes I_p.

[106] A measure ϵ of the uniqueness, i.e., indentifiability, of mobile users is calculated as the percentage of uniquely identifiable users given p spatio-temporal points. The authors show that $\epsilon \sim (vh)^{\beta}$, where $\beta \sim -p/100$, v represents the spatial aggregation in terms of number of cells merged together, and h is the level of temporal aggregation in terms of hours merged together. The relationship suggests that privacy is increasingly hard to attain by lowering the resolution of a dataset. Moreover, even a slight increment in the number of spatiotemporal points p makes users much more identifiable.

[107] See footnote 59 for a definition of differential privacy.

proposed. However, these properties are typically inferred from static graphs that lose all temporal dimensions. A most promising, although challenging, research direction is then the investigation of the dynamics and evolution of mobile call graphs at multiple time scales. Also, cross-correlating mobile call graph representations with other databases (describing, e.g., user demographics, subscriber mobility, or mobile traffic plan information) can pave the way to the practical exploitation of the knowledge on user interactions.

Cross-correlation is the key to progress also in mobile traffic analyses that target other social topics, such as those presented in Sec. 4.2, Sec. 4.3, or Sec. 4.4. Clearly, that implies the availability of external relevant databases that can be correlated to mobile traffic ones; open data initiatives that are becoming increasingly popular will be paramount in this direction. Also, much longer datasets of mobile traffic, spanning over several years (not necessarily in a continuous way) would benefit social studies, allowing the investigation of phenomena occurring at large time scales, such as urbanization, landscape evolutions, or new technology and service adoption.

7.2 Mobility

When it comes to mobility studies, there is general agreement on a number of laws that drive movement patterns in large populations over wide areas, as discussed in Sec. 5.1.1. This, together with the recognized high regularity and predictability of human trajectories, has allowed defining models that describe routinary user mobility in Sec. 5.1.2. However, we also saw that a number of factors can affect the customary mobility of users, and small attention has been paid to models that can capture such phenomena.

Concerning more fine-grained mobility analyses, such as intra-urban movement studies or the research in transportation presented in Sec. 5.2, the main question stays that of the dependability of mobile traffic as an information source for movement patterns. The literature in Sec. 5.3 is discordant on the answer, and there is a non-negligible risk that mobile traffic may bias microscopic mobility analyses.

On the positive side, there are clear indications that mobile traffic datasets featuring higher precision and sampling frequency tend to reduce or eliminate that bias [117, 125]. And, there are clear trends towards: (i) high-end mobile services that generate continuous traffic; (ii) monitoring probes that are shifted to the network edge, and collect positioning information that is both more precise and complete (see Sec. 2.2). Therefore, it is very probable that mobile traffic data will become extremely reliable in the future, even for, e.g., detailed individual trajectory mining. Yet, studies carried out on today's standard mobile traffic datasets shall not exclude the possibility of biases, especially in presence of microscopic mobility analyses.

7.3 Network

Network studies in Sec. 6.1 have unveiled clear, regular spatiotemporal dynamics in mobile traffic, and significant heterogeneity across mobile users in terms of the demand they generate. Most works in the literature have explored the temporal and spatial dimensions, and a few have focused on the type of services consumed by mobile users. What is still missing in the picture are techniques for the comprehensive characterization of mobile traffic, which can capture at a time the three dimensions of space, time, and service usage. Such a holistic approach shall answer open questions on the correlation among, e.g., land use, mobility, daily schedule and the nature of applications accessed by subscribers.

On the exploitation side, mobile traffic analysis has already led to a number of interesting proposals for novel networking approaches and solutions, outlined in Sec. 6.2. However, many subjects have been just scratched on the surface, and there is wide space for improvement in both clean-slate and incremental design of mobile network architectures and protocols. In particular, we draw the attention to the need of models of the mobile demand that can (i) process mobile traffic data on-the-fly, i.e., as soon as it is collected within the network, and (ii) predict short- (e.g., minute to hours) and medium-term (e.g., days

to weeks) evolutions in the demand. Indeed, such functions will be key to original anticipatory/cognitive networking paradigms, which are expected to be part of future 5G cellular systems.

Finally, the discussion in Sec. 6.2.3 outlines the evident need for mobile traffic data anonymization techniques that ensure subscriber privacy in a way that preserves the information utility. Their persisting absence risks to limit the availability of open datasets, and to hinder research based on mobile traffic analysis.

7.4 General

Finally, two open problems are common to all of the different research topics touched by mobile traffic analysis. They concern the reproducibility and reliability of mobile traffic analyses.

The first problem is the need for a reference set of mobile traffic datasets that can be adopted by the research community so as to favor reproducible research. Clearly, the set is to be publicly available. Moreover, it should be heterogeneous, i.e., representative of scenarios that are diverse from many viewpoints, both semantic (e.g., geographical and temporal span, subscriber density, development level of the region) and technical (e.g., nature of the probes used for the data collection), so that studies are not biased. Finally, these datasets should be in a clear and consistent format, and contain information that allow running a large variety of analyses.

The second transversal issue concerns the definition of standard best practices in the analysis of mobile traffic datasets. The overview we provided in this survey highlights how many different techniques have been proposed to perform even the most basic operations throughout all research subjects. As an example, trajectory extraction is a fundamental function for mobility analyses, yet there is no clear agreement on which is the current state of the art solution that should be adopted. There is thus a need to compare methodologies proposed in the literature, and derive a set of well-defined, validated techniques that can be safely used by the research community.

8 Conclusions

In this document we surveyed the literature on mobile traffic analyses run on operator-collected data. We proposed a hierarchical classification of studies in this recently emerged research field, and categorized a large body of relevant works accordingly. We also summarized the main features of research activities, as well as of the dataset they employ, into reference tables that provide immediate visibility on the many and varied topics concerned by mobile traffic analysis. Our review provides a comprehensive overview of the state of the art in the usage of mobile traffic data for scientific research, and allows outlining open research directions.

References

[1] C. Shang, M.C. Zhou, C. Chen, “Cellphone Data and Applications,” *International Journal of Intelligent Control and Systems*, 19(1):35–45, Mar. 2014.

[2] V.D. Blondel, A. Decuyper, G. Krings, “A survey of results on mobile phone datasets analysis,” *arXiv:1502.03406 [physics.soc-ph]*, Feb. 2015.

[3] Z. Smoreda, A.-M. Olteanu-Raimond, T. Couronne, “Spatiotemporal Data from Mobile Phones for Personal Mobility Assessment”, *Transport Survey Methods: Best Practice for Decision Making*, 41:745–767, Emerald Group, Jan. 2013.

[4] Cisco / Starent Networks, “LTE: Simplifying the Migration to 4G Networks,” *White Paper*, 2010

[5] L. Sweeney, "k-anonymity: a model for protecting privacy," *Int. J. Uncertainty, Fuzziness and Knowledge-based Systems*, 10(5):557–570, Oct. 2002.

[6] A. Machanavajjhala, J. Gehrke, D. Kifer, M. Venkitasubramaniam, "l-diversity: privacy beyond k-anonymity," *ICDE*, Atlanta, GA, Apr. 2006.

[7] N. Li, T. Li, S. Venkatasubramanian, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity," *ICDE*, Instambul, Turkey, Apr. 2007

[8] C. Dwork, F. McSherry, K. Nissim, A. Smith, "Calibrating noise to sensitivity in private data analysis," *TCC*, New York, NJ, Mar. 2006.

[9] A. Cavoukian, D. Castro, "Big Data and Innovation, Setting the Record Straight: De-identification Does Work," *White Paper*, Jun. 2014.

[10] A. Narayanan, E.W. Felten, "No silver bullet: De-identification still doesn't work," *White Paper*, Jul. 2014.

[11] A. Nanavati, S. Gurumurthy, G. Das, D. Chakraborty, K. Dasagupta, S. Mukherjea, A. Joshi, "On the Structural Properties of Massive Telecom Call Graphs: Findings and Implications", *ACM CIKM*, Arlington, VA, USA, Nov. 2006.

[12] D. Doran, V. Mendiratta, C. Phadke, H. Uzunalioglu, "The Importance of Outlier Relationships in Mobile Call Graphs", *IEEE ICMLA*, Boca Raton, FL, USA, Dec. 2012.

[13] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, M. A. de Menezes, K. Kaski, A.-L. Barabasi, J. Kertesz, "Analysis of a Large-Scale Weighted Network of One-to-One Human Communication", *New Journal of Physics*, 9(179):1–27, Jun. 2007.

[14] R. Lambiotte, V. Blondel, C. de Kerchove, E. Huens, C. Prieur, Z. Smoreda, P. Van Dooren, "Geographical Dispersal of Mobile Communication Networks", *Physica A*, 387(21):5317–5325, Sep. 2008.

[15] M. Seshadri, S. Machiraju, A. Sridharan, J. Bolot, C. Faloutsos, J. Leskovec, "Mobile Call Graphs: Beyond Power-Law and Lognormal Distributions", *ACM KDD*, Las Vegas, NV, USA, Aug. 2008.

[16] W. Reed, M. Jorgensen, "The Double Pareto-Lognormal Distribution - A New Parametric Model for Size Distribution", *Communications in Statistics - Theory and Methods*, 33(8):1733–1753, Aug. 2004.

[17] M. Karsai, N. Perra, A. Vespignani, "Time Varying Networks and the Weakness of Strong Ties", *Scientific Reports*, 4(4001):1–7, Feb. 2014.

[18] J.-P. Onnela, J. Saramaki, J. Hyvonen, G. Szabo, D. Lazer, K. Kaski, J. Kertesz, A.-L. Barabasi, "Structure and Tie Strengths in Mobile Communication Networks", *PNAS*, 104(18):7332–7336, May 2007.

[19] C.A. Hidalgo, C. Rodriguez-Sickert, "The Dynamics of a Mobile Phone Network", *Physica A*, 387(12):3017–3024, May 2008.

[20] G. Miritello, R. Lara, M. Cebrian, E. Moro, "Limited Communication Capacity Unveils Strategies for Human Interaction", *Scientific Reports*, 3(1950), Jun. 2013.

[21] G. Palla, A.-L. Barabasi, T. Vicsek, "Quantifying Social Group Evolution", *Nature*, 446, Apr. 2007.

[22] G. Siganos, S.L. Tauro, M. Faloutsos, "Jellyfish: A Conceptual Model for the AS Internet Topology", *Journal of Communications and Networks*, 8(3):339–350, sep. 2006.

[23] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, J. Wiener, "Graph Structure in the Web", *Computer Networks*, 33(6):309–320, Jun. 2000.

[24] S. Yang, B. Wu, B. Wang, "Multidimensional Views on Mobile Call Network", *Frontiers of Computer Science in PRC*, 3(3):335-346, Sep. 2009.

[25] C. Sarraute, P. Blanc, J. Burroni, "A Study of Age and Gender seen through Mobile Phone Usage Patterns in Mexico", *ASONAM*, Beijing, PRC, Aug. 2014.

[26] A. Stoica, Z. Smoreda, C. Prieur, J.-L. Guillaume, "Age, Gender and Communication Networks", *NetMob*, Boston, MA, USA, May 2010.

[27] A. Mehrotra, A. Nguyen, J. Blumenstock, V. Mohan, "Differences in Phone Use between Men and Women: Quantitative Evidence from Rwanda", *ICTD*, Atlanta, GE, USA, Mar. 2012.

[28] V. Wang, H. Zang, M. Faloutsos, "Inferring Cellular User Demographic Information using Homophily on Call Graphs", *NetSciCom*, Turin, Italy, Apr. 2013.

[29] J. Brea, J. Burroni, M. Minnoni, C. Sarraute, "Harnessing Mobile Phone Social Network Topology to Infer Users Demographic Attributes", *SNA KDD*, New York, NY, USA, Aug. 2014.

[30] V.D. Blondel, J.-L. Guillaume, R. Lambiotte, E. Lefebvre, "Fast unfolding of communities in large networks", *Journal of Statistical Mechanics: Theory and Experiment*, 10, Oct. 2008.

[31] O. Toomet, S. Silm, E. Saluveer, T. Tammaru, R. Ahas, "Ethnic Segregation in Residence, Work, and Free-Time: Evidence from Mobile Communication", *IAB Colloqium*, Nuremberg, Germany, Feb. 2012.

[32] A.J. Morales, W. Creixell, J. Borondo, J.C. Losada, R.M. Benito, "Understanding Ethnical Interactions on Ivory Coast", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[33] O. Bucicovschi, R. Douglass, D. Meyer, M. Ram, D. Rideout, D. Song, "Analyzing Social Divisions Using Cell Phone Data", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[34] V. Soto, V. Frias-Martinez, J. Virseda, E. Frias-Martinez, "Prediction of Socioeconomic Levels Using Cell Phone Records", *UMAP*, Girona, Spain, Jul. 2011.

[35] C. Smith, A. Mashhadi, L. Capra, "Ubiquitous Sensing for Mapping Poverty in Developing Countries", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[36] H. Mao, X. Shuai, Y.-Y. Ahn, J. Bollen, "Mobile Communications Reveal the Regional Economy in Cote d'Ivoire", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[37] K. Wakita, R. Kawasaki, "Estimating Human Dynamics in Cote dIvoire Through D4D Call Detail Records", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[38] A. Fajebe, P. Brecke, "Impacts of External Shocks in Commodity-Dependent Low-Income Countries: Insights from Mobile Phone Call Detail Records from Cote d'Ivoire", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[39] B. Lim, D. Doran, V. Mendiratta, M. Rodriguez, D. Klabjan, "Social Capital for Economic Development: Application of Time Series Cluster Analysis on Personal Network Structures", *NetMob D4D Challenge 2013*, Boston, MA, USA, May 2013.

[40] V. Frias-Martinez, V. Soto, J. Virseda, E. Frias-Martinez, "Computing Cost-Effective Census Maps from Cell Phone Traces", *PURBA*, Newcastle, UK, Jun. 2012.

[41] G. Krings, D. Baclin, L.J.V. Merlen, M.L. Pimenta, F. Galli, "Mobile Communication in Business Networks: Structure and Leadership", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[42] J.-P. Onnela, S. Arbesman, M. Gonzalez, A.-L. Barabasi, N. Christakis, "Geographic Constraints on Social Network Groups", *PLoS ONE*, 6(4):e16939, Apr. 2011.

[43] G. Krings, F. Calabrese, C. Ratti, V. Blondel, "Urban Gravity: A Model for Inter-City Telecommunication Flows", *Journal of Statistical Mechanics*, L07003, Jul. 2009.

[44] P. Schmitt, M. Vigil, M. Zheleva, E. Belding, "Egocentric and Population-Density Patterns of Cellphone Communication in Ivory Coast", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[45] N. Eagle, Y.-A. de Montjoye, L. Bettencourt, "Community Computing: Comparisons between Rural and Urban Societies using Mobile Phone Data", *CSE*, Vancouver, BC, Canada, Aug. 2009.

[46] S. Almeida, J. Queijo, L. M. Correia, "Spatial and Temporal Traffic Distribution Models for GSM." *IEEE VTC Fall*, Amsterdam, Netherlands, Sep. 1999.

[47] V. Soto, E. Frias-Martinez, "Automated Land Use Identification using Cell-Phone Records", *ACM HotPlanet*, Washington, DC, USA, Jun. 2011.

[48] I. Trestian, S. Ranjan, A. Kuzmanovic, A. Nucci, "Measuring Serendipity: Connecting People, Locations and Interests in a Mobile 3G Network", *ACM IMC*, Chicago, IL, USA, Nov. 2009.

[49] M. R. Vieira, V. Frias-Martinez, N. Oliver, E. Frias-Martinez, "Characterizing Dense Urban Areas from Mobile Phone-Call Data: Discovery and Social Dynamics", *IEEE SocialCom*, Minneapolis, Minnesota, USA, Aug. 2010.

[50] R.M. Pulselli, P. Romano, C. Ratti, E. Tiezzi, "Computing Urban Mobile Landscapes Through Monitoring Population Density Based on Cell-Phone Chatting." *International Journal of Design and Nature and Ecodynamics*, 3(2): 121-134, 2008

[51] D. Naboulsi, R. Stanica, M. Fiore "Classifying Call Profiles in Large-scale Mobile Traffic Datasets." *IEEE Infocom*, Toronto, Canada, Apr. 2014.

[52] F. Girardin, F. Calabrese, F. Di Fiore, C. Ratti, J. Blat, "Digital Footprinting: Uncovering Tourists with User-Generated Content," *IEEE Pervasive Computing*, 7(4):36–43, Oct. 2008.

[53] J. Candia, M. C. Gonzalez, P. Wang, T. Schoenharl, G. Madey, A.-L. Barabasi "Uncovering Individual and Collective Human Dynamics from Mobile Phone Records." *Journal of Physics A: Mathematical and Theoretical* 41(22): 224015, 2008.

[54] F. Calabrese, F. Pereira, G. Di Lorenzo, L. Liu, C. Ratti, "The Geography of Taste: Analyzing Cell-Phone Mobility and Social Events," *Pervasive Computing*, Helsinki, Finland, May 2010.

[55] M.F. Dixon, S.P. Aiello, F. Fapohunda, W. Goldstein, "Detecting Mobility Patterns in Mobile Phone Data from the Ivory Coast," *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[56] D.M. Gowan, N. Hurley, "Regional Development - Capturing a Nations Sporting Interest through Call Detail Analysis," *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[57] J. Bagrow, D. Wang, A.-L. Barabasi, "Collective Response of Human Populations to Large-Scale Emergencies", *PLoS ONE*, 6(3):e17680, Mar. 2011.

[58] S. Linardi, S. Kalyanaraman, D. Berger, "Does Conflict Affect Human Mobility and Cellphone Usage? Evidence from Cote d'Ivoire", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[59] A. Wesolowski, N. Eagle, A.J. Tatem, D.L. Smith, A.M. Noor, R.W. Snow, and C.O. Buckee, "Quantifying the Impact of Human Mobility on Malaria", *Science*, 338(6104):267–270, Oct. 2012.

[60] E. Enns, J. Amuasi, "Human Mobility and Communication Patterns in Cote d'Ivoire: A Network Perspective for Malaria Control", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[61] K. Gavric, S. Brdar, D. Culibrk, V. Crnojevic, "Linking the Human Mobility and Connectivity Patterns with Spatial HIV Distribution", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[62] N. Baldo, P. Closas, "Disease Outbreak Detection by Mobile Network Monitoring: A Case Study with the D4D Datasets", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[63] T.D. Ndie, Z. Nganmeni, S. Noutat, "Design and Implementation of a Tool for the Correlation between the Rate of Prevalence of a Pathology and the Flow of Communication between Diverse Localities", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[64] R. Chunara, E.O. Nsoesie, "Large-scale Measurements of Network Topology and Disease Spread: A Pilot Evaluation Using Mobile Phone Data in Cote d'Ivoire", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[65] A.S. Azman, E.A. Urquhart, B. Zaitchik, J. Lessler, "Using Mobile Phone Data to Supercharge Epidemic Models of Cholera Transmission in Africa: A Case Study of Cote d'Ivoire", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[66] M. Tizzoni, P. Bajardi, A. Decuyper, G.K.K. King, C.M. Schneider, V. Blondel, Z. Smoreda, M.C. Gonzalez, V. Colizza, "On the Use of Human Mobility Proxy for the Modeling of Epidemics", *PLOS Computational Biology*, 10(7):e1003716, Jul. 2014.

[67] V. Frias-Martinez, A. Rubio, E. Frias-Martinez, "Measuring the Impact of Epidemic Alerts on Human Mobility using Cell-Phone Network Data", *PURBA*, Newcastle, UK, Jun. 2012.

[68] E. Frias-Martinez, G. Williamson, V. Frias-Martinez, "An Agent-Based Model of Epidemic Spread Using Human Mobility and Social Network Information", *SocialCom*, Boston, MA, USA, Oct. 2011.

[69] M. Saravanan, P. Karthikeyan, A. Aarthi, "Exploring Community Structure to Understand Disease Spread and Control Using Mobile Call Detail Records", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[70] J.P. Leidig, Y. Kutsumi, K.A. O'Hearn, C.M. Sauer, J. Scripps, G. Wolffe, "Applying Mobile Datasets in Computational Public Health Research", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[71] M. Kafsi, E. Kazemi, L. Maystre, L. Yartseva, M. Grossglauser, P. Thiran, "Mitigating Epidemics through Mobile Micro-Measures", *NetMob*, Boston, MA, USA, May 2013.

[72] A. Lima, M. De Domenico, V. Pejovic, M. Musolesi, "Exploiting Cellular Data for Disease Containment and Information Campaigns Strategies in Country-Wide Epidemics", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[73] J. Schlaich, T. Otterstatter, M. Friedrich, "Generating Trajectories from Mobile Phone Data," *TRB 89th Annual Meeting*, Washington, DC, USA, Jan. 2010.

[74] F. Liu, D. Janssens, G. Wets, M. Cools "Profiling workers activity-travel behavior based on mobile phone data," *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[75] S. Bekhor, Y. Cohen, C. Solomon, "Evaluating Long-Distance Travel Patterns in Israel by Tracking Cellular Phone Positions," *Journal of Advanced Transportation*, 47(4):435–446, Jun. 2013.

[76] H. Wang, F. Calabrese, G. D. Lorenzo, C. Ratti, "Transportation Mode Inference from Anonymized and Aggregated Mobile Phone Call Detail Records," *IEEE ITSC*, Madeira Island, Portugal, Sep. 2010.

[77] F. Calabrese, G. Di Lorenzo, L. Liu, C. Ratti, "Estimating Origin-Destination Flows using Mobile Phone Location Data," *IEEE Pervasive Computing*, 10(4):36–44, Oct. 2011.

[78] M. Berlingerio, F. Calabrese, G. Di Lorenzo, R. Nair, F. Pinelli, M.L. Sbodio, "AllAboard: A System for Exploring Urban Mobility and Optimizing Public Transport Using Cellphone Data," *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[79] J. Ma, H. Li, Y. Huan, F. Yuan, T. Bauer, "Deriving Operational Origin-Destination Matrices from Large Scale Mobile Phone Data," *International Journal of Transportation Science and Technology*, 2(3):183–203, Nov. 2013.

[80] E. Halepovic, C.Williamson "Characterizing and Modeling User Mobility in a Cellular Data Network," *ACM PEWASUN*, Montreal, QC, Canada, Oct. 2005.

[81] U. Paul, A. P. Subramanian, M. M. Buddhikot, S. R. Das "Understanding Traffic Dynamics in Cellular Data Networks." *IEEE Infocom*, Shanghai, PRC, Apr. 2011.

[82] S. Scepanovic, P. Hui, A. Yla-Jaaski, "Revealing the Pulse of Human Dynamics in a Country from Mobile Phone Data," *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[83] E. Cho, S. A. Myers, J. Leskovec, "Friendship and Mobility: User Movement in Location-Based Social Networks," *ACM SIGKDD*, San Diego, CA, USA, Aug. 2011.

[84] G. Ranjan, H. Zang, Z.-L. Zhang, J. Bolot, "Are Call Detail Records Biased for Sampling Human Mobility?", *ACM Mobile Computing and Communications Review*, 16(3):33–44, Jul. 2012.

[85] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, A. Varshavsky, "Identifying Important Places in Peoples Lives from Cellular Network Data," *Pervasive*, San Francisco, CA, USA, Jun. 2011.

[86] M. Mamei, L. Ferrari, "Daily Commuting in Ivory Coast: Development Opportunities," *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[87] B. Cici, A. Markopoulou, E. Frias-Martinez, N. Laoutaris, "Quantifying the Potential of Ride-Sharing using Call Description Records," *ACM HotMobile*, Jekyll Island, GA, USA, Feb. 2013.

[88] C.M. Schneider, V. Belik, T. Couronne, Z. Smoreda, M.C. Gonzalez, "Unravelling Daily Human Mobility Motifs," *J.R.Soc. Interface*, 10(84), May 2013.

[89] M. Nanni, R. Trasarti, B. Furletti, L. Gabrielli, P. Van Der Mede, J. De Bruijn, E. De Romph, G. Bruil, "Transportation Planning Based on GSM Traces: A Case Study on Ivory Coast," *CitiSens*, Barcelona, Spain, Sep. 2013.

[90] V. Frias-Martinez, J. Virseda, A. Rubio, E. Frias-Martinez, "Towards Large Scale Technology Impact Analyses: Automatic Residential Localization from Mobile Phone-Call Data", *ICTD*, London, UK, Dec. 2010.

[91] B. Csaji, A. Browet, V.A. Traag, J.-C. Delvenne, E. Huens, P. Van Dooren, Z. Smoreda, V.D. Blondel, "Exploring the Mobility of Mobile Phone Users," *Physica A*, 392(6):1459–1473, Jun. 2013.

[92] C. Song, T. Koren, P. Wang, A.-L. Barabasi, "Modelling the Scaling Properties of Human Mobility," *Nature Physics*, 6(10):818–823, Sep. 2010.

[93] A. Sridharan, J. Bolot, "Location Patterns of Mobile Users : A Large-Scale Study," *IEEE Infocom*, Turin, Italy, Apr. 2013.

[94] A. Rubio, V. Frias-Martinez, E. Frias-Martinez, N. Oliver, "Human Mobility in Advanced and Developing Economies: A Comparative Analysis," *AAAI AI-D*, Palo Alto, CA, USA, Mar. 2010.

[95] D. Zhang, F. Zhang, J. Huang, C. Xu, Y. Li, T. He, "Exploring Human Mobility with Multi-Source Data at Extremely Large Metropolitan Scales," *ACM MobiCom*, Maui, HI, USA, Sep. 2014.

[96] M. Vieira, E. Frias-Martinez, P. Bakalov, V. Frias-Martinez, V. Tsortas, "Querying Spatio-Temporal Petterns in Mobile Phone-Call Databases", *IEEE MDM*, Kansas City, MO, USA, May 2010.

[97] M.C. Gonzalez, C.A. Hidalgo, A.-L. Barabasi, "Understanding Individual Human Mobility Patterns," *Nature*, 453(7196):779–782, Jun. 2008.

[98] M. Mitrovic, V. Palchykov, H.-H. Jo, J. Saramaki, "Mobility and Communication Patterns in Ivory Coast," *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[99] X. Liang, J. Zhao, L. Dong, K. Xu, "Unraveling the Origin of Exponential Law in Intra-Urban Human Mobility," *Scientific Reports*, 3(2983), Oct. 2013.

[100] H. Zang, J. Bolot. "Mining Call and Mobility Data to Improve Paging Efficiency in Cellular Networks," *ACM MobiCom*, Montreal, Quebec, Canada, Sep. 2007.

[101] C. Song, Z. Qu, N. Blumm, A.-L. Barabasi, "Limits of Predictability in Human Mobility," *Science*, 327(5968):1018–1021, Jan. 2010.

[102] X. Lu, E. Wetter, N. Bharti, A.J. Tatem, L. Bengtsson, "Approaching the Limit of Predictability in Human Mobility," *Scientific Reports*, 3(10), Oct. 2013.

[103] X. Lu, L. Bengtsson, P. Holme. "Predictability of Population Displacement after the 2010 Haiti Earthquake," *Proc. National Academy of Sciences*, 109(29):11576–11581, May 2012.

[104] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, J. Rowland, A. Varshavsky, "A Tale of Two Cities," *ACM HotMobile*, Annapolis, ML, USA, Feb. 2010.

[105] S. Isaacman, R. Becker, R. Caceres, S. Kobourov, M. Martonosi, J. Rowland, A. Varshavsky, "Ranges of Human Mobility in Los Angeles and New York," *IEEE PerCom Workshops*, Seattle, WA, USA, Mar. 2011.

[106] A. Wesolowski, N. Eagle, A.M. Noor, R.W. Snow, C.O. Buckee, "The Impact of Biases in Mobile Phone Ownership on Estimates of Human Mobility", *J.R.Soc. Interface*, 10(81), Feb. 2013.

[107] L. Bengtsson, X. Lu, A. Thorson, R. Garfield, J. von Schreeb, "Improved Response to Disasters and Outbreaks by Tracking Population Movements with Mobile Phone Network Data: A Post-Earthquake Geospatial Study in Haiti," *PLoS Medicine*, 8(8):1–9, Aug. 2011.

[108] F. Simini, M. Gonzalez, A. Maritan, A.-L. Barabasi, "A Universal Model for Mobility and Migration Patterns," *Nature*, 484(7392):96–100, Apr. 2012.

[109] S. Isaacman, R. Becker, R. Caceres, M. Martonosi, J. Rowland, V. Varshavsky, W. Willinger, "Human Mobility Modeling at Metropolitan Scales," *ACM MobiSys*, Low Wood Bay, Lake District, United Kingdom, Jun. 2012.

[110] D.J. Mir, S. Isaacman, R. Caceres, M. Martonosi, R.N. Wright, "DP-WHERE: Differentially Private Modeling of Human Mobility," *IEEE BigData*, Santa Clara, CA, USA, Oct. 2013.

[111] Y. Yang, C. Herrera, N. Eagle, M.C. Gonzalez, "A Multi-Scale Multi-Cultural Study of Commuting Patterns Incorporating Digital Traces," *NetMob*, Boston, MA, USA, May 2013.

[112] G. Rose, "Mobile Phones as Traffic Probes: Practices, Prospects and Issues," *Transport Reviews*, 26(3):275–291, May 2006.

[113] Z. Qiu, P. Cheng, "State of the Art and Practice: Cellular Probe Technology Applied in Advanced Traveler Information System." *TRB 86th Annual Meeting*, Washington, DC, USA, Jan. 2007.

[114] N. Caceres, J. Wideberg, F.G. Benitez, "Review of Traffic Data Estimations Extracted from Cellular Networks," *IET Intelligent Transport Systems*, 2(3):179–192, Sep. 2008.

[115] S.V. Wunnava, K. Yen, T. Babij, R. Zavaleta, R. Romero, C. Archilla, "Travel Time Estimation Using Cell Phones for Highways and Roadways," *Florida Department of Transportation Final Report*, Dec. 2007.

[116] H. Bar-Gera, "Evaluation of a Cellular Phone-based System for Measurements of Traffic Speeds and Travel Times: A Case Study from Israel," *Transportation Research Part C*, 15(6):380–391, Dec. 2007.

[117] A. Janecek, D. Valerio, K.A. Hummel, F. Ricciato, H. Hlavacs, "Cellular Data Meet Vehicular Traffic Theory: Location Area Updates and Cell Transitions for Travel Time Estimation," *ACM UbiComp*, Pittsburgh, PA, USA, Sep. 2012.

[118] N. Caceres, L.M. Romero, F.G. Benitez, J.M.D. Castillo, "Traffic Flow Estimation Models using Cellular Phone Data," *IEEE Transactions on Intelligent Transportation Systems*, 13(3): 1430–1441, Sep. 2012.

[119] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, C. Ratti, "Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome," *IEEE Trans. Intelligent Transportation Systems*, 12(1):141–151, Mar. 2011.

[120] R. Bolla, F. Davoli, "Road Traffic Estimation from Location Tracking Data in the Mobile Cellular Network," *IEEE WCNC*, Chicago, IL, USA, Sep. 2000.

[121] J. White, I. Wells, "Extracting Origin Destination Information from Mobile Phone Data," *IEE RTIC*, London, UK, Mar. 2002.

[122] J. Doyle, P. Hung, D. Kelly, S. McLoone, R. Farrell, "Utilising Mobile Phone Billing Records for Travel Mode Discovery," *ISSC*, Dublin, Ireland, Jun. 2011.

[123] M. Zilske, K. Nagel, "Building a Minimal Traffic Model from Mobile Phone Data," *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[124] B. Furletti, L. Gabrielli, S. Rinzivillo, C. Renso, "Identifying Users Profiles from Mobile Calls Habits", *ACM UrbComp*, Beijing, PRC, Aug. 2012.

[125] C. Iovan, A.-M. Olteanu-Raimond, T. Couronne, Z. Smoreda, "Moving and Calling: Mobile Phone Data Quality Measurements and Spatiotemporal Uncertainty in Human Mobility Studies," *Geographic Information Science at the Heart of Europe*, D. Vandenbroucke, B. Bucher, J. Crompvoets (editors), Springer, 247–265, May 2013.

[126] C. Williamson, E. Halepovic, H. Sun, Y. Wu "Characterization of CDMA2000 Cellular Data Network Traffic." *IEEE LCN*, Sydney, Australia, Nov. 2005.

[127] R. Keralapura, A. Nucci, Z.-L. Zhang, L. Gao. "Profiling Users in a 3G Network Using Hourglass Co-Clustering." *ACM MobiCom*, Chicago, Illinois, USA, Sep. 2010.

[128] M.Z. Shafiq, L. Ji, A. X. Liu, J. Wang. "Characterizing and modeling internet traffic dynamics of cellular devices." *ACM SIGMETRICS*, San Jose, California, USA, Jun. 2011.

[129] Y. Zhang, A. Arvidsson, "Understanding the Characteristics of Cellular Data Traffic," *ACM SIGCOMM CellNet Workshop*, Helsinki, Finland, Aug. 2012.

[130] E. Mucelli, A. C. Viana, K. P. Naveen, and C. Sarraute. "Measurement-driven Mobile Data Traffic Modeling in a Large Metropolitan Area," *IEEE PerCom*, St. Louis, MO, USA, Mar. 2015.

[131] Y. Wang, M. Faloutsos, H. Zang "On the Usage Patterns of Multimodal Communication: Countries and Evolution." *IEEE GI*, Turin, Italy, Apr. 2013.

[132] F. Girardin, A. Vaccari, A. Gerber, A. Biderman, C. Ratti "Towards Estimating the Presence of Visitors from the Aggregate Mobile phone Network Activity They Generate." *CUPUM*, Hong Kong, PRC, Jun. 2009.

[133] H. Hohwald, E. Frias-Martinez, N. Oliver "User Modeling for Telecommunication Applications: Experiences and Practical Implications." *UMAP*, Big Island, Hawaii, USA, Jun. 2010.

[134] J. C. Cardona, R. Stanojevic, N. Laoutaris, "Collaborative Consumption for Mobile Broadband: A Quantitative Study," *ACM CoNext*, Syndney, Australia, Dec. 2014.

[135] M.Z. Shafiq, L. Ji, A.X. Liu, J. Pang, J. Wang, "Large-Scale Measurement and Characterization of Cellular Machine-to-Machine Traffic," *IEEE/ACM Transactions on Networking*, 21(6):1960-1973, Dec. 2013.

[136] C. Ratti, R. M. Pulselli, S. Williams, D. Frenchman. "Mobile Landscapes: Using Location Data from Cell-Phones for Urban Analysis." *Environment and Planning B Planning and Design* 33(5): 727, 2006.

[137] D. Willkomm, S. Machiraju, J. Bolot, A. Wolisz "Primary Users in Cellular Networks: A Large-Scale Measurement Study." *IEEE DySPAN*, Chicago, Illinois, USA, Oct. 2008.

[138] M. Cerinsek, J. Bodlaj, V. Batagelj, "Symbolic Clustering of Users and Antennae", *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[139] S. Hoteit, S. Secci, G. Pujolle, Z. He, C. Ziemlicki, Z. Smoreda, C. Ratti "Content Consumption Cartography of the Paris Urban Region using Cellular Probe Data." *UrbaNe*, Nice, France, Dec. 2012.

[140] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, J. Wang. "Characterizing Geospatial Dynamics of Application Usage in a 3G Cellular Data Network." *IEEE Infocom*, Orlando, Florida, USA, Mar. 2012.

[141] R. Trasarti, A.-M. Olteanu-Raimond, M. Nanni, T. Couronne, B. Furletti, F. Giannotti, Z. Smoreda, C. Ziemlicki “Discovering Urban and Country Dynamics from Mobile Phone Data with Spatial Correlation Patterns.” *NetMob*, Boston, MA, USA, May 2013.

[142] B. Zong, P. Bogdanov, A. K. Singh, “Constrained Link Prediction on the D4D Dataset”, *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[143] A.-L. Barabasi, R. Albert, “Emergence of Scaling in Random Networks”, *Science*, 286(5439):509–512, Oct. 1999.

[144] F. H. Z. Xavier, L. M. Silveira, J. M. Almeida, A. Ziviani, C. H. S. Malab, H. M. Neto “Analyzing the Workload Dynamics of a Mobile Phone Network in Large Scale Events.” *UrbaNe*, Nice, France, Dec. 2012.

[145] M.Z. Shafiq, L. Ji, A.X. Liu, J. Pang, S. Venkataraman, J. Wang, “A First Look at Cellular Network Performance during Crowded Events,” *ACM SIGMETRICS*, Pittsburgh, PA, USA, Jun. 2013.

[146] F. H. Z. Xavier, L. M. Silveira, J. M. Almeida, C. H. S. Malab, A. Ziviani, H. T. Marques-Neto “Understanding Human Mobility Due to Large-Scale Events.” *NetMob*, Boston, MA, USA, May 2013.

[147] P. Paraskevopoulos, T.-C. Dinh, Z. Dashdorj, T. Palpanas, L. Serafini “Identification and Characterization of Human Behavior Patterns from Mobile Phone Data.” *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[148] D. Pastor-Escuredo, T. Savy, M. A. Luengo-Oroz “Can Fires, Night Lights, and Mobile Phones Reveal Behavioral Fingerprints Useful for Development?” *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[149] S. V. D. Elzen, J. Blaas, D. Holten, J.-K. Buenen, J. J. V. Wijk, R. Spousta, A. Miao, S. Sala, S. Chan “Exploration and Analysis of Massive Mobile Phone Data: A Layered Visual Analytics Approach.” *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[150] J. Bodlaj, M. Cerinsek, V. Batagelj, “Visualization of Traffic”, *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[151] M.V. Rodriguez, V. Mendiratta, B. Lim, D. Doran, D. Klabjan, “Interactive Visualization of Cellphone Network Data using D3: The Case of Ivory Coast”, *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[152] J. Smith, J. Stevens, M. Idris, “NVizABLE: A Web-based Network Visualization Interface”, *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[153] F. Ben Abdesslem, A. Lindgren, “Large Scale Characterisation of YouTube Requests in a Cellular Network,” *IEEE WoWMoM*, Sydney, Australia, Mar. 2014.

[154] K. Dasgupta, R. Singh, B. Viswanathan, D. Chakraborty, S. Mukherjea, A. A. Nanavati. “Social Ties and their Relevance to Churn in Mobile Telecom Networks.” *EDBT*, Nantes, France, Mar. 2008.

[155] Q. Lin, “Mobile Customer Clustering Analysis based on Call Detail Records.” *Communications of the IIMA* 7(4): 95-100, 2007.

[156] R. Becker, R. Caceres, K. Hanson, J. M. Loh, S. Urbanek, A. Varshavsky, C. Volinsky “Clustering Anonymized Mobile Call Detail Records to Find Usage Groups.” *PURBA*, San Francisco, CA, USA, Jun. 2011.

[157] T. Couronne, Z. Smoreda, A.-M. Olteanu. “Chatty Mobiles: Individual Mobility and Communication Patterns.” *NetMob*, Boston, MA, USA, Oct. 2011.

[158] Q. Xu, A. Gerber, Z.M. Mao, J. Pang. “AccuLoc: Practical localization of peformance measurement in 3G networks,” *ACM MobiSys*, Washington, DC, USA, Jun. 2011.

[159] F. Qian, Z. Wang, A. Gerber, Z.M. Mao, S. Sen, O. Spatscheck, “Characterizing Radio Resource Allocation for 3G Networks.” *ACM IMC*, Melbourne, Australia, Nov. 2010.

[160] A. Balachandran, V. Aggarwal, E. Halepovic, J. Pang, S. Seshan, S. Venkataraman, H. Yan, “Modeling Web Quality-of-Experience on Cellular Networks,” *ACM MobiCom*, Maui, Hawaii, USA, Sep. 2014.

[161] M.Z. Shafiq, J. Erman, L. Ji, A.X. Liu, J. Pang, J. Wang, “Understanding the Impact of Network Dynamics on Mobile Video User Engagement,” *ACM SIGMETRICS*, Austin, TX, USA, Jun. 2014.

[162] A. Gerber, M. Hajiaghayi, D. Pei, S. Sen, J. Erman, “To Cache or Not to Cache: The 3G Case,” *IEEE Internet Computing*, 15(2):27–34, Mar. 2011.

[163] A. Finamore, M. Mellia, Z. Gilani, K. Papagiannaki, V. Erramilli, Y. Grunenberger, “Is There a Case for Mobile Phone Content Pre-Staging?” *ACM CoNEXT*, Santa Barbara, CA, USA, Dec. 2013.

[164] F. Yu, G. Xue, H. Zhu, Z. Hu, M. Li, G. Zhang. “Cutting without Pain: Mitigating 3G Radio Tail Effect on Smartphones” *IEEE Infocom*, Turin, Italy, Apr. 2013.

[165] P. Wang, M.C. Gonzalez, C.A. Hidalgo, A.-L. Barabasi “Understanding the Spreading Patterns of Mobile Phone Viruses,” *Science 324*, no. 5930: 1071-1076, 2009.

[166] R. Agarwal, V. Gauthier , M. Becker. “Information Dissemination using Human Mobility in Realistic Environment - (E-Inspire).” *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[167] Z. Zhu, G. Cao, S. Zhu, S. Ranjan, A. Nucci, “A Social Network Based Patching Scheme for Worm Containment in Cellular Networks,” *IEEE Infocom*, Rio de Janeiro, Brazil, Apr. 2009.

[168] Y. Zhu, C. Zhang, Y. Wang “Mobile Data Delivery through Opportunistic Communications among Cellular Users: A Case Study for the D4D Challenge,” *NetMob D4D Challenge*, Boston, MA, USA, May 2013.

[169] C.-P. Wei, I.-T. Chiu. “Turning Telecommunications Call Details to Churn Prediction: A Data Mining Approach.” *Expert systems with applications* 23(2): 103-112, 2002.

[170] R. Belo, P. Ferreira “Is Social Influence Always Positive? Evidence from a Very Large Mobile Network.” *Economics of Information Technology and Digitization Workshop*, Boston, MA, USA, Jul. 2013.

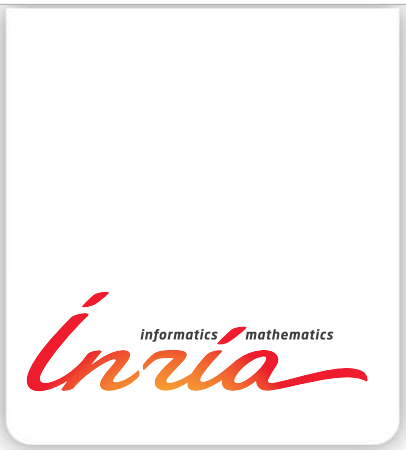
[171] G. Szabo, A.-L. Barabasi, “Network Effects in Service Usage”, *arXiv pre-print*, arXiv:physics0611177, Nov. 2006.

[172] H. Zang, J. Bolot. “Anonymization of Location Data Does Not Work: A Large-Scale Measurement Study,” *ACM MobiCom*, Las Vegas, NV, USA, Sep. 2011.

[173] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, V.D. Blondel, “Unique in the Crowd: The privacy bounds of human mobility,” *Scientific Reports* 3, 2013.

[174] Y. Song, D. Dahlmeier, S. Bressan, “Not So Unique in the Crowd: a Simple and Effective Algorithm for Anonymizing Location Data,” *ACM PIR*, Queensland, Australia, Jul. 2014.

[175] G. Acs, C. Castelluccia, "A Case Study: Privacy Preserving Release of Spatio-Temporal Density in Paris," *ACM SIGKDD*, New York, NY, USA, Aug. 2014.



RESEARCH CENTRE
GRENOBLE – RHÔNE-ALPES

Inovallée
655 avenue de l'Europe Montbonnot
38334 Saint Ismier Cedex

Publisher
Inria
Domaine de Voluceau - Rocquencourt
BP 105 - 78153 Le Chesnay Cedex
inria.fr

ISSN 0249-6399