Chapter 1
Introduction to the Data for Refugees Challenge on Mobility of Syrian Refugees in Turkey

Albert Ali Salah, Alex Pentland, Bruno Lepri, Emmanuel Letouzé, Yves-Alexandre de Montjoye, Xiaowen Dong, Özge Dağdelen and Patrick Vinck

Abstract The Data for Refugees (D4R) Challenge was a nonprofit challenge initiated to improve the conditions of the Syrian refugees in Turkey by providing a special database to scientific community for enabling research on urgent problems concerning refugees, including health, education, unemployment, safety, and social integration. The collected database was based on anonymized mobile Call Detail Record (CDR) of phone calls and SMS messages of Türk Telekom customers. It in-

A. A. Salah (✉)
Utrecht University, Utrecht, The Netherlands and Boğaziçi University, Istanbul, Turkey
e-mail: a.a.salah@uu.nl

A. Pentland
Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: sandy@media.mit.edu

B. Lepri
Fondazione Bruno Kessler, Trento, Italy
e-mail: lepri@fbk.eu

E. Letouzé
Massachusetts Institute of Technology, Data-Pop Alliance, Cambridge, MA, USA
e-mail: eletouze@datapopalliance.org

Y.-A. de Montjoye
Imperial College London, London, UK
e-mail: demontjoye@imperial.ac.uk

X. Dong
University of Oxford, Oxford, UK
e-mail: xdong@robots.ox.ac.uk

Ö. Dağdelen
Türk Telekom, Istanbul, Turkey
e-mail: ozge.dagdelen@turktelekom.com.tr

P. Vinck
Harvard University, Cambridge, MA, USA
e-mail: pvinck@hsph.harvard.edu

dicated broad activity and mobility patterns of refugees and citizens in Turkey for 1 year. The data collection period was from January 1, 2017 to December 31, 2017. The project was initiated by Türk Telekom, in partnership with the Turkish Academic and Research Council (TÜBİTAK) and Boğaziçi University, and in collaboration with several academic and nongovernmental organizations, including UNHCR Turkey, UNICEF, and International Organization for Migration. This chapter describes the Challenge in detail, providing a history of its evolution, as well as a description of the data shared with the participants of the Challenge.

1.1 Introduction

After the Syrian Civil War started in 2011–2012, civilians in increasing numbers sought refuge in neighboring countries. By May 2017, Turkey had received over 3 million refugees—the largest refugee population in the world. Initially, a significant proportion of refugees (30% at one time) lived in government-run camps near the Syrian border. Some of these camps were later shut down, and the Directorate General of Migration Management announced the number of refugees living in these camps as 177,376, as of October 10, 2018. The official number of refugees living in camps was 228,251 at the end of December 2017.[1] Many have moved to cities looking for work and better living conditions. They face problems of integration, income, welfare, employment, health, education, language, social tension, and discrimination.

The Data for Refugees (D4R) Challenge[2] was a nonprofit project to ultimately improve the conditions of the Syrian refugees in Turkey by providing a special database to the scientific community for enabling research on some urgent problems. The challenge datasets opened to the community were based on anonymized mobile Call Detail Records (CDRs) of phone calls and SMS messages of Türk Telekom customers. These data indicate communication activities and mobility patterns in Turkey for 1 year. The D4R Challenge, called the Challenge hereafter, invited research groups across the globe to submit proposals, which had been carefully evaluated. Data were then opened to selected groups on strictly regulated terms. The five focus themes of the Challenge were health, education, unemployment, safety, and social integration, respectively. The project was initiated by Türk Telekom, in partnership with the Turkish Academic and Research Council (TÜBİTAK) and Boğaziçi University and in collaboration with several academic and nongovernmental or intra-governmental organizations, including UNHCR Turkey, UNICEF, and International Organization for Migration.

A Scientific Committee of international experts guided the execution of the project. A Project Evaluation Committee (PEC)[3] was formed with representatives from academia, government (i.e., ministries related to the Challenge), and nonprofit

---

[1] https://multeciler.org.tr, accessed 13.11.2018.

[2] http://d4r.turktelekom.com.tr.

[3] http://d4r.turktelekom.com.tr/presentation/project-evaluation-committee.

organizations working in this area. This committee represented refugee interests, and its job was to ensure that the submitted research projects that are granted access to the data have clear goals, with foreseen benefits to the refugee population in Turkey and elsewhere. Access to the D4R data required PEC approval.

The general aims of the D4R Challenge were to

1. Contribute to the welfare of the refugee populations;
2. Gain insights into key issues, including safety and security, health, education, unemployment, social integration and segregation, mobility, and distribution of resources and infrastructure;
3. Help governments and international bodies model the dynamics of the refugee populations and to discover vulnerabilities (socioeconomic vulnerabilities, gaps in education and services, etc.);
4. Seed further projects, co-created with refugees, resulting in new applications, services, and innovative solutions for refugees in Turkey and elsewhere.

The lack of data on refugee mobility is a very important hurdle to the proper functioning of government services and international aid bodies. Innovative approaches attempted to overcome this included the use of satellite imagery to obtain information from the regions in crisis, with limited success [11]. The D4R project allowed, for the first time, the analysis of a large-scale mobile CDR database on refugees. The usefulness of such data has already been illustrated by the D4D Challenges [5, 14] organized previously, and the numerous projects completed on these challenges provided insights into data science for social good. Furthermore, the Challenge, by involving research groups from all around the globe, aimed to raise awareness for the refugee issues on a grand scale.

The possibilities that mobile CDR data afford for analysis of a broad set of problems are surveyed in [6]. Examples of projects conducted with similar data include analysis of disaster resilience [21], infrastructure planning [13], quantifying mobility effects on the spread of infectious diseases [4, 12], developing agent-based models for disease migrations [22], disease containment [10], analysis of community structures and sociodemographic indicators [23], detection of emergency events [9], poverty analysis [15], and mobility during holidays and religious festivals [17], to name a few.

The D4R Challenge had a distinct focus around refugee problems, and aimed to enable evaluation of refugee-related interventions and activities, including, but not limited to, educational activities, social gatherings, NGO actions, government infrastructure investments, etc. It also had the potential to provide insights into the analysis of residential segregation, population structures for specific geographical locations, and factors on social integration [19].

The rest of the present chapter is devoted to a short history of the Challenge (Sect. 1.2), to serve as an example to people who may consider running similar challenges, the description of the D4R data (Sect. 1.3), and the ethical and privacy guidelines that were made available to the participants at the onset of the Challenge (Sect. 1.4). This chapter extends the data description paper shared with the participants at the initiation of the Challenge [16].

1.2 History of the D4R Challenge

In this section, we summarize the chronological progression of the challenge its committees, and the evaluation criteria.

1.2.1 Time Line of the Challenge

The D4R Challenge was conceptually conceived in a meeting of researchers from TÜBİTAK BILGEM and Boğaziçi University on January 6, 2017. The purpose of the meeting was to develop joint project ideas for dealing with many aspects of the refugee crisis. Boğaziçi University had previously participated in the Data for Development (D4D) Challenges, and a mobile data challenge was proposed as a way of making data available to many research groups at the same time.

Following the initial idea, Türk Telekom was contacted and the possibilities of such a challenge, as well as the main approach proposed for addressing its ethical and privacy issues, were presented. A report written by the ethics committee of the Data for Development Challenge in Senegal served as a template for structuring the privacy measures [1]. At the same time, the Information and Communication Technologies Authority ICTA (in Turkish, BTK—Bilgi Teknolojileri ve İletişim Kurumu) was contacted to ensure that such a challenge could be initiated. Türk Telekom embraced the idea of the challenge, and asked for more detailed information on its accomplishment.

The Scientific Committee of the Challenge was formed between April and May 2017. A preliminary data description was prepared by this committee and presented to the legal team of Türk Telekom. It was important to ensure that no terms of the customer agreement were violated by the Challenge, and no personal information was shared in any way. In June 2017, a series of meetings were held with Türk Telekom teams to discuss the scope of the data and its anonymization. The data description was approved in July.

We have decided to collect data over 1 year to enable the observation of seasonal trends. To make the best use of the restricted refugee coverage, we requested the entire refugee base of Türk Telekom to be sampled. The refugee tag was obtained by different means (described in Sect. 1.3), which was purposefully noisy as an additional layer of protection. In September 2017, we have started forming the Project Evaluation Committee (PEC), inviting members from academia, NGOs, ministries, as well as members of the Syrian refugee community in Turkey (see Sect. 1.2.2).

A tentative schedule of the Challenge was formed in September. At the same time, the project procedure was presented to the Ethics Board of Boğaziçi University (INAREK). The Challenge, which aimed to open data of Türk Telekom to research groups outside Boğaziçi University with its own ethics committee to evaluate the project reports, was a very unusual project for the Ethics Board. The Board, chaired by Prof. Yaman Barlas, investigated the financial sourcing of the project, its partners

Fig. 1.1 The D4R logo, designed by Marcom Ajans

and supporters, the procedure for project evaluation, and the data description before making a (positive) decision.

In the meantime, Türk Telekom worked on the website design. The D4R logo (see Fig. 1.1) was designed by Marcom Ajans. The door in the logo symbolizes the open door policy of the Turkish government in the Syrian refugee crisis. The open door could be seen as a symbol for sharing the data with the "outside", a major commitment by any telecommunications company. Since the map of Turkey is roughly rectangular, it also illustrates the geographical location of Syria with respect to Turkey. The logo, D4R, and Data for Refugees are trademarked by Türk Telekom.

The dataset collection and anonymization took several months, and was completed in January 2018. Meanwhile, several organizations were contacted for their support in the project. Fondazione Bruno Kessler was the first institution to declare its official support in October 2017, followed by MIT Media Lab, Data-Pop Alliance, Istanbul and I, UNHCR, International Organization for Migration (IOM), and UNICEF. The most important concerns addressed at this stage, particularly by UNHCR, were that (1) the sharing of the data is permitted by the owners of the data, i.e., the refugees; (2) the data should be anonymized in a way to make it impossible to identify individuals in it; and (3) the data should not allow other governments to track specific refugees, even if they had other, complementary data about individuals. Our data collection ensured all these points.

The first call for proposals went out in January 2018, resulting in 102 proposals by the deadline in March, of which 33 were rejected, 59 accepted, and 10 deferred to PEC after the initial screening. The most important factor in the early rejection was the lack of scientific rigor (or content).

After the PEC evaluation, a total of 61 groups were granted access to the dataset, of which 31 groups were able to submit project reports by the Challenge deadline. The word cloud created from the titles of 61 proposals (see Fig. 1.2) illustrates the relative popularity of the main themes. Social integration was by far the most popular area selected by the project participants.

The 31 groups submitting project reports involved 163 researchers (118 male, 45 female) from three NGOs, three governmental or intergovernmental institutions, two research labs in commercial companies, and 57 academic institutions from 19 countries (see Fig. 1.3). One-third of the projects involved an expat Turkish researcher working outside Turkey. Six projects came from exclusively Turkish institutions, 18 projects from institutions outside Turkey, and seven projects involved collaborations

Fig. 1.2 A word cloud obtained from the titles of the successful project proposals
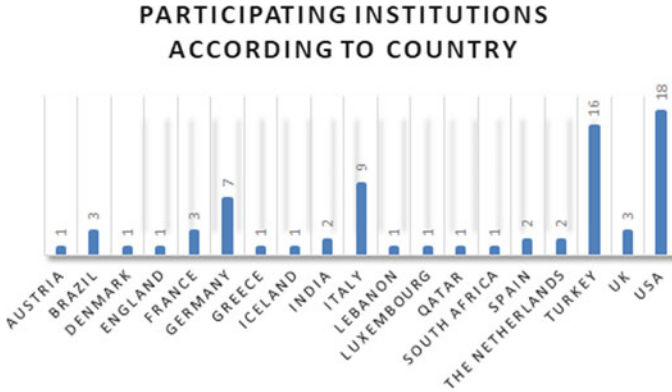

Fig. 1.3 The participating institutions according to the country

between Turkish institutions and those outside Turkey. Twenty-three projects, out of a total of 31, involved multiple institutions. Twenty-six projects were selected for presentations in the D4R Closing Workshop, and 17 groups were invited to submit revised reports for the present volume.

1.2.2 Committees

The day-to-day running of the Challenge was coordinated by the Scientific Committee, whose members were

- Albert Ali Salah (chair, Boğaziçi University and Utrecht University),
- Alex Pentland (Massachusetts Institute of Technology),
- Bruno Lepri (Fondazione Bruno Kessler),
- Emmanuel Letouzé (Massachusetts Institute of Technology and Data-Pop Alliance),
- Yves-Alexandre de Montjoye (Imperial College London),
- Xiaowen Dong (University of Oxford), and

- Patrick Vinck (Harvard Humanitarian Initiative).

This committee determined the scope of the dataset, drafted the ethical guidelines, as well as the project prescreening and evaluation criteria, and organized public dissemination activities.

The second and larger committee was the Project Evaluation Committee (PEC), whose task was to evaluate the project proposals to determine which groups will be granted access to the dataset, and to evaluate the project reports to ensure that no published content can potentially harm the refugee population. The latter is a broad criterion. In such a sensitive project, the results of analysis, even if scientifically rigorous, need to be framed carefully to make sure that they cannot be misused, intentionally or otherwise. The members of the PEC represented a wide range of interests, from academia, international nonprofit, nongovernmental organizations, refugee organizations, and the Turkish ministries:

- Senem Özyavuz (chair, Türk Telekom),
- Iyad Rahwan (Massachusetts Institute of Technology),
- Anahi Ayala Iacucci (Internews),
- Bülent Sankur (Boğaziçi University),
- Yldirm Bahadirlar (TÜBİTAK BİLGEM),
- Alex Rutherford (Massachusetts Institute of Technology),
- Claire Melamed (Global Partnership for Sustainable Development Data),
- Jean-Marie Garelli (UNHCR),
- Ahmad Garibeh (Istanbul & I),
- Geoffrey Charles Fox (Indiana University),
- Joséphine Goube (Techfugees),
- Firat Yaman Er (Türk Telekom),
- Phuong Pham (Harvard University),
- Mithat Büyükhan (Lifelong Learning General Directorate, Turkish Ministry of Education),
- Mazen AboulHosn (International Organization for Migration),
- Ömer Hakan Simsek (Turkish Medicines and Medical Devices Agency, Turkish Ministry of Health),
- Nona Zicherman (UNICEF),
- Manuel García-Herranz (UNICEF), and
- Vedran Sekara (UNICEF).

1.2.3 Evaluation Criteria

The projects submitted to D4R Challenge were screened by the Scientific Committee (SC) and the Project Evaluation Committee (PEC). The proposals that passed the initial screening were granted access to the dataset, upon submitting the signed User Agreement Form (See Appendix).

Criteria for passing the initial screening were as follows:

Table 1.1 Summary of possible project evaluation outcomes

PEC recommendation	Description
Normal publication	No restriction applied. Can be invited to present in the Challenge workshop, in related sessions and is also fully eligible for prizes
Ask to consider adjustments	Requires some wording, visual, and/or content adjustment before decision. Without further adjustment, PEC asks Not to publish it
Do not publish	PEC asks Not to publish the document. This could be due to either of the following: • The reports might be sensitive or too risky for a general publication, but could merit a presentation to the appropriate authorities. This will be evaluated on a case-by-case basis; • The methodology or conclusions do not confirm to scientifically rigorous methodology. Scientific correctness is a precondition for publication

- A project proposal is submitted (in English), and all the team members who will access the data are individually identified;
- The project's primary investigator has a permanent affiliation;
- The project uses the D4R data meaningfully;
- The project aims are aligned with the goals of the Challenge, do not represent a commercial interest, and do not endanger the privacy or well-being of individuals or groups.

The project reports were published publicly on the project website,[4] after evaluation by the D4R PEC and D4R SC. There were three possible outcomes for submitted reports (Table 1.1): (1) Normal Publication, for papers that treat the ethical issues correctly; (2) Ask to Consider Adjustments, for papers that require amendments and removal of sensitive material before publication. We asked the participants to be careful not to include statements that may harm the refugee population in any way or may promote negative sentiments about the refugee populations; and (3) Do Not Publish, either because the report is on a sensitive issue or because it is not scientifically rigorous and its conclusions are not warranted. Sensitive reports may be shared directly with related institutions or authorities. The PEC and the project proposers were designated to jointly decide on this, on a case-by-case basis. The user agreement stated that any further scientific publication based on the D4R data will be sent to the PEC/SC for evaluation, before the publication. A fixed period of 30 days is allotted to the PEC/SC to grant permission to such publications, after which tacit denial could be assumed. It was important to keep the publication options open beyond the duration of the Challenge, as rigorous scientific analysis of the data requires

---

[4]http://d4r.turktelekom.com.tr.

significant commitment by the participating groups. It is natural that these groups would require permission to publish scientific papers (and dissertations) based on their findings. However, such publications also need to be screened before reaching the public, as the risks are no less after the Challenge is finished.

The following criteria were used for the evaluation of projects with regard to the awards in each category:

- **Relevance to the focus area**: The project addresses a major problem in the focus area and outlines how its results can be used to help solve the problem;
- **Methodology**: The project is assessed for its analytical rigor and empirical backing of its conclusions, and whether additional data sources are appropriately and creatively used in conjunction with the D4R data;
- **Public communication**: It is important to raise awareness about refugee issues. The project outcomes are assessed for the clarity of presentation, and the strength of its message, for instance, through good visualizations. Additional resources (web, videos, other multimedia material, etc.) that can make the results clear to the general public are welcome and are assessed.
- **Recommendations**: The Challenge outcomes and recommendations are compiled into an accessible white paper (written in Turkish) and shared with the related government agencies, NGOs, and other stakeholders.[5] This report is one of the most direct ways for the Challenge participants to improve refugee conditions. Subsequently, the reports are assessed for the relevance of their recommendations.

1.3 Description of D4R Data

This section describes the datasets shared within the D4R Challenge.

1.3.1 Brief Summary

The D4R Challenge is based on the successful Data for Development (D4D) Challenge series [5]. Three datasets are made available to the Challenge participants, along with external helper files. The main difference from D4D is that the D4R data contain a "Refugee" flag, which indicates (with a high probability) that the CDR belongs to a refugee customer. This flag is given to customers in the database that (1) have ID numbers given to refugees and foreigners in Turkey; (2) are registered with Syrian passports; and (3) use special tariffs reserved for refugees. The proportions of these three groups in the database are not equal; the first group has 42.87%, the second group 1.06%, and the last group 56.07%, respectively. None of these groups is guaranteed to include only and exclusively refugees, which serves as a layer of

---

[5]See the D4R project website for the white paper.

protection: It is not possible to identify with certainty that a particular CDR belongs to a refugee or not, since this indicator is noisy and it is only possible to derive patterns from aggregated records. In such data collection efforts, it is prudent to have some minimal noise in the sensitive labels (e.g., refugee vs. non-refugee). We list the datasets contained in the Challenge in individual subsections.

Turkey is party to the 1951 Geneva Refugee Convention, but only acknowledges "refugee" status for people originating from Europe. Syrian refugees are officially considered "temporarily protected foreign individuals." We acknowledge that the term "refugee" is used as a blanket term in the dataset and includes migrants, asylum seekers, and foreigners (Syrian or otherwise) who have acquired a temporarily protected foreign individual ID number in Turkey (i.e., starting with 98 or 99). The dataset needs to be approached with these reservations in mind, and the analysis should carefully consider such biases in the data.

The D4R dataset is collected from 992,457 customers of Türk Telekom, of which 184,949 are tagged as "refugees", and 807,508 as Turkish citizens. A total of 1,211,839 subscriptions are included. Of these, 980,697 belong to Turkish citizens, and 231,142 belong to refugees (we refer to these customers as refugees, but as mentioned before, there exists certain level of noise in this indicator). Some of the customers had multiple phone lines; each line corresponds to a single subscription.

Out of all the refugee-tagged customers, 75% are recorded as "male", and 25% as "female". There is clearly a gender bias in the ownership of the phone line. This does not mean that 75% of the phone lines are used by men, however. We have sampled the 807K Turkish customers with the same gender distribution.

Out of all the refugee customers, 45% are registered in Istanbul. This information is obtained from the address associated with the contract. Other major cities with refugee presence are Gaziantep, Izmir, Sanliurfa, and Mersin. To simplify comparisons, we have sampled the Turkish citizen customers mainly from the cities with registered refugee presence. Tables 1.2 and 1.3 show the distribution of customers and their registered cities (only top locations are shown). The distribution over all the cities of the country is provided to the participants in a separate file. This file shows the official number of refugees registered per city, the official city population in 2017 (excluding refugees), and the percentage of refugees with respect to the city population. Additionally, it shows the number of Türk Telekom customers used for the entire D4R data collection per city, broken into "refugee" and "citizen" counts.

The numbers of registered refugees and asylum seekers in Turkey according to registration dates can be obtained from the UNHCR website.[6] Another useful source of data is the TUIK census estimates for Turkish cities, according to the year.[7] This source indicates the population size and growth of each city between 2000 and 2017.

The usage of the D4R data requires caution in interpreting the representativeness of the data for the refugee population in Turkey. At the end of March 2017, there were 75,724,413 mobile customers in Turkey across all operators (94.9% penetration rate) [2]. Excluding machine to machine (M2M) and population of the age range of

---

[6]See https://data2.unhcr.org/en/situations/syria/location/113.

[7]See http://www.tuik.gov.tr/PreIstatistikTablo.do?istab_id=1590.

Table 1.2 The distribution of customers tagged as refugees in the dataset and their registered locations. Numbers rounded to the third significant digit

Location	Number of customers	Percentage (%)
Istanbul	84,173	45.511
Gaziantep	14,898	8.055
İzmir	10,425	5.637
Sanliurfa	9,701	5.245
Mersin	9,660	5.223
Hatay	7,024	3.798
Ankara	5,580	3.017
Konya	4,718	2.551
Bursa	3,479	1.881
Outside Turkey	2,902	1.569
Other	32,440	17.540

Table 1.3 The distribution of customers tagged as Turkish citizens in the dataset and their registered locations. Numbers rounded to the third significant digit

Location	Number of customers	Percentage (%)
Istanbul	363,334	44.994
Gaziantep	80,655	9.988
İzmir	40,501	5.016
Ankara	40,443	5.009
Adana	40,415	5.005
Hatay	40,394	5.002
Konya	40,388	5.002
Antalya	40,367	4.999
Bursa	40,359	4.998
Sanliurfa	40,321	4.993
Mersin	40,242	4.983

0–9, the mobile penetration was 107%. According to data from the first 3 months of 2017, the mobile market share of Türk Telekom (Avea), from which the Challenge data were collected, was 24.7% [2].

We have used the entire refugee customer base (with the filtering conditions described previously); however, the market share of Türk Telekom also shows fluctuations according to the individual cities. Therefore, it is useful to look at official numbers of refugees distributed over the country. This is partly depicted in Fig. 1.4 according to figures[8] from the Ministry of Interior, Directorate General of Migration

---

[8]Only the top ten cities are shown. More detailed information can be obtained from http://www.goc.gov.tr/icerik/migration-statistics_915_1024.
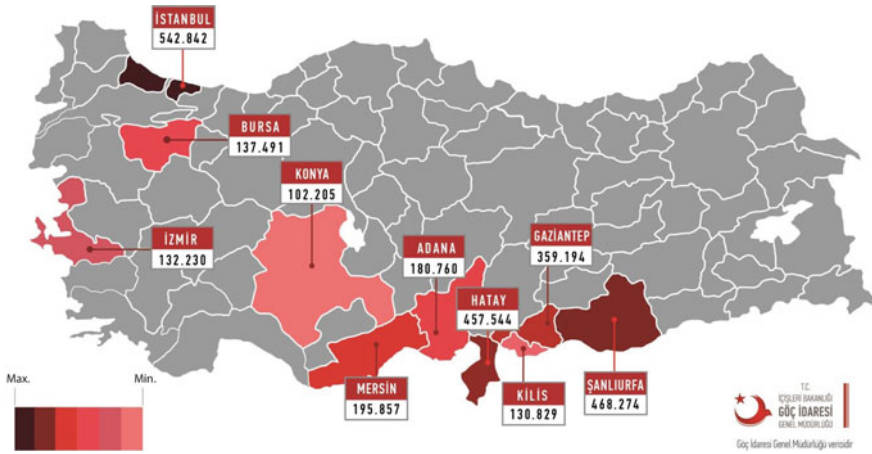
Fig. 1.4 The distribution of refugees in the country according to data from Ministry of Interior, Directorate General of Migration Management

Management in 2018. For the Challenge, we have supplied the official figures from March 2017, as mentioned before. In what follows, we describe the contents of the dataset in more detail.

1.3.2 Cell Tower Locations

The cell tower (i.e., base station) locations are provided in the file "Base_Station_Location.txt". The file contains the following fields:

- BTS_ID: The ID of the cell tower;
- MX_LAT1,MX_LAT2,MX_LAT3: DMS latitude of the cell tower;
- MX_LONG1,MX_LONG2,MX_LONG3: DMS longitude of the cell tower;
- MX_SAHAIL: The registered city of the cell tower;
- MX_SAHAILCE: The registered district of the cell tower;
- MX_POPAREA: An unofficial note about the population type around the cell tower, used internally in Türk Telekom. It takes values of RURAL, SUB_URBAN, INDUSTRIAL, SEASONAL AREAS, DENSE_URBAN, HOT SPOT, OPEN IN URBAN, AIRPORT, SUBURBANLOW, POPRURAL, and INDOOR.

In some rare cases, the precise location information of the cell tower is missing, and only the city is indicated. The interpretation of the latitude and longitude follows the degree, minutes, seconds (DMS) syntax. For example, the district of Bartin in the city of Zonguldak is represented by these six numbers as follows: (41 25 43.1184 32 4 37.9344). This corresponds to $41°25'43.1184''$ N DMS latitude, and $32°4'37.9344''$ E DMS longitude.

1.3.3 District Locations

To disambiguate the cell towers, we provide a file that contains district coordinates. This file, named "district_coordinates.csv", has the following fields:

- CITY: Name of the city;
- DISTRICT: Name of the district;
- POPULATION_2014: The official census population of the district in 2014;
- LATITUDE, LONGITUDE: The 2D (GPS) coordinates of the district;
- MX_LAT1, MX_LAT2, MX_LAT3, MX_LONG1, MX_LONG2, MX_LONG3: The DMS coordinates of the district.

A conversion script (such as https://www.latlong.net/lat-long-dms.html) can be used to convert the (latitude, longitude) variables into DMS coordinates. For example, the 2D coordinates of (41.428644 32.077204) for Zonguldak, Bartin, translate into the DMS coordinates given in the previous subsection.

1.3.4 Dataset 1: Antenna Traffic

The first database we provide includes 1-year site-to-site traffic on an hourly basis. This dataset contains the traffic between each site for a year. Calls between Türk Telekom (TT) customers and other service providers (SP) only have information about the TT side. For each record, total number and duration of calls are recorded in an aggregated fashion.

The database is split into voice and SMS partitions. For the voice partition, the file "Dataset 1_2017XX.txt" contains the data for month XX, and there are 12 such files. Each file contains the following fields:

- TIMESTAMP: Day and hour considered in format DD-MM-YYYY HH (24 h format);
- OUTGOING_SITE_ID: The ID of the site the call originated from. Unknown stations are denoted as "−99" or "9999";
- INCOMING_SITE_ID: The ID of the site receiving the call;
- NUMBER_OF_CALLS: The number of calls in the one-hour slot;
- NUMBER_OF_REFUGEE_CALLS: The number of calls involving numbers tagged as "refugee";
- NUMBER_OF_TOTAL_CALL_DURATION: The total call duration from all calls;
- REFUGEE_CALL_DURATION: The total call duration from calls involving numbers tagged as "refugee".

Note that it is possible for a call to be labeled incorrectly, as we do not know the refugee status of the other party, when the call is to a different SP than TT.

For the SMS partition, the file "Dataset 1_SMS_2017XX.txt" contains the data for month XX, and there are 12 such files. Each file contains the following fields:

- TIMESTAMP: Day and hour considered in format DD-MM-YYYY HH (24 h format);
- OUTGOING_SITE_ID: The ID of the site the SMS originated from. Unknown stations are denoted as "−99" or "9999";
- INCOMING_SITE_ID: The ID of the site receiving the SMS;
- NUMBER_OF_SMS: The number of SMS messages in the one-hour slot;
- NUMBER_OF_REFUGEE_SMS: The number of SMS messages involving numbers tagged as "refugee".

1.3.5 Dataset 2: Fine-Grained Mobility

The dataset contains cell tower identifiers used by a group of randomly chosen active users to make phone calls and send texts. The data are timestamped and a random sample of users is observed for each period of 2 weeks. At the end of each 2-week period, a fresh sample of users is drawn at random. We provide data for the entire 1-year sampling period. The users are represented by random numbers in the dataset, and no personal information is stored. To protect privacy, new random identifiers are chosen for each 2-week period. Therefore, even if a user is sampled in more than one period, records from different periods cannot be associated with each other. For missing antenna locations, a code of "−99" or "9999" is assigned. This dataset is also separated into voice and SMS partitions. Furthermore, to deal with large file sizes, it is divided into files containing incoming (IN) and outgoing (OUT) calls and SMS messages, resulting in four files per 15-day period.

The files "Dataset 2_2017XXW_In.txt", "Dataset 2_2017XXW_Out.txt", "Dataset 2_2017XXW_SMS_In.txt", and "Dataset 2_2017XXW_SMS_Out.txt" all have similar structure, where XX ranges from 01 to 26 and denotes a 15-day period (for the 52 weeks of the year), starting from January 1 to 15, 2017 in the first file. The files have the following fields:

- CALLER_ID: The randomly assigned ID of the user specifically for the 15-day period. Note that the user is not necessarily the initiator of the call, which is determined by the CALL_TYPE flag;
- TIMESTAMP: Day and hour considered in format DD-MM-YYYY HH (24 h format);
- CALLEE_PREFIX: A value that denotes "refugee" (1), "non-refugee" (2), and "unknown" (3);
- SITE_ID: The ID of the cell tower;
- CALL_TYPE: The call type is either outgoing (1) or incoming (2).

1.3.6 Dataset 3: Coarse-Grained Mobility

In this dataset, the trajectories of a randomly selected subset of users are provided for the entire observation period, but with reduced spatial resolution. We divide the entire country into the electoral prefectures (or districts) and, for each call record, only the prefecture information is provided. The IDs are randomly assigned, and two different users may have the same ID in Dataset 2 and Dataset 3. The database is split into incoming (IN) and outgoing (OUT) calls to deal with large files. The files "Dataset 3_2017XX_In.txt" and "Dataset 3_2017XX_Out.txt" have a similar structure and contain the following fields:

- CALLER_ID: The randomly assigned ID of the user (different from Dataset 2). Similarly to Dataset 2, the call initiator is determined by CALL_TYPE;
- TIMESTAMP: Day and time considered in format DD-MM-YYYY HH:MM (24 h format);
- ID: The ID of the district;
- CITY_ID: The ID of the city.

In order to obtain the mapping to the cities and prefectures (districts), two additional files are provided. In "Dataset 3_City Mapping.txt", the CITY_ID is followed by CITY_DESC, which is the name of the city. There are a total of 81 cities. In "Dataset 3_District Mapping.txt", the ID field represents the district ID, as used in Dataset 3, and the BTS_DISTRICT field gives the name of the district. There are a total of 1025 districts.

The dataset contains some missing periods due to technical issues, e.g., data were missing in particular in February and March 2017.

1.4 Ethical and Privacy Issues

In this section, we briefly discuss the ethical and privacy issues regarding the Challenge data. The collection, storage, and protection of data in the D4R Challenge comply with European Union requirements regarding the protection of personal data and the protection of privacy in the electronic communications sector. Furthermore, research on the previously conducted D4D Challenges established that the data such as offered in this project do not allow the identification of individuals [3, 7, 8, 20]. Sharad and Danezis note that providing aggregated data such as antenna traffic results in "little scope of privacy breach…since it contains no personally identifiable information about the users. It could be used to study traffic patterns during the entire period but reveals no information pertaining to the users" [18].

Definitions: Personal Data means any information relating to an identified or identifiable natural person. Personal Data does not include anonymous information, that is, information that does not relate to an identified or identifiable natural person or to data rendered anonymous in such a way that the Data Subject is not or is no

longer identifiable. Data Subject means a natural person (i.e., an individual) who can be identified directly or indirectly, in particular, by reference to Personal Data.

Consent, legitimate, and fair processing: The data in the Database come exclusively from Türk Telekom customers, who have consented to its anonymized use for research purposes through the mandatory user agreement at the time of the purchase of the phone line.[9] The content of phone activity, actual phone numbers, identities, addresses, or similar personal information are neither stored nor distributed with the Database. Subsequently, it is not possible to identify natural persons with the Database.

Transparent processing: The nature of the data, the assurance of its anonymity, as well as the ethical precautions to ensure its proper use are (at the time of the start of the Challenge) documented openly on the Challenge website. Accessible and plain language is used, and further contact information is supplied to respond to questions about the data usage.

Project Evaluation Committee (PEC): The PEC is formed with representatives from academia, government, and related NGOs. Its aim is to represent refugee interests in the Challenge, and all project proposals are prescreened by the Scientific Committee and by the PEC. The proposals that pass the initial screening are granted access to the dataset, upon submitting the signed User Agreement Form. Criteria for passing the initial screening are discussed in Sect. 1.2.3.

Data access for the Database is granted to participants during the designated Challenge period, by a mandatory user agreement prepared by Türk Telekom lawyers, and approved by ADIEK, the Ethical Conduct committee of Boğaziçi University. The agreement permits third parties to analyze the anonymized and aggregated data, summarized previously, to submit a research report at the end of the Challenge and to present the results at a dedicated workshop. A white paper is to be prepared to inform the related government bodies and NGOs about the results of the project. The project reports are published publicly on the project website, after evaluation by the D4R PEC and D4R SC. The report evaluation is discussed in Sect. 1.2.3.

Retention, destruction, and archiving: The Challenge mandates that all participants destroy the Database upon completing the Challenge. Any publication based on the Database requires the pre-approval of the Project Evaluation Committee. Any further use of the Database (for instance, to complete numerical experiments for a publication under review) will be regulated by the PEC, and extensions are to be conditionally granted for specific purposes on a case-by-case basis.

Information: Every care has been given to ensure that the information provided in the Database does not cause any harm, prejudice, or distress to customers, regardless of their refugee status. The PEC provides an additional layer of control, as it examines that the project reports comply with these maxims and suggests changes whenever necessary.

Access, correction, erasure, objection: The Database does not contain personal information, and it is not possible for individuals to request access to Personal Data. The data are anonymized and aggregated in a way to prevent identification of persons.

---

[9]See Chap. 5 of this volume on further discussions of consent.

For the same reason, correction, erasure, and objection do not apply to the Database. We note that this is a stricter protection condition than most envisioned CDR usage scenarios. Furthermore, the Database does not contain children's data, as each registered customer has to be over 18 years of age, and thus legally permitted to own and use a mobile phone line. Personal profiling (such as used for CRM applications) is not possible with the Database.

Responsibility and accountability: The responsibilities of all parties concerned are defined clearly and set out in the user agreement prepared by Türk Telekom lawyers.

Data protection by design and default: Data collection follows this principle, where any name, real phone number, or other identifying information is excluded from the design of the Database. The pseudo-random numbers representing customers are not stored anywhere along with actual phone numbers. Subsequently, the anonymization works only one way. Refugee status is indicated by purposefully noisy indicators, and no effort is spent to ensure its validity. Subsequently, only aggregate-level conclusions can be drawn from the Database. It is not possible to use the Database for the surveillance and tracking of individuals.

Limitations: Data access is not provided to institutions in a blanket permission, but to specific individuals within institutions, whose names and roles in the proposed research project are clearly defined in the user agreement.

Appendix—User Agreement

This Agreement is signed by the duly authorized representatives of the below-mentioned Parties and effective from April 20, 2018;

1. Türk Telekomünikasyon A.Ş. having its registered office at Turgut Özal Bulvari 06103 Aydinikevler, Ankara, Türkiye (hereinafter designated as "Türk Telekom"); TTNET A.Ş. having its registered office at Esentepe Mahallesi Salih Tozan Sokak No: 16 Karamancilar Iş Merkezi D Blok 34394 Şişli—Istanbul (hereinafter designated as "TTNET"); AVEA Iletişim Hizmetleri A.Ş. having its registered address at Abdi Ipekçi Cad. No: 75 34367 Maçka, Şişli—Istanbul (hereinafter designated as "AVEA") and

2. .... having its registered office at .... (hereinafter designated as "Participant").

Under this Agreement, Türk Telekom, TTNET, and AVEA together called "TTG", ... will be called as "participant", and both TTG and Participant collectively referred to as the "parties" and each individually as "Party".

TTG provides a special database to the scientific committee to enable research on refugee movements and to ultimately help the authorities in providing better conditions to refugees. The sharing requirements and the type of data that is to be shared will be determined solely by TTG. Access to this database is only possible by taking part in the "Data for Refugees" Challenge. Through the studies made with this special database, TTG aims to support the studies that will create better conditions for refugees in areas such as health, education, security, unemployment, integration and also to support the outreach programs for refugees carried by the administrative bodies.

Participants that contribute to the Challenge will aim to analyze information from the special database that is provided exclusively to them by TTG. The results of these studies will be shared with TTG in a report, with an explanation of the possible benefits of these outcomes to the refugees. The results that fulfill the Committee's criteria will be awarded.

The Participant undertakes and agrees to be bound by the following conditions in order to take part in the Challenge. Participant has to send a signed copy of this Agreement by electronic mail to the electronic mail address mentioned on the Challenge website.

By signing this Agreement, the Parties undertake to comply with the terms and conditions below.

Article 1—Definitions

1.1. "**Affiliated Scientist**" shall mean any person or entity who is employee of the Participant or who is directly affiliated to the Participant within the framework of its research and studies by contract or not, or as a student. The names of the Affiliated Scientists participating in the Challenge are listed at the end of this document.

1.2. "**Applicable Law**" shall mean the relevant laws, regulations, and other legislative acts that apply to this Agreement, such as Turkish Personal Data Protection Law No. 6698 and Turkish Electronic Communications Law No. 5809.

1.3. "**Awards**" shall mean the prizes that will be granted to Participants whose Results are selected by the Committee.

1.4. "**Challenge**" shall mean the contest in which Participants aim to create a project for benefit of the refugees and supporting administrative bodies through the analysis of a special database that consists of anonymized mobile telecommunication data of refugees, Turkish citizens, and other TTG customers.

1.5. "**Committee**" shall mean the entity composed of scientists and various personalities from national and international organizations and aims to evaluate the Results in terms of ethics, benefits, and data security. The names of the Committee members are mentioned on the Challenge website.

1.6. "**Copyrightable Results**" shall mean the elements of the Results that are or will be protected by intellectual property rights, such as software and algorithms.

1.7. "**Data**" shall mean the database containing phone calls from mobile users of TTG made between January 2016 and January 2017. Data have been anonymized and aggregated by TTG and are made available to the Participant for the sole and limited purpose of the Challenge, pursuant to criteria described in Annex-1 of this

Agreement. Under the scope of the Challenge, the anonymized data are going to be used for the purposes of research and statistics. Multiple special databases have been prepared for this Challenge.

1.8. "**Participant**" shall mean any legal person or entity, which applies to the Challenge with its authorized body to present a nonprofit study made by Affiliated Scientists. Independent natural persons can also apply as Participant.

1.9. "**Result**" shall mean the submitted written document handed over by the Participant to TTG containing the solutions proposed by the Participant within the framework of the Challenge. On the basis of this document and its content, the Committee will determine the winners of the Challenge. Results may contain tables, maps, and graphics.

Article 2—Entitlement to Participate in Challenge—Offer and Award

2.1. In order to be able to participate in the Challenge, Participant has to be (i) an academic university/faculty and/or a scientific research institution (or an entity having an equivalent status) or (ii) a nonprofit entity or (iii) a legal person representing a nonprofit organization or (iv) individual and independent research group or (v) the research division of a commercial or industrial entity that requests the data for a noncommercial purpose. If the Participants application is accepted, a unique access key will be sent to the Participant in order to enable the downloading of Data located on a web server.

2.2. Participant agrees, declares, and undertakes to use reasonable efforts and utmost care to carry out research in compliance with the objectives of the Challenge, especially regarding processing and protection of Data. Participant must submit its contribution by the deadline announced on the Challenge web page. Participant agrees that only Affiliated Scientists have access to Data, and upon completion of the Challenge, and the submission of related reports and publications, Participant agrees to destroy and completely remove Data from its servers and other storage facilities, and ensures that the Affiliated Scientists will do likewise.

2.3. By submitting to this Challenge, Participant agrees to comply with all applicable laws, including, but not limited to statutes, statutory decrees, international agreements, decisions/regulations issued by competent authorities, and all relevant announcements, statements, and briefings made by TTG. In case of noncompliance, Participant agrees, declares, and undertakes to be exclusively liable of all legal, administrative, and penal sanctions. In case of the nondestruction, storage, transmission, or processing of Data and related information, legal, administrative, and criminal liability shall be solely with the Participant.

2.4. TTG shall have the right to unilaterally terminate this Agreement at all times via sending a notification email to the Participant in case one of the following events occur:
(i) as a result of a change in legislative regulations; or
(ii) if the continuation of the Challenge is, in TTG's opinion, no longer viable; or,
(iii) if Participant breaches the terms and conditions of this Agreement.

In this case, the Participant will irreversibly destroy all Data and Data-related results from all stored spaces within 1 week of receipt of the termination notification email without further notice. Such destruction records will be submitted to TTG.

2.5. The Committee will define and announce the winners of the Challenge. Awards will be given in five categories, and TTG preserves the right not to grant an Award in any of these categories:

(i) 1st prize: Safety and Security Category,

(ii) 1st prize: Health Category,

(iii) 1st prize: Education Category,

(iv) 1st prize: Unemployment Category, and

(v) 1st prize: Integration Category.

2.6. Participant acknowledges that Committee possesses the sole authority to define the winners, according to its own selection criteria and that TTG does not possess any significant influence over the Committee's authority, and that under no circumstance it shall hold TTG responsible for Award selection process. In this respect, the Participant hereby declares that it will not object to the results of the Challenge, and that such objection will constitute material breach of this Agreement. In such a case, TTG shall keep all its statutory and contractual rights reserved.

Article 3—Confidentiality

3.1. Participant shall keep Data and all other information disclosed by TTG for the purposes of the Challenge confidential. For protection of Data, Participant undertakes to apply at least the same degree of care with which it treats and protects its own confidential information against public disclosure, but not less than a reasonable degree of care. The "reasonable care" must contain all necessary measures to provide the Data protection at the same level as predicted in Turkish Law No. 6698 and Turkish Law No. 5809. The "reasonable care" shall ensure a level of protection according to TS ISO/IEC 27001 and/ or ISO/IEC 27001 standards or current ISO safety standards.

3.2. All Data and information shall be disclosed on a need-to-know basis. Data and other information shall not be disclosed to any third party without prior written consent of TTG. Participant shall irrevocably and accurately limit the use and access to Data to Affiliated Scientists only. Participant acknowledges that any disclosure of information (including but not limited to Data) to any unauthorized person shall constitute a material breach of this Agreement and violation of the applicable law and that it shall take all reasonable measures to prevent such disclosure. Participant is authorized and commissioned to take all necessary measures to prevent any unauthorized access and therefore, in such a case, Participant is exclusively responsible from any legal, administrative, and penal liability that might occur.

3.3. Participant is obliged to inform all persons that have conducted with the Data about Data confidentiality and the limited usage of Data, meaning the usage being limited to the actualization of the purposes of Challenge. Participant agrees to be liable for violation of Applicable Law and this Agreement by all entities that have conduct with the Data. The degree of liability shall be objective (strict) liability and under the scope of commitment of a third-party's action. It is agreed by both Parties

that in such a case TTG has the right and authority to directly appeal to Participant for compensation of all damages it endures due to actions of these entities and recourse all kind of payments made to third parties and public authorities such as compensation, penalty, etc. TTG keeps all its statutory and contractual rights reserved.

3.4. Participant is responsible for making necessary notifications about confidential information to Affiliated Scientists and other authorized entities and to make sure that they shall also act in compliance with the confidentiality provisions in this Agreement. Participant shall not disclose Data or any other information partially or wholly to any other party.

3.5. Participant shall be responsible for all losses, damages, and demands and other legal, administrative, and penal liabilities that occur as a result of the actions of the Affiliated Scientists related to the Challenge and the use of Data. The Participant accepts that any kind of monetary sanction (administrative fine, compensation, etc.) that TTG confronts with due to gross negligence or unlawful intent of Participant and/or the Affiliated Scientists shall be immediately recoursed to Participant.

3.6 Participant shall be solely responsible for all losses, damages, demands, and other legal, administrative, commercial, and penal liabilities from relevant legislations that occur use of data out of purposes or use of unauthorized data.

Article 4—Consequences of Termination

4.1. This Agreement becomes effective upon its execution date and remains in force until the official announcement of the Challenge Awards. Upon the announcement of the Awards, this Agreement automatically terminates without any notice. For the avoidance of doubt, the official announcement of the Award winners may be made via project website or in a special ceremony or during an international event. Participant, without prejudice to being awarded or not, must destroy all Data that is granted within the scope of the Challenge within 1 week starting from the official announcement of the Award winners. All records regarding the destruction of Data shall be submitted to TTG. Further Data usage, for instance, for scientific publications directly related to the Challenge, will be explicitly requested from TTG and evaluated on a case-by-case basis.

4.2. The obligations of the Participant under Article 3 ("Confidentiality") shall survive the termination of this Agreement.

4.3. After the termination of this Agreement, both Parties shall immediately cease to use the confidential information (including, but not limited to Data) and intellectual property rights that belong to the other Party, if any.

Article 5—Restriction of Use—Restitution

5.1. Participant agrees to use the Data only for the purposes of the Challenge and only until the official announcement of the Award Winners. As clearly stated in the Article 4.1., Participant is obliged to destroy all Data within 1 week starting from the date of the official announcement of the Award winners. Any other use of Data of any nature outside the scope of the Challenge by Participant and/or Affiliated Scientists or any other entity is submitted to the following conditions.

5.2. Participant can apply for an extension of the right to use Data after the announcement of the Awards only in order to proceed to additional scientific analysis and researches. To do so, Participant will have to send a written application to TTG within 1 week starting from the date of the official announcement of the Awards. The written application must contain the scope of the new research Participant intends to carry out, the reason for the usage of Data, and the time period of the usage. TTG has no obligation to accept this application and Participant has no right to object to TTG's decision in this respect.

5.3. Usage of Data shall only be permitted exclusively for nonprofit purposes. In other words, Data cannot be used for commercial purposes. Participant accepts in advance that any application that indicates commercial use of Data shall not be accepted by TTG and that any kind of unauthorized commercial use of Data shall constitute material breach of this Agreement. In such case, TTG shall exercise all its statutory and contractual rights, including, but not limited to, penal clause and immediate termination without any notice. In addition to that, Participant shall be exclusively responsible for any legal, administrative, and penal liabilities that caused by commercial use of Data.

5.4. Participant is obliged to mention Data origin in all studies that use the Data. In other words, Participant must mention that Data was made available by TTG within the framework of the Data 4 Refugees Challenge, and cite the related work as described on the Challenge website.

Article 6—Intellectual Property Rights

6.1. Parties agree that all the Data made available for the purposes of this Agreement are deemed the property of TTG.

6.2. Nothing in this Agreement shall be deemed to grant a license directly or by implication of any intellectual property rights related to the Data, except the limited and nontransferable right to use such Data for till the end of Challenge.

6.3. This Agreement shall not be deemed to create any obligation for either Participant or TTG to enter into any further contractual arrangement of any kind.

6.4. Copyrights of the Copyrightable Challenge Results will be property of Participant and Affiliated Scientists.

Article 7—Results

7.1. Participant presents its Results to TTG in a project report format and on the condition of having TTG's previous written consent for each Result, Participant may publish the Results in scientific papers or conference presentations. Apart from that, TTG can also announce the Results in Award ceremony, or enable public access to the Results via publishing them in scientific and commercial broadcasts/exhibitions/conferences. TTG undertakes to refer to and give credit to Participants project reports in any such publications. The Committee evaluates the reports and can decide that the report is publishable, non-publishable due to sensitive parts, or publishable with amendments.

7.2. Upon written prior permission by TTG, Participant may use and publish the Results, including Copyrightable Results, after the official announcement of Awards.

A copy of any material proposed for publication must be submitted to TTG prior to publication. TTG shall make its best efforts to make a decision as soon as possible. In case TTG does not make any response, application is deemed to be automatically declined after 30 work days of the date of the application of permission. In case the permission is granted by TTG, Participant can start the publishing procedures.

7.3. Participant undertakes to refer to the Challenge when it uses and/or publishes the Results as follows: This study is performed using the 1-year anonymized mobile communication data made available by Türk Telekomünikasyon A.Ş. within the D4R Challenge," followed by the scientific reference indicated on the Challenge website.

7.4. For the avoidance of doubt, the intellectual property rights in all software, information, technology, or data whatsoever supplied or made available by TTG to the other party under these terms and conditions shall remain the property of TTG. Except as expressly set out in this Agreement, neither party grants to the other any license, sublicense, or other right in or to such intellectual property rights.

Article 8—Applicable Law—Jurisdiction

8.1. This Agreement shall be governed and construed as per Turkish law and the Parties agree that Istanbul (Çaglayan) Courts and Execution Offices shall have jurisdiction over the disputes arising from this Agreement.

Article 9—Miscellaneous

9.1. Participant must provide all necessary security measures required by the applicable law. TTG shall be responsible for ensuring that the transfer of Data under these conditions conforms to applicable laws.

9.2. Neither this Agreement nor any rights granted hereunder shall be assignable or otherwise transferable without the prior written consent of the other Party.

9.3. No waiver or modification of this Agreement will be binding upon the Parties unless made in writing and duly signed by an authorized representative of each Party and no failure or delay in enforcing any right, authority, or privilege will be deemed a waiver.

9.4. In the event that any of the provisions of this Agreement shall be held by a court or other tribunal of competent jurisdiction to be unenforceable, the remaining partitions hereof shall remain in full force and effect. In such case, the Parties shall cooperate, through negotiations in good faith, to replace any part of the Agreement hereto so held to be invalid or unenforceable with a legally binding, effective, and applicable provision.

9.5. Notwithstanding Article 9.3., TTG can unilaterally modify the provisions of this Agreement and Annex-1 any time where such modification is estimated necessary, due to result of a change in applicable law or decisions of competent authorities, and other situations where a modification is required as essential by TTG. For the avoidance of doubt, the Parties agree that TTG has the authority to change the type of Data that are to be shared with Participants under the scope of this Challenge at any time and Annex-1 does not constitute a commitment regarding the type of Data. Participant and Affiliated Scientists agree, declare, and undertake to act in compliance with the unilateral changes made by TTG, and to be liable of any kind of legal and

criminal liabilities, including, but not limited to any loss, damage and administrative fines and other claims made by third parties directed at TTG in case of noncompliance. In case TTG makes any payment to a third party due to noncompliance of Participant, the paid amount with all its ancillaries shall be recoursed to Participant and the Participant shall pay this amount within 2 weeks starting from the date of notification without any further notice. In the case of late, partial, or nonpayment, all statutory and contractual rights of TTG shall be reserved.

IN WITNESS WHEREOF, this Agreement is hereby duly executed by the duly authorized representatives of the Parties in 2 (two) copies on ../../... and each Party receives 1 (one) copy.

Participant title and name:

Participant signature:

Date and place of signature:

Participant email:

Participant phone number:

Names, emails, and signatures of all Affiliated Scientists with access to Data (add as many rows as necessary):

Affiliated Scientist Title and Name	Email	Signature

References

1. (April 2015) Data for Development Senegal: Report of the External Review Panel. Institute of Business Ethics
2. (June 2017) Bilgi Teknolojileri ve letisim Kurumu, Üç Aylik Pazar Verileri Raporu, 2017 Yili 1. Çeyrek. Sektörel Arastirma ve Strateji Gelistirme Dairesi Baskanli. https://www.btk.gov.tr/File/?path=ROOT%2F1%2FDocuments%2FSayfalar%2FPazar_Verileri%2F2017-Q1.pdf
3. Al-Azizy D, Millard D, Symeonidis I, O'Hara K, Shadbolt N (2015) A literature survey and classifications on data deanonymisation. In: International conference on risks and security of internet and systems, Springer, pp 36–51
4. Baldo N, Closas P (2013) Disease outbreak detection by mobile network monitoring: a case study with the D4D datasets. NetMob D4D Challenge pp 1–4
5. Blondel VD, Esch M, Chan C, Clérot F, Deville P, Huens E, Morlot F, Smoreda Z, Ziemlicki C (2012) Data for development: the d4d challenge on mobile phone data. arXiv preprint arXiv:12100137
6. Blondel VD, Decuyper A, Krings G (2015) A survey of results on mobile phone datasets analysis. EPJ Data Sci 4(1):10
7. Cecaj A, Mamei M, Zambonelli F (2016) Re-identification and information fusion between anonymized CDR and social network data. J Ambient Intell Hum Comput 7(1):83–96
8. Gambs S, Killijian MO, del Prado Cortez MN (2014) De-anonymization attack on geolocated data. J Comput Syst Sci 80(8):1597–1614
9. Gundogdu D, Incel OD, Salah AA, Lepri B (2016) Countrywide arrhythmia: emergency event detection using mobile phone data. EPJ Data Sci 5(1):25

10. Lima A, De Domenico M, Pejovic V, Musolesi M (2015) Disease containment strategies based on mobility and information dissemination. Sci Rep 5:10650
11. Machado D (2015) Analyzing geospatial patterns of Syrian refugee flows in southeastern Turkey by use of remote sensing and complementary data. Master's thesis, Institute for Geoinformatics, University of Münster
12. Mari L, Gatto M, Ciddio M, Dia ED, Sokolow SH, De Leo GA, Casagrandi R (2017) Big-data-driven modeling unveils country-wide drivers of endemic schistosomiasis. Sci Rep 7(1):489
13. Martinez-Cesena EA, Mancarella P, Ndiaye M, Schläpfer M (2015) Using mobile phone data for electricity infrastructure planning. arXiv preprint arXiv:150403899
14. Montjoye YA de, Smoreda Z, Trinquart R, Ziemlicki C, Blondel VD (2014) D4D-Senegal: the second mobile phone data for development challenge. arXiv preprint arXiv:14074885
15. Pokhriyal N, Jacques DC (2017) Combining disparate data sources for improved poverty prediction and mapping. Proc Natl Acad Sci 114(46):E9783–E9792
16. Salah AA, Pentland A, Lepri B, Letouzé E, Vinck P, de Montjoye YA, Dong X, Dağdelen Ö (2018) Data for refugees: the D4R challenge on mobility of Syrian refugees in Turkey. arXiv preprint arXiv:180700523
17. Scharff C, Ndiaye K, Jordan M, Diene AN, Drame FM (2015) Human mobility during religious festivals and its implications on public health in Senegal: a mobile dataset analysis. In: Global humanitarian technology conference (GHTC), 2015 IEEE, IEEE, pp 108–113
18. Sharad K, Danezis G (2013) De-anonymizing D4D datasets. In: Workshop on hot topics in privacy enhancing technologies
19. Silm S, Ahas R (2014) Ethnic differences in activity spaces: a study of out-of-home nonemployment activities with mobile phone data. Ann Assoc Am Geogr 104(3):542–559
20. Taylor L (2016) No place to hide? The ethics and analytics of tracking mobility using mobile phone data. Environ Plan D: Soci Space 34(2):319–336
21. Tomaszewski B (2014) Geographic information systems (GIS) for disaster management. CRC Press
22. Tompkins AM, McCreesh N (2016) Migration statistics relevant for malaria transmission in Senegal derived from mobile phone data and used in an agent-based migration model. Geospatial health 11(1s)
23. Trestian R, Shah P, Nguyen H, Vien QT, Gemikonakli O, Barn B (2017) Towards connecting people, locations and real-world events in a cellular network. Telemat Inf 34(1):244–271