USE CASE 11: Big Data frameworks for processing and analyzing mobile phone data

Reporter: Sanja Brdar (University of Novi Sad)

Current collaborators on the use case: Olivera Novović (University of Novi Sad), Apostolos Papadopoulos (Aristotle University of Thessaloniki)

cHiPSet members that expressed interest to participate in this use case: Marco Aldinucci (University of Torino), Siegfried Benkner (University of Vienna), Ari Visa (Tampere University of Technology), Edgars Celms (University of Latvia), Natalija Stojanovic (University of Nis), Luís Veiga (University of Lisboa), Pierre Kuonen (School of Engineering of Fribourg), George Suciu (Beia Consult International, Romania)

Domain description

Mobile phone service providers collect large amount of data. Every time user makes interaction by mobile phone (SMS or call), a Call Detail Record (CDR) is created in Telecom operator database. CDRs log the user activity for billing purposes and network management, but provide opportunities for different applications such as urban sensing, transport planning, social analysis and monitoring epidemics of infectious diseases, etc.

Data sources

Data are confidential, but there are few open anonymised data sets that can be used to benchmark different Big Data frameworks. For this use case we process and analyze data provided by Telecom Italia that are the result of a computation over the Call Detail Records (CDRs) generated by the cellular network. Data provide information about the telecommunication activity (SMS, Call, Internet) over the city of Milan and the Province of Trentino. Also data sets include information regarding the directional interaction strength between the different areas of the city based on the calls exchanged between users. Data encompass two month period and scale of available data is ~ half TB.

Aim of the use case

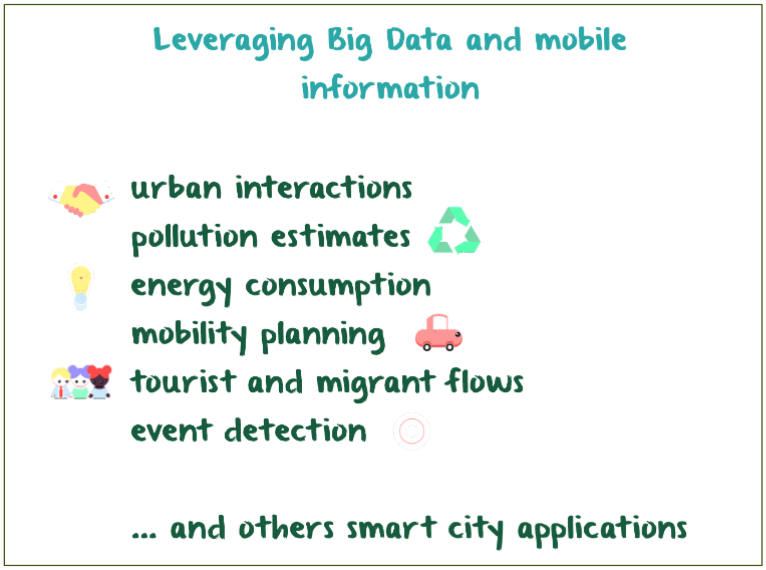
Aim of the use case is twofold:

1. Exploring possible Big Data frameworks such as Spark and Hadoop for analyzing very large data sets provided by mobile network operators; Benchmarking of different

infrastructures (Multi Core CPUs, clusters, GPUs) and parallelizing algorithms for processing and analytics of mobile phone data.

2. Exploring activity patterns and social connectivity graphs that are produced from mobile phone data; Data fusion with other sources (energy, pollution, weather…) in order to leverage mobile phone data for different applications (Table 1): urban sensing, mobility planning, energy consumption and pollution estimates, event detection and other smart city applications.

Table 1: Possible application with mobile phone data



Leveraging Big Data and mobile information

urban interactions
pollution estimates
energy consumption
mobility planning
tourist and migrant flows
event detection

... and others smart city applications

Current work

Typical workflow for processing spatio-temporal data, such as mobile phone data used in this case study, contains numerous queries across locations and timestamps of interest, spatial/time aggregations and summarization. For example, result of the spatial aggregation on the city administrative units and time aggregation on a day interval is presented in Figure 1. Figure includes two selected days (typical workday and weekend). To obtain presented interaction graphs, we first performed spatial aggregation of the data from 10000 grids to 88 city units and time aggregation from 10 minutes intervals to a day intervals. The next step in processing is the creation of pairwise matrix of interaction strengths between spatial units for each available day. Current setting with centralized MySql database and Python scripts that connect to database and run queries does not scale with increased quantity of data and complexity of analytics methods.

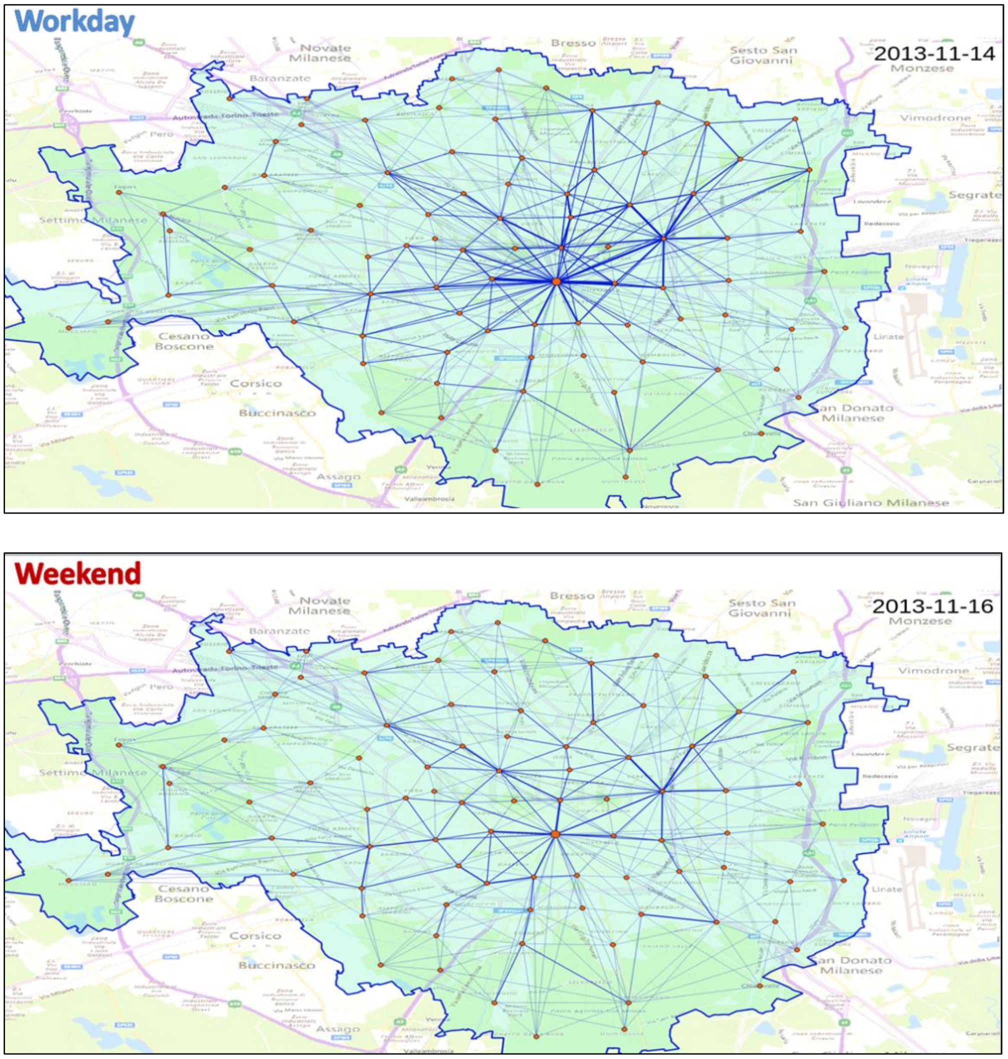


Figure 1: Significant interaction strengths across Milan city

In the current research we experiment with Apache Spark framework with aim to deliver performance evaluation for different processing flows (different spatial and time resolution, different analytics methods applied on the top of the obtained graphs, e.g. community detection, evolving graphs patterns…). The research started in February 2017 with cHiPSet STSM (University of Novi Sad, Serbia -> Aristotle University of Thessaloniki, Greece).

Our further work will include exploration of massive parallelism of the GPU and computer clusters. Besides batch processing of all available mobile phone data, we are also interested in stream processing frameworks that can provide real time analytics. The overarching goal is to evaluate different system architectures and Big Data frameworks in order to find appropriate solution for processing and analyzing mobile phone datasets in efficient and scalable way.