Complex Networks report: analysis of the network of E-roads between cities

Author: Canigó Callau i Boix*
Master en Física dels Sistemes Complexos i Biofísica.
Facultat de Física, Universitat de Barcelona, Martí i Franquès 1, 08028 Barcelona, Spain.†
(Dated: June 25, 2023)

Abstract: This report analyzes the network of E-roads between cities, that extends along Europe and part of Asia. We will use some methods to characterize the network and understand its structure, and we will also apply the configuration model to see how it works as a null model of the network. Then we will simulate SIS dynamics in the same network and see how there is a phase with a number of infections close to zero and an endemic phase with a non-zero equilibrium. Between this phases, we will observe there is a critical point, that we will study. Finally we will discuss the results and reach some conclusions.

I. INTRODUCTION

In this report we will use some methods from complex networks to analyze a real network, and then ww will SIS dynamics to it. The first part will deal with the characterization of the network: the number of nodes and links, the degree distribution, the clustering, among other magnitudes. Then we will show the results of using the configuration model on the network. In the second part, the SIS dynamics for modeling epidemics will be briefly explained, and we will analyze the role of the order parameter and the critical behaviour. At the end we will discuss the results and show what conclusions have been reached about the shape of the network and how the SIS dynamics work on it.

The network used is called Euroroads, and represents cities as nodes and their direct connections by an E-road as links. The E-road network is a system of mainly European roads, but that includes also some Western and Central Asia ones, reaching the frontier between Kazakhstan and China in the east and the one between Turkey and Syria in the south. In the majority of countries these roads can be identified by a green sign with "E-" followed to the number of the E-road, although sometimes with another local name too. As a complex network, is unipartite, undirected, unweighted, has no multiple edges and does not contain loops. The links are defined between consecutive cities on the same road, so cities connected by E-roads that have other connected cities with the same road in between are not linked. This have consequences on the structure of the network, making it short-ranged and geometrical embedded in the geography of the places for where the roads go, as we will see. The data set with the edge list of the network is public and available at [1].

* ccallabo51@alumnes.ub.edu
† master.complex.biophys@ub.edu

II. CHARACTERIZATION OF THE NETWORK

This section is based on the notes of the first half of the subject Complex Networks of the master course of Physics of Complex Systems and Biophysics of the Universitat de Barcelona [2]. The results are computed in python, in a notebook called "Assignments" available in a github repository [3].

After having cleaned the data and made the edge list doubled (so each link appear twice, in both directions, to make later computations easily), we can start with the results of assignment 1. The number of nodes is $N = 1174$, and the number of edges $E = 1417$. The average degree is defined as $\langle k \rangle = 2E/N$ and gives a value of $\langle k \rangle \approx 2.41$. In the notebook is also computed as the average of the distribution of degrees (shown in assignment 2), and give the same result, as expected. Finally, the list of degrees gives the number of neighbours for each node, going from 1 to 10, without nodes with 9 neighbours. As long as it is, it will not be shown here, but it is in the notebook as all other results.

In assignment 2 the results are presented in table I and figures 1, 2 and 3. Because there is no nodes with 9 neighbours, the results with $k = 9$ must not be considered. With that said, the magnitudes in the results use the following definitions:

$$P(k) = Prob(k_i = k) \quad (1)$$

$$P_c(k) = Prob(k_i > k) \quad (2)$$

$$\bar{k}_{nn}(k) = \sum_{k'} k' P(k'|k) \quad (3)$$

$$\kappa = \frac{\langle k^2 \rangle}{\langle k \rangle} = \frac{\sum_{k'} k' P(k') \sum_k k P(k|k')}{\sum_k k P(k)} \quad (4)$$

$$\bar{c}(k) = \frac{2}{k(k-1)N_k} \sum_{i \in Y(k)} T_i \quad (5)$$

$$\bar{c} = \sum_k P(k) \bar{c}(k). \quad (6)$$

In the plots we can observe a peaked degree distribution with the peak at $k = 2$. It has a tail that resembles a fat tail, decreasing as a power law, typical of real networks: if we plot the logarithms of $p(k)$ and k from 2 to 7 the dependence is almost linear, but because

TABLE I. Table with the degree distribution, complementary cumulative degree distribution, average nearest neighbours degree, normalized average nearest neighbours degree and degree dependent clustering as a function of the degree.

k	$p(k)$	$p_c(k)$	$\overline{k}_{nn}(k)$	$\overline{k}_{nn}/\kappa(k)$	$\overline{c}(k)$
1	0.162	0.838	2.62	0.87	0.000
2	0.521	0.317	2.86	0.95	0.008
3	0.158	0.158	3.01	1.00	0.030
4	0.096	0.062	3.19	1.06	0.050
5	0.040	0.022	3.34	1.12	0.051
6	0.013	0.009	3.37	1.12	0.040
7	0.004	0.005	3.74	1.25	0.048
8	0.004	0.001	3.33	1.11	0.014
9	0.000	0.001	0.00	0.00	0.000
10	0.001	0.000	2.40	0.80	0.000

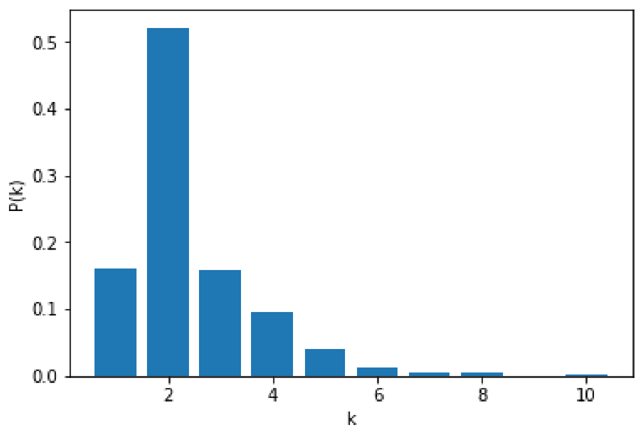

FIG. 1. Bar plot with the degree distribution of the network.

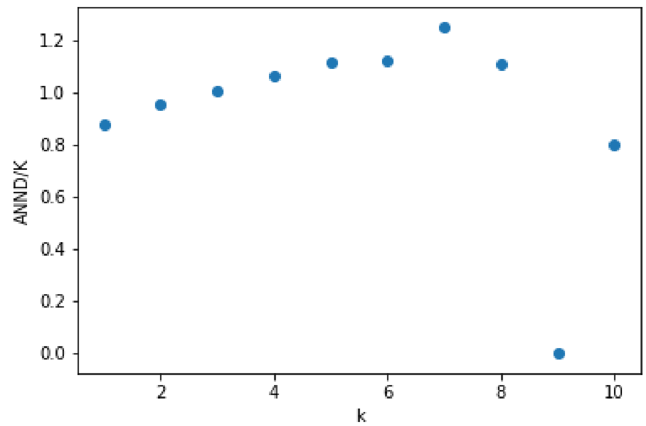

FIG. 2. Scatter plot with the normalized average nearest neighbours degree distribution.

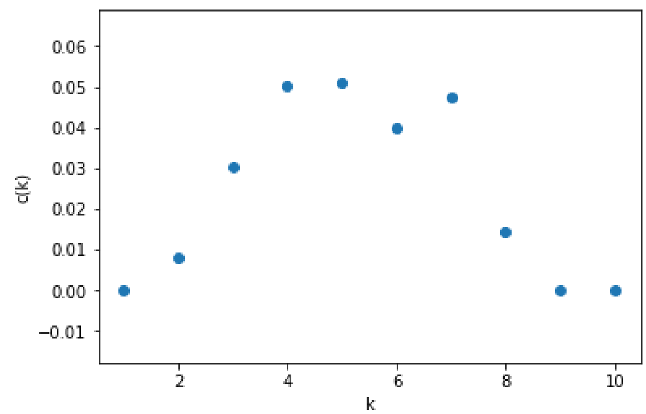

FIG. 3. Scatter plot with the degree dependent clustering distribution of the network.

of the lack of data in larger degrees it is not plotted. Regarding the nearest neighbours average degree, it has a smooth peak at $k = 7$, monotonously increasing in lower degrees and monotonously decreasing in larger ones, if we exclude the empty data of $k = 9$. The normalization factor gave a value of $\kappa \approx 3.00$ In respect to the clustering, is higher in middle, but low in all degrees if we compare it with the values that define weak and strong clustering: $\overline{c}_c(k) = 1/(k-1)$, a function that is far above all the points shown. The clustering coefficient of the network is equal to $\overline{c} \approx 0.017$.

Assignment 3 did not exist and assignment 4 was optional (and my time is limited), so we can jump into assignment 5. This assignment is about applying the configuration model to the network and compute the magnitudes from assignment 2 in the resulting network, for 1 realization and the average over 100 realizations. This model takes the nodes and their number of links and make connections between nodes randomly, without changing the degree of any, and avoiding multiple links between any pair and self-connections. Thus, the number of nodes, links, average degree and degree distribution remain the same after the model is applied.
The numerical results of both 1 and 100 realizations are shown in table II. The plot of the degree distribution is the same as in 1, so there is no need to show it, but is in the notebook anyway. The plots for the normalized average nearest neighbours degree distribution and the degree dependents clustering for 1 realization are 4, 5, and for 100 realizations are 6 and 7.
As we can see in the table, the degree distribution and complementary cumulative degree distribution are the same as in I, as expected. The values of the the normalization constant of $\overline{k}_{nn}(k)$ is $\kappa \approx 3.00$ for 1 realization and $\kappa \approx 3.02$ for 100 realizations (the value of the notebook has en error). In the normalized average nearest neighbours plots there seems to be the same value for all degrees, if we do not consider $k = 9$, what makes sense considering we have eliminated correlation when we applied the configuration model. Divergences from $\overline{k}_{nn}/\kappa = 1$ gets larger with degrees with a low number

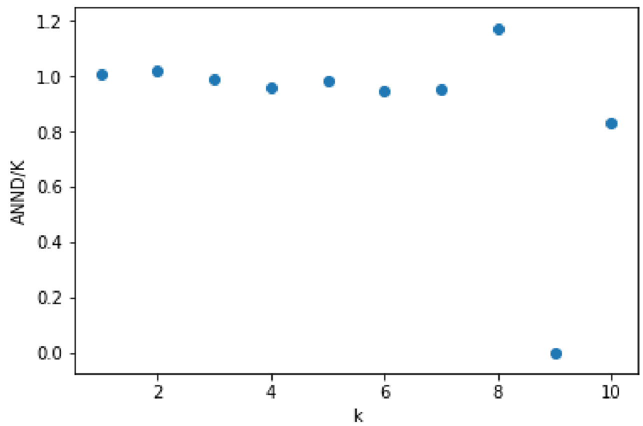

FIG. 4. Scatter plot with the normalized average nearest neighbours degree distribution for the configuration model applied on the network.

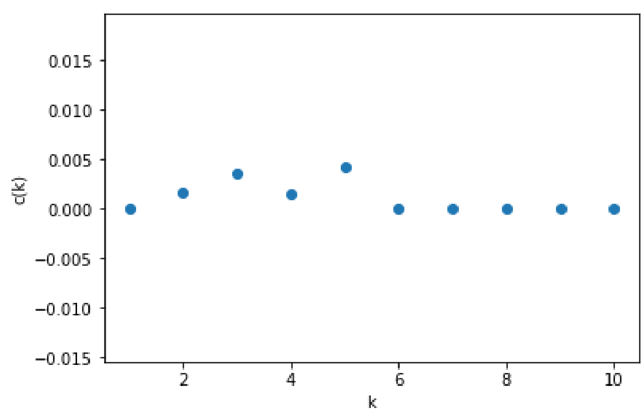

FIG. 5. Scatter plot with the degree dependent clustering distribution for the configuration model applied on the network.

of nodes, what also makes sense, statistically. Regarding the clustering plots, the configuration model had practically eliminated the already low number of triangles there were. There are some triangles created randomly, but as it is seen in 7, there are not statistically relevant, given their standard deviation. The resulting clustering coefficient are $\bar{c} = 0.0017$ for 1 realization and $\bar{c} = 0.0013 \pm 0.0014$ so, again, we can consider the clustering to be not significant.

III. SIS DYNAMICS SIMULATION

This section is about the SIS dynamics, their simulation in the chosen network, and the analysis of the results. It is based on my class notes of the second half of the subject. It is important to remark that the network used here is not exactly the same as in [1] and the last section, because it was not connected: some

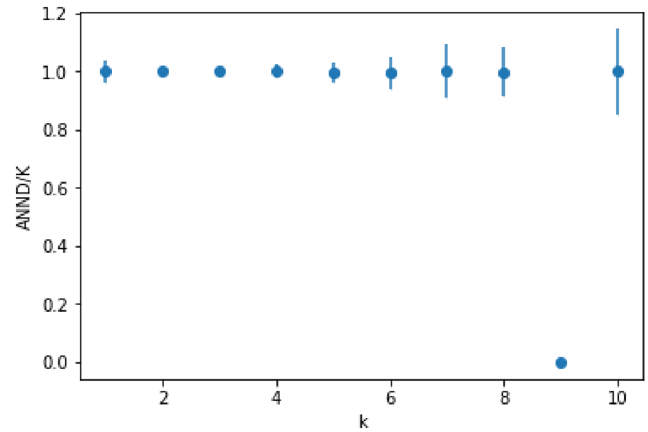

FIG. 6. Scatter plot with the normalized average nearest neighbours degree distribution for the average over 100 realizations of the configuration model. The error bars represent the standard deviation over realizations.

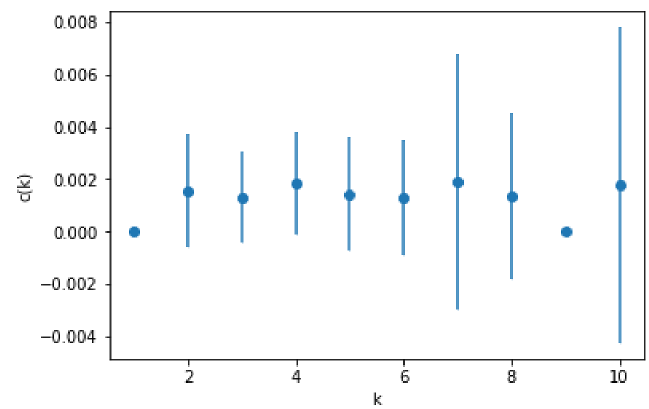

FIG. 7. Scatter plot with the degree dependent clustering distribution for the average over 100 realizations of the configuration model. The error bars represent the standard deviation over realizations.

nodes were not accessible from some other nodes through the given links. In the simulation, this would lead to clusters that can never be infected, so the network used is the subnetwork with the largest number of nodes. The process from one network to the other is in the notebook, with the connected edge list in the same GitHub at [3]. The visualization of the connected and unconnected networks are in the appendix, and for a more clear concept of them in their geometric embedding, also in the map of the Wikipedia page about the International E-road network [4].

The SIS model is used to model epidemics and considers that persons can have two states: susceptible ($n_i = 0$) or infected ($n_i = 1$). When they are susceptible they can get infected by their already infected neighbours with rate λ, and once infected they recover and turn into sus-

TABLE II. Table with the same magnitudes as in table I, but for the configuration model, applied once and for the mean over 100 realizations.

	1 realization					100 realizations				
k	$p(k)$	$p_c(k)$	$\overline{k}_{nn}(k)$	$\overline{k}_{nn}/\kappa(k)$	$\overline{c}(k)$	$p(k)$	$p_c(k)$	$\overline{k}_{nn}(k)$	$\overline{k}_{nn}/\kappa(k)$	$\overline{c}(k)$
1	0.162	0.838	3.03	1.01	0.0000	0.162	0.838	3.00	1.00	0.0000
2	0.521	0.317	3.06	1.02	0.0016	0.521	0.317	3.00	1.00	0.0016
3	0.158	0.158	2.98	0.99	0.0036	0.158	0.158	3.01	1.00	0.0013
4	0.096	0.062	2.87	0.96	0.0015	0.096	0.062	3.00	1.00	0.0019
5	0.040	0.022	2.95	0.98	0.0043	0.040	0.022	2.99	1.00	0.0014
6	0.013	0.009	2.83	0.94	0.0000	0.013	0.009	2.98	0.99	0.0013
7	0.004	0.005	2.86	0.95	0.0000	0.004	0.005	3.01	1.00	0.0019
8	0.004	0.001	3.53	1.18	0.0000	0.004	0.001	2.99	1.00	0.0014
9	0.000	0.001	0.00	0.00	0.0000	0.000	0.001	0.00	0.00	0.0000
10	0.001	0.000	2.50	0.83	0.0000	0.001	0.000	3.00	1.00	0.0018

ceptible with rate δ. The equations in the continuous time case are:

$$\frac{d\rho_i(t)}{dt} = -\delta\rho_i(t) + \lambda\sum_j a_{ij}(\rho_j(t) - \rho_{ij}(t)), \qquad (7)$$

with $i = 1, ..., N$, $\rho_i(t) = \langle n_i(t)\rangle$, $\rho_{ij}(t) = \langle n_i(t)n_j(t)\rangle$ and a_{ij} being the adjacency matrix (1 if i and j are connected by a link, 0 if not). In the mean field approximation it is assumed there are no correlations, then $\rho_{ij}(t) = \langle n_i(t)\rangle\langle n_j(t)\rangle$ and the equation that governs the evolution of the system can be simplified as:

$$\frac{d\rho_i(t)}{dt} = -\delta\rho_i(t) + \lambda(1-\rho_i(i))\sum_j a_{ij}\rho_j(t), \qquad (8)$$

that is the equation we will use in the simulations, because it only depends on $\rho_i(t)$ and not on correlations, that are more difficult to take in account.
In this approximation, the system have a fixed point at $\rho(t) = \sum_i \rho_i(t) = 0$, that changes its stability at a critical value of the parameters: $(\lambda/\delta)_c = 1$. For values below the critical recovery will dominate and $\rho(t) = 0$ will be stable, and for values above the system will evolve to another equilibrium with a non-zero value of infected. If we reorganize randomly the network at every infection or recovery, what is called a fully mixed network, this equilibrium value will be $\rho_{eq} = 1 - \frac{\delta}{\lambda\langle k\rangle}$ for $\lambda/\delta > \frac{1}{\langle k\rangle}$, and 0 for $\lambda/\delta < \frac{1}{\langle k\rangle}$.

About the simulation, it is written in fortran, and the code is available in the same GitHub repository as the other one [3], along with the resulting data that have been used in the plots and the plots themselves.
In the simulation the Gillespie algorithm has been used to choose the time between two consecutive events, assuming they are Poisson processes, with a rate that is the sum of the rates of possible events: $\lambda' = \delta N_{inf} + \lambda E_{act}$, where N_{inf} is the number of infected nodes (that is, the number of possible recoveries) and E_{act} is the number of active links, the ones that connect one infected and one susceptible (the number of possible infections). The time between consecutive processes is sorted as $\tau = -\frac{1}{\lambda'}ln(\xi)$, with ξ distributed as a uniform 0 to 1 distribution. The probabilities of the next event to be an infection and a recovery are:

$$Prob(Inf) = \frac{\lambda E_{act}}{\delta N_{inf} + \lambda E_{act}} \qquad (9)$$

$$Prob(Rec) = \frac{\delta N_{inf}}{\delta N_{inf} + \lambda E_{act}}. \qquad (10)$$

Because all results depend on the ratio between the parameters λ and δ we will set $\delta = 1$ and all results will be in units of δ when dealing with ratios of processes, and of δ^{-1} when dealing with time.
Apart of the agent based approach described, the computation of the evolution of the system had been also made with the system of differential equations in the mean field approximation 8, using the Runge-Kutta 4 algorithm.

The results are presented in 8 and 9. In the first plot we can see how for low values of λ the recovery prevails and the fraction of infected nodes tends to zero, and for high values infection prevails and the system tends to an endemic phase with a stationary non-zero fraction of infected nodes. The agent based approach is systematically below the the mean field approximation in all values of λ, what can also be seen in the second plot. In that plot the existence of a critical point is more clearly seen, with a stationary value of infected equal to zero for the order parameter λ/δ lower than the critical value $(\lambda/\delta)_c$, and a non-zero value that increases with the order parameter for $\lambda/\delta > (\lambda/\delta)_c$. The value we get from the simulation, $(\lambda/\delta)_c \approx 0.4$, is compatible with the theoretical result of fully mixed network approximation, which is $(\lambda/\delta)_c = 1/\langle k\rangle \approx 0.41$,

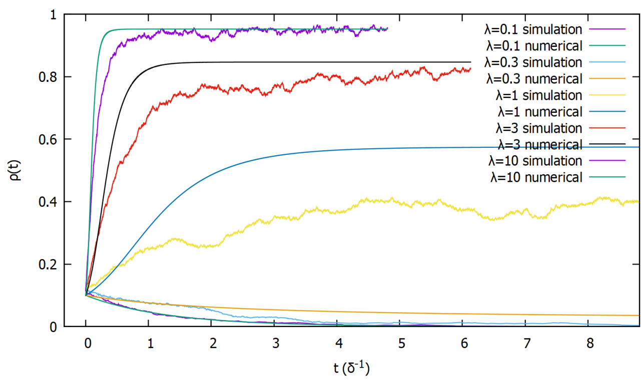

FIG. 8. Plot of the time evolution of the fraction of infected nodes in a SIS model in the chosen network. The different realizations correspond to $\lambda = 0.1, 0.3, 1, 3, 10$, with the agent based approach and the mean field approximation solved with RK4 for each value. The time is in units of δ^{-1}. The initial fraction of infected is set with a probability of 0.1 to be infected for all nodes, although the number of them can fluctuate. The number of steps are 10000 in agent based cases, and for the mean field case we have used the number of steps necessary to reach the same maximum time. The time step of the RK4 is $0.01/((\delta N + \lambda E)/2)$, so it is ≈ 100 times more frequent that the agent based time scale because it needs precision to minimize error.

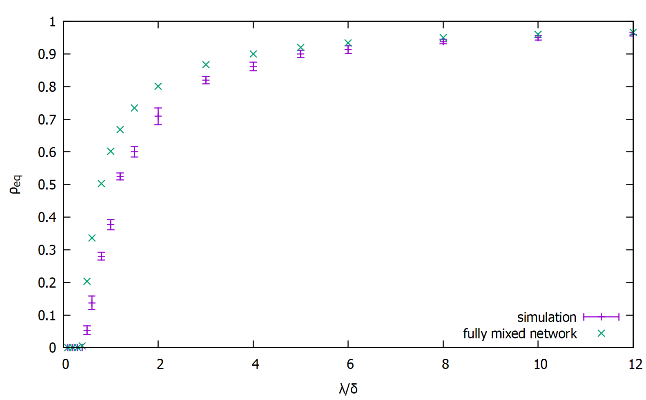

FIG. 9. Scatter plot of the equilibrium fraction of infected nodes in a SIS model as a function of λ/δ, with the result of the agent based simulation and the fully mixed network approximation. The parameters are the same as in 8. The equilibrium fraction of infected nodes has been computed as the average of the last 5000 steps (half of all of them) of the agent base model, that as can be observed in the last plot is stationary in all cases.

IV. CONCLUSIONS

About the results of assignment 1 there is little to say, apart of commenting that the average degree being between 2 and 3 are related with the fact that the majority of cities of the network simply connect two other cities.
Regarding assignment 2, the tail of the degree distribution in 1 resembles a power law when the number of nodes with a given degree are large enough, feature shared among a variety of other real networks. The distribution of the average nearest neighbours degree 2 indicates that the structure of the network has cities with degrees around 5-8 connected with cities with degrees a bit higher than the others (a weak rich club phenomena). Considering the geographical map of the network [4], this can be caused by the high density of E-roads in Central Europe, but it is only a hypothesis. The clustering of the network 3 is not significant because triangles are rarely created if the closed paths in the network have typical scales (in number of links) higher than 3.
The effect of applying the configuration model in the network is the practical destruction of correlations and clustering, with flat profiles of the average nearest neighbours degree distributions and clustering very close to zero (not zero only because of finite size effects).

In respect to the second part, we can say that the structure of the network affect the SIS dynamics on it, since the evolution for a given λ is always a bit below the mean field result in 8. This is due to the structure of the network, that is geometrical embedded and does not exhibit small world properties. This make more difficult for the infected to expand while recoveries remains working in the same way, and consequently the equilibrium fraction of infected is lower than in mean field. About the value of the critical point, the simulation and the fully mixed network results are similar, probably because the mean field approximation works well near the critical point, where the number of infected is still low and the infections are local due to the geometry of the network. The fully mixed network approximation may also work near the critical point because the network is quite homogeneous after all (despite the small peak in the average neighbours degree distribution), while it does not in the endemic phase because, again, of the spatial embedding.

APPENDIX

Here in the appendix are shown some graphs and computations made without being compulsory, just to understand the network better. In figures 10 and 11 we can see maps of the network based in its topology, using the python library "networkx" for the unconnected, original network and the biggest subnetwork used in the simulation.

Then, with the same library, the average shortest path can be computed, giving a result of ≈ 18.4. That value can be interpreted as the mean number of cities we will visit if we want to go from one random city to another. Note that does not take into account the length of the

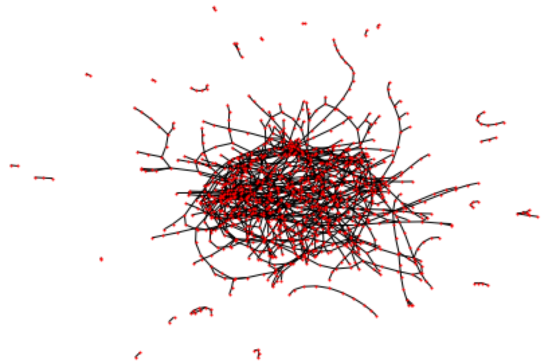

FIG. 10. Map of the E-road network.

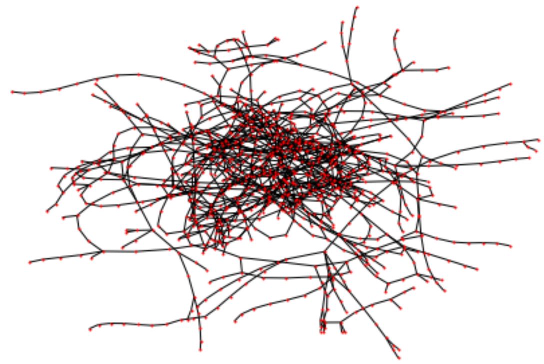

FIG. 11. Map of the biggest connected subnetwork of the E-road network.

trajectories, what can make a shortest path in the network spatially longer than the shortest path in longitude. In addition, the different density of cities that are nodes of the network can also influence the interpretation of the result. About the numerical result, it turns out to be high if we compare it to the Erdős number (≈ 6) for instance, even if the network of people has a way larger number of nodes. That is because the network is embedded in a flat geometry, and has no small world properties, what makes impossible to go to a far node using a low number of links.

[1] Euroroads network dataset – KONECT (2017).
[2] M. Serrano Moral and M. Boguña Espinal, Notes of the complex networks subject, .
[3] C. Callau i Boix, Github repository with the code, plots and data (2023).
[4] International e-road network.