Assembling and Using a Cellular Dataset for Mobile Network Analysis and Planning

Assembling and Using a Cellular Dataset for Mobile Network Analysis and Planning

Paolo Di Francesco, Francesco Malandrino, Luiz A. DaSilva

Abstract—In a world of open data and large-scale measurements, it is often feasible to obtain a real-world trace to fit to one's research problem. Feasible, however, does not imply simple. Taking next-generation cellular network planning as a case study, in this paper we describe a large-scale dataset, combining topology, traffic demand from call detail records, and demographic information throughout a whole country. We investigate how these aspects interact, revealing effects that are normally not captured by smaller-scale or synthetic datasets. In addition to making the resulting dataset available for download, we discuss how our experience can be generalized to other scenarios and case studies, i.e., how everyone can construct a similar dataset from publicly available information.

1 Introduction and motivation

Until recently, much of the research in the field of mobile wireless networks has relied on synthetic models for user mobility, network topology, and data demand. As an example, when modeling cellular networks, they were routinely assumed [1] to be made of hexagonal, regularly-spaced cells, each having the same number of users, with each user requesting the same amount of data. These simplified models are adequate to *some* topics of interest in cellular networking research, e.g., propagation and scheduling: if our goal is to divide the available spectrum resources among the existing macro- and micro-base stations, it is not so important to accurately know the location of their users.

Many emergent research topics, however, have different requirements. A prominent example is *mobile edge computing* (MEC), a paradigm that advocates moving computation and storage from the Internet (e.g., the cloud) to the mobile network itself. MEC research typically revolves around the *planning* and architecture of next-generation networks, i.e., which nodes they should include and what those nodes should do. As an example, we might have to decide whether to place virtual machines running firewalls at the base stations, at the network core nodes, or at a combination of the two; additionally, we have the option to move virtual machines between nodes over time and to disable some. It is evident that answering such questions critically depends on our knowledge of the actual network demand and deployment, and that synthetic models such as the ones used in [1] would not be of much help.

Besides realism, *scale* – in both space and time – is a key requirement for many research topics on next-generation cellular networks; in the example above, we will probably follow different virtual machine placement strategies in dense urban areas and in rural zones. This rules out many existing network traces (e.g., [2]), containing very detailed propagation and location information, but limited to some hours' traffic in a small area.

In this paper, we present – and make available for download [3], along with all the scripts we use to generate it – a real-world, large-scale dataset, collected at the time and spatial scales appropriate for research on the planning and architecture of next-generation networks. This dataset was obtained by combining demographic information, deployments of two Irish mobile networks operators (MNOs) providing cellular services throughout the entire Republic of Ireland, and their estimated data demand.

There are three main ways our effort is relevant to the *big data* community at large. First, to obtain our dataset we combine *heterogeneous information*, including demographic statistics, network deployments, and call-detail records. Furthermore, much of this information comes from *open data* efforts: in particular, cellular deployments have been made public as a reaction to concerns about "electromagnetic pollution". Finally, as a consequence of the above, we had to perform a careful *integration* work between data obtained from different sources.

Our effort generalizes easily. Indeed, as a contribution of our paper, we discuss and explain how our methodology can be applied to other topologies and to the analysis and planning of wireless networks in other locations. Researchers interested in assembling a dataset similar to ours for, say, Great Britain or Poland, can do so by using readily available information and following the very same steps we present here.

The remainder of this paper is organized as as follows. We begin by reviewing existing works using either cellular network datasets or traces in Sec. 2, highlighting how they are not entirely adequate to study network planning. We present the raw information we use in Sec. 3; in Sec. 4, we discuss how we combine such data to obtain our dataset. Sec. 5 discusses how our dataset can be used, either directly or by replicating our methodology, whose performance is explored in Sec. 6. Finally, Sec. 7 concludes the paper.

2 A brief survey of real data-based studies of cellular networks

Cellular networks cover large geographical areas and serve millions of users. In order to effectively study them, it is

All authors are with CONNECT/Trinity College, Dublin, Ireland.

of paramount importance to understand the correlation between three elements: the infrastructure, the traffic demand, and the population said demand comes from. Due to the difficulty in obtaining and combining this information, the research community seldom considers more than a couple of these aspects at the same time.

For example, Michalopoulou et al. in [4] study the interaction between cellular deployment and population density using tools from spatial statistics and spatial point processes. They take Germany as a case study, and charactarize infrastructure deployment as a spatial point process, depending on local population density. Demographic data describing population densities have always been used by operators, especially during the roll-out phase of their networks. Indeed, prior works such as [5] show that demographic data can be used to roughly estimate the spatial traffic demand assuming that higher population densities correspond to higher traffic demand that has to be served by the network operators.

While real deployment data and population data are relatively easy to find, real traffic demand datasets are more difficult to obtain. They belong to mobile operators, which typically do not release them to the public. Researchers, therefore, often resort to analytical traffic models, which have the advantage of mathematical tractability, but may lack realism. For example Boiardi et al. in [6] study the cost and energy savings of cellular networks. They rely on *coverage points*, placed on a regular grid, and *traffic points*, placed randomly and associated to a uniformly-distributed demand. The resulting scenario is not guaranteed to resemble the real-world distribution of population density or cellular infrastructure.

Occasionally, mobile operators disclose demand information to individual research groups. Among the works resulting from these cooperations, many, e.g., [7]–[10], focus on studying and characterizing user behavior. Willkomm et al. [7] were among the first to use cellular traffic traces. They characterised the usage of cellular voice services using information from a cellular operator in the US. Specifically, they studied the call arrival process and proposed a random walk model capturing the aggregate load dynamics. In [8], Keralapura et al. analyzed the browsing behavior of mobile users in an American 3G data network, by monitoring 24 hours of IP traffic. Paul et al. in [9] looked at individual subscriber behavior and traffic patterns, studying a nation-wide 3G network at the base station level. Differently, Shafiq et al. investigated application popularity and clustering [10] and device utilization [11] in a cellular network, obtaining useful insights that can be leveraged to fine tune network parameter settings such as inactivity timers of radio resource control (RRC) and the QoS profile settings and the radio network controller (RNC) admission control procedure. Other works, such as [12], [13], focus on network dynamics: in particular, Peng et al. in [13] analyze deployment and demand data to show that dynamically switching off base stations at off-peak times can deliver substantial energy savings in both urban and rural areas.

Recently, some mobile operators have made demand and deployment traces available to researchers worldwide under the form of a *challenge*: researchers submit an idea, and they receive the data to evaluate it. Prominent challenge examples include the Telecom Italia Big Data Challenge [14] and the Orange D4D challenge [15]. While the data provided by operators represent a substantial improvement over synthetic models, both datasets have some drawbacks: specifically, [15] focuses on voice and SMS, rather than data, traffic, and [14] only includes information for the city of Milan, but no suburban or rural area. Furthermore, both [15] and [14] are perforce limited to the demand experienced by the operator releasing them.

With respect to all the above works, our dataset and the processing methodology we present have the unique advantage of accounting for the distribution and characteristics of population, the demand they generate *and* the infrastructure existing to serve it. Even more important, we are able to study how two different, competing mobile operators serve the same locations at the same time – and to which extent the resulting deployments tend to resemble each other. As discussed earlier, having all this information combined in a large-scale dataset enables a dataset-driven study of cellular network planning.

3 Raw Data

In this section, we present the raw data that can be used to produce a real-world, large-scale cellular networking dataset, consisting of, (i) census information (ii) cellular infrastructure deployment; (iii) cellular data demand. We refer to the Irish datasets we use to produce our dataset; however, as detailed in Sec. 4.4, the same data is available for many countries throughout the world.

3.1 Census information

The Irish Central Statistics Office periodically releases a set of demographic and socio-economic data. They are publicly available[1] and consist of a *shapefile*, dividing the surface of the Republic of Ireland into polygons, and a *database* file, containing for each polygon such information as:

- population, number and size of households;
- job category, income distribution;
- age, ethnicity, language distribution;
- classification of the area as urban, suburban, or rural.[2]

While interesting in their own right, these data become precious when correlated with network topology and demand. As an example, we could study whether a higher data demand is associated to young people (eager consumers of multimedia content, one would expect) or to wealthy areas, owing to a higher penetration of costly, high-end, high-resolution devices.

In our case, as discussed in Sec. 4, we use the population and urban/rural area classification information to study how infrastructure deployment and per-user demand change across urban and rural areas.

1. http://www.cso.ie/en/census/census2011boundaryfiles/

2. In some cases this information is explicit. In some other cases it can be inferred by looking at the density of the population per square kilometer as it is done in [16].

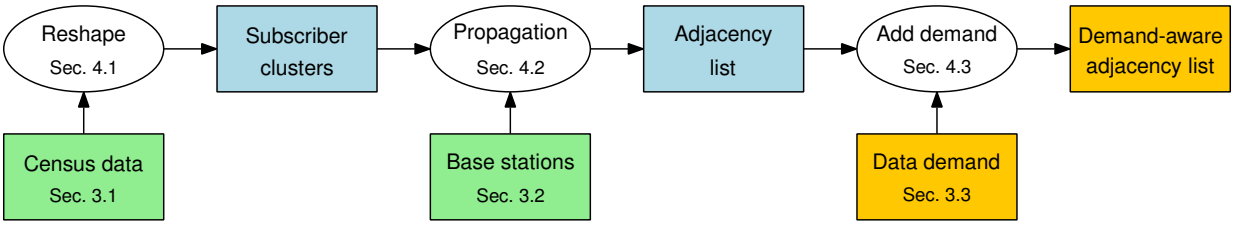


Fig. 1. From raw data to our dataset. Boxes represent traces or datasets; ovals correspond to models and algorithms. We begin from census data, processing them as described in Sec. 4.1 to obtain a list of subscriber clusters. Combining them with the location of base stations we obtain an adjacency list, as detailed in Sec. 4.2. Finally, we enhance the adjacency list by adding demand information, as shown in Sec. 4.3. Green boxes correspond to publicly available information; blue ones to information we offer for download; orange ones to information we cannot directly disclose.

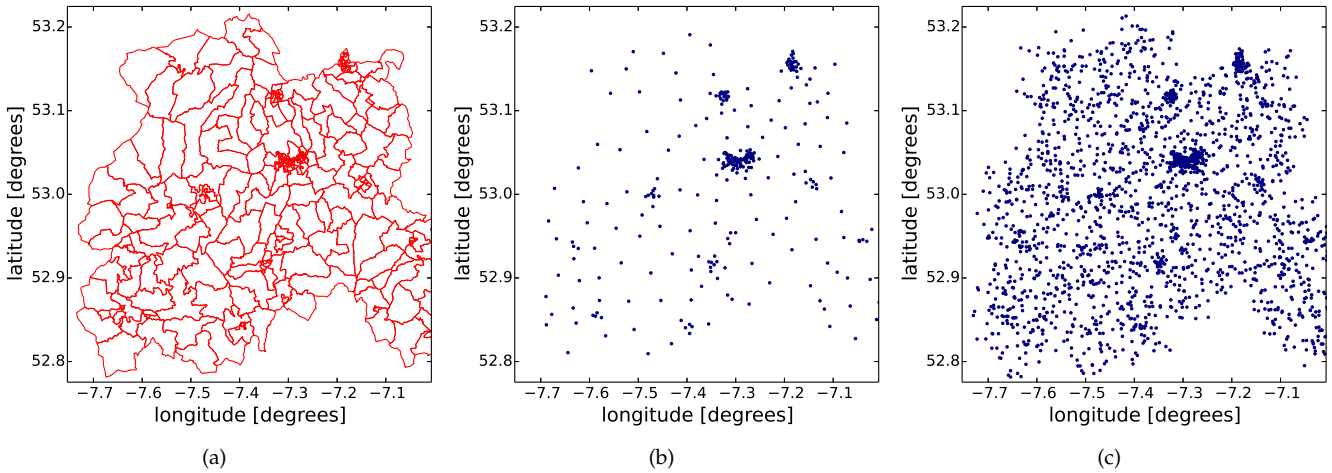


Fig. 2. County Laois, in central Ireland: census polygons (a), their barycentres (b), and the subscriber clusters we create (c). In the case of Irish census data, polygons are correlated with population, i.e., sparsely populated areas tend to be associated to bigger polygons. By having a maximum area limit in place, we are able to accurately study the coverage in both urban and rural areas.

3.2 Cellular infrastructure deployment

Our deployment data originally came from two Irish mobile operators; however, similar information is now available online from the national telecommunications regulator[3].

Our dataset includes then approximately 23,500 transmitters, distributed across around 5,000 base stations. For each of them, we know:

- coordinates and azimuth;
- technology – second generation (2G) or third generation (3G);
- mobile operator operator;
- power class and (approximate) coverage area.

As explained in Sec. 4, we use the position, technology and power class information to reconstruct the network coverage and data rates that each network can offer.

3.3 Cellular data demand

In addition to the deployment information described above, the two operators provided us with call-detail record (CDR) information about voice calls and data sessions throughout a period of two weeks in late 2013. More exactly, for each voice call and data session we know:

- time and transmitter at which it started;
- duration and amount of transferred data (for data sessions);
- identifier (TAC code) of the user device.

Regrettably, demand is also sensitive information for operators, and we are not allowed to include it in the database we share. We do, however, include the distribution of per-user demand in urban and rural areas, as detailed in Sec. 4. It can be used, as explained in Sec. 5, to generate demand samples for our own topology matching the fundamental properties of the actual demand, correlating it with demographic information if available.

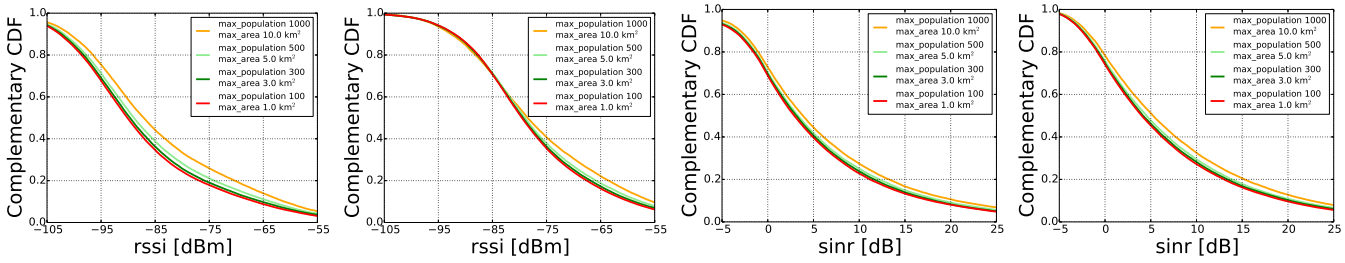
4 Creating our Dataset

As mentioned above, planning a network essentially means making sure its *infrastructure* is able to serve the *demand* of its *users*. We need to combine the raw data described in Sec. 3 into a flexible and easy to manage description of these three elements, and we obtain it through the three steps summarized in Fig. 1.

In spirit, our methodology is similar to the one adopted in the vehicular networking community: vehicular traces [17] are not the direct result of a real-world measurement campaign, but rather come from the combination of real-world topologies, high-level traffic flow statistics, and low-level mobility models.

We begin from users, abstracting their location through what we call *subscriber clusters*, as described in Sec. 4.1.

3. http://www.askcomreg.ie/mobile/siteviewer.273.LE.asp



(a) (b) (c) (d)

Fig. 3. MNO_1. Complementary CDF (CCDF) of the received signal strength indicator (RSSI) (a),(b) and signal-to-noise ratio SINR (c),(d) across different subscriber clusters, for different limits of population and area for 3G (a),(c) 2G (b),(d).

Next, we turn our attention to the infrastructure, and in Sec. 4.2 we assess the throughput that can be achieved between each element thereof (e.g., each base station) and each subscriber cluster. Finally, we assign to each subscriber cluster its demand, using operator-provided information as detailed in Sec. 4.3.

Our final dataset looks line an *enhanced adjacency list*, whose entries carry additional information besides connectivity. Specifically, for each base station and each subscriber cluster *that can communicate with each other*, we know:

- the position of both, and therefore the distance between them;
- the demand of the subscriber cluster;
- the capacity with which the base station can serve it.

4.1 From users to subscriber clusters

The format in which the census information described in Sec. 3.1 comes poses two main challenges. To begin with, its resolution is too coarse. Furthermore, polygons are complex and computationally intensive to manipulate.

To cope with these issues, we group users into *subscriber clusters*. Each subscriber cluster has a position in space, and represents a set of users that can be seen as co-located. More specifically, as shown in Fig. 2:

- we decide the maximum number of users and the maximum area each cluster can represent;
- for each polygon, we compute the number of clusters to place therein;
- we place the clusters randomly within the area of the polygon.

This solution has two main advantages. First, the number of clusters, the number of users and the area they represent are fully customizable and do not depend on the number and shape of the original polygons. Furthermore, the position of subscriber clusters is but a point in space: computing aspects such as coverage, attenuation and throughput is simple and computationally lightweight, as noted also in [5]. It is worth stressing that the placement of demand clusters is *not* distributed according to a Poisson point process. Indeed, the location and shape of the tiles is deterministic, and given by the census data we leverage. Additionally, the number of demand clusters we place in each tile is also deterministic, as explained earlier. The only random decision is where to locate the demand clusters within the tile, for which no further information is available.

On the other hand, we have to decide the right population and area limits for our subscriber clusters. As we see in Fig. 2, lower limits mean more subscriber clusters, in both densely and sparsely populated areas. Both are important; indeed, evaluating a network planning strategy often means checking that it is able to serve all the demand from urban areas, without creating coverage problems in rural ones. However, too many subscriber clusters mean more complex simulations and longer computation times – when shall we stop?

The intuitive answer is obvious – we should stop adding subscriber clusters when decreasing the area and population they represent no longer influences the distribution of the signal quality. Let us look at Fig. 3, depicting the distribution of the signal quality experienced by subscribers for MNO_1 with different technologies[4]. We start from very high limits, i.e., a situation where we place relatively few subscriber clusters. Decreasing the limits, i.e., placing more subscriber clusters, skews the distribution at first; however, after a certain level is reached, no more changes are observed. The level corresponding, in the Irish case, to a population limit of 300 people and an area limit of 3 square kilometers is arguably a good compromise between accuracy and computational complexity.

4.2 Propagation and throughput information

Thanks to the *reshaping* procedure described in Sec. 4.1, we now have the position of each subscriber cluster. We also know the position of each base station, as well as the additional information described in Sec. 3.2. Furthermore, we know which subscriber clusters correspond to urban areas and which do not. Therefore, we are now in the position to compute the throughput that each subscriber cluster can obtain from each base station. This is done in three steps:

1) compute the *attenuation*, i.e., how much the wireless signals weaken as they travel from their source to their destination;
2) compute the *signal-to-interference-and-noise ratio* (SINR) values, expressing how well each destination is able to distinguish a source from the others;
3) compute the *throughput*, i.e., how much data can be transferred between each source and each destination in a time unit.

4. Results for MNO_2 were essentially the same and are omitted for brevity.

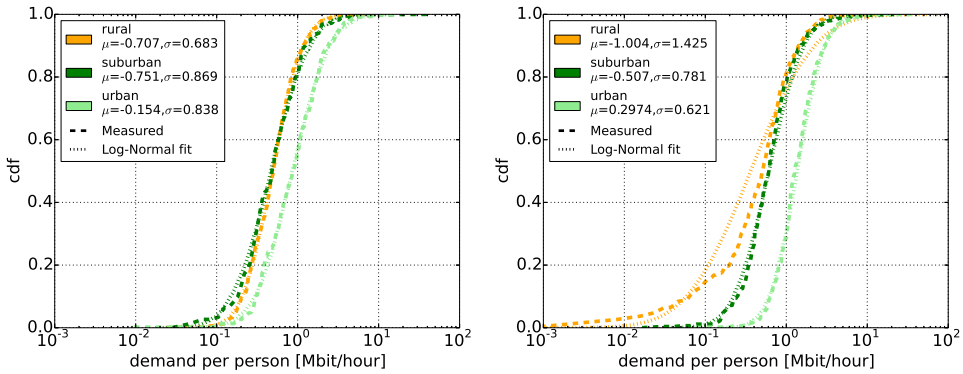


Fig. 4. Empirical CDF of per-person 3G data demand across subscriber clusters, for urban, suburban, and rural areas for two Irish operators ((a) MNO_1, (b) MNO_2)) alongside log-normal distribution fitting using the estimated parameters reported in the legend, i.e., location (μ) and scale (σ).

Attenuation: Computing the attenuation can be done using one of the many models existing in literature. In our case, we opt for the COST-231 Hata model [18]. Notice how this propagation model exploits the information we have about the power and frequency of base stations, properly rendering the heterogeneous nature of modern cellular networks. Our deployment infrastructure does not include height information; we assume the standard value of 12 meters.

It is worth stressing that our choice of propagation model and the parameters thereof is easily reversible. Indeed, our dataset includes the distance between base stations and subscriber clusters; therefore, a researcher who desires to use a different propagation model, e.g., the two-ray ground model, can simply do so, as detailed in Sec. 5.

RSSI and SINR: Our next step is to compute the received signal strength indicator (RSSI) by each subscriber cluster from each base station. This is simply the product of the transmission power of the base station and the attenuation between it and the subscriber cluster, computed as explained earlier.

The power at which base stations transmit is not always available in cellular datasets. If needed, we can simply fall back to the standard values of 47 dBm for macro-base stations and 30 dBm for micro-base stations [19].

The ratio between the power received from the desired transmitter and the power received from everyone else (plus thermal noise) is the signal-to-interference and noise ratio (SINR). We compute this value assuming the most challenging possible conditions: all base stations always have data to transmit, and they are all allowed to transmit on all frequencies (i.e., the *reuse factor* is one). The reason for this very conservative assumption lies in the nature of our problem, i.e., network planning. A properly planned network has to operate in all conditions, even when facing an exceptionally high load – the infamous "flash crowds". Daily fluctuations in the load, the fact that load peaks are unlikely to happen at the same time in different parts of the topology and similar aspects can, and indeed should, be accounted for whilst operating a network, but cannot be relied upon when planning it.

As with the propagation model discussed above, this choice can be reversed by the users of our dataset: it includes RSSI values, so it is trivial, as shown in Sec. 5, to compute the SINR under any alternative assumption if needed.

Throughput: Attenuation and RSSI can be computed, at the cost of some reasonable assumptions such as the ones we made above. Throughput, i.e., the amount of data a pair of network nodes can successfully transfer in a time unit, is either simulated or estimated. Simulation is the traditional approach: from the SINR we reconstruct the bit- or packet-error rate, and then establish whether the transmission of each packet succeeds or fails.

Owing to the scale and focus of cellular network planning, however, we adopt the other approach, and outright *estimate* the throughput from the SINR level. For example, in LTE case, we can rely on the model adopted by OFCOM, based on the Shannon bound [19, Sec. A14.90]. The throughput is 4.4 bits/Hz/s in optimal conditions, and reduces as the SINR decreases. OFCOM themselves point out that their expression is a lower-bound for cellular performance, and actual deployments may exceed it [19, Sec. A14.95]. As discussed earlier, adopting these conservative values suits our purpose and the objectives of cellular network planning.

4.3 Adding demand information

Our goal is to turn the demand information we described in Sec. 3.3 into a "demand" figure we can attach to each subscriber cluster. We proceed in four steps, as detailed next.

(i) As traditional voice is expected to represent a decreasing percentage of the traffic that future networks will face, as a first step we restrict ourselves to the demand for mobile data.

(ii) The second step is aggregating the traffic over time, as it is common practice when working with large-scale traces. For each base station, we compute the data and voice load for each one-hour period, e.g., from 7PM to 8PM of November 14th, 2013.

(iii) The third step has to do with the nature of our problem: as discussed earlier, network planning is essentially about conservative assumptions and peak load. Therefore, for each base station, we retain the load in *its own* busiest hour, even if such hours are not the same for all base stations.

(iv) Fourth and last, we need to move from a *load* associated to base stations, to a *demand* associated to subscriber clusters. We do so by:

1) ensuring that the global load we have in our raw data corresponds to the global demand of the subscriber clusters of our dataset;
2) associating each subscriber cluster to the base station that provides the highest RSSI;
3) if a base station covers multiple subscriber clusters, its load is split proportionally to the population thereof.

4.4 Creating a new dataset

Our operator data give us the *offered traffic* at each base station, at each point in time. Our key observation is that present-day and, arguably, future cellular networks will serve *all* offered traffic, by all users. In other words, each demand cluster must get at least the bitrate necessary to serve the traffic it offers during its highest-load hour, and this is what we call the *demand* the cellular network has to meet.

Now, we are able to associate to each subscriber cluster a worst-case data demand. The CDFs of *per-person* demand are summarized in Fig. 4. It is interesting to observe that people in urban areas seem to request more data than in suburban ones. There are several possible causes for this effect, from a different penetration of high-end mobile devices to the simple availability of more capacity in densely populated areas, which encourages more traffic requests; effects like this have a major impact on network planning – and are seldom captured by smaller traces and synthetic models.

As mentioned earlier, we cannot disclose the demand figures the operators shared with us, nor can we include demand information in the dataset we make available for download. However, we do include the demand characteristics presented in Fig. 4 for the two operators, i.e. the CDFs of the demand spatial distribution. The distributions show diversity depending on the operator and the area considered. However, most of the resulting distributions can be well approximated by the log-normal distribution whose parameters are obtained by parametric fitting with maximum likelihood estimates as shown in Fig. 4. This information, along with population figures from census data, can be used to reconstruct the demand at subscriber clusters level, as shown in Sec. 5 next.

Doing so does not entirely eliminate the effects of our inability to share the actual demand figures, but it does substantially mitigate them: our datasets will not yield quantitatively exact answers to research questions; however, *qualitative* features thereof are nonetheless of great interest. As an example, consider the planning and architecture of next-generation networks, mentioned in Sec. 1: through our datasets, researches will be able to devise how different levels of centralization impact network performance, or to which extent joint network planning by mobile operators can help saving on costs [12].

Furthermore, in most cases researchers are interested in testing a certain algorithm or design under a variety of *different* load conditions, all exhibiting the same features as the original one. To this end, recent works such as [20], [21] use real-world data to fit synthetic models for demand and/or deployment – performing out of choice the same step we perform out of necessity.

Finally, Fig. 4 also highlights how both the distribution parameters and the goodness-of-fit change across mobile operators and area types (e.g., urban or suburban); in particular, the fit in Fig. 4(b) is quite poor for rural areas – mostly because MNO_2 focuses most of its coverage on cities, and has few rural base stations. We are able to counter these effects by providing not only the fitted model parameters, but the complete distribution of the real-world of data demand.

5 Using our dataset

There are three ways to use our dataset: downloading it and using it as it is; customizing it to suit one's needs; or creating an entirely new dataset following our methodology.

5.1 Downloading our dataset

Our dataset is available at our GitLab repository [3], along with the code we use to generate it. It contains the files mentioned in Fig. 1 at different resolutions:
(i) a list of *subscriber clusters*, with their position, population, area, the Irish county they are in, and whether they represent an urban, suburban, or rural area;
(ii) a list of *base stations*, with their position, radio access technology (RAT), and power class;
(iii) an *adjacency list* containing, for each base station and subscriber cluster, the distance, attenuation, RSSI, and SINR computed as described in Sec. 4.2.

All datasets are in CSV format and come as compressed archives. Moreover, we are also releasing also all the Python code we have used, as well as the Irish demographic files (both in shapefile and CSV format) we based our study upon. We also provide the distribution of the per-user demand in urban, suburban, and rural areas, i.e., the information shown in Fig. 4. Finally, we include a `README` file, with a detailed explanation of the format and content of each file.

5.2 Adapting our database

As discussed in Sec. 4.2, our choices of propagation model and SINR-to-throughput mapping are not the only possible ones. In order to make enacting alternative choices as easy as possible, our adjacency list contains all intermediate data. As an example, users wishing to adopt different mapping between SINR and throughput – e.g., because they are studying a different type of RAT, from HSDPA to 5G – can use the SINR values present in the list; users needing a different propagation model can start from the `distance` values.

Needless to say, changes propagate: users changing the propagation model need to recompute the RSSI, SINR and throughput values. Also notice that the format of the adjacency list is such that all operations can be performed in a *vectorized* fashion in such environments as R and MATLAB.

5.3 Adding the demand

The data set we made available also contains the demand CDFs, i.e., the data shown in Fig. 4. Figures are expressed in megabits, are per-user, and refer to the base station's

TABLE 1
Demographic data and deployment data available to the public.

Country	Demographic data	Deployment data
England/ Wales	Office for National Statistics http://bit.ly/19HkNus	National regulator http://bit.ly/1xnPxzL
Poland	Geospatial portal http://bit.ly/1qEncEg	Office of Electronic Communications http://bit.ly/16eLVpM
Italy	National Statistic Institute http://bit.ly/1fHjFJv	Regional Environment Agency http://bit.ly/1x9PtoO

busiest hour, as explained in Sec. 4.2. The demand of each subscriber cluster can be reconstructed as follows:

1) select the appropriate scenario (i.e. urban, suburban, rural);
2) extract a realization thereof, e.g., through acceptance-rejection sampling, or by approximating the spatial distribution of the traffic with a log-normal distribution using the parameters specified in Fig. 4;
3) multiply it by the population of the subscriber cluster.

Doing so implies the assumption that demand samples are independent, which is seldom the case. However, such a simplification is sometimes unavoidable, and is also adopted in papers proposing synthetic models [22], [23]. Exploiting the correlation with socio-demographic information can further enhance the realism of the demand profiles we obtain.

Needless to say, our adjacency list can be used with an altogether different demand model, e.g., one with location-specific contents.

Information such as the one presented in Sec. 3 is increasingly easy to find. It is therefore possible to use the methodology we discussed in Sec. 4 to create an entirely new dataset, as discussed next.

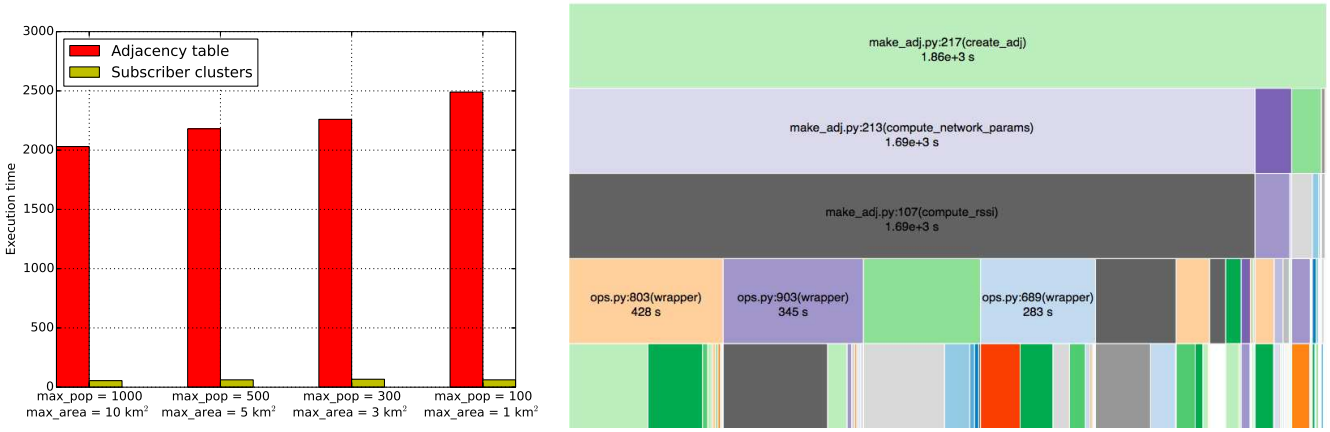
National and local statistical institutes periodically release socio-economic, geographic, and demographic data. Such data are publicly available and easly accessible online. They typically come in the form of *shapefiles* (i.e., polygons in which the territory is divided) and their companion *databases*, containing polygon-specific information. Shapefiles can be easily processed with both open-source and commercial GIS softwares, and represent the input to create the *subscriber clusters*, as explained in Sec. 4.1.

Operator deployment data can be obtained in two ways: directly from operators themselves, or through national agencies – telecommunication regulators or health authorities. They are used, along with the propagation model, to generate the adjacency list detailed in Sec. 4.2.

As an example, Tab. 1 summarizes where data similar to the one we used for our datasets can be found for England and Wales, Poland, and Italy. Demographic data come from national statistical institutes, while base station information are available through the national telecommunication regulators (in the case of Poland, and England and Wales) or the regional health department (in the Italian case).

6 Performance and Portability

The operations described in Sec. 4 and Sec. 5 are fairly complex; therefore, one could be legitimately concerned about the time and resources needed to perform them. In this section, we discuss the running time of our trace generation procedure, the main contributions thereto, as well as licensing and portability issues. All our tests are performed on commodity hardware; specifically, we used a MacBook Pro with a 2.7 GHz processor and 16 GByte of RAM, running the OSX Yosemite operating system.



(a) (b)

Fig. 5. Total time needed to create the weighed adjacency list and the subscriber clusters, for different values of the maximum population and area represented by each cluster (a); time breakdown when those limits are 1000 people and 10 km^2 respectively (b). Results have been obtained on a MacBook Pro equipped with a 2.7 GHz Core i5 processor and running OSX Yosemite.

Fig. 5(a) presents the total running time for the two main steps of our procedure, i.e., generating the subscriber clusters and computing the adjacency list. We can observe that a smaller population limit tends to increase running times – intuitively, that is because we have more elements in our adjacency list. More importantly, running times are consistently below one hour, even using the smallest possible limits: with a system configuration that would be commonplace (indeed, a little dated) on any server, applying our methodology on a nation-wide, real-world trace takes less than forty-five minutes.

Fig. 5(a) also highlights that generating the adjacency list is the most time-consuming of our steps. We thus dig a little deeper, *profile* our application and study where, i.e., in which function, the most time is spent. The results are generated with `cProfile`, summarized via `snakeviz`, and shown in Fig. 5(b). We can observe that computing the network parameters is the lengthiest task and, within that function, computing the RSSI values is the longest task. Indeed, the COST Hata model [18] we employ is very precise, but it requires computing several powers and logarithms – as we can observe from the `ops.py` calls, that are wrappers around numerical methods provided by the `pandas` and `numpy` libraries. Fig. 5(b) also represents a hint to future adopters of our methodology: if computation times are paramount, it could be worth adopting a simpler propagation model.

As far as licensing costs and portability issues are concerned, we can confidently claim that our methodology is free of both. All software we use – the Python language and its `numpy`, `pandas` and `matplotlib` libraries – is free and open source; indeed, our own software is available on GitLab [3] and shared under the MIT license. We chose the MIT license over the GPL one because it allows unrestricted usage, both commercial and non-commercial, of our software. As for portability, our software and all its dependencies run unmodified on Windows, Linux and OSX, including many less-than-recent versions thereof.

7 Conclusion

While obtaining real-world data has recently became more feasible, using such data carefully and effectively did not become any easier. Motivated by the lack of real-world traces suitable to study the *planning* (as opposed to performance) of cellular networks, we presented a large-scale, real-world dataset, including population, traffic demand and infrastructure deployment information.

As explained in Sec. 3, this information comes from different sources. It is combined, as detailed in Sec. 4, to obtain an *enhanced adjacency list*, containing information about the attenuation, SINR and attainable throughput between base stations and users.

The dataset is available for download and can be, of course, used as it is. Furthermore, as explained in Sec. 5, the information therein can be combined under different assumptions about attenuation, transmission power, and interference. Finally, our methodology can be replicated to produce a dataset similar to ours for other topologies and use cases, e.g., other nations, using publicly-available data.

Acknowledgement

This work is supported by the Science Foundation Ireland under Grant No. 10/IN.1/I3007.

References

[1] F. Malandrino, Z. Limani, C. Casetti, and C.-F. Chiasserini, "Interference-Aware Downlink and Uplink Resource Allocation in HetNets with D2D Support," *IEEE Trans. on Wireless Comm.*, 2015.
[2] F. Calabrese, M. Colonna, P. Lovisolo, D. Parata, and C. Ratti, "Real-Time Urban Monitoring Using Cell Phones: A Case Study in Rome," *IEEE Trans. on Intelligent Transportation Systems*, 2011.
[3] P. Di Francesco, F. Malandrino, and L. DaSilva. GitLab repository. https://gitlab.com/pdifranc/tbd.
[4] M. Michalopoulou, J. Riihijärvi, and P. Mähönen, "Studying the Relationships between Spatial Structures of Wireless Networks and Population Densities," in *IEEE GLOBECOM*, 2010.
[5] K. Tutschku and P. Tran-Gia, "Spatial Traffic Estimation and Characterization for Mobile Communication Network Design," *IEEE JSAC*, 1998.
[6] S. Boiardi, A. Capone, and B. Sanso, "Radio Planning of Energy-Efficient Cellular Netowrks," in *IEEE ICCCN*, 2012.
[7] D. Willkomm, S. Machiraju, J. Bolot, and A. Wolisz, "Primary Users in Cellular Networks: A Large-Scale Measurement Study," in *IEEE DySPAN*, 2008.
[8] R. Keralapura, A. Nucci, Z. L. Zhang, and L. Gao, "Profiling users in a 3G network using hourglass co-clustering," in *ACM MobiCom*, 2011.
[9] U. Paul, A. P. Subramanian, M. M. Buddhikot, and S. R. Das, "Understanding Traffic Dynamics in Cellular Data Networks," in *IEEE INFOCOM*, 2011.
[10] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang, "Characterizing Geospatial Dynamics of Application Usage in a 3G Cellular Data Netowrk," in *IEEE INFOCOM*, 2012.
[11] M. Z. Shafiq, L. Ji, A. X. Liu, and J. Wang, "Characterizing and Modeling Internet Traffic Dynamics of Cellular Devices," in *ACM SIGMETRICS*, 2011.
[12] P. Di Francesco, F. Malandrino, and L. A. DaSilva, "Mobile Network Sharing Between Operators: A Demand Trace-Driven Study," in *ACM SIGCOMM CSWS Workshop*, 2014.
[13] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, "Traffic-driven power saving in operational 3G cellular networks," in *ACM MobiCom*, 2011.
[14] Telecom Italia. Big Data Challenge. http://www.telecomitalia.com/tit/en/bigdatachallenge.html.
[15] Orange. D4D Challenge. http://www.d4d.orange.com/en/Accueil.
[16] F. Pozzi and C. Small, "Analysis of urban land cover and population density in the United States," *Photogrammetric Engineering & Remote Sensing*, vol. 71, no. 6, 2005.
[17] D. Naboulsi and M. Fiore, "On the Instantaneous Topology of a Large-scale Urban Vehicular Network: The Cologne Case," in *ACM MobiHoc*, 2013.
[18] M. Hata, "Empirical formula for propagation loss in land mobile radio services," *IEEE Trans. on Vehicular Technology*, 1980.
[19] OFCOM. LTE Technical Modelling Revised Methodology. https://www.ofcom.org.uk/__data/assets/pdf_file/0012/42312/annex14.pdf.
[20] L. Chiaraviglio, F. Cuomo, M. Maisto, A. Gigli, J. Lorincz, Y. Zhou, Z. Zhao, C. Qi, and H. Zhang, "What is the Best Spatial Distribution to Model Base Station Density? A Deep Dive into Two European Mobile Networks," *IEEE Access*, 2016.
[21] J. Ding, X. Liu, Y. Li, D. Wu, D. Jin, and S. Chen, "Measurement-driven Capability Modeling for Mobile Network in Large-scale Urban Environment," in *IEEE MASS*, 2016.
[22] D. Lee, S. Zhou, X. Zhong, Z. Niu, X. Zhou, and H. Zhang, "Spatial Modeling of the Traffic Density in Cellular Networks," *IEEE Wireless Communications*, 2014.
[23] M. Michalopoulou, J. Riihijärvi, and P. Mähönen, "Towards Characterizing Primary Usage in Cellular Networks: A Traffic-bases Study," in *IEEE DySPAN*, 2011.

Paolo Di Francesco Paolo Di Francesco is currently a Software Developer at RadarServices Smart-IT Security GmbH. Previously he was a post-doc at CTVR/CONNECT Trinity College Dublin, Ireland, where he obtained his PhD in 2015. He received the B.S., M.S. degree in telecommunications engineering from University of Bologna in 2008 and 2011 respectively. His research interests include resource sharing, network optimization, Software Defined Radios, data-analysis, and network security.

Francesco Malandrino Francesco Malandrino earned his Ph.D. in 2012 from Politecnico di Torino, Italy, where he is currently a post-doc. Before his current appointment, he held short-term positions at Trinity College, Dublin, and at the Hebrew University of Jerusalem as a Fibonacci Fellow. His interests focus on wireless and vehicular networks and infrastructure management.

Luiz A. DaSilva Luiz A. DaSilva holds the chair of Telecommunications at Trinity College Dublin, where he is a co-principal investigator of CONNECT, a telecommunications centre funded by the Science Foundation Ireland. Prior to joining Trinity College, Prof DaSilva was with the Bradley Department of Electrical and Computer Engineering at Virginia Tech. His research focuses on distributed and adaptive resource management in wireless networks, and in particular radio resource sharing and the application of game theory to wireless networks. Prof DaSilva is a Fellow of Trinity College Dublin, an IEEE Communications Society Distinguished Lecturer and a Fellow of the IEEE, for contributions to cognitive networks and to resource management in wireless networks.