Big Data in Mobile Networks

Pierdomenico Fiadino and Marc Torrent-Moreno
Data Science and Big Data Analytics Unit, EURECAT (Technology Centre of Catalonia), Barcelona, Spain

Synonyms

Big Data in cellular networks

Definition

The term **Big Data in mobile networks** refers to datasets generated by and collected from mobile cellular networks. Such datasets consist in both the actual user data transported by the network (e.g., voice or data traffic) and metadata and control messages supporting network management.

Overview

A mobile or cellular network is a telecommunication system characterized by a radio last mile and the capability of managing the mobility of end terminals in a seamless fashion (i.e., without interrupting the communication). Such a network is geographically distributed over area units called cells (from which the name *cellular* originates), each served by one or more radio transceivers called *base stations*.

Mobile networks are nowadays the most popular mean for accessing both the traditional public switched telephone network (PSTN) and the public Internet. According to a projection by GSMA, the trade association of mobile operators, 70% of people worldwide will be mobile subscribers by the end of 2017 (Iji 2017). Considering the high penetration of cellular devices, data practitioners look at mobile technologies as an unprecedented information source. Indeed, every terminal produces a large amount of meta-information that can be exploited not only for network optimization and troubleshooting tasks but also for less technical-oriented use cases, such as the study of aggregated human behaviors and trends.

The data volume transferred over an operational network, and the associated metadata such as billing records, deeply varies depending on the network size (number of customers, links, geographical distribution, etc.) and monitored interfaces and can reach rates of several tens of terabytes per day. The analysis of such massive amount of data requires specific system specifications in terms of scalability for storage and processing and ability to deal with historical and real-time data analytics, depending on the application field. Luckily, the progress in the field of Big Data analytics has facilitated the design and deployment of platforms for supporting Network Traffic Monitoring and Analysis (NTMA) applications in cellular context that meet these requirements. In particular, distributed frameworks

S. Sakr, A. Zomaya (eds.), *Encyclopedia of Big Data Technologies*,
https://doi.org/10.1007/978-3-319-63962-8_259-1

based on the MapReduce paradigm have been recently started to be adopted in the field of NTMA, cfr. (Fontugne et al. 2014). Specifically, network monitoring approaches based on Apache Hadoop have been proposed in Lee and Lee (2012), while Liu et al. (2014) focus on rolling traffic analysis.

Cellular Data Types

Operational mobile networks generate an enormous amount of data, both on the **user plane** (e.g., user traffic transferred through the packet-switched domain or voice traffic through the circuit switched) and on the **control plane** (e.g., operators' billing data, signaling for terminals' mobility management and paging), as shown in He et al. (2016). In general, one can distinguish between two macro-groups of cellular data: **billing records** (also called CDR, inherently control data) and **passive traces** (which encompass both user and control data). Figure 1 shows a summary of the cellular data types.

Call Detail Record (CDR)

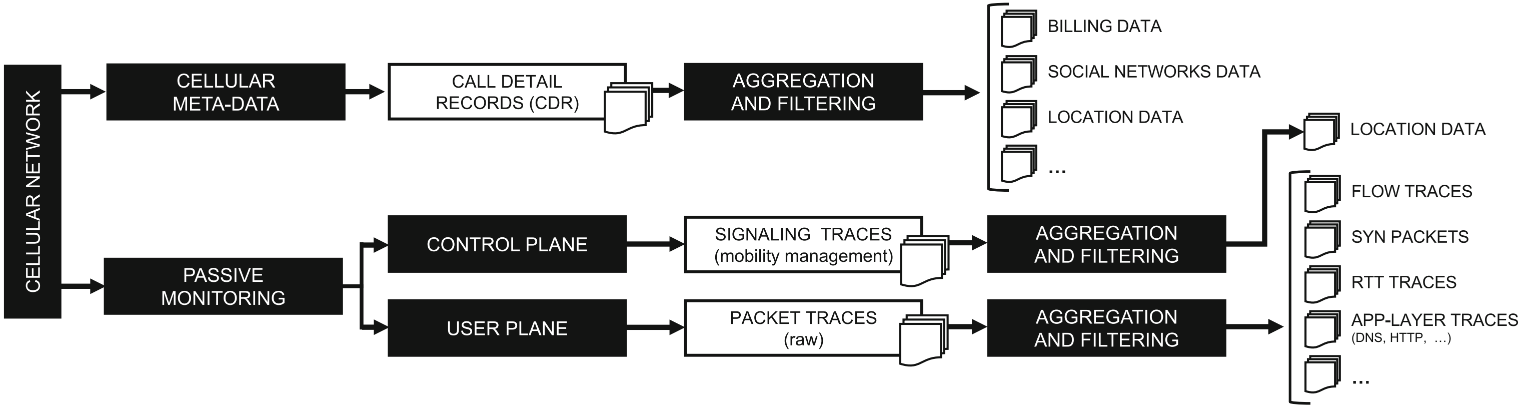
Call detail record (CDR) is the most extensively explored cellular data type. CDRs are metadata generated for supporting the billing of mobile subscribers. They consist in summary tickets of telephone transactions, including the type of activity (voice call, SMS, 2G/3G/4G data connection), the user(s) involved (the *caller* and the *callee*), a time stamp, technical details such as routing information, and the identifier of the cell offering connectivity to the terminal. The latter is especially interesting as it allows to localize the associated action in the boundaries of the cell's coverage area. In other words, CDRs carry details about subscribers' positions when they perform actions, aside from the details needed by the operator for charging them.

CDRs might not be as rich as other cellular data (e.g., signaling data including handover information), but they have been a popular study subject since the 2000s, in particular as a mean to observe human mobility and social networks at scale. The reason behind this interest should be sought in their rather high availability, as network operators are collecting these logs for billing which makes them a ready-made data source. They, in fact, do not require specialized infrastructure for the acquisition. Indeed, there is a fairly rich literature on the their exploitation for a plethora of applications, ranging from road traffic congestion detection and public transport optimization to demographic studies. Despite the initial interest, the research community gradually abandoned this data source as it has become clear that the location details conveyed by CDRs were biased by the users' activity degree. Specifically, the position of a user is observable in conjunction with an activity, which translates in the impossibility of locating a user with fine temporal granularity.

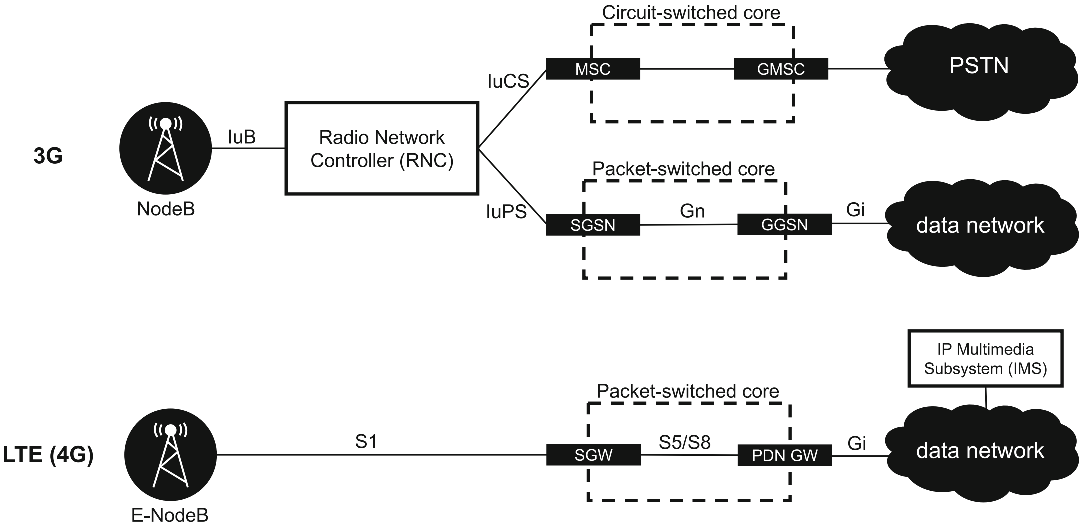
More recent studies (Fiadino et al. 2017), however, argue that the spreading of flat rates for voice and data traffic encourages users to generate more actions. In particular, competitive data plans are nowadays characterized by very high data volume limits, letting customers keep their terminals *always connected*. As a side effect, the quality of mobility data provided by CDRs improved as the average number of tickets per user has increased. A resurgence of research work tackling the study of CDRs is expected to happen in the next years. Indeed, the research community is already targeting some open issues, such as the lack of CDR standardization across operators and more accurate positioning within the cells' coverage areas (Ricciato et al. 2017).

Passive Monitoring

Passive traces are network logs collected through built-in logging capabilities of network elements (e.g., routers) or by wiretapping links through specialized acquisition hardware, such as data acquisition and generation (DAG) cards. Nowadays, high-performance DAG cards are commercially available and are largely employed by cellular network operators in order to monitor their infrastructures at different vantage points (VPs). Passive measurements allow to gain network statistics at a large scale, potentially from millions of connected devices, hence with a high degree of statistical significance. This



Big Data in Mobile Networks, Fig. 1 Summary of data types produced (or carried) by cellular networks. Billing metadata such as call detail records (CDRs) can be exploited for generating users' location data, social network information (from users' interactions), etc. Passively monitored traffic flowing in selected links allows the extraction of signaling messages of the cellular mobility management (which, in turn, provide fine-grained location and mobility data) and integral or partial user frames (packet traces) that can be further parsed to produce aggregated flow traces or filtered to application-specific traces



Big Data in Mobile Networks, Fig. 2 A simplified representation of the 3G and 4G cellular network architecture. 3G networks are characterized by the presence of two cores (the packet switched, handling data connections, and the circuit switched, handling traditional voice traffic). 4G networks have less network elements on the access network side and have just a packet-switched core, optimized for data connections

data type is intrinsically more powerful than CDRs: passive traces are exact copy of the traffic passing through a VP and open the possibility to fully reconstruct the activity of a mobile terminal through the study of both user traffic and control information. Needless to say, this poses serious challenges, in particular for what concern customers' privacy and the potential disclosure of business sensitive information. These issues have heavily limited the access to this kind of data to third parties and the research community.

The employment of passively collected traces for the purpose of network monitoring and management dates back to the 2000s: a first formalization of the techniques for passively monitoring mobile networks can be found in Ricciato (2006), where the author focuses on 3G networks and list common operator-oriented uses of the passively sniffed data. Despite the evolution of cellular networks (e.g., the passage from 3G to 4G standards), the main concept of passive monitoring has not changed much. Figure 2 illustrates a simplified cellular network architecture and highlights the links that can be wiretapped in order to collect passive traces. The cellular infrastructure is composed of two main building blocks: a radio access network (RAN), which manages the radio last mile, and a core network (CN). Fourth-generation networks (such as LTE) are characterized by less elements at the RAN, where the base stations (evolved NodeBs) are connected in a fully meshed fashion. The CN of 2G (GSM/EDGE) and 3G (UMTS/HSPA) networks is formed by the circuit-switched (CS) and the packet-switched (PS) domains, responsible for the traditional voice traffic and packet data traffic, respectively. In contrast, 4G networks are purely packet switched, as all type of traffic is carried over IP, including traditional voice calls which are managed by an IP multimedia subsystem (IMS).

In order to be reachable, mobile terminals notify the CN whenever they change location in the network topology (in other words, to which base stations they are attached to). Therefore, monitoring S1 (in LTE) or IuB/IuCS/IuPS interfaces (in 3G) would allow the collection of mobility information through the observation of network signaling (control) messages (i.e., **signaling traces**). Similarly to CDR datasets, an application for this type of traces is the study of human mobility, as specific signaling messages (such as *handovers*, *location area updates*, *periodic updates*, etc.) allow to reconstruct the trajectory of a terminal in the network topology, which, in turn, can be exploited to infer the actual mobility of the mobile customer with a certain degree of accuracy. These data could potentially provide a higher time granularity in the positioning of terminals w.r.t. CDRs, as the terminal's position is not necessarily observable in conjunction with an action (a call, SMS, data connection). The actual time granularity and position accuracy depend on multiple factors: activity state of the terminal (in general, *active* terminals provide their location changes with higher temporal granularity, even when monitoring at the CN), monitored interfaces (the closer the passive sniffing occurred to the RAN, the higher the chance to get finer-grained positioning in the cellular topology), and use of triangulation (to calculate the position within the cell's coverage area, in case the signal strength is available). The interested reader might refer to Ricciato et al. (2015) for a comprehensive discussion on the passively observed location messages, their time and space granularity, and how they compare with the more popular CDRs.

Monitoring interfaces at the core of the PS domain, such as the Gn/Gi in 3G network and S5/S8/Gi interfaces in LTE, allows the collection of large-scale **packet traces**, i.e., a copy of frames at IP level augmented with metadata such as time stamps, the ID of the link where the wiretapping occurred, and a user (IMSI) and device (IMEI) identifier. The packet traces consist of the full frames (in case the payload is kept) or the header only. Cellular packet traces are extremely informative data, as they provide a complete snapshot of what happened in the network, supporting network management and troubleshooting (e.g., detection of network impairments, data-driven and informed dimensioning of resources, etc.) but also large-scale user behavior studies (e.g., web surfing profiles, marketing analysis, etc.). For most applications, the header brings enough information, and therefore the payload can be discarded, making privacy and storage management easier. It should be noted that the payload is more and more often encrypted.

Raw packet traces can be further processed to produce more refined datasets. For example, by applying a *stateful* tracking of packets, it is possible to reconstruct **flow traces**, i.e., records summarizing *connections* (sequence of packets) between a source and a destination, aggregating by IPs and ports. Such logs could support, among others, the observation of throughput, latency, round-trip time (RTT), or other performance indicators that can be exploited for service-level agreement (SLA) verification and quality of service (QoS) studies. Another typical post-processing approach is filtering specific packet types (e.g., TCP SYN packets for the study of congestions over links and network security enforcement) or applications (i.e., **application-layer traces**, such as DNS, HTTP).

Uses of Cellular Data

In the last years, cellular network data has been broadly employed in different application areas, not necessary with technical purposes. Indeed, the progress in the development of Big Data technologies, coupled with the decrease of prices for storage, has unlocked endless possibilities for the exploitation of the massive amount of data carried and produced by mobile networks.

In general, one can distinguish between two classes of use cases, i.e., **network-related** (technical) and **human-related** (nontechnical). The use cases in the first class aim at achieving *informed* network management and optimization actions, including more specific tasks such as detection of anomalies, troubleshooting, resource dimensioning, and security monitoring. Nontechnical applications focus on the study of human behavior relying on both user and control data exchanged among user equipments, cellular network components, and remote services. Some common human-related uses of cellular data are the study of human mobility (through CDRs or signaling messages), analysis of social networks, and web surfing behavior (through passively monitoring DNS or HTTP traffic at the CN), with applications in the field of marketing and behavioral studies.

Big Data in Mobile Networks, Table 1 A non-exhaustive list of applications for mobile network data divided into two main categories (technical and nontechnical) plus a hybrid class of uses

Technical areas	Nontechnical areas	Hybrid areas
Anomaly detection	Churn prediction	Quality of Experience (QoE)
Traffic classification	Customer loyalty	
Network optimization	Web surfing	
Study of remote services	Marketing and tariff design	
Performance assessment and QoS	Human mobility	

The remainder of this section provide a non-exhaustive list of common uses of cellular data found in the scientific literature (also summarized in Table 1).

Technical Areas

The analysis of cellular data, and passive traces in particular, could support technical tasks aimed at achieving data-driven network engineering and optimization. There is an interesting research branch in this direction, targeting self-optimization, self-configuration, and self-healing based on machine learning approaches. One example is Baldo et al. (2014), where authors investigate the advantages of adopting Big Data techniques for **Self Optimized Networks (SON)** for 4G and future 5G networks. A comprehensive survey on the use of Big Data technologies and machine learning techniques in this field was edited by Klaine et al. (2017).

Cellular passive traces can be also employed for studying the functioning and behavior of **remote Over The Top (OTT) Internet services** (Fiadino et al. 2016). Indeed, the popularity of mobile applications (e.g., video streaming, instant messaging, online social networks, etc.) poses serious challenges to operators as they heavily load the access of cellular networks. Hence, understanding their usage patterns, content location, addressing dynamics, etc. is crucial for better adapting and managing the networks but also to analyze and track the evolution of these popular services.

Congestions and Dimensioning

Nowadays, most of the services have moved over the IP/TCP protocol. The bandwidth-hungry nature of many mobile applications is posing serious challenges for the management and optimization of cellular network elements. In this scenario, monitoring congestions is critical to reveal the presence of under-provisioned resources.

The use of packet traces for the detection of bottlenecks is one of the earliest applications in the field of cellular passive monitoring. Ricciato et al. (2007) rely on the observation of both global traffic statistics, such data rate, as well as TCP-specific indicators, such as the observed retransmission frequency. Jaiswal et al. (2004) target the same goal by adopting a broader set of flow-level metrics, including TCP congestion window and end-to-end round-trip time (RTT) measurements. A more recent work that targets the early detection of bottlenecks can be found in Schiavone et al. (2014).

Traffic Classification

Network traffic classification (TC) is probably one of the most important tasks in the field of network management. Its goal is to allow operators to know what kind of services are been used by customers and how they are impacting the network by studying their passive packet traces. This is particularly relevant in cellular networks, where the resources are in general scarcer and taking informed network management decision is crucial.

The field of TC has been extensively studied: complete surveys can be found in Valenti et al. (2013) and Dainotti et al. (2012). Standard classification approaches rely on deep packet inspection (DPI) techniques, using pattern matching and statistical traffic analysis (Deri et al. 2014;

Bujlow et al. 2015). In the last years, the most popular approach for TC is the application of machine learning (ML) techniques (Nguyen and Armitage 2008), including supervised and unsupervised algorithms.

The increasing usage of web-based services has also shifted the attention to the application of TC techniques to passive traces at higher levels of the protocol stack (specifically, HTTP). Two examples are Maier et al. (2009) and Erman et al. (2011), where authors use DPI techniques to analyze the usage of HTTP-based apps, showing that HTTP traffic highly dominates the total downstream traffic volume. In Casas et al. (2013a), authors characterize traffic captured in the core of cellular networks and classify the most popular services running on top of HTTP. Finally, the large-scale adoption of end-to-end encryption over HTTP has motivated the development of novel techniques to classify applications distributed on top of HTTPS traffic (Bermudez et al. 2012), relying on the analysis of DNS requests, which can be derived by filtering and parsing specific passive packet traces.

Anomaly Detection

Anomaly detection (AD) is a discipline aiming at triggering notifications when unexpected events occur. The definition of what is unexpected – i.e., what differs from a normal or predictable behavior – strictly depends on the context. In the case of mobile networks, it might be related to a number of scenarios, for example, degradation of performance, outages, sudden changes in traffic patterns, and even evidences of malicious traffic. All these events can be observed by mining anomalous patterns in passive traces at different level of granularities.

A comprehensive survey on AD for computer networks can be found in Chandola et al. (2009). An important breakthrough in the field of AD in large-scale networks was set by the work of Lakhina et al. (2005), where authors introduced the application of the principal component analysis (PCA) technique to the detection of network anomalies in traffic matrices. Another powerful statistical approach tailored for cellular networks has been presented in D'Alconzo et al. (2010). The same statistical technique has been expanded and tested in a nationwide mobile network by Fiadino et al. (2014, 2015).

The study presented in Casas et al. (2016a) is particularly interesting: it applies clustering techniques on cellular traces to unveil device-specific anomalies, i.e., potentially harmful traffic patterns caused by well-defined classes of mobile terminals. The peculiarity of this kind of studies consists in the use of certain metadata (device brand, operating system, etc.) that are observable in passive traces collected in mobile networks (in general, monitoring fixed-line networks does not provide such degree of details on the subscribers).

User-Oriented Areas

Mobile network data implicitly bring large-scale information of human behavior. Indeed, operators can mine the satisfaction of users by observing their usage patterns. This type of analysis has clear applications in the fields of marketing, tariff design, and support business decisions in a data-driven fashion. An example is the **prediction of churning users** (i.e., customers lost to competitors): the study presented in Modani et al. (2013) relies on CDRs – and on the social network inferred from them – to extract specific behaviors that lead to termination of contracts.

Another area of user-oriented research is the study of the **quality of experience (QoE)**. The term QoE refers to a hybrid network- and human-related discipline that has the goal of estimating the level of satisfactions of customers when using IP-based services. In other words, practitioners in this field try to map network performance indicators (e.g., throughput, RTT, latency, etc.) with subjective scores assigned by users in controlled environments. Some examples of QoE studies conducted over cellular network data are Casas et al. (2013b, 2016b). In these works, authors study the QoE of popular applications accessed through mobile devices from the perspective of passive cellular measurements, as well as subjective tests in controlled labs. The performance of multimedia services (VoIP, video streaming, etc.) over cellular networks is affected by a number of factors, depending on the radio access network (signal strength), core network setup (internal routing), and remote provisioning systems. This class of studies helps operators in properly dimensioning network resources not only to meet provisioning requirements from the network perspective but also taking into account the actual perception of quality by final users.

Human Mobility

The exploitation of mobile terminals and cellular networks as source of mobility information is possibly the most popular human-oriented topic in the field of mobile network data analysis. Cellular-based mobility studies have a high number of applications (e.g., city planning, analysis of vehicular traffic, public transportation design, etc.). Recently, cellular location data has been exploited to study the population density (Ricciato et al. 2015) and to infer socioeconomic indicators and study their relationship with mobility patterns (Pappalardo et al. 2016).

Mobility studies through cellular data are based on the assumption that the movements of a terminal in the context of the network topology reflect the actual trajectory of the mobile subscriber carrying the device. Location information can be extracted from CDRs or signaling messages devoted to the mobility management of terminals. Most of the existing literature is based on CDRs (Bar-Gera 2007; González et al. 2008; Song et al. 2010). However, since CDR datasets only log the position of users when an action occurs, their exploitation for the characterization of human mobility has been criticized (Ranjan et al. 2012). In Song et al. (2010), authors highlight some concerns in this direction and also show that users are inactive most of the time. The main limitation lies in the fact that the mobility perceived from the observation of CDRs is highly biased, as it strictly depends on the specific terminals' action patterns. In other words, users are *visible* during few punctual instants, which makes most of movements untraceable. Fiadino et al. (2017) claim that the situation is changing and nowadays CDRs are richer than the past, due to the increasing usage of data connections and background applications that has consequently increased the number of records. The now higher frequency of tickets allows a finer-grained tracking of users' positions. Some recent mobility studies, such as Callegari et al. (2017) and Jiang et al. (2017), are, in fact, successfully based on CDRs.

Few studies have explored more sophisticated approaches based on passively captured signaling messages – e.g., handover, location area updates, etc. Caceres et al. (2007). In Fiadino et al. (2012), authors studied the differences between CDRs and signaling data in terms of number of actions per user. Depending on the monitored interfaces, these approaches greatly vary in terms of cost and data quality (Ricciato 2006). Although the analysis of signaling is promising, there is a general lack of studies based on actual operational signaling data (some exceptions are the works by Fiadino et al. (2012) and Janecek et al. (2015), where authors focus on the study of vehicular traffic on highways and detection of congestions), as complex dedicated monitoring infrastructures for the extraction and immense data storage systems are required.

Privacy Concerns

The collection and exploitation of cellular data might raise some concerns. Indeed, given the penetration of mobile devices and the modern usage patterns, cellular networks carry an extremely detailed ensemble of private information. As seen before, by observing cellular metadata or specific network traces, it is possible to reconstruct the mobility of a mobile subscriber and infer the list of visited places, learn the acquaintances and build social networks, and even study the web surfing habits.

Customers' activity is associated to unique identifiers, such as the IMSI (International Mobile Subscriber Identity) and the IMEI (International Mobile Equipment Identity). These IDs make it possible to link all the collected logs to single users and, in turn, to their identity. To protect customers' privacy, the **anonymization** of such unique identifiers is a common practice adopted by operators before handing privacy-sensitive datasets to internal or third-party practitioners. Indeed, sharing network traces with external entities is critical in some applications, e.g., detection of attacks with federated approaches (Yurcik et al. 2008).

A typical anonymization technique consists in applying nonreversible hashing functions, such as MD5 or SHA-1, with secret salt values to IMSI and IMEI. Being these hash functions cryptographically safe, it is still possible to associate all logs and actions of a single user to an anonymous identifier, with limited risks of *collisions* (differently from other simpler approaches, such as truncation of the last digits of customer identifiers). In addition, their one-way nature prevents reconstructing the original IDs. There is a rather rich literature on advanced anonymization techniques: a recent example can be found in Mivule and Anderson (2015).

It should be noted that all the listed applications for cellular data target macroscopic behaviors, observing users' traces on the large scale. The cases in which a single user needs to be addressed are rare and restricted to specific uses (e.g., detection of malicious users) or following authorized legal warrants. Nevertheless, the industry and the research community are addressing such concerns. For example, Dittler et al. (2016) propose an approach for location management that protects users' location privacy without disrupting the basic functionalities of cellular networks. On the other hand, the increasing adoption of encryption protocols to transport user data (e.g., HTTPS, instant messaging and VoIP with end-to-end encryption, etc.) limits some potentially abusive uses of passive traces.

Cross-References

- ▶ Big Data and Network QoS/QoE
- ▶ Big Data in Computer Network Monitoring
- ▶ Big Data in Network Anomaly Detection

References

Baldo N, Giupponi L, Mangues-Bafalluy J (2014) Big data empowered self organized networks. In: European wireless 2014; 20th European wireless conference, pp 1–8. https://doi.org/10.5281/zenodo.268949

Bar-Gera H (2007) Evaluation of a cellular phone-based system for measurements of traffic speeds and travel times: a case study from Israel. Transp Res Part C Emerg Technol 15(6):380–391. https://doi.org/10.1016/j.trc.2007.06.003

Bermudez IN, Mellia M, Munafo MM, Keralapura R, Nucci A (2012) DNS to the rescue: discerning content and services in a tangled web. In: Proceedings of the 2012 internet measurement conference, IMC'12, pp 413–426

Bujlow T, Carela-Español V, Barlet-Ros P (2015) Independent comparison of popular DPI tools for traffic classification. Comput Netw 76:75–89. http://dx.doi.org/10.1016/j.comnet.2014.11.001

Caceres N, Wideberg JP, Benitez FG (2007) Deriving origin destination data from a mobile phone network. IET Intell Transp Syst 1(1):15–26. https://doi.org/10.1049/iet-its:20060020

Callegari C, Garroppo RG, Giordano S (2017) Inferring social information on foreign people from mobile traffic data. In: 2017 IEEE international conference on communications (ICC), pp 1–6. https://doi.org/10.1109/ICC.2017.7997255

Casas P, Fiadino P, Bär A (2013a) Ip mining: extracting knowledge from the dynamics of the internet addressing space. In: Proceedings of the 2013 25th international teletraffic congress (ITC), pp 1–9. https://doi.org/10.1109/ITC.2013.6662933

Casas P, Seufert M, Schatz R (2013b) Youqmon: a system for on-line monitoring of youtube QoE in operational 3G networks. SIGMETRICS Perform Eval Rev 41(2):44–46

Casas P, Fiadino P, D'Alconzo A (2016a) When smartphones become the enemy: unveiling mobile apps anomalies through clustering techniques. In: Proceedings of the 5th workshop on all things cellular: operations, applications and challenges (ATC'16), pp 19–24

Casas P, Seufert M, Wamser F, Gardlo B, Sackl A, Schatz R (2016b) Next to you: monitoring quality of experience in cellular networks from the end-devices. IEEE Trans Netw Serv Manag 13(2):181–196

Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv 41(3):15:1–15:58

Dainotti A, Pescape A, Claffy K (2012) Issues and future directions in traffic classification. IEEE Netw 26(1):35–40. https://doi.org/10.1109/MNET.2012.6135854

D'Alconzo A, Coluccia A, Romirer-Maierhofer P (2010) Distribution-based anomaly detection in 3G mobile networks: from theory to practice. Int J Netw Manag 20(5):245–269. https://doi.org/10.1002/nem.747

Deri L, Martinelli M, Bujlow T, Cardigliano A (2014) nDPI: open-source high-speed deep packet inspection. In: 2014 international wireless communications and mobile computing conference (IWCMC), pp 617–622. https://doi.org/10.1109/IWCMC.2014.6906427

Dittler T, Tschorsch F, Dietzel S, Scheuermann B (2016) ANOTEL: cellular networks with location privacy. In: 2016 IEEE 41st conference on local computer networks (LCN), pp 635–638. https://doi.org/10.1109/LCN.2016.110

Erman J, Gerber A, Sen S (2011) HTTP in the home: it is not just about PCs. SIGCOMM Comput Commun Rev 41(1):90–95. https://doi.org/10.1145/1925861.1925876

Fiadino P, Valerio D, Ricciato F, Hummel KA (2012) Steps towards the extraction of vehicular mobility patterns from 3G signaling data. Springer, Berlin/Heidelberg, pp 66–80

Fiadino P, D'Alconzo A, Bär A, Finamore A, Casas P (2014) On the detection of network traffic anomalies in content delivery network services. In: 2014 26th international teletraffic congress (ITC), pp 1–9. https://doi.org/10.1109/ITC.2014.6932930

Fiadino P, D'Alconzo A, Schiavone M, Casas P (2015) Rcatool – a framework for detecting and diagnosing anomalies in cellular networks. In: Proceedings of the 2015 27th international teletraffic congress (ITC'15), pp 194–202

Fiadino P, Casas P, D'Alconzo A, Schiavone M, Baer A (2016) Grasping popular applications in cellular networks with big data analytics platforms. IEEE Trans Netw Serv Manag 13(3):681–695. https://doi.org/10.1109/TNSM.2016.2558839

Fiadino P, Ponce-Lopez V, Antonio J, Torrent-Moreno M, D'Alconzo A (2017) Call detail records for human mobility studies: taking stock of the situation in the "always connected era". In: Proceedings of the workshop on big data analytics and machine learning for data communication networks (Big-DAMA'17), pp 43–48

Fontugne R, Mazel J, Fukuda K (2014) Hashdoop: a mapreduce framework for network anomaly detection. In: 2014 IEEE conference on computer communications workshops (INFOCOM WKSHPS), pp 494–499. https://doi.org/10.1109/INFCOMW.2014.6849281

González MC, Hidalgo C, Barabási A (2008) Understanding individual human mobility patterns. Nature 453:779–782. https://doi.org/10.1038/nature06958, 0806.1256

He Y, Yu FR, Zhao N, Yin H, Yao H, Qiu RC (2016) Big data analytics in mobile cellular networks. IEEE Access 4:1985–1996. https://doi.org/10.1109/ACCESS.2016.2540520

Iji M (2017) GSMA intelligence – unique mobile subscribers to surpass 5 billion this year. https://www.gsmaintelligence.com/research/2017/02/unique-mobile-subscribers-to-surpass-5-billion-this-year/613. Accessed 08 Feb 2018

Jaiswal S, Iannaccone G, Diot C, Kurose J, Towsley D (2004) Inferring TCP connection characteristics through passive measurements. In: IEEE INFOCOM 2004, vol 3, pp 1582–1592. https://doi.org/10.1109/INFCOM.2004.1354571

Janecek A, Valerio D, Hummel K, Ricciato F, Hlavacs H (2015) The cellular network as a sensor: from mobile phone data to real-time road traffic monitoring. IEEE Trans Intell Transp Syst. https://doi.org/10.1109/TITS.2015.2413215

Jiang S, Ferreira J, Gonzalez MC (2017) Activity-based human mobility patterns inferred from mobile phone data: a case study of Singapore. IEEE Trans BigData. https://doi.org/10.1109/TBDATA.2016.2631141

Klaine PV, Imran MA, Onireti O, Souza RD (2017) A survey of machine learning techniques applied to self organizing cellular networks. IEEE Commun Surv Tutorials 99:1–1. https://doi.org/10.1109/COMST.2017.2727878

Lakhina A, Crovella M, Diot C (2005) Mining anomalies using traffic feature distributions. SIGCOMM Comput Commun Rev 35(4):217–228

Lee Y, Lee Y (2012) Toward scalable internet traffic measurement and analysis with Hadoop. SIGCOMM Comput Commun Rev 43(1):5–13

Liu J, Liu F, Ansari N (2014) Monitoring and analyzing big traffic data of a large-scale cellular network with Hadoop. IEEE Netw 28(4):32–39. https://doi.org/10.1109/MNET.2014.6863129

Maier G, Feldmann A, Paxson V, Allman M (2009) On dominant characteristics of residential broadband internet traffic. In: Proceedings of the 9th ACM SIGCOMM conference on internet measurement conference (IMC'09). ACM, New York, pp 90–102. https://doi.org/10.1145/1644893.1644904

Mivule K, Anderson B (2015) A study of usability-aware network trace anonymization. In: 2015 science and information conference (SAI), pp 1293–1304. https://doi.org/10.1109/SAI.2015.7237310

Modani N, dey k, Gupta R, Godbole S (2013) CDR analysis based telco churn prediction and customer behavior insights: a case study. In: Lin X, Manolopoulos Y, Srivastava D, Huang G (eds) Web information systems engineering – WISE 2013. Springer, Berlin/Heidelberg, pp 256–269

Nguyen TTT, Armitage G (2008) A survey of techniques for internet traffic classification using machine learning. IEEE Commun Surv Tutorials 10(4):56–76. https://doi.org/10.1109/SURV.2008.080406

Pappalardo L, Vanhoof M, Gabrielli L, Smoreda Z, Pedreschi D, Giannotti F (2016) An analytical framework to nowcast well-being using mobile phone data. Int J Data Sci Anal 2:75–92

Ranjan G, Zang H, Zhang Z, Bolot J (2012) Are call detail records biased for sampling human mobility? SIGMOBILE Mob Comput Commun Rev 16(3):33

Ricciato F (2006) Traffic monitoring and analysis for the optimization of a 3G network. Wireless Commun 13(6):42–49

Ricciato F, Vacirca F, Svoboda P (2007) Diagnosis of capacity bottlenecks via passive monitoring in 3G networks: an empirical analysis. Comput Netw 51(4):1205–1231

Ricciato F, Widhalm P, Craglia M, Pantisano F (2015) Estimating population density distribution from network-based mobile phone data. Publications Office of the European Union. https://doi.org/10.2788/162414

Ricciato F, Widhalm P, Pantisano F, Craglia M (2017) Beyond the "single-operator, CDR-only" paradigm: an interoperable framework for mobile phone network data analyses and population density estimation. Pervasive Mob Comput 35:65–82. https://doi.org/10.1016/j.pmcj.2016.04.009

Schiavone M, Romirer-Maierhofer P, Ricciato F, Baiocchi A (2014) Towards bottleneck identification in cellular networks via passive TCP monitoring. In: Ad-hoc, mobile, and wireless networks – 13th international conference (ADHOC-NOW 2014). Proceedings, Benidorm, 22–27 June 2014, pp 72–85

Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. https://doi.org/10.1126/science.1177170

Valenti S, Rossi D, Dainotti A, Pescapè A, Finamore A, Mellia M (2013) Reviewing traffic classification. In: Biersack E, Callegari C, Matijasevic M (eds) Data traffic monitoring and analysis: from measurement, classification, and anomaly detection to quality of experience. Springer, Berlin/Heidelberg, pp 123–147

Yurcik W, Woolam C, Hellings G, Khan L, Thuraisingham B (2008) Measuring anonymization privacy/analysis tradeoffs inherent to sharing network data. In: NOMS 2008 – 2008 IEEE network operations and management symposium, pp 991–994. https://doi.org/10.1109/NOMS.2008.4575265