Theoretical challenge: low

Computational challenge: moderate; programming experience is necessary.

What is expected from me, and how can I get a high grade?

You are encouraged to work in group of two students. Please clearly identify the names of your partners in the project report. You need only submit one individual report per group of students.

Your report should be uploaded as a PDF file in the following format: lastname1_lastname2.pdf, where lastname1 and lastname2 are the last names of the students working on the project.

Your submission will consist of a report; for each question you will provide the following:

1. A technical description of the approaches that you took to solve the question;
2. Experimental results: where applicable: plots, figures, images, etc; you need to include captions, axes legends with units, colorbar, etc.
3. You will only get credit for an experiment or a result if it includes a **discussion** that connects the experiment with a theoretical analysis.
4. For the experiments, you can use the programming language of your choice: R, Python, MATLAB, etc. You should include your code in the report.
5. Use a standard variable-width font, in 12 pt (e.g., Times New Roman, Palatino, etc.) with 1-inch margins;
6. Use proper spelling and grammar.
7. Grading: for each problem, we provide the approximate percentage of points.

fmeyer@colorado.edu

1 Introduction

In this project, you will train to be a data scientist: you will help a researcher working at the Center for Disease Control (CDC) predict the spread of epidemics using face-to-face contact networks and co-presence networks.

In this project we are interested in *contact networks*. These networks provide temporal information about contacts between individuals in several environments. We model the networks using graphs. Each graph is encoded with an adjacency matrix $\boldsymbol{A}$. The set of nodes if V, and the set of edges. We describe each graph G with the pair

$$G = (V, E). \tag{1}$$

2 The Data

This project is using six datasets collected by English, French and Italian researchers. The data are available at the SocioPatterns website. These datasets are well known and are used to benchmark algorithms in network science and graph theory.

2.1 The Two Types of Networks: Face-to-Face Contact Networks and co-Presence Networks

The data provide temporal information about face-to-face contacts between individuals in several environments. Each individual was wearing a sensor capable to detect face-to-face close range proximity (1.5 m) between the sensors.

name	location	contact			co-presence		
		n	m	$\overline{d}$	n	m	$\overline{d}$
InVS13	French Institute for Public Health Surveillance	92	755	16	95	3,915	82
InVS15	French Institute for Public Health Surveillance	217	4,274	39	219	16,725	153
LH10	Hospital ward (Lyon, France)	76	1,156	30	73	1,381	38
LyonSchool	Primary school (Lyon, France)	242	8,317	69	242	26,594	220
SFHH	2009 French Society for Hospital Hygiene Conference	403	9,565	48	403	73,557	365
Thiers13	High School (Marseilles, France)	327	5,818	35	328	43,496	265

Table 1: Name, origin, number of nodes n, number of edges m, average degree $\overline{d}$ of the face-to-face contact (left) and co-presence (right) networks

In addition to this dynamic face-to-face contact network, the researchers also recorded the location – albeit with a much coarser resolution – of the participants using RFID readers located at various locations inside the environment wherein the participants interacted. There are therefore two datasets, sampled at the same exact instants:

1. a time series of face-to-face contact events between individuals,

2. a time series of co-presence events between individuals.

Table 1 provides the name, origin, number of nodes n, number of edges m, average degree $\overline{d}$ of the face-to-face contact and co-presence networks. In each case, we have access to two matrices that encode face-to-face contacts and co-presence. Each matrix is the adjacency matrix of the face-to-face contact, or co-presence.
A ZIP archive of the six datasets can be retrieved here: click to download. You simply need to load the dataset in MATLAB. For instance,

```
>> load A_LyonSchool.mat
>> whos
  Name                    Size              Bytes  Class     Attributes

  U                       242x242           468512 double
```

and to load the co-presence data,

```
>> load A_pres_LyonSchool.mat
>> whos
  Name                    Size              Bytes  Class     Attributes

  U                       242x242           468512 double
```

3 Mission challenge

Face-to-face contacts might be difficult to obtain: they require users to carry special sensors. Conversely co-presence information is easily available: it only requires that some RFID readers be installed in a building. Unfortunately, co-presence networks have poor spatial resolution, whereas contact face-to-face have excellent resolution.

In this project, you will compare the precise face-to-face contact datasets to the coarse co-presence datasets. For each dataset, you will first compare the face-to-face contact network with the co-presence networks. Both networks are defined on the same set of vertices, and therefore can be directly compared. The comparison will rely on the computation of network statistics that provide a signature of the specific geometry of each network . We will focus on simple statistics: degree, density, clustering coefficient, and largest eigenvalue.

Ultimately, you would like to answer the following question: can one replace the precise but expensive face-to-face contact data withe the vague but cheap co-presence data? You will answer this question in the context of using either one of the networks to simulate the propagation of an epidemic on the corresponding network.

To answer this question, you will compare the results of simulating the same epidemic on both networks. Using the result of the epidemic on the face-to-face network as the gold standard, you will decide whether the outcome of the the epidemic on the co-presence network could be used instead.

4 Comparing the networks using local and global statistics

Let G be weighted graph $G = (V, E, \boldsymbol{w})$, we define

- size = number of vertices = $n \stackrel{\text{def}}{=} |V|$,
- $m \stackrel{\text{def}}{=}$ number of edges,
- volume = sum of the weights along all the edges $\stackrel{\text{def}}{=} \sum_{e \in E} w(e)$.

In addition, we define the neighborhood of a vertex v as the set of vertices that are directly connected to v by and edge. Please note that v is not in the neighborhood of v.

Definition 1. *Let $v \in V$. The neighborhood of v is defined by*

$$N(v) \stackrel{def}{=} \{w \in V \setminus \{v\} : (v, w) \in E\}. \tag{2}$$

We now define six statistics that can be used to characterise the geometry of each network

1. density = ratio of the number of edges over the maximum possible number of edges (given the number of nodes)
$$\text{density} \stackrel{\text{def}}{=} \frac{2m}{n(n-1)}. \tag{3}$$
2. Dominant eigenvalue of the adjacency matrix. You will compute λ defined by
$$\lambda = \max\{|\lambda_1|, |\lambda_n|\} \tag{4}$$
where λ_1 and λ are the largest (positive) and smallest (negative) eigenvalue of the graph adjacency matrix,
$$\lambda_1 \geq \ldots \geq \lambda_n. \tag{5}$$
3. Degree distribution
$$\deg(v) \stackrel{\text{def}}{=} \sum_{u:(u,v)\in E} w_{uv}. \tag{6}$$
4. Clustering coefficient. This is the ratio of the actual number of triangles that include v as a node and for which the two other nodes are in the neighborhood of v, over the maximal possible number of such triangles (given the size of $N(v)$).

The clustering coefficient at v can be computed as follows,

$$\mathrm{cc}(v) \stackrel{\text{def}}{=} \frac{2|\Delta(v)|}{\deg(v)\,(\deg(v)-1)} \tag{7}$$

where

$$\Delta(v) \stackrel{\text{def}}{=} \{(u,w)|u \in N(v), w \in N(v), (u,w) \in E\} \tag{8}$$

is the set of edges in the neighborhood of v, $N(v)$.

For an Erdös-Rényi random graph, $G(n,p)$, the clustering coefficient is p in the limit of large n.

Because nodes tends to cluster in social networks (the friend of my friend is also my friend), the clustering coefficient can be used to quantify the "transitivity" of the connectivity.

$$\overline{\mathrm{cc}} \stackrel{\text{def}}{=} \frac{1}{n}\sum_{v\in V} \mathrm{cc}(v). \tag{9}$$

Assignment [120 = 12 x (8 + 2)]

1. For the 12 contact networks $\left(\boldsymbol{A}_f^{(i)}, \boldsymbol{A}_p^{(i)}\right), i = 1,\ldots,6$, evaluate the six statistics described above.

   When you need to compute a probability distribution (e.g., degree, clustering coefficient), please display an histogram. Please see Fig. 1 and 2 for examples of degree distributions.

   You can use a table to display all the other statistics, where you only need to compute a scalar.

2. For each of the six datasets, compare the statistics of the face-to-face networks, with those of the co-presence networks. Explain your findings.

3. For each of the 12 networks, use the statistics to decide whether the network can be modeled as

   - an Erdös-Rényi $G(n,p)$ of the same size n, if so please estimate p;
   - a preferential attachment graph the same size n.

   Please note that some statistics may match one graph model, while others may provide a poor fit with the same model.

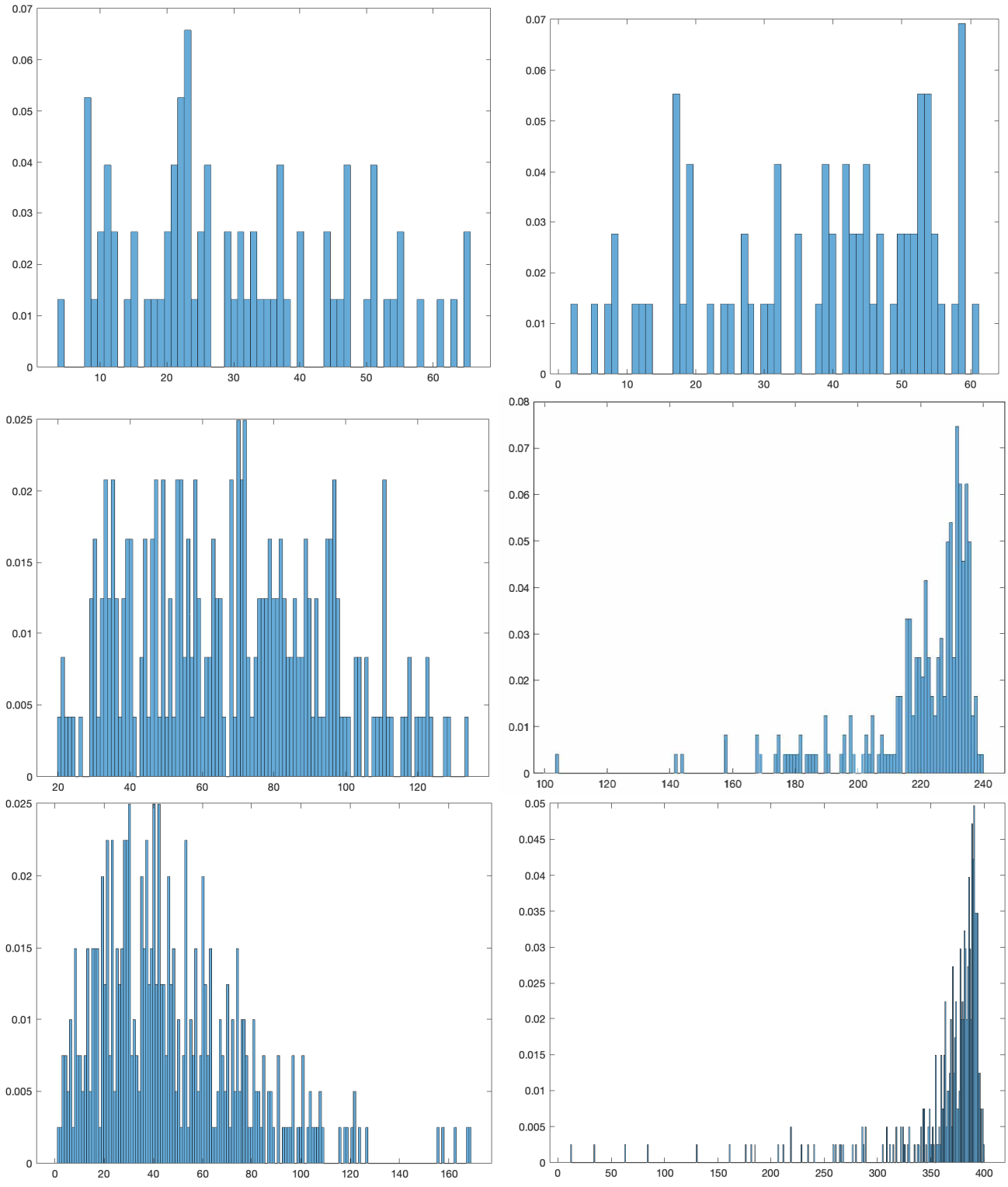

Figure 1: Top to bottom: normalized histogram of the degree distribution for the face-to-face contact (left) and co-presence (right) networks: LH10, Lyon School, and SFHH.

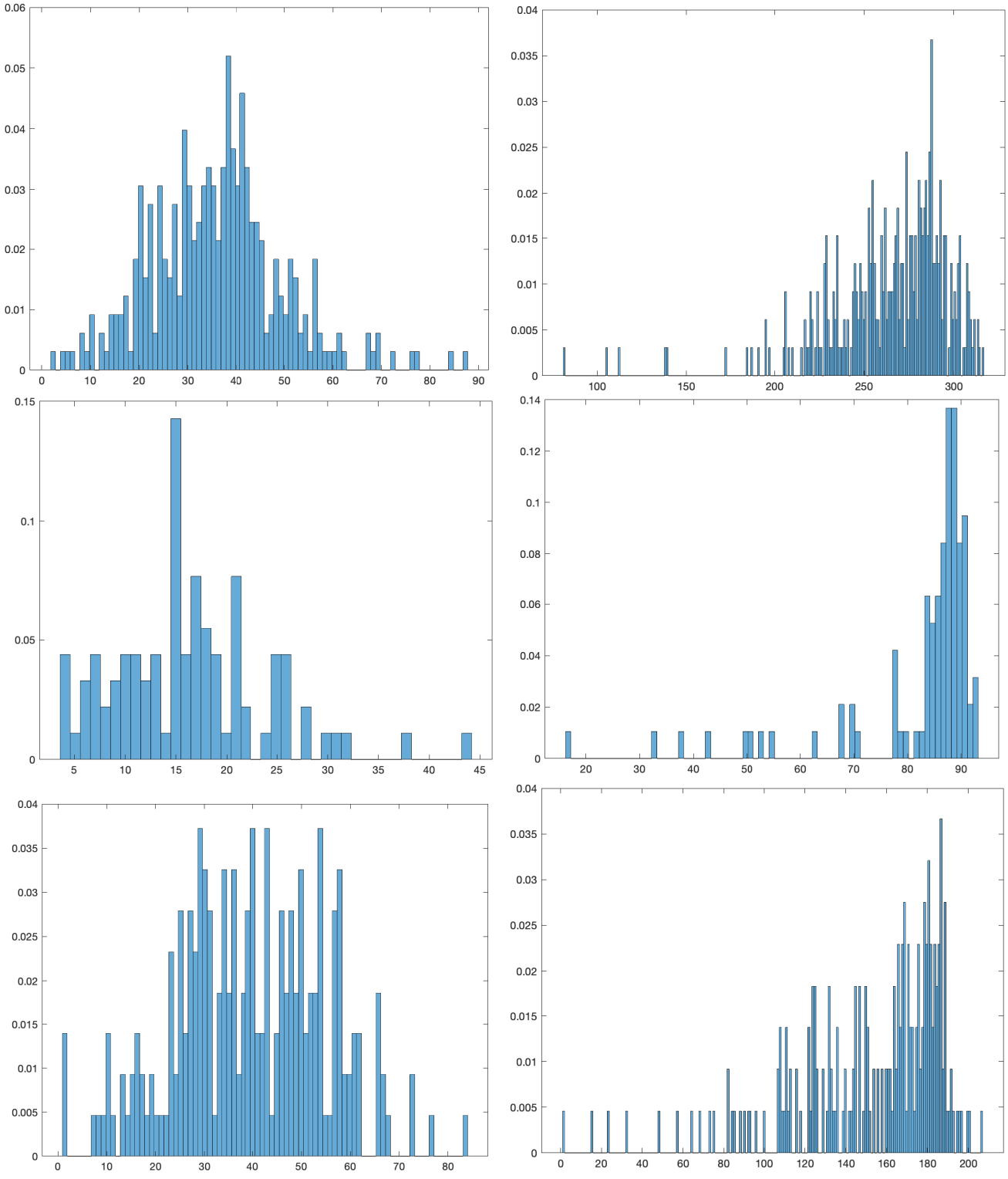

Figure 2: Top to bottom: normalized histogram of the degree distribution for the face-to-face contact (left) and co-presence (right) networks: Thiers13, InVS13, and InVS15.

5 Spread of Epidemics

We now consider a simple model of an infectious disease spreading on the network that is transmitted when two individuals are in contact (either face-to-face or co-presence). Historic examples of such epidemics include the Great Plague of London (17th century), the 1918 influenza pandemic, the Severe Acute Respiratory Syndrome (SARS) outbreak of 2003, the 2009 influenza A (H1N1) epidemic, the 2019 SARS-CoV-2 (COVID-19) pandemic, etc.

Similar mathematical models can be used to model gossips on social networks, or the propagation of computer viruses in the context of cybersecurity.

5.1 The susceptible-infective-recovered (SIR) model on a network

We describe the Kermack-McKendrick epidemic model, also known as the susceptible-infective-recovered (SIR) model. In this model each individual is a node (vertex) of the network. The total number of nodes n is divided into three subsets,

$$n = S(t) + I(t) + R(t), \tag{10}$$

where

1. $S(t)$ = number of susceptible nodes who have not been infected, but have no immunity and thus can be infected if exposed to the disease;
2. $I(t)$ = number of infected nodes who can transmit the infection to susceptible nodes by contact;
3. $R(t)$ = number of recovered (or immune) nodes who were previously infected, but have since then recovered and gained immunity against the infection. These nodes can neither become infected, nor transmit the disease.

The random variables $S(t)$, $I(t)$ and $R(t)$ evolve according to the following dynamics.

- If v is susceptible at time t and it is the neighbor of an infected node u, then v can become infected at time $t + \Delta t$ with a probability proportional to

$$\beta A_{uv} \Delta t, \tag{11}$$

  where A_{uv} is the entry in the adjacency matrix of the contact network, which accounts for the amount of contacts that happened between u and v.

- If v is infected at time t, then v can recover at time $t + \Delta t$ with a probability proportional to

$$\mu \Delta t. \tag{12}$$

- If v is recovered (immune) at time t, then its status no longer changes.

The parameter β (measured in inverse of time units) controls the infection rate: if v is the neighbor of an infected node, then the average time interval for v to become infected is $1/\beta$.

Similarly, the parameter μ (also measured in inverse of time units) is the recovery rate: it takes a time interval of $1/\mu$ for node v to recover once it has been infected, or equivalently the

mean infectious period is $1/\mu$.

To understand the roles of β and μ and their effect on the dynamic of the infection, let us consider the situation where all the vertices are initially healthy but susceptible.

Once a single node is infected, it will infect about $\beta\overline{d}\Delta t$ nodes over the time step Δt, where $\overline{d}$ is the average degree of the graph G. If we integrate this number of infections over the mean infection period, we obtain about $\beta\overline{d}/\mu$ nodes infected by node v.

Now, each infected node will itself infect about $\beta\overline{d}/\mu$ nodes during the mean infectious period. The number

$$\rho_0 = \frac{\beta}{\mu}\overline{d} \tag{13}$$

where $\overline{d}$ is the average degree, is known as the basic reproduction number. ρ_0 determines if an epidemic will occur: if $\rho_0 < 1$, then the infection will die out, whereas if $\rho_0 > 1$ the number of infected vertices will increase exponentially.

In this model, an epidemic will always come to an end when sufficiently many nodes are protected because they have recovered, and gained immunity.

5.2 Numerical simulation of the SIR model on a network

We consider the problem of simulating an epidemic on a graph $G = (V, E)$ with adjacency matrix $\boldsymbol{A}$. To carry out the simulations we need the following variables:

- β: infection rate
- μ: recovery rate
- Δt: time step
- T: time interval for the simulation
- $\boldsymbol{A}$: adjacency matrix

The pseudo-code in Algorithm 1 describes the simulation of the SIR infection on the graph $G = (V, E)$. The source of the infection is a random node v_0.

All epidemics are initiated with a source node chosen at random uniformly amongst the set of vertices. The simulation stops at $t = t_\infty$ when the set of infectious nodes is empty, to wit all vertices are either immune (recovered) or were never infected (susceptible).

The impact of the epidemic is quantified using the fraction of the number of vertices n_r that recovered (and therefore were infected),

$$n_r \overset{\text{def}}{=} \frac{R(t_\infty)}{n}. \tag{14}$$

Specifically, for each simulation of an epidemic, we assess whether $n_r > 0.2$, and if this is the case we record n_r. Such simulations where at least 20% of the population was infected are considered to be large outbreaks. Finally, in all simulations, we use $\beta = 4 \times 10^{-4}$, and we vary μ.

Algorithm 1 numerical simulation of the SIR model

procedure SIR($A, \beta, \mu, \Delta t, v_0, V$)

// Initialize all the variables

    nTimeSteps $\leftarrow T/\Delta t$
    $q := \mu \Delta t$

    susceptibleNodes $\leftarrow V$
    infectiousNodes $\leftarrow \{v_0\}$
    recoveredNodes $\leftarrow \emptyset$

// main loop

    for $t := 1$, nTimeSteps **do**

        for all $v \in$ infectiousNodes **do**

// neighbors of v become infected with probability p

            for all $u \in$ neighbors(v) **do**
                if $u \in$ susceptibleNodes **then**
                    $p := \beta A(u, v) \Delta t$
                    infectiousNodes $\leftarrow v_0$ with probability p
                end if
            end for

// v becomes recovered with probability q

            recoveredNodes $\leftarrow v_0$ with probability q
        end for
    end for
end procedure

In Section 4 you have compared the precise face-to-face contact datasets to the co-presence datasets. Face-to-face contacts might be difficult to obtain; conversely co-presence information is easily available.

The comparison of network statistics is interesting, but does not necessarily yield reliable information about the possibility of using the co-presence datasets as surrogates for the face-to-face contact. In this last part of the project, you study the numerical simulation of an epidemic on the detailed aggregated face-to-face contact, and compare it to a simulation obtained on the co-presence network. Furthermore, you will study if vaccinations policies devised using the coarser co-presence data can be as effective as those that are designed using the precise face-to-face contact network data.

Assignment [720 = 6 x 3 x 20 x 2]

For each of the six face-to-face datasets, perform 100 simulations of the SIR epidemic, using the followings parameters:

- the unique vertex that is the source of the infection is chosen uniformly at random;
- $\beta = 4 \times 10^{-4}$;
- $\mu = 100\beta/k, k \in \{1, 2, 3, 4, 5\}$;
- $\Delta t \propto 1/(\beta \times \text{median}\,(\boldsymbol{A}))$, where $\text{median}\,(\boldsymbol{A})$ is median entry in the adjacency matrix of the network of interest.

7. Plot as a function of ρ_0 (given by (13)) the distributions of recovered nodes.
8. Plot as a function of ρ_0 the fraction of epidemics where the final fraction of recovered nodes, n_r, given by (14), is greater than 20 %.
9. Plot as a function of ρ_0 the average number of recovered nodes for those epidemics where the final fraction of recovered nodes is greater than 20 %.

5.3 Effect of Vaccination

We study a simple model of vaccination where a subset of nodes acquire an immunity to the infection. These nodes cannot transmit the disease and are never infected. You will study two different public health strategies:

1. vaccination of 20 nodes at random;
2. vaccination of the 20 nodes with the highest degrees.

Assignment [720 = 6 x 3 x 20 x 2]

For each of the six face-to-face datasets, perform 10 simulations of the SIR epidemic, using the followings parameters:

- the unique vertex that is the source of the infection is chosen uniformly at random;
- $\beta = 4 \times 10^{-4}$ and $\mu = \beta$;
- 20 nodes are vaccinated at random or
- the 20 nodes with the highest degrees are vaccinated.

11. Plot the ratio between the fraction of the total number of outbreaks that reach at least 20% of the population with and without vaccination.
12. Plot the ratio between the median outbreak size (that reach at least 20% of the population) with and without vaccination.
13. Comment on the best vaccination policy.