Project Title:

Mobile Phone Activity in Milan

Abstract:

Human activities such as using mobile devices to connect to Internet and other humans produce unprecedented amount of digital records. These types of data can be used in a wide range of problems including mobility planning, tourist flows, urban structures and interactions, event detection, urban well-being and many others. Furthermore, these data can be used for the purpose of cellular network diagnostics and maintenance. In this project, one of the richest open source dataset ever released on two geographical areas is explored and visualized. The dataset is composed of telecommunications data from the city of Milan and the Province of Trentino [1]. One of the main goals of this project is to provide a visualization tool to wireless network engineers to enable them to easily track mobile phone user activities that can in turn helps to diagnose cellular network issues and fix them. To achieve this goal, a graph representation of the mobile phone activity was chosen where nodes represent the cellular towers and the edges represent the activity of the users. Furthermore, this tool should also provide a user friendly representation of the mobile phone user activities for presentation of results to nontechnical audiences. To achieve this goal, a Great Circles representation of the data in a geographic layout was chosen.

As the dataset is large and has many dimensions (domestic vs. international, 7 days of a week, different hours of a day, different types of transmitted data, ...), the visualization tool should be highly interactive. For this purpose, an application was developed using Shiny (R) which provides several filters to focus on the subset of the data of interest. This app also provides the two main types of visualization of the mobile phone activities (Graph and Great Circles). This highly interactive tool can be used to easily spot congested areas/cells of the city (event detection), derive the pattern of mobile phone activities through different days of the week and hours of the day, spot inactive cells,... which in turn are valuable information to main a healthy cellular network. As mentioned before, this tool can also be used to visualize the phone activities on a geo layout (using great circles) for the purpose of presentation and also to see the mobile phone usage patterns more clearly.

Introduction:

The rapid and universal adoption of mobile phones and the exponential increase in the use of Internet services is generating an enormous amount of data that can be used to provide new fundamental and quantitative insights on socio-technical systems [1]. The Call Detail Records (CDRs) can be used to extract human mobility patterns [2], social interactions [3], estimates population densities [4], models cities structures [5], and models the spread of diseases [6]. These CDRs can furthermore be used to gain insights into the health of the wireless cellular networks based on mobile phone usages and detect any potential issues.

The current standard way of analyzing network performance and diagnosing issues are done by cellular system engineers by looking the the CDRs and manually (using Excel) identifying the issues (for example the highly overloaded cells based on the number of active users). This process is very cumbersome and time consuming due to the very high volume of the data. This issue can be alleviated by providing the user friendly representation of the data that can complement (or even replace) the manual analysis. One of the easy to understand representation of the CDR data is graph where nodes represent the cell cites and the edges represent the human activities using their cell phones. This representation can help cellular system engineers to readily and visually identify the troubled cells (for example highly overloaded cell) by looking at the structure of the graph and the number of the connections of each cell cites (number of edges connected to each node). This graph representation can also be used to detect the mobile phone user usage pattern across different

geography and time and this can be used to schedule major cell site maintenance (when the cell site services have to be shut off). This graph representation can also be used to identify the dormant cell sites (cell sites with minimum user activities) and changing their configurations to serve more users. As the graph representation might not be very easy to understand, a great circles representation of the data on a geo layout can also help to visualize the mobile phone activities in an easier to digest way. This visualization can be used for presentation purposes to nontechnical audiences. In the next section, the dataset will be described in details.

Dataset:

The Mobile Phone Activity dataset is composed by one week of Call Details Records (CDRs) from the city of Milan and the Province of Trentino (Italy) [7]. The Mobile phone activity dataset is a part of the Telecom Italia Big Data Challenge 2014, which is a rich and open multi-source aggregation of telecommunications, weather, news, social networks and electricity data from the city of Milan and the Province of Trentino (Italy). The original dataset has been created by Telecom Italia in association with EIT ICT Labs, SpazioDati, MIT Media Lab, Northeastern University, Polytechnic University of Milan, Fondazione Bruno Kessler, University of Trento and Trento RISE. The dataset available on Kaggle is a subset of this telecommunications data. The complete version of the dataset is available online [8].

Every time a mobile phone user engages in a telecommunication interaction (sms, voice call, Internet session), a Radio Base Station/Cell (RBS) is assigned by the operator and delivers the communication through the network. Then, a new CDR is created recording the time of the interaction and the RBS which handled it. The following activities are present in the dataset:

- Datetime: Date in yyyy-mm-dd HH:ii format
- CellID (Source): identification string of a given square of Milan GRID
- countrycode (Target): the phone country code of the target destination
- received SMS: activity proportional to the amount of received SMSs inside a given square id and during a given Time interval. The SMSs are sent from the nation identified by the country code
- Sent SMS: activity proportional to the amount of sent SMSs inside a given square id during a given Time interval. The SMSs are received in the nation identified by the country code
- Incoming calls: activity proportional to the amount of received calls inside the square id during a given Time interval. The calls are issued from the nation identified by the country code
- Outgoing calls: activity proportional to the amount of issued calls inside a given square id during a given Time interval. The calls are received in the nation identified by the country code
- Internet activity: number of CDRs generated inside a given square id during a given time interval. The Internet traffic is initiated from the nation identified by the Country code

In particular, Internet activity is generated each time a user starts an Internet connection or ends an Internet connection. Moreover, during the same connection a CDR is generated if the connection lasts for more than 15 min or the user transferred more than 5 MB. The dataset includes both domestic (from Milan to other provinces of Italy) and international (from Milan to other countries) data. The domestic data only includes the incoming/outgoing calls while international data has all the aforementioned activity records.

The datasets is spatially aggregated in a square cells grid. The area of Milan is composed of a grid overlay of 10,000 (squares with size of about 235×235 meters. This grid is projected with the WGS84 (EPSG:4326) standard [1]. The data provides CellID, CountryCode, Province name and all the aforementioned telecommunication activities aggregated every 60 minutes.

Data Wrangling:

The raw dataset (shown below) comprises of a set of source and target ids with the respective activities between these two. This dataset needs to be converted to a set of edges and nodes.

datetime	CellID	countrycode	smsin	smsout	callin	callout	internet
2013-11-01 00:00:00	1	0	0.3521			0.0273	
2013-11-01 00:00:00	1	33					0.0261

International CDR Sample

datetime	CellID	provinceName	cell2Province	Province2cell
2013-11-01 00:00:00	1	MILANO	0.1894	0.0541
2013-11-01 00:00:00	1	PAVIA	0.0273	

Domestic CDR Sample

The nodes dataset was created by concatenating the CellID and countrycode (province for domestic dataset) columns. A label was also added to each node based on the type of the node (Milan cells, domestic cells, international cells). The edges dataset was created based on the raw activity dataset (shown above). The datetime was parsed to derive the day and the hour of the activity per each CDR (row). The value of the activity of each CDR is treated as the edge weight for the purpose of the visualization.

Other Sources of Data:

The countrycode provided in the dataset had to be converted to the country name for the ease of understanding. This was done by merging the node and edges with the country code dataset in [9]. Furthermore, for the Great Circle visualizations, the latitude and longitude of the countries, Italy provinces, and the cells located in Milan is needed. The latitude and longitude of the countries and the Italian provinces can be derived from the maps library in R. The latitude and longitude of the cell grids in Milan is provided as a Geojson file in [7]. These two data sources are merged with nodes and edges to get the (lat,long) pairs.

Tasks:

As mentioned before, there are two major goals for this projects. One is to provide a visualization tool to wireless network engineers to help to diagnose cellular network issues and fix them. The second goal is to provide a user friendly representation of the mobile phone user activities for presentation of results to nontechnical audiences.

To achieve the first goal a node-link diagram is chosen. Therefore the raw data should be transformed to a set of nodes and edges. This node-link diagram has the following properties:

1. Several filters to choose the type of the data; domestic vs international, sms vs call vs Internet.
2. Zoom in/out capability.
3. Interaction with the graph by selecting nodes and highlighting their connected edge as well as ability to move around the nodes to desired coordinates.
4. Adding time sliders to choose the time interval within a day, day of the week and animation across time.
5. Adding labels to nodes and edges based on user input as well as ability to choose the desired graph layout.
6. Ability to show only the most significant edges based on a user input threshold for congested cell analysis
7. Ability to hide nodes and edges on drag for easy interactions.

The audience for this visualization are cellular network engineers trying to debug and maintain a cellular network. It will be mainly used for:

1. Identifying the troubled (congested) cells based on their incoming/outgoing flows of data.
2. Identifying the idle (inactive) cells based on the user activities and optimizing them to serve more users.

3. Identifying the cell use pattern across days of the weeks and hours of the day to schedule maintenance time (when the cell has to be shut down) in low traffic time intervals.
4. Identifying the data usage patterns across time and geography to optimize the cells dynamically based on mobile phone user's usage.

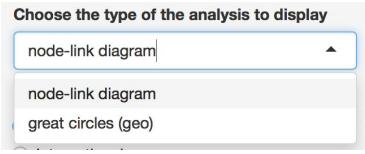
To achieve the second goal, a Great Circles visualization on a geo layout was chosen. This visualization also exploits the transformed nodes and edges data and has the following properties:

1. Several filters to choose the type of the data; domestic vs international, sms vs call vs Internet.
2. Adding time sliders to choose the time interval within a day, day of the week and animation across time.
3. Ability to show only the most significant edges based on a user input threshold for congested cell analysis

The audience for this visualization is mainly nontechnical users who are seeking to find mobile phone user's usage patterns across time and geography. This patterns can be studied alongside other sources of the data (like census data) for socio-technical analysis (like targeted marketing).

Solutions (and Some Results):

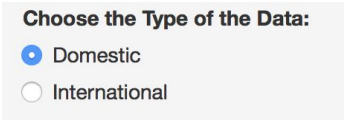
To visualize the mobile phone activities, the node-link diagram was chosen (over adjacency matrix) because the graph is a star-shaped graph with its central nodes connecting to outer layer nodes and in this layout, node-link diagram gives a more complete pictures where comparing the connection pattern of central nodes are easier. Also the great circles on a geo layout can serve as an easier to digest visualization for the purpose of presentation. The user can choose between these two types of visualizations using a drop down menu at the beginning of the app:



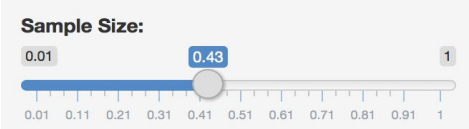
Depending on the chosen type of visualization, the rest of the options gets updated (different filters, labels,...). Next we will go over the details of each of these two visualizations.

Node-Link Diagram Visualization Design:

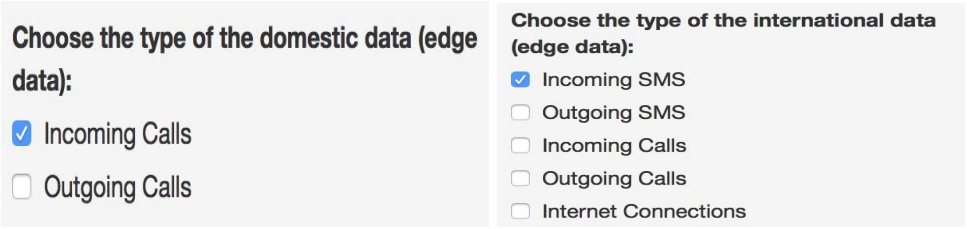
The two main category of the data are domestic (from Milan to other provinces of Italy) and international (from Milan to other countries). These two types of data have different features. The domestic data contains only the incoming/outgoing call while the international data has incoming/outgoing sms and Internet activities as well. Therefore, a radio button is offered for the user to choose the type of the data. Depending on the chosen data type, the rest of the layout will be different (due to different features of the data):



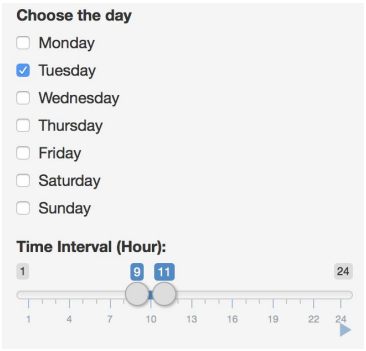
A big challenge with this dataset is the large volume of it. This can cause slow user interactions. As in many situations, a carefully chosen subset of the data can gives the same insights into the data as the whole and makes it much easier and faster to interact with the data. For this purpose, we reduce the items using a sampling slider window in the app where the user can choose any fraction between 0.01 and 1. This fraction will be used to uniformly and at random sample the data for each day separately. This will help to get a subset of the data that has almost the same distribution of the whole dataset:



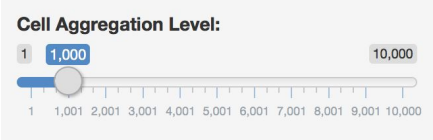
As mentioned before, the user activity (incoming sms, outgoing call,...) is shown as edges. As the network diagnostics are usually done per distinct type of data (sms, call, Internet) a checkbox group is offered on the app for the user to filter the data by reducing the attributes of interest. The width of the edges is proportional as the magnitude channel of the data type. If multiple data type is chosen, the edge width will be the sum of the magnitude channel of all the chosen data types:



The dataset includes the hourly data for 7 days of a week and is dynamic in nature. As the study per day and per hour of the day is the focus of a cellular system engineer (to diagnose network issues at specific time intervals and days), a checkbox group is offered where for the user to reduce the items to the days of the interest. For the view change over time of the day, a time slider is added where the user can select the time interval of the choice. There's also an animation function added to this time slider (play button on the lower right corner) where the user can see the animated transitions and change of the graph over time:

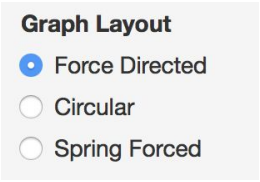


There are 10000 cells in the city of Milan and this can translate to 10000 nodes in the node-link diagram. This can cause severe occlusion and a dense graph that is hard to analyze. Furthermore, the cellular network studies are usually done per areas and not necessarily per cell. To overcome this, the concept of aggregation is used to reduce the nodes. The nodes are reduced by aggregating them into groups of cells (changing the CellID to CellID%aggregation_level). In this way, multiple cells can be visualized as only one node and study them as an area (instead of a single cell). The choice of the aggregation ratio is extended to the user as a slider window it can very well depend on the specific case under study:

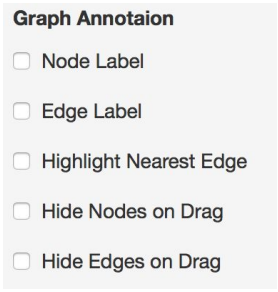


The graph layout can also be selected by the user using a radio button. A force directed graph can gives insight into the highly connected nodes and push separate the inactive cells. A circular layout is helpful when it is desired to have fixed coordinates for the nodes. A spring forced layout is also available to observe the dynamics of the graph into equilibrium state. In the force directed and

spring forced layout, the force is the weights of the edges:



Nodes and edges are labeled based on the user interaction. The name of the country/province/cell number are used for labeling nodes. The edge labels is their corresponding weight. Interaction with the graph by selecting nodes and highlighting their connected edge (based on user input) is also possible. The user is able to select nodes and move them to desired coordinates. An option is available to hide the edges or nodes when dragging the graph (helps in placing the nodes in desired coordinates without having occlusion caused by edges):



In cellular network analysis, the edge weight is of significant importance as it is a measure of the amount of activity and the load of the cell. Sometimes, it is important to only consider the most active edges (with the most amount of activity.weight). For this purpose, a slider window is introduced where the user can choose to only visualize the edges with the most significant weight. This window slider is based on the percentile of the distribution of the edges:

The user is also able to zoom in and out of the graph by scrolling.

Great Circles Visualization Design:

Other than the graph layouts and graph annotation sections, all the other parts of the visualization design of the node-link diagram applies to the Great Circles visualizations. In this visualization the edges and nodes have the same meaning as in node-link diagram. However the coordinates of the nodes are the latitude and longitude of the cells (integrated using the third source of data). The edges are arcs drawn between two nodes. This visualization is also on top of a geographic map. In the domestic dataset, this map is the choropleth map of the Italy provinces and in the international dataset, this map is the choropleth map of the world. Different color was used for different data types (incoming sms, Internet,...) using Rcolorbrewer (qualitative color). If more than one data type is chosen, the back color is used for the edges.

Implementation:

As mentioned earlier, this project was implemented in R. Shiny library was used to create an application. The app is consisted of two main files. The ui.R defines the layout of the app and create an HTML file with the defined elements (filters, sliders, plots,...). The server.R is where all the main implementation and design choices happen. server.R gets the user input from ui.R and do all the data preprocessing the produces the visualizations and pass them as input to ui.R. ui.R in turns render the visualizations as an HTML file.

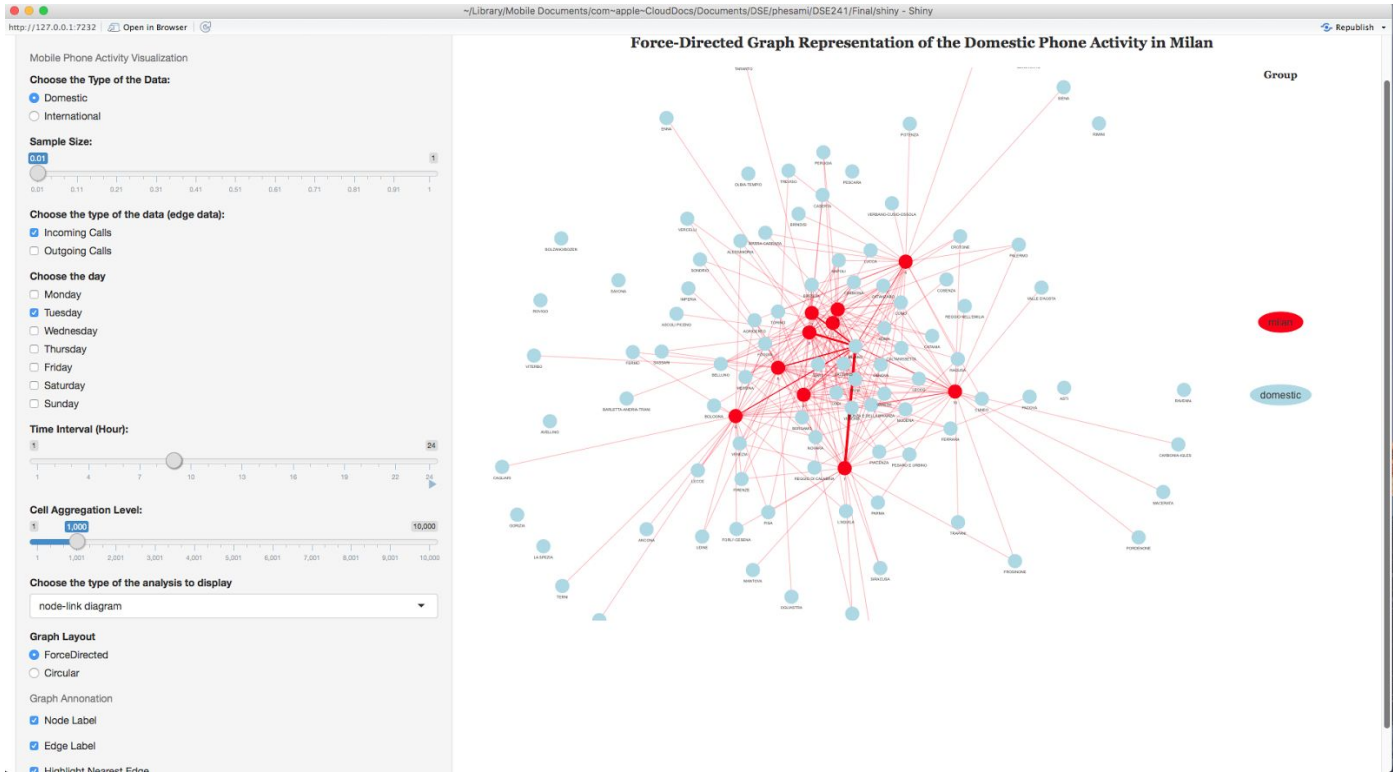
Software Usage:

There are a couple of ways to run this application. The app has been deployed on the shiny

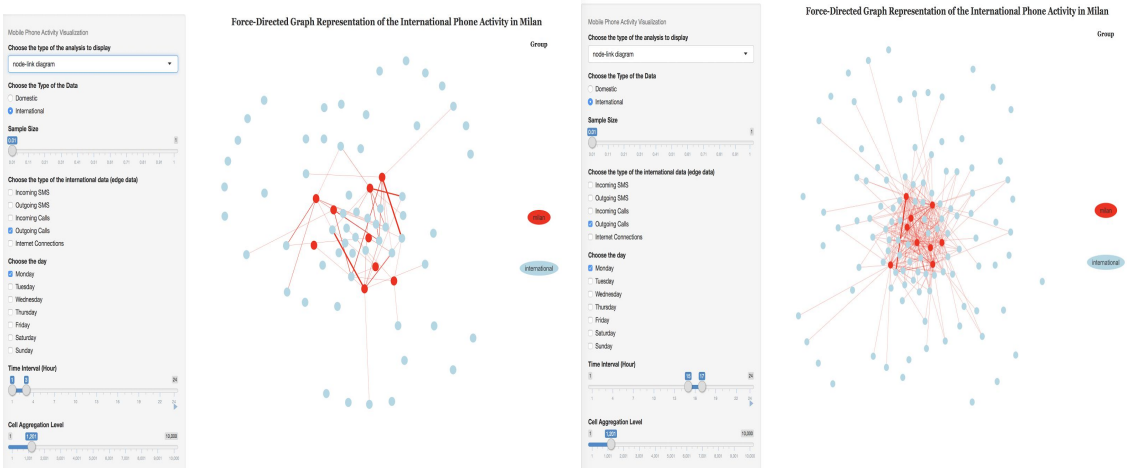
server and the simplest way is to open the app at : https://peymanshiny.shinyapps.io/milan_phone_activity_shiny_dse241_peyman_hesami/. However, the shiny server connections (free tier) are not stable and for high volume apps like this it gets disconnected very often. The second easiest way to run the app is to open either one of ui.R or server.R files in Rstudio and run the app using the *Run App* buttons on the upper right corner of the main window. This will launch the app in a separate window. The app can also be opened in the default browser by hitting the *Open in Browser* button on the upper left side of this window.

Results:

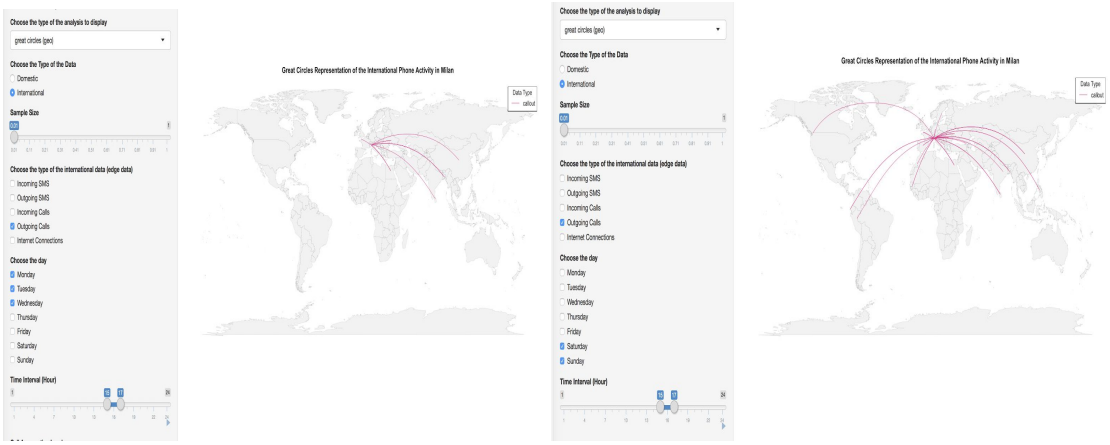
Here is the overall layout of the app:



The node-link representation of the data can help in identifying the congested cells based on the amount of activities at any given time interval. For example by animating the change over time cellular network engineers can identify the congested and highly loaded cells at any given time and optimize the network for optimum usage. In the 00:25 to 00:50 section of the demo video, one can readily see the activities slows down and cells become more inactive after midnight and after 6am the cells starts getting more activities and load. This pattern can change from day to day (weekend vs weekdays) and this can help engineers to identify the idle time intervals for cells and schedule maintenance during these time intervals. It also can be used to detect events (areas with highly loaded cells) during any time intervals and make appropriate network configuration changes to uniformly distribute loads/users across different cells. The following two snapshots shows the difference in network loads and connectivity between after midnight (left) and afternoon (right) on a Monday where the more load and congested cells can easily (and as expected) be seen in the afternoon timeframe.



The great circles visualizations can also be used to identify the pattern of mobile phone user's usage across time and geography. The snapshots below shows the difference between the outgoing calls usage between the weekdays (left) and weekend (right), where most of the significant communications in the afternoon on weekdays are happening with eastern asia while on the weekends is more uniformly distributed across the world.



Challenges:

The reading of the CSV files into memory and extracting the hour from the datetime was the main source of processing delay in this app causing very slow user interactions. To overcome this, an efficient R library (data.table) was used to read in the data (instead of the standard read.csv) and also the extraction of hours and day was done using regular expressions (instead of standard extraction from POSIXct time format). Another way of making the app faster was by used only a portion of data through random sampling (as explained before).

Shortcomings and Potential Improvements:

Although the great circles visualization provides very meaningful insight into the data, it is not interactive (where user can move the objects around) and in some cases where the data density is high, the overlap of the edges might cause occlusion. This can be alleviated by making the visualization interactive where the user is able to modify the the location of different objects (source and target) to be able to overcome any possible occlusion. The great circle visualization also does not offer zoom (although a zoom implementation was tried out but shiny is not supporting this for static images).

Furthermore, the presented datasets can be enriched by using census data provided by the Italian National Institute of Statistics (ISTAT) [10]. This dataset is composed of four parts: Territorial Bases, Administrative Boundaries, Census Variables, and data about Toponymy.

Bibliography:

1. Barlacchi, Gianni, Marco De Nadai, Roberto Larcher, Antonio Casella, Cristiana Chitic, Giovanni Torrisi, Fabrizio Antonelli, Alessandro Vespignani, Alex Pentland, and Bruno Lepri. "A multi-source dataset of urban life in the city of Milan and the Province of Trentino." Scientific data 2 (2015).
2. Song, C., Qu, Z., Blumm, N. & Barabasi, A. Limits of predictability in human mobility. Science 327, 1018–1021 (2010).
3. Miritello, G., Rubén, L., Cebrian, M. & Moro, E. Limited communication capacity unveils strategies for human interaction. Scientific Reports 3 (2013).
4. Deville, P. et al. Dynamic population mapping using mobile phone data. PNAS 111, 15888–15893 (2014).
5. Louail, T. et al. From mobile phone data to the spatial structure of cities. Scientific reports 4 (2014).
6. Wesolowski, A. et al. Quantifying the impact of human mobility on malaria. Science 338, 267–270 (2012).
7. Mobile Phone Activity Dataset: https://www.kaggle.com/marcodena/mobile-phone-activity/version/4
8. Mobile Phone Activity (Full) Dataset: http://go.nature.com/2fz4AFr
9. Country Phone Code Dataset: https://www.worlddata.info/downloads/
10. Italy Censu Dataset: http://www.istat.it/it/archivio/104317