Big Data Processing, Analysis and Applications in Mobile Cellular Networks*

Sanja Brdar[1], Olivera Novović[1], Nastasija Grujić[1], Horacio González–Vélez[2], Ciprian-Octavian Truică[3], Siegfried Benkner[4], Enes Bajrovic[4], and Apostolos Papadopoulos[5]

[1] BioSense Institute, University of Novi Sad, Novi Sad, Serbia
{sanja.brdar, novovic, n.grujic}@biosense.rs
[2] Cloud Competency Centre, National College of Ireland, Dublin 1, Ireland
horacio@ncirl.ie
[3] Computer Science and Engineering Department, Faculty of Automatic Control and Computers, University Politehnica of Bucharest, Bucharest, Romania
ciprian.truica@cs.pub.ro
[4] Faculty of Computer Science, University of Vienna, Austria
{siegfried.benkner, enes.bajrovic}@univie.ac.at,
[5] Department of Informatics, Aristotle University of Thessaloniki, Greece
papadopo@csd.auth.gr

Abstract. When coupled with spatio-temporal context, location-based data collected in mobile cellular networks provide insights into patterns of human activity, interactions, and mobility. Whilst uncovered patterns have immense potential for improving services of telecom providers as well as for external applications related to social wellbeing, its inherent massive volume make such 'Big Data' sets complex to process. A significant number of studies involving such mobile phone data have been presented, but there still remain numerous open challenges to reach technology readiness. They include efficient access in privacy-preserving manner, high performance computing environments, scalable data analytics, innovative data fusion with other sources–all finally linked into the applications ready for operational mode. In this chapter, we provide a broad overview of the entire workflow from raw data access to the final applications and point out the critical challenges in each step that need to be addressed to unlock the value of data generated by mobile cellular networks.

Keywords: Data Analysis · HPC · Big Data · Cellular Networks

1 Mobile Cellular Networks - From Data to Applications

There is a tremendous growth of new applications that are based on the analysis of data generated within mobile cellular networks. Mobile phone service providers

* This article is based upon work from COST Action IC1406 High-Performance Modelling and Simulation for Big Data Applications (cHiPSet), supported by COST (European Cooperation in Science and Technology)

collect large amounts of data with potential value for improving their services as well as to enable social good applications [7]. As an example, every time a user makes via mobile phone interaction (SMS, call, internet), a *call detail record* (CDR) is created and stored by a mobile network operator. CDRs not only log the user activity for billing purposes and network management, but also provide opportunities for different applications such as urban sensing [5], transport planning [3, 28], disaster management [38, 46, 64] socio-economic analysis [45, 57] and monitoring epidemics of infectious diseases [36, 62, 11, 10].

Several studies have reviewed applications to analyse CDRs, however most focus on specific aspects such as data analytics for internal use in telecom companies [26], graph analytics and applications [7], or public health [44]. This survey aims to cover the entire workflow from raw data to final application, with emphasis on the gaps to advance technology readiness. Figure 1 depicts our main concept which shall be used to summarise the state of the art work and identify open challenges.

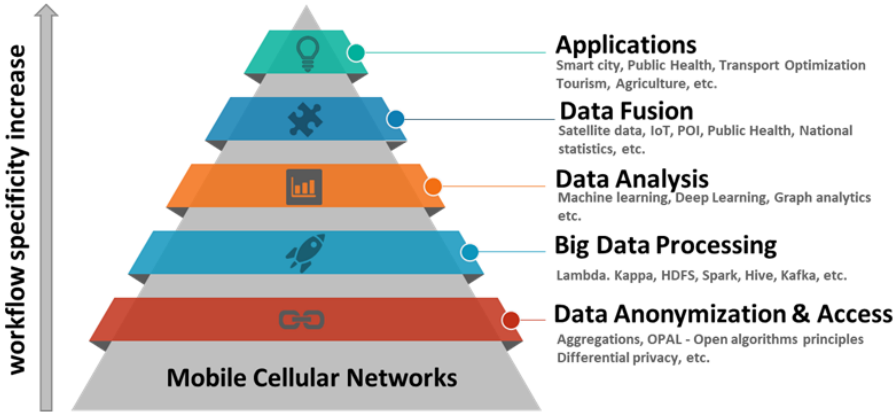

Fig. 1. Mobile Cellular Networks - From Location Data to Applications.

The rest of this paper is structured as follows. Section 2 provides some background on mobile cellular networks and the nature of the data sets available. It also sets the basis for different approaches to anonymization. Section 3 presents a discussion of data-intensive approaches and architectures to deal with the computationally-demanding nature of detecting patterns from telecom data. Then, Section 4 discusses approaches to analyze mobile operators data sets via graph analysis and machine learning. Section 5 enumerates some relevant external data sources that can complement mobile phone data, while Section 6 elaborates on diverse pertinent applications. Finally, Section 7 furnishes the summary and objectives for future research efforts.

2 Data Anonymization and Access

With the pervasive adoption of smartphones in modern societies, in addition to CDRs, there is now a growing interest in xDRs, Extended Data Records. They enclose information on visited web sites, used applications, executed transactions, etc. Coupled with cell-tower triangulation, applications can infer fine-grain phone locations [29], thus making data volumes even larger. Telecom data typically include spatial and temporal parameters to map device activity, connectivity, and mobility.

Telecom operators follow rigorous procedures for data anonymization to preserve privacy such that anonymized records cannot be linked to subscribers under any normal circumstances. Furthermore, before releasing any data to third parties, data sets are usually aggregated on temporal and/or spatial scales. For example, the numbers of calls as well as the duration of calls between any pair of antennas are aggregated hourly and movement trajectories are provided with reduced spatial resolution [1]. Differential privacy paradigm adds noise to original data up to the level not affecting the statistics significantly to preserve users' privacy. Another approach, suggested by the Open Algorithms (OPAL) initiative, proposes moving the algorithm to the data [35]. In their model, raw data are never exposed to outside parties, only vetted algorithms run on telecom companies' servers.

An example of preserving privacy of users by releasing only pre-aggregated data is Telecom Italia Big Data Challenge [4]. Opened data sets accumulated activity and connectivity across defined spatial cells of the city of Milan and in the Province of Trentino in 10 min resolution. Despite aggregation, data sets are still rich source of information, especially when fused with other data such as weather, news, social networks and electricity data from the city. To get some useful insight about the data we further describe and visualize activity and connectivity maps from Telecom Italia data sets and mobility from Telekom Srbija data set.

2.1 Activity

The activity data set consists of records with square id, time interval, sms-in activity, sms-out activity, call-in activity, call-out activity, internet traffic activity and country code, for each square of grid network. The data is aggregated in ten minutes time slots. We did further aggregation on daily level to gain overall insight into daily base activity. Figure 2 illustrates an aggregated activity of mobile phone users in the city of Milan. We observe that areas with highest activity refer to urban core of the city, whereas areas with lower activity levels refer to peripheral parts of the city. The same analysis is performed for the Province of Trentino and corresponding results are presented in Figure 3. Although the inspected area of the Trentino Province exceeds significantly the urban area of the city of Trentno, the same pattern in distribution of mobile phone activity is present - high activity in urban area along lower activity in rural areas. From

the visual inspection of Figure 3 we observe that higher activity areas spatially refer to transit areas with main roads, which was expected.

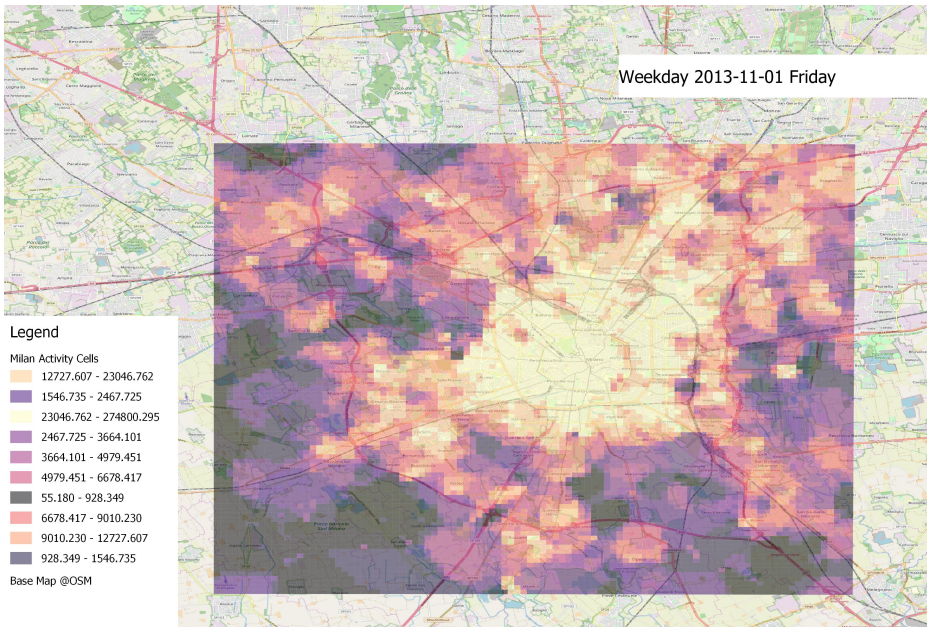

Fig. 2. Aggregated activity over spatial area of the city of Milan.

2.2 Connectivity

Connectivity data provides directional interaction strength among the squares (cells) of the grid network. Records consist of *timestamp*, *square id1*, *square id2* and *strength* which represents the value (weight) of aggregated telecom traffic multiplied with a constant k to hide exact number of calls and sms recorded by single base station [4]. As in [43] we performed additional spatial aggregation, and analyzed connectivity patterns between different city zones of Milan through the lens of graph theory. For illustration purposes we created a single undirected, weighted graph for a typical working day from the data set. In Figure 4 we present the obtained spatial graph of connectivity links. During the work week, the city center acts as a hub, the strongest links are gathered close to the city center, while on weekends and holidays the opposite pattern occurs [43].

The second type of connectivity data presents connectivity from the city of Milan to other Provinces in Italy. Additional aggregation is applied to extract daily base connectivity patterns. Figure 5 presents connectivity links from different areas of the city of Milan to Provinces in Italy. We may conclude that the distribution of connectivity links is regular to all Provinces, and that the majority of links start from central areas of the city of Milan.

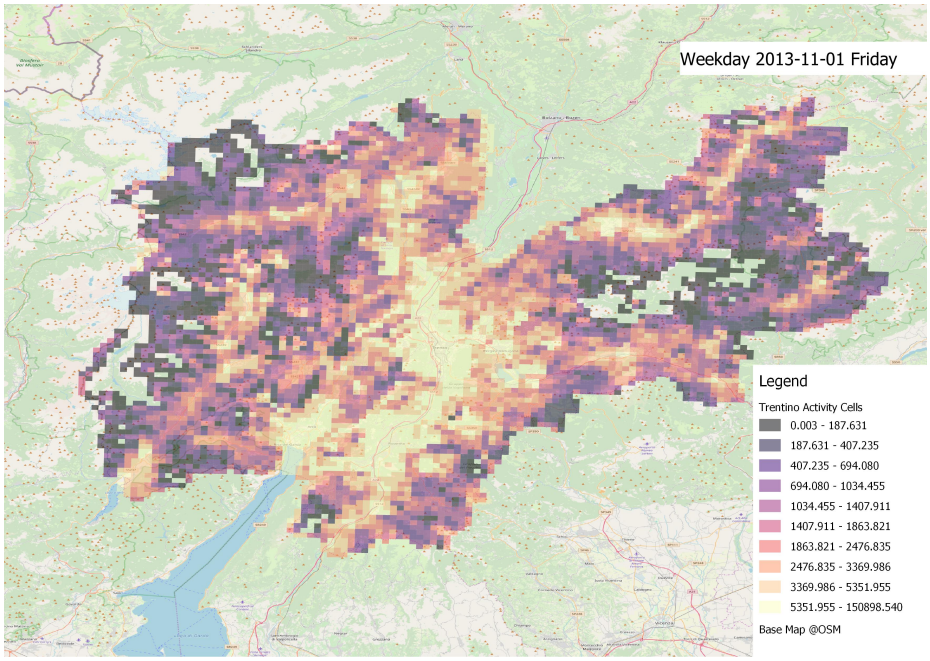

Fig. 3. Aggregated activity over spatial area of Trentino Province

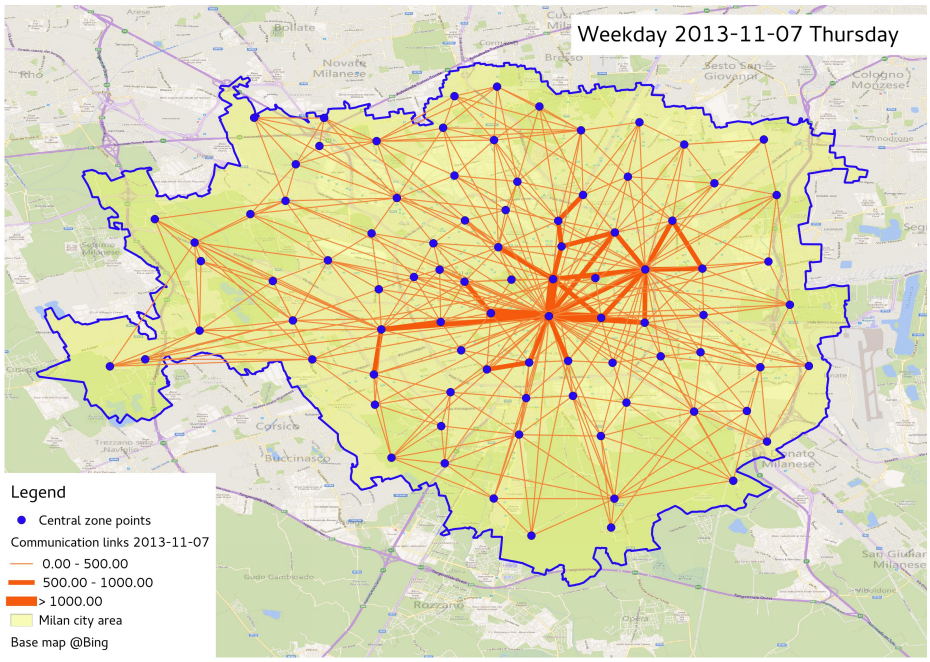

Fig. 4. Mobile Cellular Networks - From Location Data to Applications

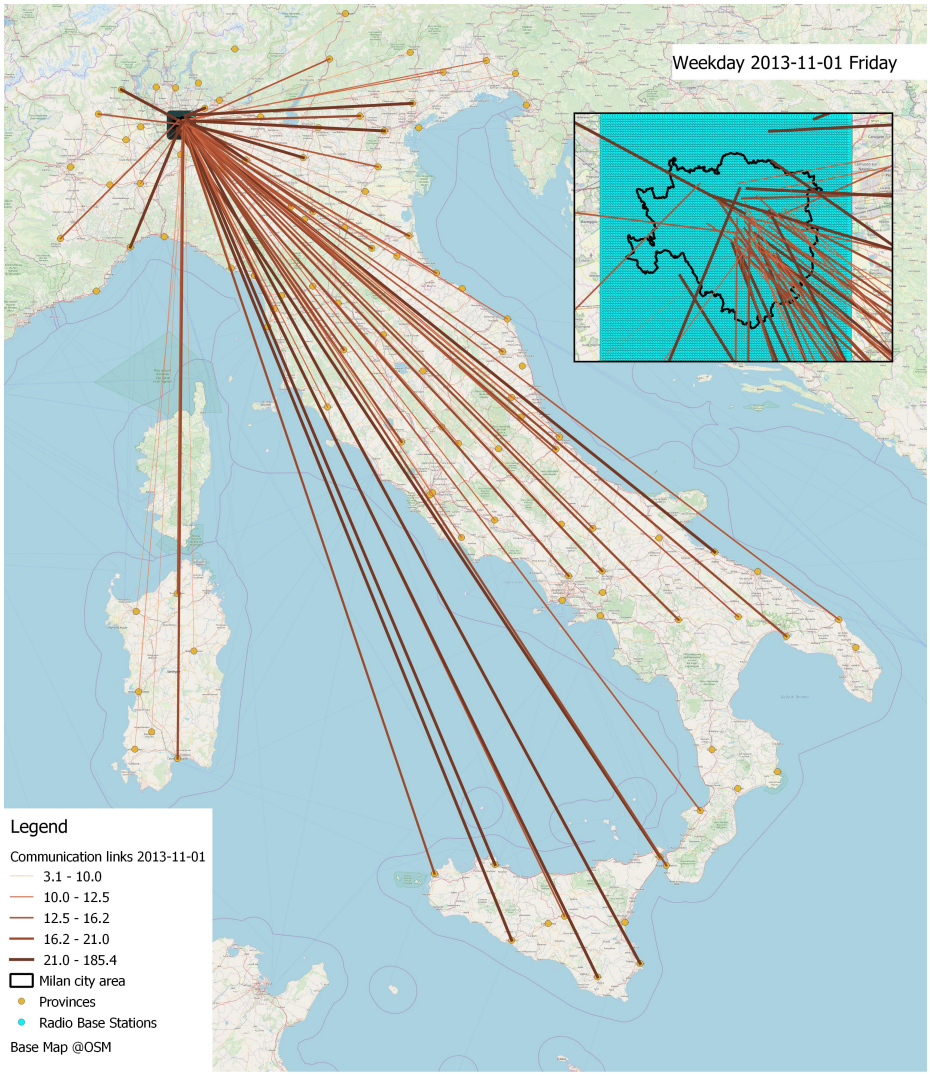

Fig. 5. Connectivity from the city of Milan to Provinces

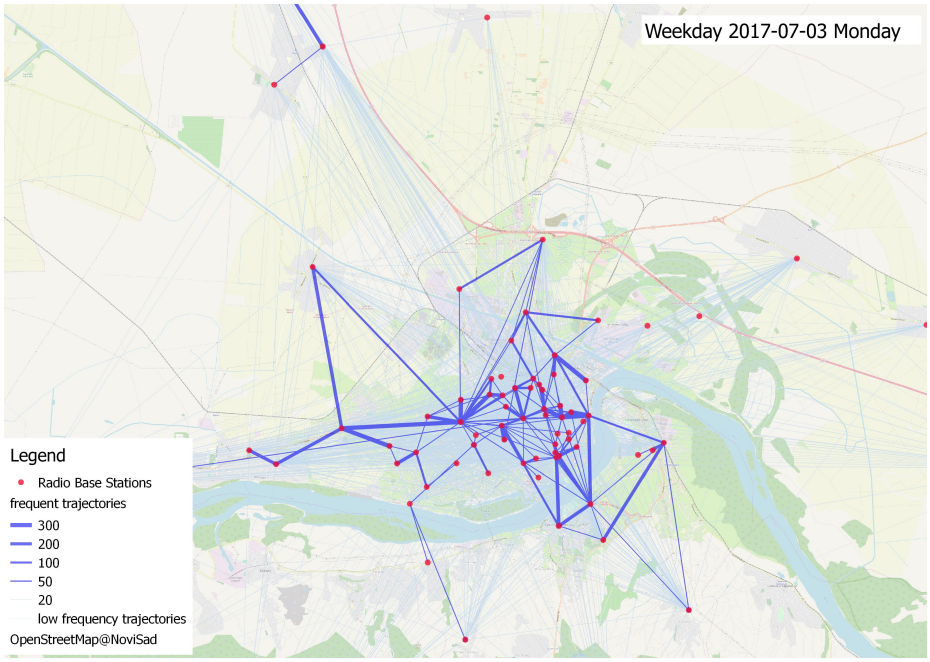

Fig. 6. Mobility across the city of Novi Sad, Serbia

2.3 Mobility

Mobile phone data can reveal the approximate location of a user and its mobility trace based on geographical location of the Radio Base Stations which registered the traffic. In [16] the authors proposed a novel computational framework that enables efficient and extensible discovery of mobility intelligence from large-scale spatial-temporal data such as CDR, GPS and Location Based Services data. In [25] the authors focus on usage of Call Detail Records (CDR) in the context of mobility, transport and transport infrastructure analysis. They analyzed CDR data associated with Radio Base Stations together with Open Street Map road network to estimate users mobility. CDR data can provide generalized view of users mobility, since data is collected only when the telecom traffic happens. To illustrate mobility data set we created Figure 6 that presents a map with mobility traces across the city of Novi Sad on 3rd July 2017, for the time interval between 6am and 12pm extracted from raw CDR data through aggregation of visited locations' sequences of anonymous users. Data originate from Serbian national operator, Telekom Srbija, released under non-disclosure agreement. From mobility traces we can detect few locations in the city that acts as trajectory hubs.

3 Big Data Processing

The typical workflow applied for processing spatio-temporal data, such as mobile phone data used in this case study, contains numerous queries across locations and timestamps of interest, spatial/time aggregations and summarization. Existing solutions are rarely focusing on the execution time, scalability, and throughput that are of high importance for the implementation and near real-time settings. In this section, we present briefly some important concepts and architectural issues related to processing Big Data.

3.1 Big Data Architectures

Over the last decade we have witnessed a tremendous progress and innovation in large-scale data processing systems and the associated data-driven computation. Among many others, these include MapReduce-based computational systems, data streaming technologies, and NoSQL database systems. A major challenge is to build systems that on the one hand could handle large volumes of batch data and on the other hand offer the required scalability, performance and low latency required for integration and real-time processing of massive, continuous data streams. In the following paragraphs, we discuss some of the architectural principles underlying Big Data systems that address this challenge, in particular the Lambda and the Kappa architectural alternatives.

Lambda Architecture Big Data systems often face the challenge of how to integrate processing of "new" data that is being constantly ingested into a system with historical (batch) data. Newly arriving (real-time) data is usually processed using stream-based processing techniques, while historical data is periodically reprocessed using batch processing. The Lambda architecture [40] is a blueprint for a Big Data system that unifies stream processing of real-time data and batch processing of historical data.

The Lambda architecture pursues a generalized approach to developing Big Data systems with the goal of overcoming the complexities and limitations when trying to scale traditional data systems based on incrementally updated relational databases. In an incremental database system, the state of the database (i.e. its contents) is incrementally updated, usually when new data is processed. In contrast to incremental database systems, the Lambda architecture advocates a functional approach relying on immutable data, i.e., new data is added on top of the immutable historical data (batch data) already present in the system.

As opposed to traditional distributed database systems, e.g., where distribution of tables across multiple machines has to be explicitly dealt with by the developer, a key underlying principle of the Lambda architecture is to make the system aware of its distributed nature so that it can automatically manage distribution, replication and related issues. Another key aspect of the Lambda architecture is its reliance on immutable data as opposed to incrementally updated data in relational database systems. Reliance on immutable data is essential for achieving resilience with respect to human errors.

The Lambda architecture promises to tackle many important requirements of Big Data systems, including scalability, robustness and fault tolerance (including fault-tolerance with respect to human errors), support for low-latency reads and updates, extensibility, easier debugging and maintainability. At a high-level of abstraction, the Lambda architecture is comprised of three layers, the batch layer, the serving layer, and the speed layer.

The batch layer stores the raw data (also often referred to as batch data, historical data, or master data set), which is immutable. Whenever new data arrives, it is appended to the existing data in the batch layer. The batch layer is responsible for computing batch views taking into account all available data. The batch layer periodically recomputes the batch views from scratch so that also the new data that has been added to the system since the computation of the last batch views is processed.

The serving layer sits on top of the batch layer and provides read access to the batch views that have been computed by the batch layer. The serving layer usually constitutes a distributed database, which is populated with the computed batch views, and ensures that the batch views can be randomly accessed. The serving layer is constantly updated with new batch views once these become available. Since the serving layer only needs to support batch updates and random reads, but no random writes (updates), it is usually significantly less complex than a database that needs to support random reads and writes. While the serving layer enables fast read-only access to the pre-computed batch views, it must be clear that these views may not be completely up-to-date, since data that has been acquired since the latest batch views have been computed have not been considered.

The speed layer is provided on top of the serving layer in order to support real-time views on the data. The speed layer mitigates the high latency of the batch layer by processing the data on-the-fly, as it arrives in the system, using fast, incremental algorithms to compute real-time views of the data. As opposed to the batch layer, which periodically recomputes the batch views based on all historical data form scratch, the speed layer does not compute real-time views from scratch. To minimize latency, it only performs incremental updates of the real-time views taking into account just the newly arrived data. The real-time views provided by the speed layer are of temporary nature. Once the new data has arrived at the batch layer and has been included in the latest batch views, the corresponding real-time views can be discarded.

Figure 7 depicts the main architectural aspects of the Lambda architecture. Data streamed in from data sources (sensors, Web clients, etc.) is being fed in parallel both into the batch layer and the speed layer, which compute the corresponding batch views and real-time views, respectively.

The lambda architecture can be seen as a trade-off between two conflicting goals: speed and accuracy. While computation of real-time views is being done with very short latencies, computation of batch views is typically a very high-latency process. On the other hand, since the speed layer does not take into account all of the available data, real-time views are usually only approximations,

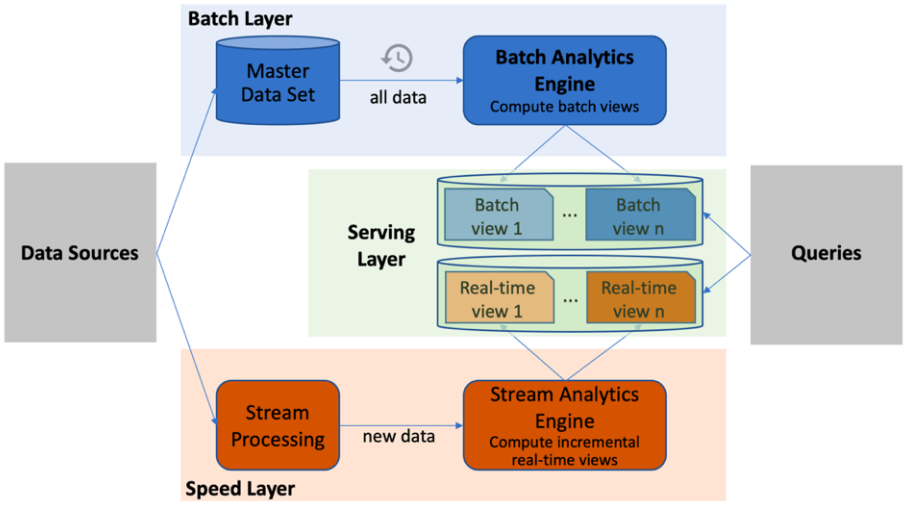

Fig. 7. The Lambda Architecture.

while batch views provide accurate answers considering all data available in the master data store at a certain point in time. In order to get a view of all the available data (batch data and new data) queries have to be resolved such that they combine the corresponding batch-views and real-time views, which can either be done in the serving layer or by the client applications.

The Lambda architecture has been widely recognized as a viable approach to unifying batch and stream processing, by advocating real-time stream processing and batch re-processing on immutable data. There are, however, some potential drawbacks associated with the Lambda architecture. Although a major objective of the lambda architecture is to reduce the complexity as compared to traditional distributed database systems, this goal often cannot be fully realized. While the batch layer usually hides complexity from the developers, typically by relying on some high-level MapReduce framework (e.g., Hadoop), the speed layer may still exhibit significant complexities to the developers of Big Data solutions. In addition, having to develop and maintain two separate data processing components, the stream layer and the batch layer, adds to the overall complexity. Another potential issue with the Lambda architecture is that constantly recomputing the batch views from scratch might become prohibitively expensive in terms of resource usage and latency.

Kappa Architecture A limitation of the Lambda architecture is that two different data processing systems, i.e., the stream layer and the batch layer, have to be maintained. These layers need to perform the same analytics, however realized with different technologies and tools. As a consequence, the system

becomes more complex and debugging and maintenance become more difficult. This drawback is being addressed by the Kappa architecture [31].

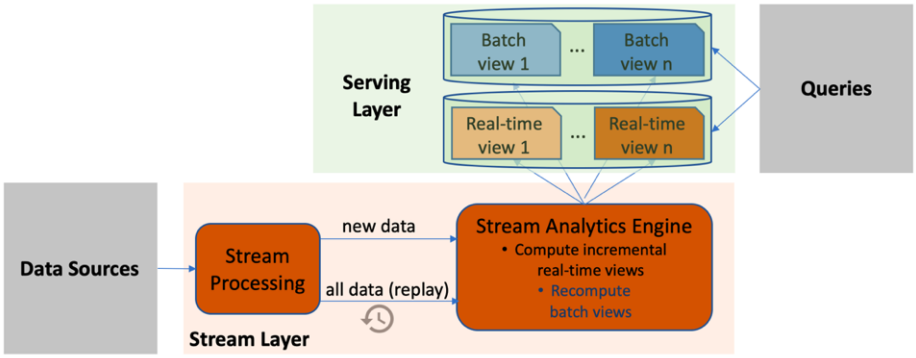

Fig. 8. The Kappa Architecture

The Kappa architecture constitutes a simplification of the Lambda architecture by uniformly treating real-time data and batch data as streams. Consequently, batch processing as done in the lambda architecture, is replaced by stream processing. The Kappa architecture assumes that (historical) batch data can also be viewed as a (bounded) stream, which is often the case. What is required, however, is that the stream processing component also supports efficient replay of historical data as a stream. Only if this is the case, batch views can be recomputed by the same stream analytics engine that is also responsible for processing real-time views. Besides the ability to replay historical data, the order of all data events must be strictly preserved in the system in order to ensure deterministic results.

Instead of a batch layer and a speed layer, the Kappa architecture relies on a single stream layer capable of handling the data volumes for computing both real-time views and batch views. Overall system complexity decreases with the Kappa architecture as illustrated in Figure 8. However, it should be noted that the Kappa architecture is not a replacement of the Lambda architecture, since it will not be suitable for all use cases.

3.2 Big Data Frameworks

There is a plethora of Big Data frameworks and tools that have been developed in the past decade. As a result, both the Lambda architecture and Kappa architecture can be implemented using a variety of different technologies for the different system components. In the following, we briefly discuss a few frameworks that are most typically used to implement Big Data systems based on the Lambda or Kappa architecture.

Hadoop The Apache Hadoop ecosystem is a collection of tools for developing scalable Big Data processing systems [63]. The Hadoop File System (HDFS) is a distributed file system for storing large volumes of data on distributed memory machines (clusters) transparently handling the details of data distribution, replication and fail-over. The Hadoop MapReduce engine utilizes HDFS to support transparent parallelism of large-scale batch processing that can be formulated according to the MapReduce programming model. Hadoop is often used to implement the batch layer in data processing systems that implement the Lambda Architecture.

Spark Apache Spark introduces *Resilient Distributed Data sets* (RDDs) and *Data Frames* (DFs) [65, 66]. Spark can work nicely within the Hadoop ecosystem, although this is not mandatory, since Spark is self-contained with respect to task scheduling and fault tolerance. Moreover, it supports a large collection of data sources, including HDFS. Spark supports iterative MapReduce tasks and improves performance by explicitly enabling caching of distributed data sets. A wide range of functions support categorization of application components into data transformations and actions. In addition, Spark provides stream processing functionality, a rich machine learning library, a powerful library for SQL processing on top of Data Frames and also a library specifically designed for graph processing (GraphX). Spark is often used for implementing the speed layer in a Lambda or the stream layer in a Kappa architecture.

Kafka Apache Kafka [30, 60] is a scalable message queuing and log aggregation platform for real-time data feeds. It provides a distributed message queue and a publish/subscribe messaging model for streams of data records, supporting distributed, fault-tolerant data storage. The framework is run as a so-called *Kafka cluster* on multiple servers that can scale over multiple data centers. Kafka supports efficient replay of data streams and thus it is often used to implement systems that resemble the Kappa architecture.

Samza Apache Samza [42] is a scalable, distributed real-time stream processing platform that has been developed in conjunction with Apache Kafka and that is often used for implementing Big Data systems based on the Kappa architecture. Samza can be integrated easily with the YARN resource management framework.

Resource Management Frameworks YARN is a resource negotiator included with Apache Hadoop. YARN decouples the programming paradigm of MapReduce from its resource management capabilities, and delegates many scheduling functions (e.g., task fault-tolerance) to per-application components. Apache Mesos is a fine-grained resource negotiation engine that supports sharing and management of a large cluster of machines between different computing frameworks, including Hadoop, MPI, Spark, Kafka, etc. The main difference between YARN and Mesos is the resource negotiation model. Whereas YARN

implements a push-based resource negotiation approach, where clients specify their resource requirements and deployment preferences, Mesos uses a pull-based approach, where the negotiator offers resources to clients which they can accept or decline.

4 Data Analysis

Data Analysis is the scientific process of examining data sets in order to discover patterns and draw insights about the information they contain. In the case of data collected by mobile phone providers, typically in the form of CDRs, the analysis focuses in two main directions: i) graph analysis and ii) machine learning. Moreover, the data analysis must incorporate the spatial-temporal characteristics of such data.

4.1 Graph Analytics

Graph mining is a heavily active research direction with numerous applications [2, 15] that uses novel approaches for mining and performing useful analysis on datasets represented by graph structures. Current research directions can be categorized into the following groups [52]: i) Graph clustering used for grouping vertices into clusters; ii) Graph Classification used for classifying separate, individual graphs into two or more categories; iii) Subgraph mining used for producing a set of subgraphs occurring in at least some given threshold of the given input example graphs.

One of the core research directions in the area of graph clustering is the discovery of meaningful communities in a large network [20] from the perspective of spatial-temporal data that evolves over time. In the majority of real-life applications, graphs are extremely sparse usually following power-law degree distribution. However, the original graph may contain groups of vertices, called *communities*, where vertices in the same community are more well-connected than vertices across communities. In the case of CDR data, the graph corresponds to user interactions and communities correspond to groups of people with strong pair-wise activity within the group delimited by spacial-temporal boundaries. To enable efficient community detection in potentially massive amounts of data, the following problems must be tackled [58]: *i*) the algorithmic techniques applied must scale well with respect to the size of the data, which means that the algorithmic complexity should stay below $\mathcal{O}(n^2)$ (where n is the number of graph nodes), and *ii*) since these techniques are unsupervised, the algorithms used must be flexible enough to be able to infer the number of communities during the course of the algorithm. Moreover, the temporal dimension of the data must be taken into account when detecting communities to better understand the natural evolution of user interactions. Some algorithms that qualify for this task are Louvain [8], Infomap [54], Walktrap [50], FastGreedy [14], etc.

The result of community detection analysis is a set of grouped vertices that have very strong inner connectivity. The results could be presented on the map,

since telecom data is georeferenced. In Figure 9 we present geographical map of Milan city with wide suburban area overlayed with the results of community detection analysis in 3D. Communities that have smaller overall area are presented with higher bars. From visual inspection of Figure 9 we can notice that the dense urban area of the city has a larger number of small communities, while in the sparsely populated suburban area there are a few very large communities. High number of communities within small spatial area is reflecting dynamic nature of telecom traffic in urban areas, which is strongly related to people flow and its dynamic across the city.

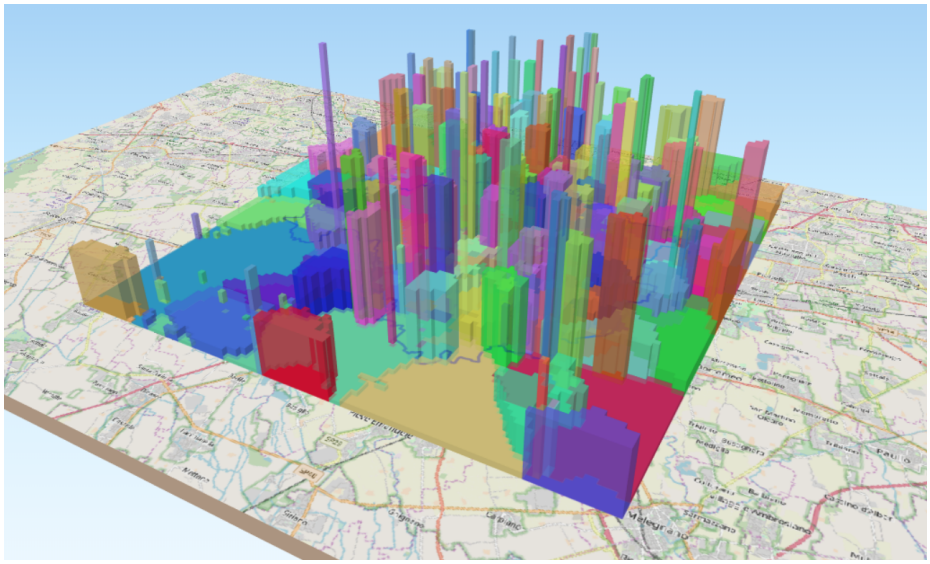

Fig. 9. Communities over the city of Milan in 3D.

Collective classification and label propagation are two important research directions in the area of graph classification for vertex classification. Iterative classification is used for collective classification to capture the similarity among the points where each vertex represents one data point either labeled or unlabelled [55]. Label propagation is a converging iterative algorithm where vertices are assigned labels based on the majority vote on the labels of their neighbors [67]. In the case of CDR data, these algorithms can be used to draw insights about users and their neighborhoods by finding the correlations between the label of a user and i) its observed attributes, ii) the observed attributes (including observed labels) of other users in its neighborhood, iii) the unobserved labels of users in its neighborhood. The spatial-temporal dimension of the data also plays an important role as the correlations will bring new insight into the propagation of labels and the way user neighborhood is built.

Subgraph mining deals with the identification of frequent graphs and subgraphs that can be used for classification tasks, graph clustering and building indices [51]. In the case of CDR data, subgraph mining can help to detect hidden patterns in active user communities delimited into spatial-temporal boundaries by contrasting the support of frequent graphs between various different graph classes or to classify user interaction by considering frequent patterns using the spatial-temporal dimensions as a cardinal feature.

4.2 Machine learning

Spatial-temporal data analysis is an important and evolving domain of machine learning. The main direction when dealing with such data is forecasting and prediction in support of the decision-making process.

Classical machine learning techniques, from simple ones for sequential pattern mining (e.g., Apriori, Generalized Sequential Pattern, FreeSpan, PrefixSpan, SPADE) to more complex ones (e.g., Linear, Multilinear, Logistic, Poisson or Nonlinear Regression), can be used to capture the dependencies between spatial and temporal components and help with making accurate predictions into the future and extract new knowledge about the evolution of users and their interests.

With the increasing evolution and adoption of neural networks, new deep learning architectures are developed for the analysis of spatial-temporal data and used for making and quantifying the uncertainty associated with predictions [56]. These techniques can be employed in the process of making accurate predictions for spatial-temporal data when working in both big data and data scarce regimes managing to quantify the uncertainty associated with predictions in a real-time manner.

5 Data Fusion

Identified patterns from telecom data reach true value when combined with other sources. As illustrated in Figure 10 processed and analyzed telecom data can be fused with diverse data sources in context of various applications. We summarized several fusion scenarios in Table 1. The list is not exhaustive, only highlights diversity of the combinations, and some of the examples might integrate mobile phone data with more than one external source. Satellite data, environmental data, IoT, Points-of-Interests (POI), National statistics and other sources can add to the value of mobile phone data. For example, satellite data can provide information on land cover types and changes and IoT can collect valuable ground truth measurements.

Bringing together heterogeneous datasets with mobile phone data and using them jointly is challenging due to typical mismatch in the resolutions of data, multimodal and dynamic nature of data. Some applications on mobile phone data demand external sources only for training and validation (e.g. learning model to predict socio-economic indicators based on features extracted from

Table 1. Data fusion scenarios - mapping external data sources with telecom data.

External Data Source	Examples
Satellite data	NASAs Tropical Rainfall Measurement Mission (TRMM) satellite ⟷ anomalous patterns of mobility and calling frequency [38] Landsat-7 for deriving impact map of floods ⟷ aggregated activity by day and by antenna [46] SPOT-Vegetation satellite for calculating vegetation index ⟷ average number of calls between all market pairs [27]
Environmental data	The air quality estimated by regional model ⟷ staying at home and travel patterns [17] Availability of environmental freshwater measured as the total length of the rivers in each spatial unit ⟷ estimate of mobility obtained from CDRs [39] Logs of the climatic conditions: temperature, relative humidity, air pressure and wind speed from weather stations ⟷ inferring the social network for each subject [49]
POI	Events on famous POIs across city ⟷ users presences in the area [21] POIs from Google Earth for land use inference ⟷ aggregated number of calls managed by each of base transceiver station towers [48] Pokémon POIs ⟷ city-level aggregated distributions of number of connected devices and downloaded information from xDR records [24]
IoT	Inductive loop vehicle detectors ⟷ mobility, rush hours trafic [28]
Census, Surveys	Travel surveys ⟷ daily commuting from mobility traces patterns [3] Census on journey to work ⟷ activity and connectivity around laborshed area [5] Demographic and health surveys ⟷ connectivity and mobility across country [11] National statistics on socio-economic development ⟷ human mobility patterns [45] Household income and expenditure survey ⟷ top up credit amounts, mobility and social network features [57]
Infrastructure	The street network (highways and primary streets) from OpenStreetMap, metro network, bus routes ⟷ xDR data aggregated into origin-destination (OD) matrices [23] Customer sites of each power line per grid square and line measurement indicating the amount of flowing energy ⟷ aggregated people dynamics features from the mobile phone network activity [9]

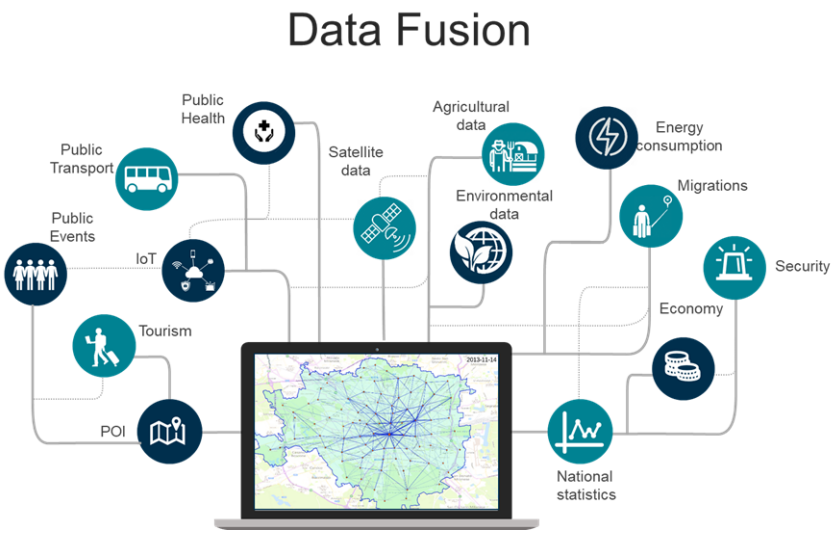

Fig. 10. Fusion of mobile phone data with other sources.

telecom data). Here special attention is needed to understand the bias and avoid spurious correlations. Other scenarios demand continuous information flow from external source and dynamic integration (e.g. air quality measurements fused with aggregated mobility from telecom data). The main challenge here is the timely processing of external data and proper alignment with mobile phone data.

Fusion scenarios reported in Table 1 illustrate heterogeneity of external data sources, all having an important role in unlocking the value of mobile phone data coming from telecom operators. The quality of final application depends on the availability of external sources, efficiency of data processing and the quality of delivered information and its integration.

6 Applications

A plethora of research work has been published related to the usage of telecom data for a multitude of purposes. Telecom data contains rich user behaviour information, and it can reveal mobility patterns, activity related to specific locations, peak hours or unusual events. Extracting frequent trajectories, home and work location detection, origin destination matrices are further examples of knowledge that may be mined from rich telecom data. Telecom operators have a great interest to analyze collected data for optimizing their services. For example, time-dependent pricing schemes can maximize operators profit, as well as users grade of service. Dynamic data pricing frameworks combining both spatial

and temporal traffic patterns [18] allow estimating optimal pricing rewards given the current network capacity.

Telecom data significantly enriched many different fields and boosted external social good applications. Studies in transportation, urban and energy planning, public health, economy and tourism have benefited most from this valuable new resource that surpasses all alternative sources in population coverage, spatial and temporal resolution.

Transportation planning applications need information on different modes of trips, purposes, and times of day. With telecom data transportation models can effectively utilize mobility footprints at large scale and resolution. This was validated by an MIT study [3] on the Boston metropolitan area where the authors demonstrated how CDR data can be used to represent distinct mobility patterns. In another example, origin destination matrices inferred from mobile phone data helped IBM to redesign the bus routes [6] in the largest city of Ivory Coast - Abidjan.

Mobility patterns derived from telecom data could be very valuable for public health applications, in particular epidemiology. Surveillance, prioritization and prevention are key efforts in epidemiology. Mobile phone data demonstrated utility for dengue [62], HIV [11, 22], malaria [61], schistosomiasis [39], Ebola epidemic [47], and cholera outbreaks [19]. Another suitable public health application is concerned with air quality where recent studies embraced telecom data to better quantify individual and population level expose to air pollution. In [17] the authors highlighted the need to dynamically assess exposure to NO_2 that has high impact on peoples health. Their method incorporated individual travel patterns.

Urban studies highly explored the potential of mobile phone data and discovered that it can be used for urban planning [5], detecting social function of land use [48], in particular residential and office areas as well as leisure-commerce and rush hour patterns [53], and extracting relevant information about the structure of the cities [37]. Recent applications propose an analytical process able to discover, understand and characterize city events from CDR data [21] and a method to predict the population at a large spatio-temporal scale in a city [13]. All urban studies fit into the wider context of smart city applications and therefore more breakthroughs on the usage of mobile phone data are expected.

With the growing role of tourism there is increased interest to investigate utility of mobile phone data to understand tourists experiences, evaluate marketing strategies and estimate revenues generated by touristic events. Mobility and behaviour patterns have been recently used to derive trust and reputation models and scalable data analytics for the tourism industry [59, 33]. The Andorra case study has proposed indicators in high spatial and temporal resolutions such as tourist flows per country of origin, flows of new tourists, revisiting patterns, profiling of tourist interests to uncover valuable patterns for tourism [34]. Special attention is given to large scale events that attract foreign people [12]. Arguably, tourists via their mobile devices have quickly become data sources for crowdsourced aggregation with dynamic spatial and temporal resolutions [32].

Other high impact applications include social and economical development [45, 57], disaster events management such as cyclones landfall [38] or earthquakes [64], and food security [68, 27].

Although many studies demonstrated utility of mobile phone data in various applications, reaching the operational level is still not that close. If we recall the summary of workflow's steps provided in Figure 1, all further described in the previous sections, we can realize that technologies used in each step need to match with specific application.

7 Summary and Vision

This chapter provided an overview of all steps in discovering knowledge from raw telecom data in the context of different applications. Knowledge about how people move across a city, where they are gathering, what are home, work and leisure locations along with corresponding time component are valuable for many applications. The biggest challenges in this process are privacy and regulation, real-time settings and data fusion with external sources.

Efforts directed toward providing access to telecom large-scale human behavioral data in a privacy-preserving manner [41] are necessary. Real-time settings raise critical issues concerning computational infrastructure, big data frameworks and analytics. There is a lack of research and benchmark studies that evaluate different computational architectures and big data frameworks. Only a few studies tackled issues of parallelization and distributed processing. In [16] authors proposed mobility intelligence framework based on Apache Spark for processing and analytics of large scale mobile phone data. Another example is the study [58] that provided computational pipeline for the community detection in mobile phone data, developed in Apache Hive and Spark technology, and benchmarked different architectures and settings. More of these studies are needed to choose the right architecture and processing frameworks. Graph analytics together with machine learning have become indispensable tools for telecom data analytics, but the streaming nature of data demands for change detection and online adaption. External data sources mentioned in the data fusion section are also advancing (e.g., new satellites launched, enhanced IoT ecosystems) and will help us to understand spatio-temporal context better.

Future research must address all critical aspects to reach technology readiness for operational environment. This will enable applications based on mobile phone data to have high impact on decision making in urban, transport, public health and other domains and will certainly open opportunities for new applications.

References

1. Acs, G., Castelluccia, C.: A case study: Privacy preserving release of spatio-temporal density in Paris. In: 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining–KDD '14. pp. 1679–1688. ACM, New York (2014). https://doi.org/10.1145/2623330.2623361

2. Aggarwal, C.C., Wang, H.: Managing and Mining Graph Data. Springer (2010). https://doi.org/10.1007/978-1-4419-6045-0
3. Alexander, L., Jiang, S., Murga, M., González, M.C.: Origin–destination trips by purpose and time of day inferred from mobile phone data. Transportation Research Part C: Emerging Technologies **58**, 240–250 (2015). https://doi.org/10.1016/j.trc.2015.02.018
4. Barlacchi, G., De Nadai, M., Larcher, R., Casella, A., Chitic, C., Torrisi, G., Antonelli, F., Vespignani, A., Pentland, A., Lepri, B.: A multi-source dataset of urban life in the city of milan and the province of trentino. Scientific data **2**, 150055 (2015)
5. Becker, R.A., Caceres, R., Hanson, K., Loh, J.M., Urbanek, S., Varshavsky, A., Volinsky, C.: A tale of one city: Using cellular network data for urban planning. IEEE Pervasive Computing **10**(4), 18–26 (2011). https://doi.org/10.1109/MPRV.2011.44
6. Berlingerio, M., Calabrese, F., Di Lorenzo, G., Nair, R., Pinelli, F., Sbodio, M.L.: Allaboard: a system for exploring urban mobility and optimizing public transport using cellphone data. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 663–666. Springer (2013)
7. Blondel, V.D., Decuyper, A., Krings, G.: A survey of results on mobile phone datasets analysis. EPJ Data Science **4**(1), 10 (2015). https://doi.org/10.1140/epjds/s13688-015-0046-0
8. Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008**(10), P10008 (2008). https://doi.org/10.1088/1742-5468/2008/10/P10008
9. Bogomolov, A., Lepri, B., Larcher, R., Antonelli, F., Pianesi, F., Pentland, A.: Energy consumption prediction using people dynamics derived from cellular network data. EPJ Data Science **5**(1), 13 (2016). https://doi.org/10.1140/epjds/s13688-016-0075-3
10. Bosetti, P., Poletti, P., Stella, M., Lepri, B., Merler, S., De Domenico, M.: Reducing measles risk in turkey through social integration of syrian refugees. arXiv preprint arXiv:1901.04214 (2019)
11. Brdar, S., Gavrić, K., Ćulibrk, D., Crnojević, V.: Unveiling spatial epidemiology of hiv with mobile phone data. Scientific reports **6** (2016). https://doi.org/10.1038/srep19342
12. Callegari, C., Garroppo, R.G., Giordano, S.: Inferring social information on foreign people from mobile traffic data. In: Communications (ICC), 2017 IEEE International Conference on. pp. 1–6. IEEE (2017)
13. Chen, J., Pei, T., Shaw, S.L., Lu, F., Li, M., Cheng, S., Liu, X., Zhang, H.: Fine-grained prediction of urban population using mobile phone location data. International Journal of Geographical Information Science pp. 1–17 (2018)
14. Clauset, A., Newman, M.E., Moore, C.: Finding community structure in very large networks. Physical review E **70**(6), 066111 (2004)
15. Cook, D.J., Holder, L.B.: Mining Graph Data. John Wiley & Sons (2006). https://doi.org/10.1002/0470073047
16. Dang, T.A., Deepak, J., Wang, J., Luo, S., Jin, Y., Ng, Y., Lim, A., Li, Y.: Mobility genome-a framework for mobility intelligence from large-scale spatio-temporal data. In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA). pp. 449–458. IEEE (2017)
17. Dewulf, B., Neutens, T., Lefebvre, W., Seynaeve, G., Vanpoucke, C., Beckx, C., Van de Weghe, N.: Dynamic assessment of exposure to air pollution using mobile phone data. International journal of health geographics **15**(1), 14 (2016)

18. Ding, J., Li, Y., Zhang, P., Jin, D.: Time dependent pricing for large-scale mobile networks of urban environment: Feasibility and adaptability. IEEE Transactions on Services Computing (2017)
19. Finger, F., Genolet, T., Mari, L., de Magny, G.C., Manga, N.M., Rinaldo, A., Bertuzzo, E.: Mobile phone data highlights the role of mass gatherings in the spreading of cholera outbreaks. Proceedings of the National Academy of Sciences **113**(23), 6421–6426 (2016)
20. Fortunato, S.: Community detection in graphs. Physics Reports **483**(3), 75–174 (2010). https://doi.org/10.1016/j.physrep.2009.11.002
21. Furletti, B., Trasarti, R., Cintia, P., Gabrielli, L.: Discovering and understanding city events with big data: The case of rome. Information **8**(3), 74 (2017)
22. Gavric, K., Brdar, S., Culibrk, D., Crnojevic, V.: Linking the human mobility and connectivity patterns with spatial hiv distribution. NetMob D4D Challenge pp. 1–6 (2013)
23. Graells-Garrido, E., Caro, D., Parra, D.: Inferring modes of transportation using mobile phone data. EPJ Data Science **7**(1), 49 (2018)
24. Graells-Garrido, E., Ferres, L., Caro, D., Bravo, L.: The effect of pokémon go on the pulse of the city: a natural experiment. EPJ Data Science **6**(1), 23 (2017)
25. Gundlegård, D., Rydergren, C., Breyer, N., Rajna, B.: Travel demand estimation and network assignment based on cellular network data. Computer Communications **95**, 29–42 (2016)
26. He, Y., Yu, F.R., Zhao, N., Yin, H., Yao, H., Qiu, R.C.: Big data analytics in mobile cellular networks. IEEE access **4**, 1985–1996 (2016)
27. Jacques, D.C., Marinho, E., d'Andrimont, R., Waldner, F., Radoux, J., Gaspart, F., Defourny, P.: Social capital and transaction costs in millet markets. Heliyon **4**(1), e00505 (2018)
28. Järv, O., Ahas, R., Saluveer, E., Derudder, B., Witlox, F.: Mobile phones in a traffic flow: a geographical perspective to evening rush hour traffic analysis using call detail records. PloS one **7**(11), 1–12 (2012). https://doi.org/10.1371/journal.pone.0049171
29. Jiang, S., Fiore, G.A., Yang, Y., Ferreira Jr, J., Frazzoli, E., González, M.C.: A review of urban computing for mobile phone traces: current methods, challenges and opportunities. In: Proceedings of the 2nd ACM SIGKDD international workshop on Urban Computing. p. 2. ACM (2013)
30. Kreps, J.: Kafka : a distributed messaging system for log processing. In: Proceedings of teh 6th International Workshop on Networking Meets Databases (NetDB) (2011)
31. Kreps, J.: Questioning the Lambda architecture. Online article (Jul 2014), https://www.oreilly.com/ideas/questioning-the-lambda-architecture, (Last accessed: 15/Dec/2018)
32. Leal, F., Malheiro, B., González-Vélez, H., Burguillo, J.C.: Trust-based modelling of multi-criteria crowdsourced data. Data Science and Engineering **2**(3), 199–209 (Sep 2017). https://doi.org/10.1007/s41019-017-0045-1
33. Leal, F., Veloso, B.M., Malheiro, B., Gonzlez-Vlez, H., Burguillo, J.C.: Scalable modelling and recommendation using wiki-based crowdsourced repositories. Electronic Commerce Research and Applications **33**, 100817 (2019). https://doi.org/https://doi.org/10.1016/j.elerap.2018.11.004
34. Leng, Y., Noriega, A., Pentland, A., Winder, I., Lutz, N., Alonso, L.: Analysis of tourism dynamics and special events through mobile phone metadata. arXiv preprint arXiv:1610.08342 (2016)

35. Lepri, B., Oliver, N., Letouzé, E., Pentland, A., Vinck, P.: Fair, transparent, and accountable algorithmic decision-making processes. Philosophy & Technology pp. 1–17 (2017)
36. Lima, A., De Domenico, M., Pejovic, V., Musolesi, M.: Disease containment strategies based on mobility and information dissemination. Scientific reports **5** (2015). https://doi.org/10.1038/srep10650
37. Louail, T., Lenormand, M., Ros, O.G.C., Picornell, M., Herranz, R., Frias-Martinez, E., Ramasco, J.J., Barthelemy, M.: From mobile phone data to the spatial structure of cities. Scientific reports **4**, 5276 (2014)
38. Lu, X., et al.: Detecting climate adaptation with mobile network data in Bangladesh: anomalies in communication, mobility and consumption patterns during cyclone Mahasen. Climatic Change **138**(3-4), 505–519 (2016)
39. Mari, L., Gatto, M., Ciddio, M., Dia, E.D., Sokolow, S.H., De Leo, G.A., Casagrandi, R.: Big-data-driven modeling unveils country-wide drivers of endemic schistosomiasis. Scientific Reports **7**(1), 489 (2017)
40. Marz, N., Warren, J.: Big Data Principles and best practices of scalable realtime data systems. Manning (2006)
41. de Montjoye, Y.A., et al.: On the privacy-conscientious use of mobile phone data. Scientific Data **5**, 180286 EP– (Dec 2018). https://doi.org/doi: 10.1038/sdata.2018.286
42. Noghabi, S.A., Paramasivam, K., Pan, Y., Ramesh, N., Bringhurst, J., Gupta, I., Campbell, R.H.: Samza: Stateful scalable stream processing at linkedin. Proc. VLDB Endow. **10**(12), 1634–1645 (2017)
43. Novović, O., Brdar, S., Crnojević, V.: Evolving connectivity graphs in mobile phone data. In: NetMob, The main conference on the scientific analysis of mobile phone datasets. pp. 73–75. Vodafone (2015)
44. Oliver, N., Matic, A., Frias-Martinez, E.: Mobile network data for public health: opportunities and challenges. Frontiers in public health **3**, 189 (2015)
45. Pappalardo, L., Pedreschi, D., Smoreda, Z., Giannotti, F.: Using big data to study the link between human mobility and socio-economic development. In: Big Data (Big Data), 2015 IEEE International Conference on. pp. 871–878 (2015). https://doi.org/10.1109/BigData.2015.7363835
46. Pastor-Escuredo, D., et al.: Flooding through the lens of mobile phone activity. In: Global Humanitarian Technology Conference (GHTC), 2014 IEEE. pp. 279–286. IEEE (oct 2014). https://doi.org/10.1109/GHTC.2014.6970293
47. Peak, C.M., Wesolowski, A., zu Erbach-Schoenberg, E., Tatem, A.J., Wetter, E., Lu, X., Power, D., Weidman-Grunewald, E., Ramos, S., Moritz, S., et al.: Population mobility reductions associated with travel restrictions during the ebola epidemic in sierra leone: use of mobile phone data. International journal of epidemiology **47**(5), 1562–1570 (2018)
48. Pei, T., Sobolevsky, S., Ratti, C., Shaw, S.L., Li, T., Zhou, C.: A new insight into land use classification based on aggregated mobile phone data. International Journal of Geographical Information Science **28**(9), 1988–2007 (2014)
49. Phithakkitnukoon, S., Leong, T.W., Smoreda, Z., Olivier, P.: Weather effects on mobile social interactions: a case study of mobile phone users in lisbon, portugal. PloS one **7**(10), e45745 (2012)
50. Pons, P., Latapy, M.: Computing communities in large networks using random walks. Journal of Graph Algorithms and Applications **10**(2), 191–218 (2006)
51. Ramraj, T., Prabhakar, R.: Frequent subgraph mining algorithms – a survey. Procedia Computer Science **47**, 197–204 (2015). https://doi.org/10.1016/j.procs.2015.03.198

52. Rehman, S.U., Khan, A.U., Fong, S.: Graph mining: A survey of graph mining techniques. In: International Conference on Digital Information Management (ICDIM 2012). pp. 88–92 (2012). https://doi.org/10.1109/ICDIM.2012.6360146
53. Ríos, S.A., Muñoz, R.: Land use detection with cell phone data using topic models: Case santiago, chile. Computers, Environment and Urban Systems **61**, 39–48 (2017)
54. Rosvall, M., Bergstrom, C.T.: Maps of random walks on complex networks reveal community structure. Proceedings of the National Academy of Sciences **105**(4), 1118–1123 (2008). https://doi.org/10.1073/pnas.0706851105
55. Sen, P., Namata, G., Bilgic, M., Getoor, L., Galligher, B., Eliassi-Rad, T.: Collective classification in network data. AI magazine **29**(3), 93 (2008)
56. Senanayake, R., Jean, N., Ramos, F., Chowdhary, G.: Modeling and decision-making in the spatiotemporal domain. In: Conference on Neural Information Processing Systems (2018)
57. Steele, J.E., et al.: Mapping poverty using mobile phone and satellite data. Journal of The Royal Society Interface **14**(127) (2017). https://doi.org/10.1098/rsif.2016.0690
58. Truică, C.O., Novović, O., Brdar, S., Papadopoulos, A.N.: Community detection in who-calls-whom social networks. In: International Conference on Big Data Analytics and Knowledge Discovery. pp. 19–33. Springer (2018)
59. Veloso, B., Leal, F., González-Vélez, H., Malheiro, B., Burguillo, J.: Scalable data analytics using crowdsourced repositories and streams. J. Parallel Distrib. Comput. **122**, 1–10 (2018). https://doi.org/10.1016/j.jpdc.2018.06.013
60. Wang, G., Koshy, J., Subramanian, S., Paramasivam, K., Zadeh, M., Narkhede, N., Rao, J., Kreps, J., Stein, J.: Building a replicated logging system with apache kafka. Proc. VLDB Endow. **8**(12), 1654–1655 (Aug 2015). https://doi.org/10.14778/2824032.2824063, http://dx.doi.org/10.14778/2824032.2824063
61. Wesolowski, A., Eagle, N., Tatem, A.J., Smith, D.L., Noor, A.M., Snow, R.W., Buckee, C.O.: Quantifying the impact of human mobility on malaria. Science **338**(6104), 267–270 (2012)
62. Wesolowski, A., et al.: Impact of human mobility on the emergence of dengue epidemics in pakistan. Proceedings of the National Academy of Sciences **112**(38), 11887–11892 (2015). https://doi.org/10.1073/pnas.1504964112
63. White, T.: Hadoop: The Definitive Guide. O'Reilly, 4 edn. (2015)
64. Wilson, R., et al.: Rapid and near real-time assessments of population displacement using mobile phone data following disasters: the 2015 Nepal Earthquake. PLoS Currents **8** (2016). https://doi.org/10.1371/currents.dis.d073fbece328e4c39087bc086d694b5c
65. Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M.J., Shenker, S., Stoica, I.: Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In: Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation. pp. 2–2. NSDI'12, USENIX Association, Berkeley, CA, USA (2012), http://dl.acm.org/citation.cfm?id=2228298.2228301
66. Zaharia, M., Xin, R.S., Wendell, P., Das, T., Armbrust, M., Dave, A., Meng, X., Rosen, J., Venkataraman, S., Franklin, M.J., Ghodsi, A., Gonzalez, J., Shenker, S., Stoica, I.: Apache spark: A unified engine for big data processing. Commun. ACM **59**(11), 56–65 (Oct 2016). https://doi.org/10.1145/2934664, http://doi.acm.org/10.1145/2934664

67. Zhu, X., Ghahramani, Z.: Learning from labeled and unlabeled data with label propagation. Tech. Rep. CMU-CALD-02-107, Carnegie Mellon University (Jun 2002)
68. Zufiria, P.J., Pastor-Escuredo, D., Úbeda-Medina, L., Hernandez-Medina, M.A., Barriales-Valbuena, I., Morales, A.J., Jacques, D.C., Nkwambi, W., Diop, M.B., Quinn, J., et al.: Identifying seasonal mobility profiles from anonymized and aggregated mobile phone data. application in food security. PloS one **13**(4), e0195714 (2018)