

Systems ad methods for big and unstructured data

Project name

Group Members:

Avci Oguzhan

Gentile Nicole - 10594355

Rigamonti Davide

Singh Raul

Tagliaferri Mattia

# Problem Specification

We represented a population of XXX people in the USA.

The database receives data coming from an application that uses sensor to understand whether two people using the application had a contact; data includes date and time of the contact.

Some places, like restaurants, theaters and hospitals, collect date and name of visiting people. Moreover, hospitals also record people who are hospitalized for covid reasons.

The idea is to use this database so that, if a person gets positive, we can understand who are all people who had a contact with him/her. Data are recorded from 02/2020 and can be used for analytical purposes.

# Hypothesis

Vaccine date, date of the last contagion, date of the last negative test fields are optional fields: a person is positive if the contagion date is after the date of the last negative test; people can do tests without being infected; people can decide not to get the vaccine; vaccinated people can get infected.

Members of a family are all the people who live together, including, for example, roommates.

When adding data about people going to places, we do not consider distances between the city where they live in and the location they go to. So, they can reach different places in different cities during the same day, assuming they do this in different moments.

Regarding places we have: people who got covid after 18/10 went to places from 18/09 to 17/10, since we assume that later they are in quarantine; all the others went to visit locations from 18/09/2021 to 17/11/2021. We don’t retain previous data since it is useless for the queries.

For simplicity, for people who are in hospital date of hospitalization and contagion coincide.

3) Ogni edificio può avere al massimo 20 visite da persone diverse (range(6,10) + range(6,10)

4) Non si considerano le visite in ospedale, solo i ricoveri

6) Persone diverse ed edifici diversi non possono essere in più città diverse

# ER

….

# Dataset description

We used three types of nodes: Person, Location and City.

People are characterized by their name and surname, birthdate, city, vaccine date, date of the last contagion, date of the last negative test. People can be linked by family relationships.

Locations have a name and a type (restaurant, theater, hospital); type is used so that the dataset can be easily expanded with new types of location.

Possible relationships are:

* WENT\_TO to track people who visited a certain location;
* IN\_FAMILY;
* IS\_IN between locations and cities;
* LIVES\_IN to link people to the city they live in;
* HAS\_MET to indicate contacts between people who use the application;
* IS\_HOSPITALIZED\_IN to indicate people are/were hospitalized for covid reasons (if they are currently negative, then they are not in the hospital).

# How we created

To create the nodes, we used some csv files containing data about people, name of places and cities. People’s csv contains recovered and infected people, as well as people who have never got coronavirus. Then, we randomly added relationships among nodes: family members, contacts and visits.

# Write basic Queries (minimum 5) and Commands (minimum 3)

…